Untersuchungen zum Antwortverhalten und zu Modellen der Skalierung bei der Messung psychologischer Konstrukte

Diplom Psychologe Jörg-Henrik Heine

Vollständiger Abdruck der von der Fakultät für Humanwissenschaften der Universität der Bundeswehr München zur Erlangung des akademischen Grades eines

Doktors der Philosophie (Dr. phil.)

genehmigten Dissertation.

Gutachter:

1. Univ.-Prof. Dr. phil. habil. Christian Tarnai
2. Prof. Dr. phil. Karl-Heinz Renner

Die Dissertation wurde am *5. Juni 2019* bei der Universität der Bundeswehr München eingereicht und durch die Fakultät für Humanwissenschaften am *5. Februar 2020* angenommen.
Die mündliche Prüfung fand am *5. März 2020* statt.

„Models should not be true, but it is important that they are applicable, and wether they are applicable for any given purpose must of course be investigated“

Georg Rasch (1960, S. 38)

Für Karin und Clara

Abstract

The present thesis deals with the different reaction patterns of individual persons or groups of persons to psychodiagnostic questionnaires. The empirical data basis for the analyses in a total of five studies is formed by the answers of two samples of students from the University of the Federal Armed Forces in Munich, who answered questions on a total of 15 psychometric scales from three constructs (*personality*, *vocational interests* and *preferences of musical taste*). The theoretical part of the thesis deals in four chapters with the aspects *scaling of questionnaire data and index formation*, *theory of the investigated constructs*, *theory of reaction patterns in the response to questionnaires* as well as *psychometric modeling of questionnaire data*. The empirical part of the work is divided into two sections, in which analyses of the scaling and psychometric modelling of questionnaire data are carried out according to two different scaling principles. As an overarching finding from this work, it can be seen that two fundamentally different (implicit) response models are effective in respondents when answering psychodiagnostic questionnaires. These show themselves in a consistent manner as a phenomenon that spans scales and constructs. The two response models follow a proximity-distance response process or a dominance response process and indicate the application of two different principles for index formation in psychological diagnostics for the evaluation of questionnaire data. These in turn correspond to two different groups of psychometric models for scaling, explaining and modeling questionnaire data.

response pattern, response style, response set, bias, measurement models, scaling

Zusammenfassung

Die vorliegende Arbeit befasst sich mit den unterschiedlichen Reaktionsmustern einzelner Personen oder Personengruppen auf psychodiagnostische Fragebogenverfahren. Die empirische Datenbasis zu den Analysen in insgesamt fünf Studien bilden die Antworten von zwei Stichproben von Studierenden der Universität der Bundeswehr in München, welche Fragen zu insgesamt 15 psychometrischen Skalen aus drei Konstrukten (*Persönlichkeit*, *berufliche Interessenorientierungen* und *Präferenzen des Musikgeschmacks*) beantwortet haben. Der theoretische Teil der Arbeit behandelt in vier Kapiteln die Aspekte *Skalierung von Fragebogendaten und Indexbildung*, *Theorie zu den untersuchten Konstrukten*, *Theorie zu Reaktionsmustern bei der Beantwortung von Fragebogen* sowie *psychometrische Modellierung von Fragebogendaten*. Der empirische Teil der Arbeit gliedert sich in zwei Abschnitte in denen jeweils Analysen zur Skalierung und psychometrischen Modellierung von Fragebogendaten, nach zwei unterschiedlichen Skalierungsprinzipen, durchgeführt werden. Als übergeordneter Befund aus dieser Arbeit zeigt sich, dass bei der Beantwortung von psychodiagnostischen Fragebogenverfahren zwei fundamental unterschiedliche (implizite) Antwortmodelle bei den antwortenden Personen wirksam sind. Diese zeigen sich in konsistenter Weise als skalen- und konstruktübergreifendes Phänomen. Die beiden Antwortmodelle folgen einem *Nähe–Distanz*-Antwortprozesses oder einem *Dominanz*-Antwortprozesses und indizieren für die Auswertung von Fragebogendaten die Anwendung von zwei unterschiedlichen Prinzipien zur Indexbildung im Rahmen der psychologischen Diagnostik. Diese korrespondieren wiederum mit zwei unterschiedlichen Gruppen von psychometrischen Modellen zur Skalierung, Erklärung und Modellierung von Fragebogendaten.

Antwortmuster, Antwortstil, Messfehler, Messmodelle, Skalierung

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung, Überblick und Einführung in die Dissertation

1.1 Einleitung in die übergreifenden Fragestellungen

Die vorliegende Arbeit befasst sich mit der unterschiedlichen Art und Weise mit der einzelne Personen oder Personengruppen auf psychodiagnostische Fragebogenverfahren reagieren. Die Inhalte der einzelnen Fragen in solchen Fragebögen beziehen sich auf unterschiedliche Konstrukte die verborgene Merkmale oder Einstellungen umfassen. Die Antworten, als manifest beobachtbare Reaktionen der Personen, in dem in der Regel vorgegebenen Kategoriensystem der einzelnen Fragen, werden als Indikatoren für diese Merkmale oder Einstellungen angesehen. Das Resultat der Beantwortung der gesamten Fragen eines Fragebogens stellt, personenbezogen betrachtet, ein individuelles Reaktionsmuster dar, welches entweder Gemeinsamkeiten mit anderen, ähnlichen Antwortmustern oder aber auch idiosynkratische Züge aufweisen kann. In dieser Arbeit werden diese unterschiedlichen Reaktionsmuster einerseits klassifiziert und andererseits deren Implikationen auf die Fragebogendiagnostik behandelt – so beispielsweise die Auswirkungen unterschiedlicher Antwortmuster auf Befunde zu Zusammenhängen zwischen verschiedenen Konstrukten. Es wird der Frage nachgegangen, ob die unterschiedlichen Antwortmuster eher als konsistente, möglicherweise konstruktübergreifende Verhaltensdispositionen anzusehen

sind, oder sich aber spezifisch für unterschiedliche Konstrukte (vgl. Kapitel 2) und möglicherweise als Ergebnis deren jeweiliger Form der Operationalisierung, ergeben. Es wird überprüft, ob sich unterschiedliche Antwortmuster durch eine individuell unterschiedliche Interpretation der Fragen bei deren Beantwortung erklären lassen. Damit einher geht die Frage, ob das bei der Testentwicklung implizit angenommene *Antwortmodell*, welches in ein entsprechendes *psychometrisches* Modell zur *Skalierung* bzw. Indexbildung (vgl. Kapitel 4) mündet, in gleicher Weise von allen antwortenden Personen antizipiert wird und damit eine universelle Gültigkeit aufweist. In diesem Sinne wird auch der Frage nachgegangen, inwieweit die geübte Praxis der variablen- und dimensionsorientierten Auswertung der Daten aus der Anwendung von Fragebogenverfahren unter Annahme einer *latenten*, kontinuierlichen Merkmalsvariablen, nicht zusätzlich um eine personenzentrierte Auswertung solcher Daten ergänzt werden muss. Solch eine personenzentrierte Auswertung birgt die Möglichkeit der Entdeckung, Interpretation und Deutung idiosynkratischer Sichtweisen und Reaktionen auf die einzelnen in den Messinstrumenten vorgegebenen Fragen. Die daraus resultierenden Erkenntnisse müssten dann bei der Anwendung unterschiedlicher Skalierungsmodelle zu Indexbildung berücksichtigt werden.

In dieser kurzen Beschreibung der nachfolgend gegebenen Inhalte deutet sich an, dass die manifest vorgefundenen Reaktionen der befragten Personen nicht nur durch die über das jeweilige Konstrukt definierte Messintention bestimmt werden. Vielmehr können hier unterschiedlichste Aspekte während der Beantwortung mit einfließen. Diese Aspekte können allgemein oder individuell, zeitlich stabil oder variabel wirksam werden und sich aus der Person oder der Situation ergeben. Die Systematisierung und Untersuchung derartiger, aus messtheoretischer Perspektive eigentlich störender Aspekte, hat in der entsprechenden Literatur eine lange Historie (vgl. Kapitel 3). Aus der Perspektive der Skalierung fällt bei der Durchsicht der Literatur auf, dass sich die Grundlage für die empirisch begründete Systematisierung solcher Aspekte bei der Messung psychologischer Konstrukte, meist auf ein summativ, kumulatives Verrechnungsmodell zur Indexbildung für die Personenantworten auf die einzelnen Fragen in einem Fragebogenverfahren stützt. Dies erscheint vor dem Hintergrund, dass diese summative Verrechnung historisch betrachtet erst später Einzug in die Praxis der Skalierung von Daten aus psychodiagnostischen Fragebogenverfahren gefunden hat umso erstaunlicher.

In den entsprechenden Studien zu den unterschiedlichen Einflussfaktoren auf das Antwortverhalten werden meist, je nach Forschungsinteresse der Autoren, einzelne Aspekte (isoliert) in den Vordergrund gestellt und entsprechend untersucht. Dementsprechend existiert eine recht große Anzahl von spezifischen Begrifflichkeiten für das beobachtete Phänomen abweichender Antwortmuster – mit untereinander teils disjunkten und teils sich überschneidenden Definitionen. In der vorliegenden Arbeit werden diese unterschiedlichen Begrifflichkeiten daher auch systematisch dargestellt und diskutiert. Um eine derartige Systematisierung vorzunehmen, ist es sinnvoll zunächst einen neutralen (Ober-)Begriff einzuführen. Der auch im Titel der vorliegenden Arbeit verwendete, vielleicht zunächst unspezifisch erscheinende, Begriff „*Antwortverhalten*“ wird insofern bewusst gewählt. Der Begriff „*Antwortmuster*“ oder der englische Begriff „*pattern*“ [engl.: *pattern* ≡ deut.: *Muster*] wird dagegen in den folgenden Abschnitten immer dann synonym verwendet, wenn auf die empirischen Daten Bezug genommen wird, welche das Resultat des Antwortverhaltens der untersuchten Personen sind. Diese Antwortmuster reflektieren möglicherweise nicht nur quantitative Unterschiede, sondern sind, im Sinne einer (qualitativen) Typologie, möglicherweise Ausdruck des individuellen Erlebens- und Verhaltensmusters der antwortenden Personen. Der Vorteil dieser beiden unspezifisch erscheinenden Begriffe liegt darin, dass sich diese zunächst lediglich auf das (besondere) empirisch vorgefundene Ergebnis aus der Datenerhebung beziehen. Es wird dabei keine Deutung oder Wertung hinsichtlich der möglicherweise unterschiedlichen, zu vermutenden Entstehungsursachen vorweg genommen. Die vorgefundenen „Antwortmuster“, welche in Folge eines mehr oder weniger durchschaubaren internen psychischen Prozesses innerhalb der antwortenden Person entstehen, sollen so zunächst ohne vorweggenommene Interpretationen „nur“ beobachtet werden. Darüber hinaus bietet der Begriff des „Antwortverhaltens“ im Gegensatz zu dem in der deutschsprachigen Literatur auch verwendeten Begriff des „Antwortstils“ [engl. *response style*] den Vorteil, dass er keinerlei a priori Konnotationen hinsichtlich der Konsistenz, eben im Sinne einer überdauernden, stilistischen Besonderheit bestimmter Personen (-gruppen), vorwegnimmt oder unterstellt. Auch der oft verwendete Begriff der „Antwortverzerrung“ [engl. *response set* oder *response bias*] kann, trotz dieser fehlenden „stilistischen Konnotation“, letztlich eine

a priori Interpretation darstellen. Die gewählten Begriffe „Antwortverhalten“ und „Antwortmuster“ nehmen demgegenüber keine inhaltlichen Interpretationen vorweg und bleiben somit neutral. Die Interpretation der unterschiedlichen Antwortmuster bleibt somit zunächst offen.

Das individuell unterschiedliche Antwortverhalten bei der Anwendung von Fragebogenverfahren kann in Analogie zu individuell unterschiedlichem zwischenmenschlichen Kommunikationsverhalten betrachtet werden (Borg & Staufenbiel, 2007; Bortz & Döring, 2006). So merken Borg und Staufenbiel (2007) in einer Fußnote zur potentiellen Mehrdeutigkeit von Fragebogen-Items an, dass „*... die Befragung einer Person mit einem Item als ein Kommunikationsprozess zwischen Forscher und Befragtem verstanden werden kann.*“ (Borg & Staufenbiel, 2007, S. 21). Im Zusammenhang mit einer kritischen Diskussion der Begriffe *Selbstdarstellung* vs. *Testverfälschung*, weisen Bortz und Döring (2006) darauf hin, dass die Bearbeitung von Tests oder Fragebögen als (besondere) Form der Kommunikation zwischen Testleiter und Testperson aufgefasst werden kann:

> Aus Sicht der Probanden wird das Ausfüllen von Tests oder Fragebögen als Kommunikation erlebt. Testpersonen wissen, dass sie anderen Menschen durch den Test etwas über sich mitteilen und machen sich Gedanken darüber, wer sie sind, was sie mitteilen wollen und was nicht, bei wem die Informationen ankommen, wie der Empfänger auf sie reagieren könnte und was mit ihnen geschieht. (Bortz & Döring, 2006, S. 232)

Als Ergebnis einer solchermaßen individuell unterschiedlichen, besonderen Form der Kommunikation zwischen Testleiter und Testperson, ergeben sich in den Daten dann potentiell *idiosynkratische* Antwortmuster. In Anlehnung an Paul Watzlawicks Metakommunikatives Axiom – ***Man kann nicht nicht kommunizieren!*** (vgl. z. B., Watzlawick, Beavin & Jackson, 2000), kann man bezüglich der unterschiedlichen Reaktionen der Personen auf das jeweils vorgelegte Fragebogenverfahren argumentieren, dass sich ein bestimmtes Antwortmuster (in den Daten) einfach *immer* als manifestes Ergebnis der Bearbeitung oder auch des Auslassens einzelner Fragen eines psychodiagnostischen Fragebogenverfahrens ergibt. Da nach Watzlawicks Axiom Verhalten kein Gegenteil hat, man sich also nicht ***nicht verhalten*** kann, ist es auch unmöglich, nicht zu kommunizieren, womit schließlich jedes Verhalten kommunikativen Charak-

ter hat. Bezogen auf die (indirekte) Kommunikation zwischen Testperson und Testleiter bei der Anwendung von psychodiagnostischen Fragebogenverfahren kann diese Kommunikation auch darin bestehen, entweder nur auf bestimmte, im Extremfall auf keine, oder in einer konstanten, stereotypen Art und Weise auf die vorgelegten Fragen zu antworten. Solch ein Verhalten, da es nicht der üblichen Erwartung (des Testentwicklers) entspricht, stellt dann zunächst einmal eine besondere – *idiosynkratische* – Form eines abweichenden Antwortmusters dar. Die Frage ob es sich bei dem Antwortverhalten und den daraus resultierenden Antwortmustern um einen *Stil*, im Sinne einer überdauernden, möglicherweise vom Inhalt der einzelnen Fragen abhängigen oder unabhängigen Eigenschaft der antwortenden Personen handelt, muss auf der Grundlage empirischer Untersuchungen geklärt werden.

Die vorliegende Arbeit widmet sich daher der Untersuchung des Antwortverhaltens bei der Messung psychologischer Konstrukte mittels Fragebogenverfahren.

1.2 Fragebogen zur Messung psychologischer Konstrukte

Fragebogen sind ein ubiquitäres Phänomen und durchdringen in der heutigen Zeit unseren Alltag. Fast jeder dürfte bereits irgendwann in seinem Leben mindestens einmal mit einem Fragebogen „konfrontiert“ gewesen sein. Fragebogen sind ein praktisches und ökonomisches Instrument zur Datenerhebung und werden insbesondere in den Sozialwissenschaften zur Erfassung von persönlichen Einstellungen, kognitiven Fähigkeiten, politischen Ausrichtungen und sonstigen Meinungsbildern eingesetzt. In der *Sozialpsychologie*, der *Differentiellen Psychologie* und der *Persönlichkeitspsychologie* ist der Einsatz von Fragebogen zur Selbstbeschreibung bei der Messung psychologischer Konstrukte und interindividuell unterschiedlicher Einstellungen eine gängige Praxis.

Den an der jeweiligen Untersuchung beteiligten Personen werden dabei zunächst eine Reihe von Fragen oder Aussagen vorgelegt, welche von diesen eingeschätzt und beantwortet werden müssen. Die einzelnen Fragen oder Aussagen werden, auch in der deutschsprachigen Literatur, oft mit dem Begriff *Item* bezeichnet. Borg (1992, S.62) merkt zu dem Begriff „Item“ kritisch an, dass dieser in der sozialwissenschaftlichen Literatur oft keine feste Bedeutung habe. Demgegenüber definiert Osterlind (1990) ein *Item* im Zusammenhang mit der Erfassung individueller Merkmale über Fragebogenverfahren wie folgt:

> A test item in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form for answering; and, it is intended to yield a response from an examinee from which performance in some psychological construct (such as an ability, predisposition, or trait) may be inferred. (Osterlind, 1990, S. 3)

Das *Item*, also die einzelne Frage, wird nach dieser allgemeinen Definition im Rahmen der Fragebogendiagnostik hier als (kleinste) Messeinheit definiert. Auf der strukturellen Ebene gliedert sich das *Item* in zwei Bestandteile. Zum einen in ein Frageteil, der als *Stimulus* bzw. *Reiz* bezeichnet wird, welcher eine Antwort auslösen soll. Zum anderen ein Antwortteil, welcher meist aus einem fest vorgegebenen Antwortformat besteht. Die mit der Fragestellung initiierten Antworten der Personen werden dabei als Indikatoren für einen gewissen

Ausprägungsgrad bezogen auf ein Merkmal in den so zu messenden psychologischen Konstrukten angesehen. Das Vorhandensein eines fest vorgegebenen Antwortformats ist nach dieser Definition von Osterlind (1990), ein wichtiges Charakteristikum des *Items*. Bei diesem, in der Fragebogendiagnostik meist eingesetzten, *geschlossenem Antwortformat*, müssen die Fragen oder Aussagen über ein vorgegebenes Kategoriensystem[1] beantwortet werden. Dieses Kategoriensystem zur Beantwortung der einzelnen Fragen eines Fragebogens (oder auch *Inventars*) wird als *Antwortskala* bezeichnet. In Abgrenzung dazu wird eine Menge von Fragen oder Items, welche **ein** spezifisches Merkmal oder eine spezifische Merkmalsdimension messen sollen, als *Skala* oder auch *psychometrische Skala* bezeichnet. Im einfachsten Fall muss den vorgelegten Fragen oder Aussagen entweder zugestimmt oder diese abgelehnt werden. Werden mehrere Antwortkategorien in der Antwortskala eingesetzt, besteht die Aufgabe darin, das Ausmaß der Zustimmung (oder Ablehnung) zu der betreffenden Frage über die Wahl einer entsprechenden Kategorie auszudrücken, wobei die Antwortkategorien als abgestufte Skala der Zustimmung oder Ablehnung zu interpretieren sind. Die zweistufigen, geschlossenen Antwortformate werden in der psychometrischen Literatur meist als *dichotom* und die mehrstufigen Formate meist als *polytom* bezeichnet. Ein Beispiel für ein polytomes Antwortformat ist die von Likert (1932) vorgeschlagene fünfstufige Antwortskala, welche auch in den heute verbreiteten Fragebogen meist von den Antwortkategorien „*starke Ablehnung*“ bis „*starke Zustimmung*“ reicht. Solche Antwortskalen werden nach dem englischen Begriff „*rating*“ für „Einschätzen“ auch als *Ratingskalen* bezeichnet (Bortz & Döring, 2006). Die Anwendung solcher Ratingskalen in Fragebogen ist weit verbreitet und blickt auf eine lange Historie zurück (Ellson & Ellson, 1953). Trotz der weiten Verbreitung von Ratingskalen ist deren Konstruktion, deren Einsatz und die anschließende Auswertung hinsichtlich der Qualität der resultierenden Messwerte nicht unumstritten (McReynolds & Ludwig, 1987; Saal, Downey & Lahey, 1980). So gilt es zum Beispiel die grundsätzliche Frage zu klären, ob sich die Skalenpunkte der vorgegebenen Ratingskala beim Einsatz von verbalen Wortmarken auf *Häufigkeiten*, *Intensitäten*,

[1] Antworten, welche im Rahmen eines offenen Antwortformates erhoben werden, müssen vor deren weiterer (statistischer) Auswertung einer entsprechenden Kategorisierung unterzogen werden.

Wahrscheinlichkeiten, *Prozentangaben* oder aber allgemeine *Bewertungen* beziehen (vgl. Bortz & Döring, 2006). Solche Fragen sind mehr oder weniger direkt mit den Inhalten der einzelnen Items bzw. der Art der Operationalisierung des zu erfassenden Konstruktes verbunden. Es stellt sich dabei zunächst die grundlegende Frage nach der theoretischen Definition des zu erfassenden Merkmals. Damit verbundenen stellen sich dann unterschiedliche Fragen wie die nach der sprachlichen Formulierung der Items (Horan, DiStefano & Motl, 2003; McPherson & Mohr, 2005; Nye, Newman & Joseph, 2010; Schriesheim & Eisenbach, 1995), der konnotativen Bedeutung der einzelnen Items (Chang, 1995), der Polarität des erfassten Merkmals (Reise & Waller, 2009; Russell & Carroll, 1999a, 1999b; Watson & Tellegen, 1999), sowie in Verbindung mit letzterem die Polarität der eingesetzten Items (Lam & Stevens, 1994) und deren Antwortskalen (Dalal & Carter, 2015b; Schriesheim, Eisenbach & Bailey, 1991; van Sonderen, Sanderman & Coyne, 2013).

Es stellt sich übergreifend also die Frage nach der Art und Weise, in der die antwortenden Personen einerseits das Merkmal und andererseits die Antwortskalen interpretieren und welche Art von Einstellungen bei der Beantwortung der einzelnen Fragen und auch der gesamten Skala, assoziiert und aktiviert werden. Ferner besteht die Frage, inwieweit diese Assoziationen bei allen antwortenden Personen in gleicher Weise wirksam werden und ob diese mit den Vorstellungen des Testentwicklers übereinstimmen. Segura und González-Romá (2003) merken hierzu beispielsweise an, dass sich für einige der befragten Personen unerwartete Antwortmuster nicht etwa ergeben, weil deren Antworten falsch wären, sondern weil diese nicht mit den Vorstellungen des Testentwicklers übereinstimmen. In diesem Sinne zeigt sich beispielsweise für Fragebogen im Bereich der *Persönlichkeit* (vgl. Kapitel 2 *Theorie zu den untersuchten Konstrukten*) nicht selten, dass sich bei deren Beantwortung immer wieder Antwortmuster ergeben, welche den Erwartungen des Testkonstrukteurs widersprechen. Die hier kurz angerissenen Aspekte werden daher im Kapitel 3 *Theoretischer Hintergrund zu Antwortmustern* eingehender dargestellt.

Neben diesen, mehr oder weniger am Inhalt der Items orientierten Fragestellungen, stellt sich insbesondere bei den polytomen Antwortformaten auch die Frage nach dem Skalenniveau der Rating- bzw. Antwortskala der einzelnen

Items. Theoretisch denkbar sind hier zunächst die Skalenniveaus *nominal-*, *ordinal-*, *intervall-* und *verhältnisskaliert*, wie sie von Stevens (1946) vorgeschlagen wurden.

Nominalskalen enthalten Informationen über die Unterschiedlichkeit von Dingen, Personen oder Ereignissen im Hinblick auf ein bestimmtes Merkmal. Ein Beispiel könnte hier das Merkmal „bevorzugte Eissorte“ sein, das man mit den vorgegebenen Kategorien „Schokolade“, „Erdbeere“, „Vanille“, „Waldmeister“ sowie einer Restkategorie „Sonstige Sorten“ erfassen könnte. Die vorgegebenen Kategorien drücken hier lediglich eine *Unterschiedsrelation* in Bezug auf das Merkmal aus.

Ordinalskalen enthalten Informationen über die Ordnungsreihenfolge von Dingen, Personen oder Ereignissen in Bezug auf ein bestimmtes Merkmal. So werden beispielsweise bei Sportwettbewerben die Rangplätze 1., 2. und 3. Platz, usw. vergeben, welche sich zum Beispiel bei Laufwettbewerben auf das eigentlich kontinuierliche „Merkmal“ Geschwindigkeit bzw. Zeit beziehen. Entscheidend bei solchen *Ordnungsreihenfolgen* ist, im Vergleich zur in diesem Beispiel tatsächlich gemessenen Zeit, dass hier lediglich Informationen über die Reihenfolge enthalten sind - nicht aber Informationen über die Abstände zwischen den Rangplätzen. Solche Ordinalskalen werden daher manchmal auch als Rangskalen bezeichnet. Ordinalskalen weisen, trotz des Fehlens von belastbaren Informationen bezüglich der Abstände der Skalenpunkte, eine wichtige *quantitative* Eigenschaft auf. So ergibt sich aus den „*größer als*“ oder „*kleiner als*“ Relationen der einzelnen Skalenpunkte der Ordinalskala die Forderung der Eigenschaft der *Transitivität* (Coombs, 1951; Gulliksen, 1946). Transitivität bedeutet dabei, dass hinsichtlich der Relationen von drei Skalenpunkten einer Ordinalskala (A, B, C) aus den gesicherten Relationen $A > B$ und $B > C$ auf die Relation zwischen A und C geschlossen werden kann, welche bei Transitivität der Skala dann $A > C$ lauten muss (Coombs, Raiffa & Thrall, 1954; Diekmann, 2014). Die Transitivität stellt in der quantitativen sozialwissenschaftlichen Forschung eine wichtige Eigenschaft von Relationen und Ordnungen im Rahmen der Messung und dem Vergleich von Merkmalsausprägungen bei Personen und der Klassifikation von Aufgaben dar (Borg, 1992).

Intervallskalen zeigen auch die Ordnungsreihenfolge von Dingen, Personen oder Ereignissen in Bezug auf ein bestimmtes Merkmal an, wobei zusätzlich

die Intervalle zwischen den einzelnen Skalenpunkten entweder als gleich, zumindest aber im Hinblick auf die Relation der Intervalle als „vergleichbar“, angesehen werden. Die Intervalle (Abstände) zwischen der Reihe der natürlichen Zahlen 1, 2, 3, 4 und 5 beträgt zum Beispiel immer 1 und ist damit, auch im mathematischen Sinne nicht nur vergleichbar, sondern auch *gleich*.

Verhältnisskalen unterscheiden sich von den Intervallskalen dadurch, dass sie einen natürlichen Nullpunkt aufweisen und damit die Verhältnisse (engl. *ratios*) zwischen einzelnen Skalenpunkten sinnvoll interpretiert werden können. Das Merkmal beziehungsweise die Variable Alter hat zum Beispiel einen sinnvoll zu interpretierenden Nullpunkt und das Verhältnis zwischen den Skalenpunkten „11 Jahre alt“ zu „44 Jahre alt“ kann inhaltlich sinnvoll als „viermal älter“ oder „viermal jünger“ interpretiert werden. Grundsätzlich ist bei der Beantwortung von Fragebogen-Items auch noch die *Absolutskala* als Skalenniveau für die Antworten auf einzelne Items denkbar. So könnte zum Beispiel auf die Frage - *„wie häufig gehen Sie in einer Woche aus?“* - die Angabe der absoluten Häufigkeit außerhäuslicher Aktivität erwartet werden - welche theoretisch einen gültigen Wertebereich von null bis zu einer Obergrenze haben könnte, die lediglich durch das Zeitbudget der jeweils befragten Personen definiert wäre. Wie das Beispiel allerdings schon nahelegt, werden Fragen nach absoluten Häufigkeiten sinnvollerweise eher über ein offenes Frageformat erfasst.

Liegen die erhobenen Daten vor, stellt sich also die Frage nach der angemessenen Methode zu deren Auswertung. Nach Coombs (1956) lassen sich die so aus Fragebogen gewonnenen Daten(-Punkte) als Realisation einer Verhaltensbeobachtung der empirischen Relation zwischen den Items beziehungsweise Stimuli und den antwortenden Personen interpretieren. Die Items (mit geschlossenem Antwortformat) sollen dabei einen bestimmten Ausprägungsgrad in Bezug auf die zu messende Eigenschaft oder Einstellung repräsentieren. Die Personen werden bei der Bearbeitung der Fragen implizit dazu aufgefordert ihre persönliche Ausprägung mit der des jeweiligen Items zu vergleichen – *„. . . the individual was, in effect, asked to compare his ability level with the difficulty level of the item.“* (Coombs, 1956, S. 317). Bei der Auswertung der Daten gilt es auch zu berücksichtigen, ob die einzelnen Items eines Fragebogens nur einer oder mehreren Merkmalsdimensionen eines Konstrukts zugeordnet werden. So umfasst zum Beispiel der in der vorliegenden Arbeit eingesetzte Fragebogen

zum Konstrukt *Persönlichkeit* fünf Merkmalsdimensionen, denen jeweils nur vier der insgesamt 20 Items zugeordnet sind. Diese einzelnen Merkmalsdimensionen eines Fragebogens, welche durch eine Reihe von Items repräsentiert sind, werden oft auch als *psychometrische Skalen* bezeichnet. In der Regel werden die einer Merkmalsdimension zugeordneten Items jeweils getrennt zu einem Index zusammengefasst (Latcheva & Davidov, 2014). Nach einer solchen Zusammenfassung der Antworten auf die einzelnen Items, die nach einem bestimmten Verfahren bzw. nach einer bestimmten *Verrechnungsvorschrift* erfolgt, steht der Index mit einem einzigen Wert für die Eigenschaftsausprägung jeder der antwortenden Personen. Im Kern geht es nun darum zunächst theoriegeleitet zu begründen und später empirisch zu belegen, *welche* Verrechnungsvorschrift für die einzelnen Fragen einer oder mehrerer *psychometrischer Skalen* in einem Fragebogen zur *Indexbildung* angemessen ist. Lässt sich die angewendete Verrechnungsvorschrift bei der Indexbildung empirisch rechtfertigen, kann der resultierende Messwert für die Ausprägung einer jeden Person in Bezug auf die zu erfassende Eigenschaft angesehen werden.

Bei einer solchen Indexbildung sind im Wesentlichen zwei Teilschritte zu unterscheiden, welche sich aus den einzelnen Elementen als Bestandteile eines Fragebogens mit mehreren psychometrischen Skalen ergeben (vgl. z. B. Greving, 2007). Zunächst muss bezogen auf die Antwort- beziehungsweise Ratingskala festgelegt werden, wie die empirischen Personenantworten zu den einzelnen Items angemessen in ein „*numerisches Relativ*“ (vgl. Fischer, 1974, S. 115, 116) übersetzt werden können. Nach Pfanzagl (1959, 1971) besteht das generelle Ziel dabei darin, die Antworten der Personen innerhalb des vorgegebenen (Antwort-)Kategoriensystems in einer solchen Weise auf eine Menge von Zahlen abzubilden, dass sich die empirischen Relationen in den numerischen Relationen abbilden.

Als Beispiel sei hier das in den Sozialwissenschaften oft im Rahmen einer „*Per-fiat-Messung*“ (Bortz & Döring, 2006, S. 70) vorgenommene Zuordnungsschema genannt: $1 \equiv ja; 0 \equiv nein$ oder aber auch $0 \equiv starke\ Ablehnung$ bis $4 \equiv starke\ Zustimmung$. Die Anwendung einer solchen oder ähnlichen Zuordnungsregel – „*scoring rule*“ (Bechger, Maris, Verstralen & Béguin, 2003, S. 320) – wird, soweit in entsprechenden empirischen Untersuchungen überhaupt explizit erwähnt, als *scoring* bezeichnet (vgl. Leunbach, 1961), sowie

Abschnitt 1.3.1 in dieser Arbeit. Im zweiten Schritt geht es dann darum, ob und wie die so übertragenen Antworten auf geeignete Art und Weise zu einem, oder auch mehreren Gesamtmesswerten – den Indizes – für die jeweiligen psychometrischen Skalen, zusammengeführt werden können. Die dahinterliegende Idee besteht darin, die einzelnen Items als *manifeste Indikatoren* für nicht direkt zu beobachtende *latente Eigenschaften* der befragten Personen anzusehen (Kromrey, 1994). In der sozialwissenschaftlich empirischen Forschung stellen solche *latente Variablen* einzelne Dimensionen theoretischer *Konstrukte* (mit unterschiedlicher Anzahl von Dimensionen) dar, welche dazu eingeführt werden, um die beobachteten Zusammenhänge zwischen den manifest beobachteten Indikatoren (die einzelnen Fragen in einem Fragebogen) zu erklären (Borsboom, 2008; Rost & Langeheine, 1997). Das grundlegende Ziel besteht darin, die unterschiedlichen Ausprägungen in der *latenten Variablen* in Zahlen in einem Index abzubilden, sodass die entsprechende Merkmalsausprägung gemessen wird (Narens, 1981; Narens & Luce, 1986).

Der empirische Ansatz zur Indexbildung, welcher sich in den Sozialwissenschaften durchgesetzt hat, bedient sich dabei dimensionsreduzierender Verfahren, wie zum Beispiel der (linearen) Hauptkomponenten- und Hauptachsenanalyse (z. B. Galton, 1888; Hotelling, 1933; Pawlik, 1971; Pearson, 1901; Spearman, 1904; Überla, 1977), welche oft auch übergreifend als Faktorenanalyse (FA) bezeichnet werden. Als Variablen orientiertes Verfahren analysiert die FA Zusammenhänge einzelner Variablen (z. B. Galton, 1888) und hat ihren Ursprung in den Arbeiten von Pearson und Spearman (Pearson, 1901; Spearman, 1904). Die Faktorenanalyse ist ein multivariates statistisches Verfahren das häufig in der Psychologie und anderen Sozialwissenschaften eingesetzt wird (vgl. Pawlik, 1971) und gilt als Standardverfahren bei der Analyse und Interpretation von Fragebogen zur Erfassung von Eigenschaften und Persönlichkeitsmerkmalen (Cattell & Saunders, 1954a; Pawlik, 1971; Überla, 1977). Die FA führt eine große Anzahl von Variablen als manifeste Indikatoren auf einen kleineren Satz von Faktoren zurück (Pawlik, 1971). Die FA wird eingesetzt, um die Items einzelnen, inhaltlich interpretierbaren Skalen zuzuordnen (Borg & Mohler, 1993). Einen historischen Überblick zu den Anfängen der Faktorenanalyse oder der Hauptkomponenten- und Hauptachsenanalyse gibt Burt (1949). Die methodischen Grundlagen zur FA und ihre Varianten werden bei Überla (1977) behandelt. Pawlik (1971) referiert die Bedeutung der FA für die

psychologische Forschung.

Demgegenüber bedient sich der facettentheoretische Ansatz zur Indexbildung substanzwissenschaftlicher Theorien (Borg & Mohler, 1993). Dabei werden einerseits gezielt Itemformulierungen gesucht und andererseits die Zuordnung einzelner, bereits bestehender Fragen beziehungsweise Items zu den jeweiligen Indizes oder Skalen theoretisch begründet (Borg, 1992; Borg & Staufenbiel, 1993). Borg (1992) beschreibt den Ansatz der Facetten Theorie dahingehend, dass nicht der über einen Index zu messende Sachverhalt direkt definiert wird, sondern dieser über das Universum aller möglichen Items (z. B. Mohler, 2006), welche sich auf diesen zu erfassenden Sachverhalt beziehen, definiert wird – „*... So wird z. B. nicht eigentlich Intelligenz selbst definiert, sondern das Universum der Intelligenzitems festgelegt als Menge aller Fragen zum Verhalten eines Individuums, ...*“ Borg (1992, S. 63). Ein solches systematisiertes Vorgehen soll dabei die Verbindung von Theorie (Itementwicklung) und Empirie (Überprüfung der Skalierbarkeit) sicherstellen.

Im Zusammenhang mit der allgemeinen Frage nach der Verbindung zwischen Theorie und Empirie sprechen Steyer und Eid (2000) hier vom sogenannten *Überbrückungsproblem*, welches jeder psychologischen Messung inhärent ist. Diese Grundfrage kann letztlich auch als der eigentliche Leitgedanke bei jedweder psychometrischen Modellbildung angesehen werden (vgl. Kapitel 4 *Psychometrische Modellierung*). Nach Browne (2000) befasst sich in diesem Sinne die *Psychometrie* mit der Quantifizierung und Analyse von interindividuellen Differenzen und dabei speziell mit der Entwicklung von geeigneten Methoden zur Überprüfung der Operationalisierung bei der Erfassung von psychologischen Konstrukten, sowie der Auswertung von daraus resultierenden (Beobachtungs-)Daten. Zu der daher bei der Datenauswertung notwendigerweise einhergehenden Modell- und Skalenbildung, stellt Coombs (1967) fest, dass letztlich jedes Mess- oder Skalierungsmodell bereits eine bestimmte Theorie über das [menschliche] Verhalten impliziert – „*A measurement or scaling model is actually a theory about behavior ...*“ (Coombs, 1967, S. 5). Je nach gewählter Operationalisierung zur Lösung des Überbrückungsproblems zwischen Theorie und Empirie und je nach gewähltem psychometrischen (Antwort-) Modell ergeben sich nämlich unterschiedliche Erwartungen hinsichtlich des Verhaltens der getesteten Personen, also deren beobachteter Antwortmuster. Ein

bestimmtes Antwortmuster, welches im Rahmen einer bestimmten psychometrischen Modellvorstellung durchaus den „normalen“ Erwartungen entspricht, kann unter der Annahme eines anderen psychometrischen Antwortmodells, bereits als abweichend klassifiziert werden. Diese unterschiedliche Art und Weise der psychometrischen Modellbildung und die damit zusammenhängende Indexbildung im Rahmen der Datenauswertung wird in der psychometrischen Literatur unter dem Begriff *Skalierung* subsumiert, welcher in der englischsprachigen Literatur oft als *scaling* bezeichnet wird (vgl. Coombs, 1950, 1956, 1967; Torgerson, 1967). Das Ziel der Skalierung besteht dabei in der Quantifizierung der erhobenen Daten (Young, 1984).

1.3 Skalierung von Fragebogendaten

Im Verlaufe der psychometrischen Forschung sind zu der Problematik der *Skalierung* eine ganze Reihe von Methoden und Modellen vorgeschlagen worden. Bei den meisten dieser (Mess-) und Skalierungsmethoden geht es darum entweder Personen oder aber „äußere Einflüsse“ auf Personen, meist als *Reize* oder *Stimuli* bezeichnet, zu ordnen – also zu skalieren. Bei der Auswertung von Fragebogendaten stellen diese *Reize* oder *Stimuli* die Fragen beziehungsweise die Items dar, die als Kommunikationsanlass für die befragten Personen, und damit als Ursache von deren Reaktion angesehen werden (z. B. Kraut, 1995). Im Zusammenhang mit der Messung von individuellen Einstellungen teilt Torgerson (1967) die Skalierungsmethoden in drei grundsätzliche Herangehensweisen ein. Zum einen die personenorientierte Skalierung „*The Subject-Centered Approach*“; die damit verbundenen Fragestellungen beziehen sich auf die Unterschiedlichkeit von Personen. Die Ergebnisse aus der Anwendung dieser Methoden resultieren in der Regel direkt in Skalenwerten für die zu messenden Personen. Als ein prominentes Beispiel sei hier die im folgenden Abschnitt dargestellte Skalierung nach Likert (1932) genannt.

Zweitens, die reiz- und indikatororientierten Methoden „*The Stimulus - Centered or Judgement Approach*“. Dabei geht es zunächst um die Bestimmung des *Ausprägungsgrades* oder der *relativen Stärke* von Reizen oder Indikatoren, welche erst später zur Messung von Personeneigenschaften herangezogen werden können. Im Falle von Fragebogenverfahren soll dabei die relative *Schwierigkeit* der einzelnen Items in Bezug auf eine Merkmalsdimension bestimmt werden. Als Ergebnis der Anwendung solcher Methoden resultieren Skalenwerte für die Indikatoren, also die Reize beziehungsweise die Items. Als ein Beispiel seien hier die Verfahren der Skalenbildung nach Thurstone und Chave (1929) genannt, welche wiederum auf den von Fechner (1860a, 1860b) entwickelten psychophysischen Verfahren basieren.

Drittens, reaktionsorientierte Verfahren „*the response approach*“ bei denen sowohl den Personen als auch den Indikatoren oder Reizen, beziehungsweise bei Fragebogen den Items, Skalenwerte (gleichzeitig) zugeordnet werden. Diesen Skalierungsverfahren liegt, wie auch den anderen beiden Herangehensweisen, zunächst die Annahme einer eindimensionalen, kontinuierlichen Eigenschafts-

oder Einstellungsdimension zugrunde, auf der sich sowohl die Personen hinsichtlich ihrer Ausprägung, als auch die Items als Indikatoren hinsichtlich ihrer *Schwierigkeit* in dieser Dimension, welche sie repräsentieren, anordnen lassen. Im Zusammenhang mit psychometrischen Antwortmodellen der *Item-Response-Theory* (IRT), welche in den folgenden Abschnitten noch eingehend dargestellt werden, werden solche Eigenschafts- oder Einstellungsdimensionen auch als *latent traits* bzw. *latente Variablen* bezeichnet (vgl. Borsboom, 2008; Bortz & Döring, 2006, S. 206). In Abgrenzung zu den beiden anderen Verfahren nach dieser Klassifikation von Torgerson (1967), implizieren diese reaktionsorientierten Verfahren explizit eine bestimmte, empirisch an den Daten überprüfbare, Modellformulierung zum Antwortprozess. Als Beispiel für solche Skalierungsverfahren und Modellformulierungen sind hier im Vorgriff auf Kapitel 4 *Psychometrische Modellierung*, zum Beispiel das Guttman-Modell (Guttman, 1950), das Rasch-Modell (Rasch, 1960) mit seinen Erweiterungen (Masters, 1982; Muraki, 1992) und auch das erstmals unter dem Begriff *Unfolding* konzeptionell vorgestellte Modell von Coombs (1950, 1967), sowie die Modelle von Andrich (1988), Andrich (1989) und Hoijtink (1990), sowie von Verhelst und Verstralen (1993), Andrich und Luo (1993), Andrich (1996) sowie Roberts und Laughlin (1996) und Roberts, Donoghue und Laughlin (2000) zu nennen.

In den folgenden Abschnitten sollen nun zwei wichtige Vertreter der klassischen Skalierungsverfahren, welche für das Verständnis der vorliegenden Arbeit relevant sind, näher dargestellt werden. Die beiden Skalierungsverfahren werden zunächst beschreibend, aus der Perspektive einer praktischen Durchführung dargestellt, ohne dabei die diesen Verfahren zumindest bereits implizit gegebenen *psychometrischen* (Antwort-)Modellannahmen zu vertiefen. Eine genauere Darstellung dieser Antwortmodelle und ihrer Annahmen folgt dann, jeweils mit Rückbezug zum entsprechenden Skalierungsverfahren, bei der Diskussion der damit verbundenen psychometrischen Modelle zum Antwortverhalten im Kapitel 4 *Psychometrische Modellierung*.

1.3.1 Skalierung nach Likert

Wie bereits weiter oben dargestellt wird in der psychometrischen Literatur das Skalenniveau für polytome Antwortskalen kontrovers diskutiert (z. B. Ca-

rifio & Perla, 2007; Jamieson, 2004). Ungeachtet dieser Debatte, auf die später im Zusammenhang mit den verschiedenen *psychometrischen Antwortmodellen* und den davon abweichenden idiosynkratischen Antwortmustern noch detaillierter eingegangen wird, wird bei der Anwendung solcher Antwortskalen in Fragebogen in der Praxis meist ein relativ einfaches Auswertungsschema zugrunde gelegt. So sieht die Auswertung solcher psychometrischen Skalen, mit polytomen (aber auch dichotomen) Ratingskalen, im einfachsten Falle – insbesondere bei der Diagnostik im Einzelfall – meist vor, die Summe oder auch den Mittelwert der zugeordneten numerischen Werte der jeweils gewählten Antwortkategorien über alle Items einer Dimension des betreffenden Inventars zu bilden (z. B. Nunnally, 1978). Dabei wird vorausgesetzt, dass zwischen der Summation der einzelnen Itemscores und der latenten Eigenschaft der befragten Personen eine monoton, lineare Beziehung besteht – „*It assumes only that individual items are monotonically related to underlying traits and that a summation of item scores is approximately linearly related to the trait.*“ (Nunnally, 1978, S. 531). Dieses Vorgehen stützt sich auf das von Rensis Likert begründete Konzept zur Erfassung psychologischer Eigenschaften oder individueller Einstellungen (Likert, 1932; Likert, Roslow & Murphy, 1934). Likert, suchte nach einer Methode, mit der individuelle Einstellungen auf wissenschaftliche Weise erfasst werden können, um die daraus resultierenden Messwerte im Sinne einer metrischen Skala, d. h. eines (eigentlich) intervallskalierten Messwertes, zu interpretieren. Die von Likert (1932) vorgeschlagene Methode besteht zunächst darin, die befragten Personen eine Reihe von Items auf einer fünfstufigen Ratingskala einschätzen zu lassen. Wie Borg und Staufenbiel (2007, S. 21) betonen, handelt es sich bei „Likert-ltems“ um Fragen deren Frageteile als allgemeine Aussagen formuliert sind. Anstatt also zum Beispiel direkt zu fragen „*Wie sehr mögen Sie Erdbeereis?*“ und zur Beantwortung der Frage eine unipolare Antwortskala (z. B. von 0 bis 10) vorzugeben, würde das entsprechende Likert-Item als allgemeine Aussage formuliert werden – in diesem Beispiel also „*Ich esse ganz gerne Erdbeereis*“. Die von Likert dann zur Beantwortung vorgeschlagene Antwortskala reicht dabei von „*starke Ablehnung*“ bis „*starke Zustimmung*“ und enthält eine neutrale Mittelkategorie. Über die vorgegebene Antwortskala wird in diesem Beispiel, im Likert'schen Sinne, also das Ausmaß einer Zustimmung zu der *allgemeinen Aussage* als solcher und nicht etwa

direkt das Ausmaß der Präferenz für Erdbeereis, erfasst. Die Quantifizierung dieser (latenten) Präferenz ergibt sich erst *indirekt* aus der anschließenden Verrechnung mehrerer „ähnlicher“ Items im Rahmen der hier durch Summierung erfolgenden Indexbildung.

Der hier am Beispiel dargestellte Unterschied in den „Formulierungen“ mag zwar auf den ersten Blick kaum beachtenswert erscheinen, allerdings weisen Borg und Staufenbiel (2007) darauf hin, dass eine befragte Person die allgemein formulierte (Likert) Aussage – in diesem Beispiel „*Ich esse ganz gerne Erdbeereis*“ – letztlich aus zwei Gründen ablehnen könnte, was „*... zu einer unerwünschten Mehrdeutigkeit der Antworten führen*“ (Borg & Staufenbiel, 2007, S. 21) kann. Eine Ablehnung der Aussage könnte entweder damit begründet werden, dass die befragte Person Erdbeereis „nicht gerne isst“, oder aber damit, dass die befragte Person „ganz ausschließlich *nur* Erdbeereis bevorzugt und isst“ – ihre Ausprägung auf dem Merkmal „Vorliebe für Erdbeereis“ also über der von ihr möglicherweise als „mittelmäßig“ eingestuften Aussage „*Ich esse ganz gerne Erdbeereis*“ liegt. Der entscheidende Punkt, welcher anhand dieses einfachen Beispiels verdeutlicht werden soll, liegt also darin, dass eine als Likert-Item allgemein formulierte Aussage letztlich von „zwei Seiten“ her abgelehnt werden kann. Weil die (latente) Merkmalsausprägung (hier die Präferenz für Erdbeereis) der antwortenden Person entweder *über* oder auch *unter* der durch das jeweilige Item repräsentierten Merkmalsintensität liegt. Ein vielleicht noch deutlicheres Beispiel gibt van Schuur (2011) mit der beispielhaften Formulierung eines hypothetischen Items zur fragebogenbasierten Erfassung der Körpergröße. So könnte die Aussage „*Ich bin ungefähr 1,7 Meter groß*“ letztlich aus zwei möglichen Gründen abgelehnt werden – entweder weil die antwortende Person eher 1,8 Meter groß ist oder weil sie eher 1,6 Meter groß ist (van Schuur, 2011, S. 2-3). Liegt ein solcher Fall einer prinzipiell aus zwei Gründen abzulehnenden Aussage vor, so widerspricht dies natürlich der eigentlich intendierten Methodik der *summativen Verrechnung* der einzelnen Itemantworten bei der Auswertung. Diese praktischen Überlegungen und beispielhaften Betrachtungen zum Antwortprozess sollen an dieser Stelle bereits kritisch angemerkt und zunächst nur als exemplarisches Beispiel erwähnt sein, um zu verdeutlichen, dass sich aufgrund von vermeintlich eindeutigen Formulierungen und Fragestellungen oftmals ganz überraschende, den Erwartungen

widersprechende, Antwortreaktionen ergeben können. Liegt nämlich der Beantwortung eines Likert-Items (bei manchen Personen) tatsächlich ein solcher, wie im diesem Beispiel konstruierter, Antwortprozess vor, so würde dies dem im folgenden dargestellten Auswertungsverfahren nach Likert widersprechen. Im weiteren Verlauf der Darstellung verschiedener Antwortmodelle wird sich zeigen, dass solche unterschiedlichen Antwortreaktionen Äquivalenzen zu den verschiedenen, alternativen psychometrischen Modellen der Skalierung aufweisen (vgl. Kapitel 4 *Psychometrische Modellierung*).

Zur weiteren Auswertung einer psychometrischen Skala nach Likert (1932) und Likert et al. (1934), werden den gewählten Kategorien der Antwortskalen der einzelnen Items nach einer bestimmten, meist inhaltlich begründeten, Zuordnungsregel ganzzahlige, aufsteigende, numerische Werte zugeordnet, was als *Scoring* bezeichnet wird (vgl. Bechger et al., 2003, S. 320). Diese Zuordnungsregeln leiten sich in den meisten Fällen von den die Skalenpunkte der Antwortskalen überschreibenden Wortmarken ab, wie z. B. „*starke Ablehnung*“, „*Ablehnung*“, „*neutral*“, und so fort. Zur systematischen Untersuchung und Entwicklung solcher Regeln zur Zuordnung von einzelnen Wortmarken und Punkten der Antwortskalen sei hier für den deutschsprachigen Raum z.B. auf die Untersuchung von Rohrmann (1978) verwiesen. Etwas neuere Untersuchungen beispielsweise von Wright, Gaskell und O'Muircheartaigh (1994) zeigen, dass sich die befragten Personen durchaus hinsichtlich ihrer Einschätzung solcher unscharf quantifizierenden Wortmarken [*vague quantifiers*] systematisch unterscheiden, was in Folge negative Einflüsse auf die objektive Vergleichbarkeit der Messung haben kann.

Likert ging weiter davon aus, dass die einzelnen Items entweder eindeutig positiv oder negativ formulierte Aussagen über den zu messenden Sachverhalt darstellen. Je nach Richtung oder *Polarität* der Formulierung (positiv / negativ) muss daher beim *Scoring* nach Likert nur darauf geachtet werden, die Zuordnung der numerischen Werte zu den vorgegebenen Antwortkategorien entsprechend in Richtung des zu erfassenden Merkmals vorzunehmen. Also bei negativ formulierten Items entsprechend umgekehrt. Unter diesen Voraussetzungen ergibt sich der (vorläufige) Skalenwert einer Person als Summe der Zahlenwerte der gewählten Antwortkategorien über alle Items, weshalb dieses Verfahren nach Spector (1992) auch als *summierte Ratingskalierung* bezeich-

net wird (vgl. auch Borg & Staufenbiel, 2007). Dieser Art der Bildung von Summenwerten liegen einige axiomatische Annahmen zugrunde, welche eng verbunden sind mit denen der Klassischen Testtheorie (KTT), welcher daher ein eigener kurzer Abschnitt weiter unten gewidmet ist. Zur Analyse und Selektion der zunächst vorläufigen Auswahl von eingesetzten Items im Rahmen der summierten Ratingskalierung wird im Wesentlichen auf zwei Itemkennwerte zurückgegriffen. Diese Kennwerte lassen sich direkt aus den empirischen Daten bestimmen. Dies sind zum einen die *psychometrische Schwierigkeit* und zum anderen die *Trennschärfe* des jeweiligen Items.

Die *psychometrische Schwierigkeit* eines Items ist hier als Wahrscheinlichkeit der Zustimmung zu einem Item auf seiner mehrstufig polytomen, oder auch zweistufig dichotomen Antwortskala, definiert. Die psychometrische Schwierigkeit p_i eines Items mit mindestens zweistufiger Antwortskala bestimmt sich dabei als Quotient aus der Summe der beim Scoring zugeordneten numerischen (Kategorie-)Werte x_{iv} für das Item i über alle antwortenden Personen v und dem Produkt aus dem numerischen Wert der höchsten Kategorie $max(m_i)$ und der Anzahl der Personen n (vgl. Gleichung 1.1).

$$p_i = \frac{\sum_{v=1}^{n} x_{iv}}{max(m_i) \cdot n} \tag{1.1}$$

Zu beachten ist dabei, dass die den Antwortkategorien der Ratingskala des jeweiligen Items zugeordneten numerischen Werte für die unterste Kategorie bei dem Wert $m = 0$ beginnen und, beispielsweise bei einer Likert–Antwortskala mit fünf Kategorien, für die höchste Kategorie bei dem Wert $m = 4$ enden (vgl. Bortz & Döring, 2006, S. 219). Der so gebildete Schwierigkeitsindex p_i variiert innerhalb eines Wertebereiches von $p_{i_{min}} = 0$ bis $p_{i_{max}} = 1$. Umgangssprachlich „leichte“ Items, denen viele Personen zustimmen, weisen daher einen hohen *psychometrischen Schwierigkeitsindex* p_i auf. Für die endgültige psychometrische Skala werden nach dem Schwierigkeitskriterium die Items möglichst so ausgewählt, dass sie insgesamt einen breiten Bereich der psychometrischen Schwierigkeit abdecken (vgl. Bortz & Döring, 2006, S. 222).

Die *Trennschärfe* eines Items soll Auskunft darüber geben, wie gut das Item zwischen Personen mit hoher und niedriger Merkmalsausprägung diskriminiert. Zur Bestimmung der Trennschärfe eines Items hatte Likert ursprüng-

lich die Analyse der gewählten Antwortkategorien bzw. der Item-Mittelwerte der, nach dem Summenwert der Gesamtskala gebildeten Extremgruppen (hohe vs. niedrige Merkmalsausprägung), vorgeschlagen (Likert, 1932, S. 51). Items deren Mittelwerte sich für die beiden Gruppen kaum oder nicht unterscheiden, weisen demnach eine geringe Diskriminationsfähigkeit hinsichtlich der Merkmalsausprägung auf. Items mit solchen Eigenschaften tragen also im Sinne einer Differenzierung wenig Information zur Messung bei. In aktuellen Anwendungen dieser Itemanalyse nach dem Kriterium der Trennschärfe werden allerdings Trennschärfeindizes verwendet, welche auf der bivariaten Korrelation des jeweiligen Items mit dem Summenwert der gesamten Skala über die Personen (t) basieren. Diese Trennschärfeindizes weisen damit einen (theoretischen) Wertebereich zwischen $r_{it} = -1$ und $r_{it} = 1$ auf – Bortz und Döring (vgl. 2006, S. 219) für eine ausführliche Darstellung. Für die endgültige psychometrische Skala werden nach dem Kriterium Trennschärfe nun diejenigen Items ausgewählt deren Trennschärfen in einem mittleren positiven Bereich zwischen $r_{it} = 0$ und $r_{it} = 1$ liegen.

1.3.2 Die Klassische Testtheorie

Die Klassische Testtheorie (KTT) stellt im Vergleich zu den praktischen Vorgaben Likerts zur Konstruktion von psychometrischen Skalen insofern eine Verallgemeinerung dar, als dass darin theoretische Annahmen über die Zusammensetzung der (summierten) Testwerte aufgestellt werden (vgl. Borg & Staufenbiel, 2007, S. 313), und so ein Messfehlermodell für die Testwerte etabliert (z. B. Bühner, 2006, S. 24, ff.). Die insgesamt fünf Axiome der KTT (oder auch drei Axiome und zwei Zusatzannahmen vgl. Moosbrugger, 2012, S. 104), welche hier nicht in aller Einzelheit diskutiert werden sollen, beziehen sich dabei, außer dem ersten Existenzaxiom, letztlich alle auf Aussagen zum Messfehler. Eine zentrale Annahme hinsichtlich des beobachteten Testwertes X einer Person besteht darin, dass sich dieser zusammensetzt aus einem wahren Wert T (für **T**rue score) und einem Fehleranteil E (für **E**rror), siehe Gleichung 1.2.

$$X = T + E \tag{1.2}$$

Der beobachtete Messwert X ist dabei als eine diskrete Zufallsvariable definiert, welche die jeweilige Personenantwort repräsentiert (Bechger et al., 2003). Die Funktion von X wird durch die beim Scoring angewendete Zuordnungsregel definiert. Der wahre Wert T ist definiert als Erwartungswert der Verteilung von X gegeben die latente, nicht beobachtbare Merkmalsausprägung θ einer zu messenden Person. Formal also als $T = \mathbb{E}(X|\theta)$. Dies ist unter Annahme der Normalverteilung der Fehler mit dem Mittelwert null und der Varianz $Var(X|\theta)$, derjenige Testwert, welcher sich als Mittelwert aller, durch theoretisch unendliches Testen einer Person mit demselben Messinstrument, beobachteten Messwerte ergibt. Oder aber praktisch, der Mittelwert aller beobachteten Messwerte von Personen mit derselben Merkmalsausprägung, wobei die gleichen Annahmen wie oben für die Verteilung der Fehler gelten. Für eine detaillierte Darstellung der einzelnen Axiome der KTT sei hier auf die entsprechenden einführenden Lehrbücher verwiesen (z. B. Borg & Staufenbiel, 2007; Bortz & Döring, 2006; Bühner, 2011; Fischer, 1974; Moosbrugger & Kelava, 2012; Steyer & Eid, 2000). Diese auch als Verknüpfungsaxiom der KTT bezeichnete formale Beschreibung der Zusammenhänge zwischen der beobachteten Größe Testwert X und der angenommenen, nicht beobachtbaren Größen wahrer Wert T und des Fehleranteils E, sowie die KTT im Allgemeinen sind nicht ohne Kritik geblieben. So wird zum Beispiel das Existenzaxiom, also die Annahme eines wahren Wertes X einer Person insofern kritisiert, als dass dieser lediglich eine theoretische Annahme sei, welcher in der Praxis, außer in seiner Definition als Erwartungswert der Messung, eher irreführend sei (Steyer & Eid, 2000). Ein weiterer Kritikpunkt an der KTT, welcher regelmäßig vorgebracht wird, bezieht sich darauf, dass die KTT im eigentlichen Sinne eine reine *Messfehlertheorie* sei (z. B. Bühner, 2006) welche, wie Fischer (1974, S. 124) schreibt „*... am eigentlichen Problem des Messens von Testleistungen [...] vorübergeht.*“ und weiter: „*... sie verabsäumt es, das* ***Zustandekommen*** *der Testleistung [...] zum Gegenstand der Betrachtung zu machen*“ (Fischer, 1974, S. 124). Diese fundamentale Kritik fußt auf der Tatsache und Überlegung, dass die einzelnen Zuordnungs- und Rechenschritte nach Bechger et al. (2003, S. 320), im Rahmen der Fragebogenauswertung nach der KTT ein Intervallskalenniveau voraussetzen (siehe dazu auch Rost, 1999), was allerdings meist nur angenommen wird und nicht empirisch überprüft wird. Darüber hin-

aus besteht für die Trennschärfe- und Schwierigkeitsindizes zur Itemanalyse die Problematik der Stichprobenabhängigkeit. Je nachdem wie die interessierende Merkmalsausprägung in der den Analysen zugrunde liegenden Personenstichprobe verteilt ist, können sich durchaus unterschiedliche Befunde zur Itemcharakteristik ergeben (Moosbrugger, 2012). Ein großer Vorteil der Klassischen Testtheorie mag in ihrer einfachen Anwendbarkeit begründet sein, was auch ein Grund dafür sein dürfte, warum nach Rost (1999) „*... 95% der Testentwicklungen nach der klassischen Testtheorie erfolgen*“ (Rost, 1999, S. 140). Die insbesondere von Fischer (1974), vorgebrachte Kritik an der KTT wurde Anfang der 1960er Jahre mit der Entwicklung des logistischen Testmodells von Rasch (1960) im Rahmen der Item-Response-Theory (IRT) aufgegriffen. Diese, inzwischen deutlich erweiterte, Modellfamilie liefert ein psychologisch plausibles Erklärungsmodell für das Zustandekommen der empirisch beobachtbaren Personenantworten auf der Grundlage von mindestens zwei Modellparametern – dem *Personenparameter*, also der (erst noch zu messenden) Ausprägung der Person auf der jeweiligen Eigenschaftsdimension und analog dazu dem *Itemparameter*. Auf diese, hier nur kurz im Zusammenhang mit der Darstellung und Kritik an der KTT erwähnten Modelle, wird weiter unten noch detaillierter eingegangen werden. Trotz aller Kritik an der KTT muss abschließend festgestellt werden, dass diese, wenn auch als reine „Messfehlertheorie“, gerade als solche, wertvolle Zusammenhänge zwischen dem resultierenden Messwert und seinem dazugehörigen Messfehler formalisiert. Insofern kann die von Fischer (1974) geäußerte Kritik an der KTT eher in deren „fälschlicher“ Interpretation als psychologisches Modell für den Antwortprozess verstanden werden. Bezieht man die hier bisher nur kurz angesprochenen Testmodelle der IRT in eine abschließende Betrachtung der Methodik bei der Auswertung von Likert Fragebogendaten mit ein, kann man feststellen, dass die Testmodelle aus dem Bereich der IRT überprüfbare Annahmen zum Zustandekommen der Messwerte machen und die KTT dann ergänzend eine Theorie zur Zusammensetzung dieser Testwerte (aus Messwert + Fehler) liefert. In diesem Sinne vertritt zum Beispiel Rost (1999) die Ansicht, dass es sich bei der KTT und der IRT und deren logistischen Testmodellen nicht um konkurrierende, sondern um sich ergänzende, komplementäre Theorien handelt (vgl. dazu auch Kubinger, 2000). Die Verbindung zwischen KTT und IRT wird auch bei Bechger et al. (2003) aus theoretischer und praktischer Perspektive diskutiert.

1.3.3 Skalierung nach Thurstone und Fechner

Im Gegensatz zu der von Likert vorgeschlagenen Methode der „*summierten Ratingskalierung*“ verfolgt die von Thurstone und Chave (1929), historisch gesehen, bereits früher vorgeschlagene Methode zur Messung von individuellen Einstellungen ein anderes Prinzip. Im Gegensatz zu Likerts Methode geht es dabei im ersten Schritt zunächst darum, für die zur späteren Messung eingesetzten Items (Stimuli) den Ausprägungsgrad beziehungsweise deren *relative Schwierigkeit* in Bezug auf die zu messende Eigenschaft zu bestimmen. Torgerson (1967) klassifiziert dieses Vorgehen daher als den reiz- und indikatororientierten Methoden „*The Stimulus-Centered or Judgement Approach*“ (Torgerson, 1967, S. 46) zugehörig. Zur Bestimmung dieser relativen Schwierigkeit der Items haben Thurstone und Chave (1929) wiederum verschiedene Vorgehensweisen vorgeschlagen, welche sich alle auf Expertenurteile stützen. Bei der *Methode der gleich erscheinenden Intervalle* [„*method of equal-appearing intervals*“] wird zunächst eine Reihe von Items gesammelt und diese den Experten zu Beurteilung vorgelegt. Die Aufgabe der Experten besteht dann darin, die Items direkt auf einer elfstufigen Skala anzuordnen, sodass einerseits der Wertebereich der Skala komplett abgedeckt wird und andererseits die Abstände zwischen den elf Skalenpunkten gleich erscheinen – „*... the intervals between successive piles should be apparently equal shifts of opinion as judged by the subject.*“ (Thurstone & Chave, 1929, S. 30). Der entscheidende Unterschied zu Likerts Methode besteht darin, dass die Urteile der Experten hier nicht deren persönliche Einstellung zu der zu messenden Eigenschaft darstellen, sondern lediglich eine mehr oder weniger objektive Einschätzung der relativen Schwierigkeit der Items repräsentieren (vgl. Borg & Staufenbiel, 2007, S.309). Zur Überprüfung der Qualität der Expertenurteile muss im Anschluss das Ausmaß der Übereinstimmung bei der Bewertung der Items bestimmt werden. Berechnet man zum Beispiel die Streuung der Expertenurteile für jedes einzelne Item, so sollte diese im Idealfall sehr gering ausfallen. Als Maß für die relative Schwierigkeit des jeweiligen Items wird dann der Mittelwert oder Median der Expertenurteile zu diesem Item als Skalenwert verwendet. Den eigentlich zu testenden Personen werden dann für die Endform des Tests ausgewählte Items vorgelegt, mit der Bitte nur diejenigen Items anzukreuzen, denen sie zustimmen. Der Messwert für die zu erfassende Eigenschaft einer so getesteten

Person ergibt sich dann als Median oder Mittelwert aus den zuvor bestimmten Skalenwerten derjenigen Items, denen die Person zugestimmt hat (vgl. Borg & Staufenbiel, 2007, S.313). Als Kritik an diesem Vorgehen kann angeführt werden, dass es sich bei der aus den Expertenurteilen resultierenden elfstufigen Skala im strengen Sinne nur um eine Skala mit mit ordinalem Skalenniveau handelt. Die Anweisung an die Experten die Items den elf Skalenpunkten mit „gleich erscheinenden“ Abständen einzuteilen impliziert zwar in gewisser Hinsicht ein Intervallskalenniveau – allerdings wird dieses Skalenniveau dabei auch nur angenommen und nicht überprüft oder durch das methodische Vorgehen gesichert. Zur Bestimmung der relativen Itemschwierigkeiten hat Thurstone (1927b) daher eine weitere Methode vorgeschlagen, welche in interpretierbaren Abständen der relativen Itemschwierigkeiten im Sinne einer Intervallskala resultieren. Diese Methode hat ihren Ursprung in den Anfängen der Allgemeinen und Experimentellen Psychologie von Gustaf Theodor Fechner (Fechner, 1860a, 1860b). Fechner befasste sich mit der Erforschung der Zusammenhänge zwischen menschlichem, mentalem Erleben und den quantitativ messbaren physikalischen Größen, welche diese Wahrnehmungen über die verschiedenen Sinnesmodalitäten auslösen. Fechner interessierte sich bei seinen Forschungen zur Reizunterscheidung für die Schwelle der Unterschiedsempfindlichkeit beim paarweisen Vergleich von zwei Reizen (einer Wahrnehmungsmodalität) mit unterschiedlicher Intensität. Fechner (1860a) definierte dazu drei Methoden, von denen die „*Methode der richtigen und falschen Fälle*“ (Fechner, 1860a, S. 72) hier insofern von Bedeutung ist, als dass sich aus ihr eine probabilistisch, monotone Beziehung zwischen der Wahrscheinlichkeit zur Wahrnehmung eines Reizunterschiedes und der Größe des *psychologischen Abstandes* der beiden Reize ableiten lässt (Fechner, 1860a). Formal nimmt danach die Wahrscheinlichkeit zur Wahrnehmung, dass Reiz i stärker ist als Reiz j, $p(i \succ j)$, mit steigender (physikalischer) wahrer Differenz der beiden reizauslösenden Stimuli s_i und s_j, also $s_i - s_j$ zu. Nähert sich dabei die (wahre) Differenz der beiden reizauslösenden Stimuli dem Wert $s_i - s_j \to 0$ so nähert sich die Wahrscheinlichkeit zur Wahrnehmung eines Reizunterschiedes dem Wert $p(i \succ j) \to 0,5$.

Diese Dominanzwahrscheinlichkeiten $p(i \succ j)$ von jeweils zwei Reizen unterschiedlicher Intensität, lassen sich empirisch über einen vollständigen Paarvergleich aller Reize bestimmen. Diese von Fechner entwickelte Methode des

paarweisen Stimulusvergleichs übertrug Thurstone auf das Problem der Bestimmung der relativen Schwierigkeiten von Fragebogen-Items (Thurstone, 1927b, 1927c, 1929). Bei diesem Vorgehen werden zur Bestimmung der relativen Itemschwierigkeiten die einzelnen Items den urteilenden Personen jeweils in allen möglichen Paarungen zum Vergleich vorgelegt. Für einen vollständigen Paarvergleich müssen bei k Items $\binom{k}{2}$ Vergleiche von jeder urteilenden Person durchgeführt werden. Die resultierenden absoluten Dominanzhäufigkeiten lassen sich für k Stimuli oder Items in einer symmetrischen $k \times k$ Dominanzmatrix D anordnen, wie sie in Tabelle 1.1 (oben) am Beispiel für vier Stimuli, welche in einem vollständigen Paarvergleich von 26 Personen bewertet wurden, dargestellt sind (Beispiel entnommen aus: Borg & Staufenbiel, 2007, S. 100).

Tabelle 1.1 Matrix Anordnungen für $k = 4$ Stimuli aus vollständigem Paarvergleich.

Transformation	Matrixanordnung
Dominanzhäufigkeiten	$D_{f_{i\succ j;j\succ i}} = \begin{matrix} . & 19 & 23 & 20 \\ 7 & . & 18 & 15 \\ 3 & 8 & . & 12 \\ 6 & 11 & 14 & . \end{matrix}$
Wahrscheinlichkeiten	$P_{p_{i\succ j;j\succ i}} = \begin{matrix} .50 & .73 & .88 & .77 \\ .27 & .50 & .69 & .58 \\ .12 & .32 & .50 & .48 \\ .23 & .42 & .54 & .50 \end{matrix}$
z-Werte	$Z_{p_{i\succ j;j\succ i}} = \begin{matrix} 0 & +0.62 & +1.20 & +0.74 \\ -0.62 & 0 & +0.50 & +0.19 \\ -1.20 & -0.50 & 0 & -0.10 \\ -0.74 & -0.19 & +0.10 & 0 \end{matrix}$

Anmerkungen: Matrix Anordnungen der Dominanzhäufigkeiten (oben), Wahrscheinlichkeiten (Mitte) und z-Werte (unten) für $k = 4$ Stimuli aus vollständigem Paarvergleich; $n = 26$ beurteilende Personen; (Beispiel entnommen aus: Borg & Staufenbiel, 2007, S. 100).

Die in Tabelle 1.1 (oben) in Matrixanordnung dargestellten Werte oberhalb der Hauptdiagonalen geben die Dominanzhäufigkeiten $i \succ j$ und die Werte

unterhalb der Hauptdiagonalen die Dominanzhäufigkeiten $j \succ i$ wieder, wie sie von den 26 Personen bewertet wurden (i entspricht den Zeilen und j entspricht den Spalten). Die entsprechenden Wahrscheinlichkeiten in Tabelle 1.1 (Mitte) ergeben sich durch Division der absoluten Häufigkeiten durch die Anzahl der bewertenden Personen. Thurstone (1927a) konnte nun zeigen, dass sich unter bestimmten Annahmen den Wahrscheinlichkeiten zu einem Dominanzurteil jeweils z-Werte der Normalverteilung zuordnen lassen, was er im *Law of comparative judgement (case five)* formulierte. Die z-transformierten Skalenwerte der vier Stimuli aus obigem Beispiel ergeben sich nach Thurstone (1927a) als Zeilenmittelwerte der Matrix $Z_{p_{i\succ j;j\succ i}}$ (vgl. Tabelle 1.1, unten), welche zur Vermeidung von negativen Werten so verschoben werden können, dass der kleinste z-Wert dem transformierten Skalenwert 0 entspricht (vgl. Borg & Staufenbiel, 2007, S. 99).

Diese, unter dem Begriff *LCJ-Skalierung (**L**aw of **C**omparative **J**udgement)* bekannt gewordene Methode der Skalierung zielt, wie bereits weiter oben erwähnt, darauf ab zunächst die relative Schwierigkeit der Items zu bestimmen, anhand derer dann die eigentliche Messung der zu testenden Personen vorgenommen werden kann. Thurstone (1928) schlug zur praktischen Durchführung der Einstellungsmessung vor, den zu testenden Personen eine Auswahl von ungefähr 25, bereits skalierten Items vorzulegen, mit der Aufforderung diejenigen Items denen sie zustimmen mit einem „+“ Zeichen und diejenigen die sie ablehnen mit einem „−“ Zeichen zu versehen. Der Messwert einer Person ergibt sich dann als mittlerer Skalenwert derjenigen Items denen die Person zugestimmt hat – „*The score for each person is the average scale value of all the statements that he has indorsed*“ (Thurstone, 1928, S. 553).

1.4 Zusammenfassung zur Skalen- und Indexbildung

Betrachtet man die beiden in den vorangegangenen Abschnitten vorgestellten Prinzipien zur Skalierung vergleichend, so lässt sich zusammenfassen, dass sich die von Likert (1932); Likert et al. (1934) und Thurstone (1928); Thurstone und Chave (1929) jeweils vorgeschlagenen Ansätze zur Skalierung bzw. Einstellungsmessung, hinsichtlich ihrer konzeptionellen Annahmen zum datengenerierenden Anwortprozess bzw. Antwortmodell grundlegend unterscheiden (vgl. auch Andrich, 1996).

Bei Likerts Prinzip der summierten Rating-Skalierung wird (implizit) ein *kumulatives Antwortmodell* bei der Beantwortung der einzelnen Items postuliert, welches eine *Dominanz-Relation* zwischen der Ausprägung der Person und der relativen Schwierigkeit der Items auf der zu messenden Eigenschaftsdimension beschreibt. Der Messwert einer Person auf der zu erfassenden Merkmalsdimension ergibt sich als Summe aus den beim Scoring den Antwortkategorien zugeordneten Zahlenwerten (vgl. Nunnally, 1978, S. 531). Je mehr Items einer Skala eine Person (stark) zustimmt (oder bei Leistungstests diese richtig löst), desto höher fällt ihr Messwert auf der zu erfassenden Merkmalsdimension aus.

Bei der Skalierung nach Thurstone wird explizit ein *Präferenz*-Antwortmodell postuliert, welches eine *Nähe–Distanz-Relation* zwischen der Ausprägung der Person und der relativen Schwierigkeit der Items auf der zu messenden Eigenschaftsdimension annimmt. Der Messwert einer Person auf der zu erfassenden Merkmalsdimension ergibt sich unmittelbar aus den zuvor bestimmten Schwierigkeiten derjenigen Items, welchen die Person zugestimmt hat. Die von Thurstone (1928) vorgeschlagene Methode zur Einstellungsmessung, bei der bereits durch Experten skalierte Items eingesetzt werden, setzt dabei, zumindest implizit, eine solche *Nähe–Distanz-Relation* voraus. Dabei repräsentieren die Items einzelne Messpunkte auf dem latenten Einstellungskontinuum und die Personen stimmen nur denjenigen Items zu, die auf diesem Kontinuum nahe bei ihrer eigenen Ausprägung auf diesem Kontinuum liegen. Je besser das von einem oder mehreren Items repräsentierte Ausmaß der Merkmalsausprägung mit der Ausprägung der zu testenden Person übereinstimmt, desto höher wird die Zustimmungswahrscheinlichkeit der Person zu diesem Item ausfallen.

Das Item mit der höchsten Zustimmungswahrscheinlichkeit repräsentiert somit den Punkt auf dem (latenten) Einstellungskontinuum, welcher die Merkmalsausprägung einer Person am besten, oder auch *ideal* beschreibt. Da die Personen mit ihrer jeweiligen Eigenschaftsausprägung sozusagen an ihrem jeweiligen *Idealpunkt* auf der gemeinsamen Skala der latenten Einstellungsdimension verortet sind und nur denjenigen Items zustimmen, welche in der Nähe dieses Idealpunktes liegen, werden die diesen Antwortprozess modellierenden psychometrischen Modelle auch *Idealpunktmodelle* genannt (Brady, 1985, 1989, 1990; Gediga, 1998). Wandert man entlang des latenten Kontinuums der Einstellungsdimension von der niedrigsten bis zur höchsten Ausprägung, so steigt bei diesen Antwortmodellen die Wahrscheinlichkeit zur Zustimmung zu einem Item mit zunehmender Eigenschaftsausprägung bis zu einem gewissen Punkt (dem Idealpunkt der jeweiligen Person) zunächst an, um dann jenseits dieses Punktes wieder zu sinken (vgl. auch Abbildung 1.1 oben). Im Gegensatz dazu wird bei der von Likert (1932); Likert et al. (1934) vorgeschlagenen Methode der summierten Rating-Skalierung, wie bereits erwähnt, implizit ein kumulatives Antwortmodell vorausgesetzt. Beim kumulativen Antwortmodell wird (zumindest implizit) die Annahme getroffen, dass es sich bei dem vorliegenden Antwortprozess um eine *Dominanz-Relation* zwischen dem Grad der (latenten) Eigenschaftsausprägung der Person und den *psychometrischen Schwierigkeiten* der einzelnen Items, in Bezug auf das zu erfassende Merkmal, handelt. Aus dieser Annahme einer Dominanz-Relation folgt, dass Personen, welche einen bestimmten Ausprägungsgrad auf der latenten Eigenschaftsdimension aufweisen, ein typisches monotones Antwortmuster auf den nach deren Schwierigkeiten geordneten Items erzeugen. Die Wahrscheinlichkeit zur Zustimmung zu einem Item steigt danach monoton mit zunehmender Eigenschaftsausprägung einer Person an (vgl. Abbildung 1.1 unten). Denjenigen Items, deren Schwierigkeit zum Beispiel unterhalb der Eigenschaftsausprägung einer bestimmten Person liegen, sollte aufgrund der Dominanz-Relation von dieser Person eher zugestimmt werden, wohingegen alle schwierigeren Items von dieser Person eher abgelehnt würden.

Auch wenn die in diesem Abschnitt beschriebenen, von Thurstone (1927a, 1927b, 1927c, 1928, 1929) vorgeschlagenen Methoden der Itemskalierung und die daran anschließende Messung der Personen zunächst eher mit dem Ideal-

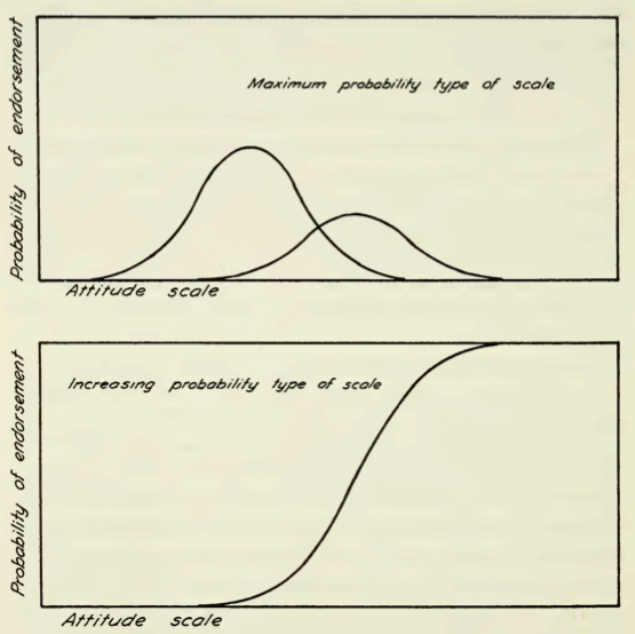


Abbildung 1.1 Darstellung der Kurven zur Antwortwahrscheinlichkeit in Abhängigkeit der latenten Eigenschaftsausprägung; Nähe–Distanz-Antwortprozess (oben) und Dominanz-Antwortprozess (unten); Reproduktion der Abbildung aus Thurstone und Chave (1929, S. 94).

punktantwortmodell assoziiert sind, muss betont werden, dass bereits Thurstone und Chave (1929) beide Kurvenverläufe der Wahrscheinlichkeit zur Zustimmung zu einem Item, in Abhängigkeit der jeweiligen Eigenschaftsausprägung, in Erwägung gezogen haben (vgl. Abbildung 1.1).

Die beiden hier in Kapitel 1 vorgestellten Techniken zur Skalierung korrespondieren mit jeweils (fundamental) unterschiedlichen Antwortprozessen. In diesen beiden Antwortprozessen ist die Beziehung zwischen Personen und Items entweder als *Dominanz* oder als *Nähe–Distanz*-Relation charakterisiert. Dies bildet die Grundlage für eine ganze Reihe unterschiedlicher *psychometrischer Mess- und Antwortmodelle*, welche jeweils mit einer der beiden Skalierungstechniken verbunden sind (Stokman & van Schuur, 1980). Solche psychometrischen Antwortmodelle können dann die Grundlage zur Untersuchung von typischen (vom Modell abweichenden) Antwortmustern bilden. Einige dieser Modelle werden daher in Kapitel 4 *Psychometrische Modellierung* vorgestellt werden.

Zur Verdeutlichung der Bedeutung der beiden in Abbildung 1.1 dargestellten möglichen Kurvenverläufe und den damit jeweils verbundenen unterschiedlichen Antwortmodellen, sei bereits an dieser Stelle auf die sich daraus ergebenden Implikationen im Hinblick auf die Auswertung von Fragebogendaten hingewiesen. Ist beispielsweise für eine Reihe von Items sichergestellt, dass der Prozess zu deren Beantwortung, für alle Personen einer Stichprobe einem kumulativen Antwortmodell folgt, so besteht zwischen den Itemschwierigkeiten und den Merkmalsausprägungen der Personen eine Dominanz-Relation. Diese Dominanz-Relation wiederum stellt dann eine (minimale) notwendige – aber nicht hinreichende – Bedingung für die sinnvolle Verrechnung der Anzahl der zugestimmten oder gelösten Items dar (bei dichotomen Items). Der so gebildete Summenwert kann, unter bestimmten weiteren Bedingungen, zumindest approximativ, als Messwert für die zu erfassende Merkmalsausprägung der einzelnen Personen interpretiert werden, welcher mindestens ein ordinales Skalenniveau aufweist.

Demgegenüber wäre diese minimal notwendige Voraussetzung zur Verrechnung der Itemantworten zu einem Summenwert beim Vorliegen eines Idealpunkt Antwortmodells mit einer Nähe–Distanz-Relation nicht gegeben. Würde dennoch unter diesen (fehlenden) Voraussetzungen ein Summenwert gebildet, so könnte dieser im Extremfall für jede Person einer Stichprobe (bei dichotomen Items) den Wert „1“ aufweisen, obwohl die Personen möglicherweise tatsächlich höchst unterschiedliche Grade der Merkmalsausprägungen aufweisen könnten. Liegt eine derartig Situation für einen empirischen Datensatz vor, so ist offensichtlich, dass in solch einem Fall die Validität der Messung, zumindest lokal für manche Personen oder Personengruppen, deutlich eingeschränkt ist (vgl. Dalal & Carter, 2015a; Roberts, Laughlin & Wedell, 1999).

Im Hinblick auf diese beiden bereits von Thurstone und Chave (1929) postulierten Verläufe der Antwortwahrscheinlichkeiten, muss ferner angemerkt werden, dass sich diese auf der Basis theoretischer Überlegungen als Funktion der Art der Item- beziehungsweise Skalenkonstruktion (vgl. Abschnitt 1.2) ergeben sollten. So sollte sich der eingipflige Verlauf der Antwortwahrscheinlichkeiten immer dann ergeben, wenn es sich bei den einzelnen Items um Aussagen zur Präferenz handelt. Insbesondere auch dann, wenn die einzelnen Items von den Personen direkt in eine individuelle Rangreihe eingestuft werden sollen. Bei

allgemeinen Aussagen, wie sie bei Likert Items anzustreben sind, sollte sich dagegen in Abhängigkeit der Merkmalsausprägung der Personen eher ein monoton steigender Verlauf der Antwortwahrscheinlichkeiten ergeben. Allerdings stützen sich diese hier dargestellten Zusammenhänge letztlich lediglich auf logische Ableitungen auf Basis theoretischer Überlegungen und bedürfen zur Überprüfung letztlich einer empirischen Untersuchung. Denkbar wäre nämlich auch, dass sich diese beiden Verläufe der Antwortwahrscheinlichkeit nicht (nur) in Abhängigkeit der Art der Items bzw. der psychometrischen Skalen ergeben, sondern sich in Abhängigkeit von interindividuell verschiedenen Interpretationen beziehungsweise Rezeptionen der Items bzw. der gesamten Skala ergeben (vgl. hierzu auch das Beispiel „Eis-Präferenz“ und „Körpergröße“ in Abschnitt 1.3.1).

Die in der vorliegenden Arbeit analysierten Skalen lassen im Hinblick auf die beiden unterschiedlichen Antwortmodelle, zumindest theoretisch, verschiedene Interpretationen zu – beziehungsweise bieten verschiedene Möglichkeiten der Rezeption durch unterschiedliche Personengruppen an (vgl. dazu auch z. B. Hardy & Ford, 2014). Wobei es sich bei der Rezeption dieser Skalen als Präferenzskalen nicht notwendigerweise um ein über alle Personen hinweg konsistentes Phänomen handeln muss. Vielmehr ist es nicht unplausibel anzunehmen, dass im Hinblick auf die Rezeption der Items beziehungsweise der gesamten Skala interindividuelle Unterschiede bestehen können. Solche Unterschiede in der Wahrnehmung und der anschließenden Beantwortung der Items manifestieren sich in unterschiedlichen Formen typischer, idiosynkratischer Antwortmuster in den analysierten Antwortdaten. Die Theorie sowie empirische Befunde aus der Literatur zu solchen Antwortmustern werden in Kapitel 3 *Theoretischer Hintergrund zu Antwortmustern* vorgestellt und diskutiert. Daneben kann auch angenommen werden, dass die in der vorliegenden Arbeit eingesetzten Skalen in unterschiedlichem Ausmaß mit den beiden Antwortmodellen assoziiert sind. So ist es beispielsweise nicht unplausibel anzunehmen, dass es sich bei den in Abschnitt 5.1 dargestellten Skalen zu beruflichen Interessen und zur Musikpräferenz eher um Präferenzmodelle handelt, welche daher von einer größeren Anzahl von Personen eher nach einem Nähe–Distanz-Antwortprozess antizipiert werden. Eine derartige Argumentation ließe sich aber, insbesondere auf der Ebene einzelner Items, auch für die hier ebenfalls eingesetzten Skalen zur

Persönlichkeit begründen. Allerdings unterscheiden sich die drei hier untersuchten Konstrukte (*Persönlichkeit*, *berufliche Interessenorientierungen* und *Musikpräferenzen*), trotz ihrer gemeinsamen Verortung im Bereich der Psychologie interindividueller Unterschiede, dennoch im Hinblick auf ihre theoretische Fundierung einerseits und andererseits im Hinblick auf die damit verbundenen methodischen Traditionen bei deren Skalierung. Im folgenden Kapitel 2 *Theorie zu den untersuchten Konstrukten* werden daher diese drei Konstrukte vorgestellt.

Kapitel 2

Theorie zu den untersuchten Konstrukten

Mit dem Einsatz psychodiagnostischer Fragebogeninventare zur Selbstbeschreibung werden oft Forschungsfragen untersucht, welche sich im Fachbereich der *Differentiellen Psychologie* und *Persönlichkeitspsychologie* verorten lassen. Allgemein wird die Begründung der *Differentiellen Psychologie* William Stern (1900, 1911) zugeschrieben, dessen Interesse sich insbesondere auf die Etablierung einer wissenschaftlich vertretbaren Auffassung der menschlichen Person richtete. Wie Lamiell (2006) in seinem Aufsatz zum Werk von William Stern allerdings darlegt, entwickelte sich Stern aber auch zu einem Kritiker der *Differentiellen Psychologie*, indem diese „*immer stärker von quantitativen Messverfahren und statistischen Begriffen geprägt wurde, wobei Individuen als Exemplare von Kategorien betrachtet werden mussten, drohte auch diese Disziplin, Sterns Auffassung nach, aus Personen bloße Sachen zu machen*“ (Lamiell, 2006, S. 253).

Der Forschungsgegenstand der *Differentiellen Psychologie* lässt sich, im Vergleich zur *allgemeinen Psychologie*, durch ihren Fokus auf die interindividuellen Unterschiede von menschlichen Merkmalen definieren. Demgegenüber befasst sich die allgemeine Psychologie eher mit universellen Gesetzmäßigkeiten, welche oft im Bereich psychophysiologischer Funktionen von Menschen allgemein verortet sind. Derartige Ansätze haben zum Beispiel ihren Ursprung in den von Fechner (1860a, 1860b) entwickelten psychophysischen Verfahren zur Erforschung allgemeiner Prinzipien der Wahrnehmung. Bei der Suche nach solchen

allgemeinen Gesetzmäßigkeiten werden dann die interindividuellen Unterschiede eher als Messfehler oder eigentlich unerwünschte Variation angesehen.

Diese beiden unterschiedlichen Perspektiven auf menschliche Eigenschaften lassen sich den beiden von Cronbach (1957) identifizierten historisch, methodischen Strömungen der Psychologie zuordnen. Cronbach (1957) benannte dabei einerseits die *experimentelle Psychologie* und andererseits die *Korrelationspsychologie* (vgl. Renner, Heydasch & Ströhlein, 2012, S. 89). Die wissenschaftliche Anwendung und Auswertung von Fragebogendaten ist innerhalb dieser beiden Hauptströmungen der Psychologie in der Regel mit der *Korrelationspsychologie* verbunden. Die Korrelationspsychologie ist innerhalb der vier von Stern (1911) identifizierten Disziplinen der Differentiellen Psychologie die variablenorientierte Sichtweise auf personenbeschreibende Merkmale wie sie mit Fragebogenverfahren erhoben werden. Die mit diesem Ansatz untersuchten Fragestellungen beziehen sich auf die psychologische Theoriebildung zu Zusammenhängen zwischen unterschiedlichen Konstrukten. Die in der vorliegenden Arbeit behandelten Konstrukte (*Persönlichkeit*, *Präferenzen des Musikgeschmacks* und *berufliche Interessenorientierungen*) und die zu ihrer Operationalisierung eingesetzten Fragebogenverfahren lassen sich, trotz ihrer augenscheinlichen Unterschiedlichkeiten, alle zu verschiedenen Aspekten im Bereich der Differentiellen Psychologie zuordnen und werden im Folgenden beschrieben.

2.1 Persönlichkeit und interindividuelle Unterschiede

Bereits Ende der Dreißigerjahre des vorigen Jahrhunderts beginnen Allport und Vernon (1930) ihren Aufsatz zum psychologischen Forschungsfeld der *Persönlichkeit* mit der Feststellung, dass es nicht mehr möglich sei, in einer einzelnen Überblicksarbeit eine vollständige Darstellung aller Aspekte dieses Forschungsbereiches darzustellen. Angesichts der stetig wachsenden Anzahl an Publikationen allein zu diesem Teilgebiet der Differentiellen Psychologie, dürfte sich die Situation in den letzten gut 80 Jahren Persönlichkeitsforschung hinsichtlich der Übersichtlichkeit nicht verbessert haben. Darüber hinaus beklagen die Autoren bereits damals das Fehlen kritischer und vollständiger Zusammenfassungen des Forschungstandes: *„Summaries are not lacking, but none is in itself complete, and few attempt a critical classification and interpretation.“* (Allport & Vernon, 1930, S. 677). In dem folgenden Abschnitt soll der Forschungsgegenstand *Persönlichkeit* daher anhand einiger exemplarisch dargestellten Forschungszugänge und kritischer Positionen skizziert werden, um dann auf den in dieser Arbeit verfolgten Zugang im Rahmen des *Eigenschaftparadigmas* zu fokussieren. Ausgangspunkt soll dazu eine allgemeine Definition von *Persönlichkeit* sein, welche möglichst umfassend die unterschiedlichen wissenschaftlichen Forschungszugänge zum Konstrukt *Persönlichkeit* integriert. Eine solche Definition gibt zum Beispiel bereits Ross (1987), welcher Persönlichkeit wie folgt definiert:

> Persönlichkeit ist ein zusammengesetztes Konstrukt, das für die Gesamtheit der Handlungen, Denkprozesse, emotionalen Reaktionen und motivationalen Bedürfnisse von Personen steht, durch die sie als genetisch programmierte biologische Organismen mit ihrer Umwelt interagieren, wobei sie ihre Umwelt beeinflussen und von ihr beeinflusst werden
>
> [let us consider personality as a composite construct that stands for the sum total of people's actions, thought processes, emotional reactions, and motivational needs, through which they, as genetically programmed biological organisms, interact with their environment, influencing it and being influenced by it.]
>
> (Ross, 1987, S. 7).

Aus dieser allgemeinen und breiten Definition lassen sich verschiedene wissenschaftlich Perspektiven und Forschungszugänge zum Konstrukt Persönlichkeit ableiten. So beschreibt Funder (2001) neben den vier, von ihm als „klassisch“ bezeichneten psychologisch wissenschaftlichen Perspektiven, dem Psychoanalytischen-, dem Behavioristischen-, dem Humanistischen- und dem Eigenschafts-(Trait-) Paradigma, zusätzlich die vom Autor als „neu“ bezeichnete Paradigmen bzw. wissenschaftlich Zugänge, nämlich den sozial–kognitiven, den biologischen und den evolutionären Zugang. Einige dieser Forschungszugänge sollen im Folgenden beschrieben werden.

2.1.1 Psychoanalytisches Paradigma der Persönlichkeit

Als Beginn des psychoanalytischen Paradigmas der *Persönlichkeit* kann sicherlich Sigmund Freuds (1856–1939) erstes topografisches Modell der Seele gelten, welches er bereits in einem seiner frühen Werke (*Die Traumdeutung* – Freud, 1911) formulierte. Eine grundlegende Basis bei der Erklärung menschlichen Erlebens und Verhaltens sowie auch interindividueller Unterschiede, besteht nach Freud in der von ihm angenommenen Existenz der *psychischen Energie*. Jeder Mensch verfügt danach über eine bestimmte Menge psychischer Energie, deren Ursprung er in angeborenen Trieben sah (vgl. z. B. Asendorpf & Neyer, 2012, S. 9). Die unterschiedliche Verarbeitung oder der unterschiedliche Fluss der Energie manifestiere sich in menschlichen Phänomenen wie Wahrnehmen, Fühlen, Denken, Erinnern oder Träumen. Diese Phänomene klassifiziert Freud zunächst innerhalb drei (primärer) mentaler Prozesse – das *Bewusste* das *Vorbewusste* und das *Unterbewusste*. Eine weitere wichtige Grundlage von Freuds Theorie zur (individuellen) Psychodynamik und Symptomgenese und damit des psychoanalytischen Paradigmas der *Persönlichkeit* allgemein bildet dabei das *Unterbewusste*, dessen Inhalte durch verschiedene Mechanismen der Verdrängung unterdrückt werden. Aus diesem ersten Modell, welches die mentalen Prozesse zunächst nach ihrer Qualität kategorisierte, entwickelte Freud in späteren Jahren sein bekanntes Strukturmodell bestehend aus den Instanzen *ES*, *ICH* und *ÜBER-ICH* (Freud, 1923, 1933), welches die mentalen Prozesse nach deren Funktion kategorisiert und die (primären) mentalen Prozesse integriert.

Der Ausgangspunkt der psychoanalytischen Theorie lag dabei zunächst in einer am jeweiligen Symptom orientierten Therapie, sodass sich der Gegen-

stand der analytischen Untersuchungen zunächst fast ausschließlich auf Kranke bezog. Wie bereits Fromm (1932) darstellt, entstanden aber im Verlauf der Entwicklung der psychoanalytischen Theorie zusätzlich Fragen nach dem Ursprung und der Bedeutung individueller psychischer Unterschiede auch bei Gesunden. Im Vergleich zu der eher am pathologischen Symptom orientierten Therapie, liegt der Fokus der frühen psychoanalytischen *Charakterologie* dabei auf den sich durch Verdrängung, Sublimierung und (anschließender) Reaktionsbildung herausbildenden, individuellen psychischen Eigenarten (vgl. Fromm, 1932). Als individuelle Persönlichkeit oder Charakter sieht Freud diese, bei jeder Person individuell unterschiedlich gelagerte, Triebdynamik und deren Steuerung durch die drei von ihm beschriebenen Instanzen.

Freuds psychoanalytisches Paradigma lenkte damit (erstmals) die wissenschaftliche Aufmerksamkeit auf interindividuelle Differenzen im Wahrnehmen und Erleben. Im Vergleich zu den damals bis dahin eher vorherrschenden Ansätzen im Rahmen psychophysisch experimenteller Ansätze (z. B. Fechner, 1860a, 1860b), schuf er damit die Grundlage zur Berücksichtigung idiosynkratischer Muster des Erlebens und Verhaltens (Rosenzweig, 1951, 1985). An dieses frühe Modell der *psychoanalytischen Persönlichkeit* mit seinen distinkten Elementen *ES*, *ICH* und *ÜBER-ICH* knüpften dann später verschiedene psychoanalytische Richtungen an.

Ein ebenso prominenter Vertreter wie Freud innerhalb des psychoanalytischen Paradigmas ist Carl Gustaf Jung (1875–1961), welcher später als Begründer einer eigenen psychoanalytischen Schulrichtung angesehen wurde, die sich der *aktiven Imagination* als Zugang zum Unterbewussten bediente (Graf-Nold, 2005). Jung, welcher in seinen Anfangsjahren zunächst eine enge Beziehung zu Freud pflegte und gemeinsame Ansichten in Bezug auf die Psychoanalyse teilte, trennte sich 1911 als Anhänger von Freud (Federn, 2005). Diese Trennung begründete sich im Wesentlichen in der im Vergleich zu Freuds unterschiedlichen Auffassung vom Unbewussten (Graf-Nold, 2005). In den Jahren nach dieser Trennung entwickelte Jung, in Auseinandersetzung mit den entstandenen unterschiedlichen psychoanalytischen Auffassungen von Freud und anderen seiner ehemaligen Anhänger, seine psychoanalytische Typenlehre (Jung, 1921). Dieses Werk ist insofern auch für das innerhalb der Differentiellen Psychologie und Persönlichkeitspsychologie heute vorherrschende *Eigenschaftsparadig-*

ma von Bedeutung, als dass Jung in diesem Werk erstmals die beiden typischen psychischen Grundfunktionen *Extraversion* und *Introversion* begründete. Dieses maßgeblich von Jung eingeführte Konzept fand einige Akzeptanz und Verbreitung, wobei sich die unterschiedlichen weiterentwickelten Definitionen der beiden Typisierungen allerdings nicht immer deckten (Freyd, 1924; Guilford, 1934). Jung selbst kombinierte seine beiden psychischen Grundfunktionen mit vier typischen Einstellungen des Bewusstseins – *Denken*, *Fühlen*, *Intuiren* und *Empfinden* – zu einer acht Typen umfassenden Typologie.

Mitte bis Ende des 20 Jahrhunderts entwickelten sich auch auf dem amerikanischen Kontinent eigenständige psychoanalytische Richtungen mit eigener psychoanalytischen Tradition (Friedman, 2011). So entwickelte beispielsweise Kohut (1971, 1977) seine erfahrungsnahe Theorie der *Selbstpsychologie* der psychoanalytischen Persönlichkeit. Die Selbstpsychologie fokussiert dabei nicht isoliert auf die innere Psychodynamik der Person, sondern insbesondere auf deren Relation zu einem Objekt oder auf deren Beziehung zu anderen Personen. Solche individuell unterschiedlich gelagerten *Objektbeziehungen* drücken dabei die individuelle Persönlichkeit aus.

Ein weiterer typischer Vertreter einer neueren, sehr eigenständigen psychoanalytischen Theorie der *Persönlichkeit* ist Charles Brenner, welcher die Aufgabe der aus seiner Sicht „artifiziellen“, mentalen Prozesse zugunsten ganzheitlich, holistisch geprägter Vorstellungen der *psychoanalytischen Persönlichkeit* propagierte und sich damit am weitesten von Freuds psychoanalytischem Persönlichkeitsmodell entfernte (Brenner, 1994) – vergleiche auch Friedman (2011) für eine Zusammenfassung der aktuelleren Entwicklungen.

Es muss festgestellt werden, dass sich das psychoanalytische Paradigma in den letzten Jahrzehnten zunehmender Kritik stellen musste (z. B. Eagle, 2007; Kernberg, 1993; Summers, 2008). Diese Kritik konzertiert sich nicht zuletzt auf den Vorwurf der (bislang) unzureichenden empirisch untermauerten Überprüfung der verschiedenen Theorien und Modelle im Rahmen des psychoanalytischen Paradigmas. Die Positionen einzelner Kritiker sind dabei sehr unterschiedlich und reichen bis hin zur kompletten Ablehnung der psychoanalytischen Theorie mit der damit verbundenen Negierung ihrer Wissenschaftlichkeit (Grünbaum, 1988). Anderseits wird der heuristischer Wert und das reichhaltige Theoriegebäude der psychoanalytischen Theorie im Hinblick auf

die Entwicklung unterschiedlicher Formen therapeutischer Intervention (und deren Erfolge) betont und anerkannt (Kernberg, 1993). Auch wenn zum Beispiel Sigmund Freud keine absolute Persönlichkeitstheorie im Sinne eines universellen verhaltensprädiktiven Modells beschrieb, so entwickelte er jedoch mit seinem Instanzenmodell ein funktionales Modell von der menschlichen Persönlichkeit, wobei er sich in hohem Maße mit kindlichen Erfahrungen und Krisen auseinander setzte, das wiederum andere Persönlichkeitstheoretiker beeinflusste. Die unter anderem von Grünbaum (1988) dargestellte Kritik an der Wissenschaftlichkeit des psychoanalytischen Theoriegebäudes muss aus heutiger Sicht insofern eingeschränkt werden, als dass sie die neueren, seit Anfang der Neunzigerjahre etablierten Bemühungen der psychoanalytischen Schule um eine objektive und reliable Klassifikation psychodynamischer Konzepte im Rahmen der *Operationalisierten Psychodynamischen Diagnostik – OPD* noch nicht berücksichtigt (Arbeitskreis OPD, 1996). Bei diesem Instrumentarium handelt es sich um eine psychodynamisches Diagnosesystem, welches auf einer Klassifikation psychodynamischer Hintergrundkonstrukte basiert, mit dem Ziel die wesentlichen Variablen für psychodynamische Theorien – unter anderem *Übertragungsmuster*, innere *Konfliktkonstellationen*, *Objektbeziehungen* und *strukturelle Bedingungen* der individuellen Persönlichkeit – zu identifizieren und zu operationalisieren (z. B. Dahlbender & Tritt, 2011). Dadurch soll den allgemeinen Gütekriterien psychodiagnostischer Instrumente wie Objektivität, Reliabilität und Validität in der Diagnose entsprochen werden und insgesamt eine eher evidenzbasierte Orientierung der psychoanalytischen Theorie angestrebt werden (S. H. Gray, 2002). Neuere Befunde weisen in diesem Sinne darauf hin, dass beispielsweise die insbesondere in der neueren psychoanalytischen Theorie postulierten *Objektbeziehungs-Theorien* ein vielversprechender Ausgangspunkt für die Untersuchung aktueller kognitiver Konzepte wie z. B. Lernprozesse sind, welche die (individuelle) Persönlichkeitsentwicklung erklären können (Imbasciati, 2003). Insofern lässt sich im psychoanalytischen Paradigma zunehmend eine empirisch evidenzorientierte Wende feststellen (z. B. Fonagy, 2015; Fonagy et al., 2015; Goodyer et al., 2011; Suszek, Holas, Wyrzykowski, Lorentzen & Kokoszka, 2015; Taylor et al., 2012).

Insgesamt lässt sich feststellen, dass sich der aktuelle psychoanalytische Blick auf die Persönlichkeit eher durch holistische Betrachtungsweisen aus-

zeichnet (z.B. Brenner, 1994). Zusätzlich erscheint die Klassifikation und Definition von Typen gegenüber der Quantifizierung bestimmter Merkmalsausprägungen innerhalb des psychoanalytischen Paradigmas vorherrschend zu sein. So werden auch bereits bei Jung (1921) die acht von außen beobachtbaren *Typen* durch im Unterbewussten wirkende *Gegentypen* kompensatorisch ergänzt (Graf-Nold, 2005). Dieses ganzheitliche Konzept der menschlichen Psyche wird dabei zum Beispiel von Jung einerseits durch Gegensätzlichkeit und andererseits durch ihre Einheitstendenz charakterisiert. Bezüglich der beiden psychischen Grundfunktionen *Extraversion* und *Introversion* wenden Guilford und Braly (1930) allerdings in einer kritischen Betrachtung des Konzeptes ein, dass neben Carl Gustav Jung auch andere Autoren wie zum Beispiel Otto Gross (1902), Ludwig Klages (1910; 1926) sowie Wiliam Stern (1900), die den psychischen Grundfunktionen Jungs ähnlichen Konzepte propagierten. Das Typenkonzept von Otto Gross sei hier nur insofern erwähnt, da es wegen seiner biologisch, physiologischen Fundierung Parallelen aufweist zu der später von Kretschmer (1977) und Sheldon (1963) verfolgten Idee einer physiologisch, morphologisch orientierten Annäherung an das Konstrukt *Persönlichkeit*.

2.1.2 Biologisch, konstitutionstypologische und evolutionäre Zugänge

Auch wenn biologisch, konstitutionstypologische und vor allem evolutionäre Zugänge zu individuellen Unterschieden und Persönlichkeit innerhalb des vergleichsweise jungen Faches Psychologie beispielsweise von Funder (2001) eher den „neuen“ Paradigmen zugerechnet werden, so können doch solche Ansätze letztlich auf eine vergleichsweise lange Historie zurückblicken. So lässt sich die Idee, unterschiedlich ausgeprägte menschliche Verhaltensdispositionen über evolutionäre Prozesse zu erklären, bereits in den Schriften des Begründers der Evolutionstheorie Charles Darwin (1809–1882) finden (Darwin, 1859, 1872). Die von Darwin (1872) in ihren ersten Ansätzen formulierten Thesen zum evolutionären Ursprung menschlicher (und tierischer) Ausdrucksformen emotionaler Dispositionen und Verhalten allgemein wurden Mitte des 20. Jahrhunderts wieder aufgegriffen und begründeten ein eigenes Paradigma – die *Evolutionspsychologie* (Ghiselin, 1973). Auch die Idee, dass physiologische Zustände oder

Eigenschaften gewissermaßen als Ursachen für ein unterschiedlich ausgeprägtes Temperament bzw. Ausprägungen der Persönlichkeit anzusehen sind, lässt sich historisch sehr weit bis zum Begründer der westlichen Medizin, Hippokrates (460 – 377 v. Chr.) zurückverfolgen. Hippokrates unterschied in seiner Typologie der Temperamente vier Typen: Den *Sanguiniker*, *Phlegmatiker*, *Choleriker* und den *Melancholiker*, welche er jeweils auf das Vorherrschen eines der vier Körpersäfte Blut, Schleim, gelbe und schwarze Galle zurückführte. Diese frühe Klassifikation der Temperamente oder Persönlichkeitsausprägungen wurde später von Galen (131 – 201 v. Chr.) auf neun Typen des Temperaments ausgeweitet, welche von ihm als jeweils eng verbunden mit einer bestimmten physiologischen Konstitution angesehen wurden (vgl. Ross, 1987, S. 53). Im 18. und 19. Jahrhundert wurde diese Idee, über körperliche Erscheinungsformen auf Temperament und Charakter eines Menschen zu schließen, in der Lehre der Physiognomik wieder aufgegriffen (Fisseni, 1998). Ausgehend von der Annahme, dass das Gehirn der eigentliche Sitz aller geistigen Tätigkeit des Menschen sei,[1] entwickelte der Arzt Franz Joseph Gall (1758–1828) seine gehirntopologische Lehre der Phrenologie. Dabei sollten einzelnen Hirnarealen typische charakterliche Eigenschaften und Dispositionen zugeordnet werden. Fisseni (1998) merkt zu diesem biologisch orientierten Ansatz zur *Persönlichkeit* an, dass aber bei der Phrenologie, obwohl zunächst von richtigen Annahmen ausgehend, letztlich nicht haltbare Schlüsse zur strukturellen Organisation des Gehirns gezogen werden; „*Fälschlicherweise, weil simplifizierend, argumentiert sie, im Gehirn als zentralem Organ des Geistes seien die wichtigsten 'Einheiten' der Persönlichkeit abgrenzbar angelegt.*“ Fisseni (1998, S. 113).

Obwohl derartige biologisch begründete Typologien aus heutiger Perspektive entweder aufgrund fehlender empirischer Fundierung oder aufgrund zwischenzeitlich aufgedeckter, methodischer Fehler in der Datenauswertung als nicht mehr zeitgemäß erscheinen, können solche Ansätze dennoch als Vorläufer moderner neuro- und evolutionswissenschaftlicher Zugänge zur Persönlich-

[1] Diese aus heutiger Sicht eventuell „trivial“ beziehungsweise selbstverständlich erscheinende Annahme, stellte bezogen auf die antiken Vorstellungen zur Seele insofern eine Neuerung dar, als dass beispielsweise der griechische Philosoph Aristoteles in seiner zoologischen Schrift „Historia Animalium“ die These vertrat, dass das Gehirn aufgrund seiner „Blutlosigkeit“ im Gegensatz zum Herzen kaum als Sitz der Seele in Frage komme.

keit angesehen werden. So lassen sich bis Mitte des 20. Jahrhunderts Ansätze zur Typologie der Persönlichkeit finden, welche auf beobachteten physiologischen Merkmalen aufbauen. Dabei wird versucht, die angenommene Beziehung zwischen bestimmten Varianten der menschlichen Physiologie und korrespondierenden Verhaltensmerkmalen empirisch zu untermauern. Solche Bestrebungen mit unterschiedlich stark ausgeprägter empirischer Fundierung, lassen sich auch bereits Anfang des 20. Jahunderts finden (z. B. Gross, 1902). Zwei etwas „aktuellere“ und prominente Vertreter solcher konstitutionstypologischen Ansätze waren der Psychiater und Neurologe Ernst Kretschmer (1888–1964) und im amerikanischen Raum der Mediziner und Psychologe Wiliam H. Sheldon (1898–1977). In Verbindung mit psychiatrischen Untersuchungen war Kretschmer aufgefallen, das bestimmte psychische Erkrankungen *scheinbar* jeweils mit bestimmten Körperbauformen einhergehen. Angeregt durch seine Beobachtungen entwickelte Kretschmer eine Typologie, welche den *pyknischen* (gedrungen, untersetzt), den *leptosomen* (schmal, spitz) und den *athletischen* Typus unterschied. Diese drei Typen wurden von Kretschmer als jeweils unterschiedlich assoziiert angesehen mit den beiden vom Psychiater Kraepelin (1856 – 1926) unterschiedenen, psychopathologischen Formkreisen – dem „manisch depressiven Irresein“ und der auch als Schizophrenie bezeichneten „Dementia praecox“ (Kraepelin, 1983). Im Vergleich zu anderen früheren biologisch, typologischen Ansätzen und in Abgrenzung zu naiven Populärphysiognomien, bezieht sich Kretschmer dabei ausdrücklich auf (scheinbar) empirisch belegte Zusammenhänge. So wird beispielsweise in dem von seinem Sohn Wolfgang Kretschmer 1977 in der 26. Auflage herausgegebenen Hauptwerk, basierend auf einer Beobachtung mit 260 Fällen, der Zusammenhang zwischen dem *pyknischem* Typ und manisch depressiver Pathologie begründet (Kretschmer, 1977, S. 32). Die später auf dieser Basis entwickelte Typologie der normalen (nicht pathologischen) Temperamente stützt sich auf die von Kretschmer vertretene *Kontinuitätshypothese*, bei der er einen kontinuierlich, fließenden Übergang zwischen gesunder und pathologischer psychischer Verfassung annahm (Matz, 2002, S. 21). Einen ähnlichen, insbesondere aber methodisch leicht unterschiedlichen Ansatz der Typologie propagierte Wiliam H. Sheldon. Vergleichbar mit Kretschmers oben beschriebenem Ansatz sieht Sheldon Unterschiede menschlichen Verhaltens auf der Grundlage der individuell unterschiedlichen biologisch,

genetischen Disposition begründet. Bezugnehmend auf Befunde aus der Embryologie entwickelte Sheldon eine Klassifikation, welche drei so genannte *Somatotypen* umfasste (Sheldon, 1963). Diese bezeichnete er, in ihrer dominanten Ausprägung, als *Endomorphie*, *Mesomorphie* und *Ektomorphie*. Im Gegensatz zu Kretschmers Typologie postuliert Sheldon aber, dass es sich bei diesen drei somatisch begründeten Aspekten um jeweils quantitativ unterschiedlich ausgeprägte Dimensionen handelt (Schneewind, 1982, S.125 ff.). Sheldon schlägt für jede der drei Dimensionen eine siebenstufige Skala vor (1 ≡ „geringe Ausprägung“ bis 7 ≡ „starke Ausprägung“), sodass sich im Rahmen seiner Typologie theoretisch 343 ($7 \times 7 \times 7$) individuell unterschiedliche ausgeprägte Somatotypen unterscheiden ließen. Allerdings ergeben sich aus Sheldons, anhand der physischen Morphologie begründeten, Definitionen der drei somatischen Aspekte Einschränkungen bezüglich der Kombinationsmöglichkeiten der jeweils drei Skalenwerte. So sind extreme Ausprägungen auf allen drei „Skalen“ gleichzeitig, z. B. (7,7,7) oder (1,1,1), definitionsgemäß und aus logischen Gründen zur Beschreibung des individuellen Somatotypen nicht möglich. Wie auch Kretschmer versuchte Sheldon, systematische Zusammenhänge zwischen der von ihm aufgestellte Klassifikation konstitutioneller Typen einerseits und beobachtetem Verhalten andererseits aufzuzeigen. Über empirische Daten zum Verhalten, welche er an zuvor (selbst) „typisierten“ College Studenten über Interviews und Beobachtungen erhoben hatte, konnte er zunächst recht starke Zusammenhänge zwischen seinen Somatotypen und den drei von ihm anhand der Interview- und Beobachtungsdaten erfassten Temperament-Typen nachweisen (Schneewind, 1982, S.128, 131). Trotz seiner akribisch durchgeführten und empirisch fundierten Untersuchungen wurden auch gegenüber Sheldons Ansatz der Typologie, insbesondere methodisch begründete, kritische Einwände entgegen gebracht. Diese bezogen sich zum einen auf die Tatsache, dass sowohl die somatische Typisierung als auch die Typisierung der Temperamente nicht von unabhängigen Beurteilern, sondern von Sheldon selbst durchgeführt wurde. Andererseits wurden die logischen und definitorischen Abhängigkeiten, welche zwischen den drei somatotypischen „Skalen“ bestehen, bemängelt (vgl. Schneewind, 1982, S. 135 ff., für eine umfangreichere Darstellung der kritischen Einwände). Als Fazit der hier am Beispiel der beiden Vertreter Kretschmer und Sheldon kurz vorgestellten biologisch konstitutionstypologischen Zugangs zum

Konstrukt *Persönlichkeit* sollen im Sinne einer kritischen Würdigung noch einige Punkte aus heutiger Perspektive zu beiden Ansätzen aufgeführt werden. Sowohl Kretschmer als auch Sheldon begründeten ihre Typologien zunächst auf empirischen Daten. Des Weiteren begründen Kretschmer als auch Sheldon die Validität ihrer Persönlichkeitsmodelle dann mit den bei der statistischen Analyse der empirischen Daten gefundenen Zusammenhänge zwischen ihren physischen Merkmalstypologien und, im Falle von Kretschmers psychopathologischen und im Falle von Sheldons „normalen", menschlichen Verhaltensvariationen. Allerdings weisen Asendorpf und Neyer (2012, S. 134) darauf hin, dass es sich bei diesen, teils recht starken, korrelativen Zusammenhängen eher um Scheinkorrelationen handeln dürfte, welche sich z. B. bei Kretschmers Ansatz leicht auf eine Konfundierung durch die nicht berücksichtigte Variable Alter zurückführen lassen. Auch wenn sich daher die von Kretschmer und Sheldon gefundenen Zusammenhänge nachträglich eher als nicht haltbar erweisen, so ist dennoch positiv hervorzuheben, dass sich beide Ansätze empirischer Beobachtungsdaten bedienen und damit die an empirischen Daten orientierte Entwicklung von Persönlichkeitsmodellen mitbegründeten. Boerner (2015) weist in einer kritischen Diskussion von Kretschmers Konstitutionsgedanken darauf hin, dass die „*... Thematik des Zusammenhangs von Körperbau, Persönlichkeit bzw. Temperament und psychischen Störungen allenfalls noch eine historische Rolle ...*" (Boerner, 2015, S. 140-141) spiele. Einschränkend bezüglich solcher biologisch, physiologischer Persönlichkeitsmodelle weisen Asendorpf und Neyer (2012, S. 134) drauf hin, dass psychologische Analysen von physischen Merkmalen durch die, hinsichtlich der statistisch, methodischen Datenauswertung, eher „naiven" Vorgehensweisen insgesamt in Verruf geraten wären.

Neuere Ansätze im Rahmen biologischer oder evolutionärer Persönlichkeitsmodelle stützen sich demgegenüber nicht auf mehr oder weniger subjektive Einschätzungen körperlicher Merkmale. In aktuellen Arbeiten werden heute eher medizinisch, physikalische Messwerte oder aber auch Typologien auf der Basis genetischer Grundlagen herangezogen. Dabei wird eher nicht nach biologischen Merkmalen zu einer deterministisch, kausal orientierten Prädiktion des Verhaltens oder des Charakters gesucht. Vielmehr wird z. B. im Bereich der neurobiologisch, psychiatrischen Forschung eher nach neurobiologischen Korrelaten von psychischen Störungen oder mentalen Zuständen gesucht (z. B.

Förstl & Förstl-Hautzinger-Roth, 2006). Dabei ergeben sich immer wieder Befunde zu den Wechselwirkungen zwischen biologischen Mechanismen und Verhalten, welche, im Sinne meist probabilistisch formulierter Zusammenhänge, in Verbindung zu unterschiedlichen Persönlichkeitsmerkmalen stehen (z. B. Knutson et al., 1998; Roth, 2001). Eine der am meisten akzeptierten Theorien zu biologischen Modellen in der Persönlichkeitspsychologie ist die biopsychologische Theorie der *Persönlichkeit*, die von Jeffrey Alan Grey im Jahr 1970 (z. B. J. A. Gray, 1970) vorgeschlagen wurde. Gray nimmt hypothetisch zwei Systeme an, welche die Verhaltensaktivität bestimmen. Einerseits das Verhaltenshemmungssystem (BIS) [Behavioral Inhibition System] und das Verhaltensaktivierungssystem (BAS) [Behavioral Aktivation System]. Das BIS wird dabei mit der Empfindlichkeit gegenüber einer Bestrafung und der damit verbundenen Vermeidungsmotivation verknüpft, während das BAS eher mit der Empfindlichkeit gegenüber Belohnung und der Annäherungsmotivation verknüpft ist. Gray stützte seine Theorie auf den Befund, dass psychodiagnostische Skalen, die Attribute dieser hypothetischen Systeme abbilden, in systematischem Zusammenhang mit Eigenschaften der Persönlichkeit stehen. Danach korreliert beispielsweise die Dimension *Neurotizismus* (vgl. Abschnitt 2.1.3) positiv mit der BIS-Skala und ist gleichzeitig negativ mit der BAS–Skala assoziiert (J. A. Gray, 1970). Die moderne (kognitive) Neurowissenschaft zielt darauf ab, die Prinzipien der Beziehungen zwischen neuroanatomischen Funktionen, Strukturen und Besonderheiten im Gehirn und beobachtbarem Verhalten zu untersuchen, indem sowohl gemeinsame Mechanismen als auch individuelle Unterschiede zwischen Individuen identifiziert werden (z. B. Spada, 1992). Ein klassischer und viel beachteter Beitrag von Douglas (1967) befasst sich dabei mit der Rolle der Hippocampus Struktur im Gehirn in Bezug auf Verhaltensdispositionen. Als weitere Beispiele zu solchen neuroanatomische Korrelaten zu Persönlichkeits- und Verhaltensdispositionen individueller Eigenschaften sollen hier, stellvertretend für eine ganze Reihe von Studien aus diesem Bereich, einige prominente Befunde erwähnt werden. So belegen beispielsweise die Untersuchungen von Scholz, Klein, Behrens und Johansen-Berg (2009), dass das erlernen bestimmter spezifischer motorischer Fertigkeiten (hier das Erlernen des Jonglierens) mit strukturellen Veränderungen im Gehirn einhergeht. Ferner zeigt die auch in der nicht wissenschaftlichen Öffentlichkeit

prominent rezipierte Studie von Woollett und Maguire (2011), dass bei Taxifahrern mit einem längeren Lernprozess zu räumlichen Orientierungsaufgaben (Beschäftigung mit Straßenkarten der City of London) eine Vergrößerung der Hippocampus Struktur im Gehirn einhergeht. Ein prominenter Vertreter des neuropsychologischen Ansatzes ist der Nobelpreisträger Eric Kandel, der mit seinen Arbeiten diese Forschungsrichtung bemerkenswert bereichert hat und dadurch das Verständnis von Verhaltensänderungen und Korrelaten auf molekularer Ebene mit weitreichenden Implikationen für Lernen und Gedächtnis erweitert (E. Kandel, Schwartz & Jessel, 2000; E. R. Kandel, 2005). Neben solchen grundlegenden Arbeiten und allgemeinen Befunden existiert eine Vielzahl von einzelnen Studien, die neuroanatomische Korrelate zu spezifischen Persönlichkeitsmerkmalen untersuchen, wie sie nach dem Big-Five-Modell im Rahmen des *Eigenschaftsparadigmas* (vgl. nächster Abschnitt 2.1.3) definiert sind. So ist beispielsweise die Tendenz eher nach neuen Erfahrungen zu suchen – ein Konzept das inhaltlich mit der Persönlichkeitsdimension *Offenheit* assoziiert werden kann – nach Befunden aus der neuropsychologischen Forschungsrichtung mit bestimmten genetische bedingten Variationen des Dopaminergen-Systems und neuroanatomischen Besonderheiten assoziiert (S. B. Martin et al., 2007). Canli, Sivers, Whitfield, Gotlib und Gabrieli (2002) finden in ihrer Untersuchung, dass bestimmte Prozesse der Amygdala – einer wichtigen Struktur des menschlichen Gehirns für die Verarbeitung emotionaler Signale – als Funktion individuell unterschiedlich ausgeprägter *Extraversion* mit positiven Emotionen und annäherungsbezogenem Verhalten assoziiert ist, welches mit dem sozial interaktiven Stil von extravertierten Personen konsistent ist. Westlye, Bjornebekk, Grydeland, Fjell und Walhovd (2011) finden Zusammenhänge zwischen angstbezogenen Persönlichkeitseigenschaften und spezifischen Mikrostrukturen und der strukturellen Integrität der weißen Substanz im Gehirn (vgl. auch Montag, Reuter, Weber, Markett & Schoene-Bake, 2012). Mit derartigen Befunden übereinstimmend finden Aghajani et al. (2014), dass *Neurotizismus* und *Extraversion* mit der funktionalen Konnektivität der Amygdala Gehirnstruktur (vgl. auch Markett et al., 2013), assoziiert ist. So fällt die Vernetzung innerhalb der Gehirnregion Insula zwischen unterschiedlich ängstlichen Personen verschieden aus, wobei hoch ängstliche Personen danach eine eher wenig effiziente Vernetzungsstruktur haben (Markett, Montag, Melchers,

Weber & Reuter, 2016). Solche Befunde erweiternd zeigen W.-Y. Liu et al. (2013), dass *Gewissenhaftigkeit* mit einer weniger ausgeprägten Verschaltung einzelner Gehirnareale begleitet ist.

Übergreifend argumentiert Nettle (2006), wiederum aus evolutionspsychologischer Perspektive, dass jede der Big-Five-Dimensionen der menschlichen Persönlichkeit als Ergebnis eines Optimierungsprozesses zwischen unterschiedlichen Fitnesskosten und -nutzen im Rahmen evolutionärer Prozesse gesehen werden kann. Da bei so einem komplexen Prozess kaum ein einzelner optimaler Wert, im Sinne einer optimalen Konfiguration von Persönlichkeitsausprägungen, existieren dürfte, ist zu erwarten, dass die genetische Vielfalt und Unterschiedlichkeit in der Population erhalten bleibt. Speziell im Hinblick auf die unterschiedliche Art und Weise der Beantwortung von Fragebogen untersuchen beispielsweise Waller und Reise (1992) den genetischen Einfluss auf die Skalierbarkeit der Antwortmuster im Rahmen der *Item-Response-Theory* (IRT) anhand der Analyse einer großen Stichprobe von Zwillingsdaten (vgl. Lykken, Bouchard, McGue & Tellegen, 1990). Waller und Reise (1992) finden dabei, dass sich etwa 20 % der Variabilität der Skalierbarkeit auf genetische Faktoren der antwortenden Personen zurückführen lassen und folgern demzufolge, dass genetisch, biologische Determinanten eine substanzielle Rolle bei der Entstehung von Antwortmustern und deren Skalierbarkeit spielen können.

2.1.3 Das Eigenschaftsparadigma und das Fünf-Faktoren-Modell der Persönlichkeit

Die weiteren empirischen Inhalte der vorliegenden Arbeit stützen sich nach einer Taxonomie unterschiedlicher Forschungszugänge zum Konstrukt *Persönlichkeit* von Funder (2001) auf das *Eigenschaftsparadigma* (*Trait-Paradigma*), in dessen Rahmen Persönlichkeit meist mittels psychodiagnostischer Fragebogeninventare erfasst wird. Das zentrale Konzept dieses Zugangs zur Persönlichkeit ist das Merkmal – und bei dessen zeitlicher Stabilität die *Eigenschaft* – in der (meist englischsprachigen) Literatur als *Trait* bezeichnet (z. B. Allport, 1927; Burt, 1939; Filter, 1921; Lanning, 1991; Thurstone, 1934).

Nach Schneewind (1982, S. 109 ff.) ist das heute unter dem Schlagwort „Big-Five“ bekannte Fünf-Faktoren-Modell (FFM) der *Persönlichkeit* dem Ei-

genschaftsparadigma (oder auch Trait-Paradigma) zuzuordnen. Das Big-Five-Modell ist eng verknüpft mit dem faktorenanalytischen Ansatz der Persönlichkeitspsychologie innerhalb der statistisch-mathematischen Tradition der psychologischen Persönlichkeitsforschung. Es entwickelte sich in Folge der Anwendung, der von Spearman (1904) im Zusammenhang mit der Intelligenzforschung propagierten Faktorenanalyse (vgl. Pawlik, 1971; Überla, 1977) auf andere menschliche Verhaltensbereiche (z. B. Thurstone, 1935, 1938), woraus sich ein dimensionales Faktorensystem der Persönlichkeit zur Erklärung der vielfältigen menschlichen Verhaltensweisen ergab (Schneewind, 1982, S. 112). Bei aktuellen Persönlichkeitstests werden daher meist verschiedene Aspekte bzw. Dimensionen des Konstrukts Persönlichkeit mit Hilfe entsprechender Skalen operationalisiert. Über die verschiedenen Formen der Operationalisierung und Fragebogen hinweg wird dabei eine unterschiedliche Anzahl von Dimensionen zugrunde gelegt, oder bei der Testkonstruktion mittels faktorenanalytischer Methoden empirisch ermittelt. (Cattell, 1946; Cattell & Saunders, 1954a; Guilford, 1975; Thurstone, 1934). Auch wenn es teilweise kontroverse Ansichten über die Definition und Anzahl der einzelnen Dimensionen im Konstrukt Persönlichkeit herrschen (vgl. z. B. A. H. Buss & Finn, 1987; Cattell, 1944; Costa & McCrae, 1992a, 1992b; Eysenck, 1992a, 1992b, 1993; John & Srivastava, 1999), so besteht weitgehend Konsens darüber, dass diese Dimensionen der Persönlichkeit als relativ dauerhafte psychologische Eigenschaften von Personen definiert sind. Das Konzept dieser Definition basiert auf der Erweiterung des von Stern (1911) eingeführten Schemas vom Komparations- und Korrelations Forschung zur Unterschiedlichkeit zwischen *Personen* und *Merkmalen*, um eine zeitliche Komponente durch Cattell (1946). Aus einem *Merkmal* einer Personen wird im Sinne des *Kovariationswürfels* nach Cattell (1946) dann eine *Eigenschaft* [*trait*], wenn dieses sich als zeitlich stabil erweist (vgl. auch Cattell, 1988a). Solche Eigenschaften eigenen sich dann mehr oder weniger gut dazu, um Personen und ihr Verhalten zu beschreiben, vorherzusagen, zu erklären oder zu verstehen.

Die theoretische Grundlage des Eigenschaftsparadigmas bilden dabei bereits früher angestellte Überlegungen zur Analyse der menschlichen Sprache als methodischer Zugang zum psychologischen Konstrukt der *Persönlichkeit* (Baumgarten, 1933; Galton, 1884; Klages, 1926), welche schließlich in die For-

mulierung der *Sedimentationshypothese* mündeten. Dabei werden personenbeschreibende Adjektive als sprachliche „Sedimente“ realer menschlicher Merkmalsunterschiede angesehen (John, Angleitner & Ostendorf, 1988). Ausgehend von dieser theoretischen Basis führten Allport und Odbert (1936) unter Anwendung der bereits von Thurstone (1935, 1938) für den Bereich der Einstellungsmessung angewendeten Faktorenanalyse eine der ersten psycholexikalischen Untersuchungen durch. Das bis heute angewendete, als Grundlage dienende Prinzip besteht darin, zunächst eine umfangreiche und möglichst umfassend vollständige Anzahl von personenbeschreibenden Adjektiven zusammenzustellen. Diese werden dann Testpersonen vorgegeben und sollen von diesen hinsichtlich ausgewählter Zielpersonen eingeschätzt werden. Dazu wurden z. B. von Allport und Odbert (1936) 17953 adjektivische Eigenschaftsbezeichnungen als Ergebnis einer vorangegangenen inhaltlichen Auswertung von Webster's New International Dictionary als Basis verwendet. Die aus einer solchen „Personen-Einschätzung“ entstehende Datenmatrix wird dann einer oder mehreren aufeinander aufbauenden Faktorenanalysen unterzogen, mit dem Ziel, die anfänglich der Anzahl der Adjektive entsprechende Dimensionalität in den Antwortdaten schrittweise zu reduzieren (z. B. Cattell, 1945, 1988b; Guilford & Guilford, 1939a, 1939b; Thurstone, 1934). Dieses methodische Vorgehen begründete innerhalb des Eigenschaftsparadigmas der Persönlichkeitspsychologie den so genannten *lexikalischen Ansatz* bzw. das *lexikalische Eigenschaftsparadigma* (John et al., 1988; Piedmont & Aycock, 2007; Saucier & Goldberg, 2001). Als Ergebnis einer Vielzahl von weiteren empirischen Untersuchungen nach diesem lexikalischen Ansatz, ergaben sich dabei immer wieder in konsistenter Weise fünf orthogonale (Haupt-)Dimensionen der Persönlichkeit. Diese begründeten das Fünf-Faktoren-Modell (FFM – Big-Five-Modell) der Persönlichkeit (Costa & McCrae, 1985, 1992a; McCrae & Costa, 1985, 1987), welches heute die theoretische Grundlage der meisten Persönlichkeitstests bildet (Digman, 1990; Goldberg, 1990, 1992; John & Srivastava, 1999). Das Big-Five-Modell nach McCrae und John (1992) umfasst demnach die folgenden (Haupt-)Dimensionen bzw. Faktoren: *Neuroticism* [*Neurotizismus*, *Ängstlichkeit*], *Extraversion* [*Extraversion*], *Openness to Experience* [*Offenheit*], *Agreeableness*[*Verträglichkeit*] und *Conscientiousness* [*Gewissenhaftigkeit*]. Die Tabelle 2.1 gibt eine Definition dieser fünf Dimensionen oder Faktoren über

(aus dem Englischen übersetze), personenbeschreibende Adjektive nach McCrae und John (1992, S. 178-179, dort Tabelle 1).

Tabelle 2.1 Dimensionen des Big-Five-Modells und beschreibende Adjektive.

Dimension	beschreibende Adjektive
Neurotizismus	ängstlich, selbstbemitleidend, angespannt, empfindlich, labil, bedauernd
Extraversion	aktiv, durchsetzungsfähig, energetisch, enthusiastisch, aufgeschlossen / kontaktfreudig, gesprächig
Offenheit	künstlerisch, neugierig, einfallsreich, einfühlsam, originell, breit interessiert
Verträglichkeit	anerkennend / verständnisvoll, nachsichtig, großzügig, freundlich, mitfühlend, vertrauensvoll
Gewissenhaftigkeit	leistungsfähig, organisiert, planvoll, verlässlich, verantwortungsvoll, gründlich

Anmerkungen: Personenbeschreibende Adjektive in Anlehnung an McCrae und John (1992, S. 178-179, Tabelle 1).

In den unterschiedlichen, fragebogenbasierten Operationalisierungen für das Konstrukt *Persönlichkeit* werden dabei innerhalb der jeweiligen (Haupt-)Dimension manchmal eine unterschiedliche Anzahl von Subdimensionen als *Facetten* der übergeordneten fünf Faktoren angenommen (z. B. Costa, McCrae & Dye, 1991). Diese basieren teilweise auf empirischen Ergebnissen aus faktorenanalytisch ausgewerteten Untersuchungen oder wurden zusätzlich theoriebasiert postuliert (Cattell, 1968; Edwards, 1983; Edwards & Abbott, 1973; John & Srivastava, 1999).

Das Eigenschaftsparadigma und die damit verbundene faktorenanalytische Perspektive im Rahmen des Big-Five-Modells der Persönlichkeit ist gut vereinbar mit dem in der Psychometrie entwickelten Konzept der *latenten Variablen* (vgl. Borsboom, 2008; Bortz & Döring, 2006, S. 206; sowie Abschnitte 1.2 und 1.3). Dabei wird üblicherweise davon ausgegangen, dass eine einfache

lineare (oder additive) Beziehung zwischen der jeweiligen Dimension der Persönlichkeit und den entsprechenden Items besteht, die zur Erfassung des Merkmals verwendet werden (z. B. Cattell & Saunders, 1954a; McCrae & Costa, 1987). Dementsprechend werden daher lineare statistische Modelle - wie das Hauptachsenmodell [Principal Axis Factor Analysis Model – PFA] oder das Hauptkomponentenmodell [Principal Components Model – PCA] (vgl. Pawlik, 1971; Überla, 1977), zur Analyse der Dimensionalität von Fragebogen-Items zur Persönlichkeit, herangezogen. Wie Waller, Tellegen, McDonald und Lykken (1996) betonen, hat diese Praxis der am Eigenschaftsparadigma orientierten Forschung im Bereich der Persönlichkeitspsychologie einige Fortschritte und Erkenntnisse beschert, nicht zuletzt weil solche linear additive Modelle mathematisch leicht handhabbar, in verbreiteter Software leicht implementierbar sind und darüber hinaus psychologisch bedeutsame Ergebnisse liefern. Allerdings propagieren Waller et al. (1996) aber auch den komplementär, ergänzenden Einsatz nichtlinearer Modelle zur Analyse des Antwortverhaltens bei Persönlichkeitsinventaren.

Vor dem Hintergrund der meist auch bei der Konstruktion der Fragebogenverfahren im Bereich Persönlichkeit zugrunde gelegten linear additiven Modelle, ist aus der Perspektive der Indexbildung bzw. Skalierung (vgl. Abschnitt 1.3) die Anwendung der summierten Ratingskalierung (vgl. Borg & Staufenbiel, 2007; Spector, 1992, S. 313; sowie Abschnitt 1.3.1) naheliegend. Auch bei der Skalierung entsprechender Fragebogenskalen im Rahmen der Item Response Theory (IRT; vgl. Kapitel 4 *Psychometrische Modellierung*), wird daher davon ausgegangen, dass zumindest auf der Ebene der einzelnen (Sub-)Skalen jeweils Eindimensionalität besteht (z. B. Rost, 2000; Rost, Carstensen & von Davier, 1999), da ohne diese Annahme die Bildung von einfachen Summenwerten als Indizes für die einzelnen Dimensionen bei der Auswertung der Tests aus messtheoretischer Sicht nicht zu rechtfertigen wäre (Rost, 2002). Bei einer derartigen Summenbildung muss je nach Inventar und konkreter Operationalisierung auf der Ebene einzelner Items darauf geachtet werden, dass die meist mehrstufigen Antwortoptionen der Antwortskalen sämtlicher Items einer Persönlichkeitsdimension eine einheitliche Polung aufweisen. Ausgehend von deren jeweiliger sprachlicher Formulierung müssen vor der Summenwertbildung hier unter Umständen Umkodierungen vorgenommen werden (z. B. Likert, 1932;

Osterlind, 2002) sowie Abschnitt 3.2.2 in dieser Arbeit. Der Einsatz von Items, die hinsichtlich ihrer sprachlichen Formulierung unterschiedlich gepolt sind, lässt sich bei Skalen zur Erfassung von Dimensionen der Persönlichkeit (bspw. für die Dimension *Extraversion*) unterschiedlich begründen. Einerseits besteht die Intention darin, dadurch negative Auswirkungen von Antworttendenzen auf die Messung zu verringern (vgl. ausführlicher dazu in Abschnitt 3.2.2). Andererseits wird der gemischte Einsatz sprachlich unterschiedlich formulierter Items bei Persönlichkeitsskalen mit deren antagonistisch, bipolaren Natur begründet, wie sie beispielsweise für die Dimension *Extraversion* bereits von Jung (1921) in seiner psychoanalytischen Typenlehre angelegt wurde. Jung (1921) definierte dabei *Extraversion* und *Introversion* als zwei antagonistische psychischen Grundfunktionen, welche im Rahmen des Eigenschaftsparadigmas als jeweilige Endpunkte einer bipolaren Dimension (*Extraversion*) aufgefasst werden können (vgl. Abschnitt 2.1.1 im Kapitel 2.1 *Persönlichkeit und interindividuelle Unterschiede*). Eine detailliertere Diskussion solcher Fragen der Operationalisierung auf der Ebene einzelner Items, deren sprachlicher Formulierung und deren (unterschiedlich gepolter) Antwortskalen wird in Kapitel 3 *Theoretischer Hintergrund zu Antwortmustern* und dort insbesondere in Abschnitt 3.2.2 gegeben.

Verbreitete Operationalisierungen des Big-Five-Modells der Persönlichkeit in Form von (umfangreichen) Fragebogeninventaren sind im deutschsprachigen Raum das *NEO-Persönlichkeits-Inventar* (NEO-PI-R – Ostendorf, 2004) oder das *NEO-Fünf-Faktoren-Inventar* (NEO-FFI – Borkenau & Ostendorf, 1993, 2008). Beide Fragebogenverfahren erfassen mit über hundert Items neben den fünf Hauptdimensionen weitere, untergeordnete Persönlichkeitsfacetten nach dem Eigenschaftsparadigma. In der vorliegenden Arbeit wird ein Variante (vgl. Schmolck, 2003, 2004, 2005, 2006a, 2006b) des von Rammstedt und John (2005) publizierten *BFI–K* eingesetzt, welcher eine aus dem Englischen übersetze Kurzversion (Rammstedt, 1997) des ursprünglich von John, Donahue und Kentle (1991) entwickelten, ursprünglich 44 Items umfassenden *Big-Five-Inventory* ist. Eine detaillierte Darstellung dieses Fragebogeninventars wird in Kapitel 5 *Stichproben und Instrumente* in Abschnitt 5.1.1 gegeben.

2.1.4 Kritische und integrative Perspektiven auf die Psychologie der Persönlichkeit

Die behavioristischen Perspektiven der Psychologie in den Sechziger und Siebziger Jahren (z. B. Skinner, 1965) lieferten die Basis für einen kritischen Blick auf die Psychologie der Persönlichkeit. Als Gegenkonzept zu den, über introspektive Zugänge erlangten, teils vagen Persönlichkeitskonzepten der Anfang des zwanzigsten Jahrhunderts noch vorherrschenden psychoanalytischen Sichtweise auf menschliches Verhalten (Asendorpf & Neyer, 2012, S. 40) propagierte Skinner eine auf (objektive) empirische Beobachtung begründete Perspektive zur Erklärung menschlichen Verhaltens (Skinner, 1965). Skinner stützte seine Untersuchungen und Theorien zu menschlichem Verhalten und Erleben auch auf Verhaltensexperimente an Tieren. Skinner entwickelte so seine Theorie zur operanten Konditionierung weiter (Skinner, 1963), die sich an die frühen Untersuchungen von Iwan Petrowitsch Pawlow (1849–1936) zur klassischen Konditionierung und einfachen Reiz-Reaktions Schemata (vgl. z. B. Babkin, 1949), anschloss. Innere psychische Prozesse oder eben auch nicht direkt beobachtbare, das Verhalten steuernde Konstrukte wie *Persönlichkeit*, wurden im Rahmen der streng behavioristischen Sichtweise als Teil einer „Black Box“ betrachtet, deren wissenschaftlich empirische Untersuchung (mangels objektivierbarer Zugänge) daher wenig sinnvoll erschien (Asendorpf & Neyer, 2012, S. 40). Daneben vertraten führende Theoretiker des sozialen Lernens wie beispielsweise Mischel (1968, 1973) und Mischel und Shoda (1995) in einflussreichen Publikationen die Position, dass die Beurteilung und Klassifikation von Aspekten der Persönlichkeit wenig prädiktive Validität im Hinblick auf beobachtbares Verhalten mit sich bringe. Nach einer sicher extrem überspitzten Formulierung dieser kritischen Perspektive gibt es gar keine Persönlichkeit (vgl. z B. Goldberg, 1993) – wie sich Menschen verhalten und was sie tun, hängt vielmehr von den jeweiligen Situationen ab, in denen sie sich befinden und nicht von irgendwelchen überdauernden Dispositionen oder Eigenschaften. Allerdings muss zu dieser überspitzten Sichtweise einschränkend angemerkt werden, dass Mischels (1968) zentrale These letztlich auch als Gegenposition zu der radikal behavioristisch anmutenden These „keine Persönlichkeit“, aufgefasst werden kann. Bei der Überprüfung der empirischen Literatur kam Mischel zu dem Schluss,

dass beispielsweise soziales Verhalten von Situation zu Situation erhebliche Varianz oder auch Diskrepanz zeigt. Mischels (1968) Schlussfolgerung war, dass die menschliche Persönlichkeit so reich und komplex ist, dass es daher keine einfache Menge von universellen Merkmalsdimensionen geben kann, die zur Beschreibung der menschlichen Persönlichkeit angemessen ist (Shadel & Cervone, 1993). So zieht Mischel (1968) am Ende seines Buches „Personality and Assessment“ bezüglich eigenschaftsorientierter Persönlichkeitsmodelle das folgende kritische Fazit (vgl. auch Mischel, 2004, S. 18):

> The traditional trait-state conceptualizations of personality, while often paying lip service to man's complexity and to the uniqueness of each person, in fact lead to a grossly oversimplified view that misses both the richness and the uniqueness of individual lives. (Mischel, 1968, S. 301)

Daneben kritisierten humanistische Psychologen aus einer anderen theoretischen Perspektive die moralischen Aspekte der Verwendung von Persönlichkeitsmodellen zur Klassifizierung von Personen. Das *Humanistische Paradigma* der psychologischen Persönlichkeitstheorien bildete sich Mitte der Sechziger Jahre als Reaktion auf die damals vorherrschende analytische und behavioristische Perspektive der Psychologie. Die humanistischen Perspektiven auf Persönlichkeit leiten sich hauptsächlich aus den Schriften von Maslow (1962, 1969) und C. R. Rogers (1946, 1959, 1961) ab, die propagierten, dass Menschen nur dadurch verstanden werden können, dass sie lernen, wie sie sich selbst erfahren, und nicht durch äußere Beobachtungen dessen, was sie sagen und tun. Auch in Abgrenzung zu Freuds eher universellem Strukturmodell der psychoanalytischen Persönlichkeit (vgl. Freud, 1923, 1933), betont C. R. Rogers (1946) im Rahmen seines klientenzentrierten Therapie Ansatzes die Einzigartigkeit des Individuums und dessen individuelle Realität (C. R. Rogers, 1946, 1947). Aus seinem therapeutisch orientierten Ansatz leitete C. R. Rogers (1961) die Grundgedanken seiner personzentrierten Persönlichkeitstheorie ab, wonach sich die individuelle Persönlichkeit aus dem grundlegenden menschlichen Bedürfnis nach Autonomie, Selbstverwirklichung und Selbstaktualisierung ergibt. Aus solch einer humanistischen Sichtweise war die Klassifizierung von Personen nach Persönlichkeits- oder Verhaltensmerkmalen, die sie mit anderen Menschen teilen, nicht nur „Zeitverschwendung“, sondern auch ein entmenschlichendes

Verfahren, das die Menschen ihrer individuellen Würde beraubt.

Die kritischen Perspektiven des Behaviorismus und Humanismus stellten insgesamt eine große Herausforderung für die eigenschaftsorientierte Persönlichkeitspsychologie dar. Wie von Carlson (1975) beschrieben, führten diese kritischen Sichtweisen auf den Bereich der Persönlichkeitspsychologie dazu, dass dieses Untersuchungsgebiet in den Sechziger Jahren „*praktisch verschwand*“ (Carlson, 1975, S. 393). Ein interessanter Ansatz zur Integration der am Verhalten orientierten, behavioristischen Sichtweise mit dem weiter oben beschriebenen *Eigenschaftsparadigma* der Persönlichkeit wird 1983 von D. M. Buss und Craik unter dem Begriff *act frequency approach* [dt. etw. *Handlungshäufigkeitsansatz*] propagiert. D. M. Buss und Craik (1983) definieren Dispositionen oder Eigenschaften dabei als Zusammenfassungen von Handlungshäufigkeiten, wobei die einzelnen Handlungen (das einzelne Verhalten) selbst keinen erklärenden Status besitzt. Eigenschaften, Traits oder Dimensionen der Persönlichkeit werden nach diesem theoretischen Modell als soziokulturelle Emergenzen und als natürliche kognitive Kategorien mit einzelnen Handlungen als deren Entitäten angesehen. Die Kategoriegrenzen bleiben dabei eher unscharf und teilweise überlappend und die einzelnen Vertreter der Kategorien (einzelne Handlungen / Verhalten) unterscheiden sich im Hinblick auf ihre Typikalität für die jeweilige Kategorie bzw. Eigenschaft (D. M. Buss & Craik, 1983). Dieses von D. M. Buss und Craik (1983) propagierte Modell lässt sich auch als untermauerndes Erklärungsmodell für die in Ansätzen bereits von Galton (1884) formulierte Sedimentationshypothese ansehen (vgl. auch Abschnitt 2.1.3), welche den lexikalischen Ansatz als empirische Grundlage des Eigenschaftsparadigmas der Persönlichkeit begründet (John et al., 1988).

In der Auseinandersetzung zwischen der eher behavioristisch orientierten Sichtweise auf menschliches Verhalten und des eigenschaftsorientierten Paradigmas der Persönlichkeit entwickelte sich die *Person – Situation Debatte* (vgl. Mischel, 1968, 2009). Kern dieser inzwischen, auch zugunsten interaktionistischer Persönlichkeitsmodelle (vgl. Cattell, 1980; Endler & Magnusson, 1976; Murtha, Kanfer & Ackerman, 1996; Rushton & Endler, 1977; Shoda, 1999) beendeten Debatte (z. B. Fleeson & Noftle, 2008; Funder, 2001; Kenrick & Funder, 1988), war die Frage, ob eher stabile Personeneigenschaften und Dispositionen (*Traits*) oder aber situative Variablen bei der Vorhersage

menschlichen Verhaltens bestimmend sind (Fleeson & Noftle, 2008; Rauthmann, 2014). Ausgangspunkt dieser Debatte war die empirische Beobachtung, dass menschliches Verhalten eine vergleichsweise niedrige transsituative Konsistenz aufweist (Asendorpf & Neyer, 2012, S. 28) aus der Mischel (1968) ursprünglich folgerte, dass das menschliche Verhalten eher durch Situationen als durch Persönlichkeitseigenschaften bestimmt wird (vgl. auch Mischel & Peake, 1982). Allerdings lieferten beispielsweise Funder und Colvin (1991) einige Evidenz für eine relativ hohe situationsübergreifende Konsistenz von Verhalten unter der Voraussetzung, dass dieses reliabel und auf einem angemessenen Abstraktionsgrad erfasst wird. Frühere Studien hatten Verhalten demnach (zu) sehr kleinteilig auf der Ebene einzelner konkreter Verhaltensmanifestationen im Gegensatz zur Ebene der angemessenen psychologischen Bedeutung erfasst. In diesem Sinne stellt der bereits weiter oben erwähnte *act frequency* Ansatz von D. M. Buss und Craik (1983) ein integrierend, erklärendes Modell dar, mit dem sich letztlich auch die Konstruktion und der Einsatz *mehrerer* (verhaltensnah formulierter) Items zur Erfassung einer Eigenschaftsdimension im Rahmen des Big-Five-Modells begründen lässt. Darüber hinaus zeigten Shoda, Mischel und Wright (1994), dass eine niedrige transsituative Konsistenz dennoch mit einer hohen Stabilität von individuellen Situationsprofilen vereinbar ist. Demnach zeigen Menschen ein zwar nicht notwendigerweise konsistentes, aber ein dennoch individuell typisches Muster von situationsübergreifender Variabilität (Shoda & Mischel, 2000). Ferner zeigte sich, dass Personen individuelle und stabile Muster bei *Verhalten*×*Situation*–Interaktionen, aufweisen (Mischel, Shoda & Mendoza-Denton, 2002). Basierend auf den Ergebnissen aus einer Untersuchung an eineigen Zwillingen erweiterten Borkenau, Riemann, Spinath und Angleitner (2006) das *Person*×*Situation*–Interaktionsmodell zusätzlich um eine genetische Komponente. Auf Basis ihrer Befunde schlussfolgern Borkenau et al. (2006), dass die genetisch Komponente 25% der Varianz der *Person*×*Situation*–Profile erklärt. Solch ein Befund steht weitgehend im Einklang mit dem von Waller und Reise (1992) berichteten Befund, dass sich etwa 20 % der Variabilität der Skalierbarkeit von Antwortreaktionen auf Fragebogen-Items auf genetische Faktoren zurückführen lassen.

2.2 Berufliche Interessenorientierungen und das Modell von Holland

Die Entwicklung des Modells der *beruflichen Interessenorientierungen* von J. L. Holland (1959; 1997) fußt auf der Grundlage der sich Anfang der Zwanziger Jahre entwickelnden berufspsychologischen Forschungsrichtung, in deren Rahmen die Entwicklung diagnostischer Verfahren zur Erfassung von Interessenprofilen und deren Anwendung in der Berufsberatung als pragmatische Zielsetzungen standen (vgl. Prenzel, 1988; Strong, 1943). Das Modell *beruflicher Interessenorientierungen* wurde vor diesem Hintergrund zunächst auf Basis von J. L. Hollands Erfahrungen aus seiner Zeit bei den Amerikanischen Streitkräften als Interviewer für Rekruten und verschiedener weiterer Tätigkeiten im Kontext der Berufsberatung (vgl. Wilson et al., 2008) vor über 50 Jahren als Theorie zur Berufswahl entwickelt (Holland, 1959, 1963; Holland, Krause, Nixon & Trembath, 1953). Dabei schlug J. L. Holland als Ergebnis seiner persönlichen Erfahrungen im Bereich der Berufsberatung sowie empirischen Untersuchungen an College-Studenten ein System der Berufsklassifikation mit schließlich sechs Dimensionen vor – *Realistic*, *Investigative*, *Artistic*, *Social*, *Enterprising* und *Conventional* (Holland, Whitney, Cole & Richards, 1969).

Holland selbst entwickelte als empirische Grundlage seiner Theorie das *„Vocational Preference Inventory“* (VPI – Holland et al., 1953) dessen 160 Items im Grunde zunächst eine Auswahl von Berufsbezeichnungen waren. Die empirische Anwendung einer ersten Fassung des VPI erzielte Ergebnisse, die vergleichbar waren mit denen anderer bereits existierender Inventare (Wilson et al., 2008). So zum Beispiel dem in den Zwanziger und Dreißiger Jahren entwickelten *„Strong Vocational Interest Blank“* (SVIB – Strong, 1943), einem Inventar zur Erfassung von beruflichen Orientierungen, welches, analog zu dem oben beschriebenen Ansatz im Rahmen des Eigenschaftsparadigmas für den Konstruktbereich Persönlichkeit, eine faktorenanalytisch, empirische Grundlage hat (vgl. D. P. Campbell, Borgen, Eastes, Johansson & Peterson, 1968; D. P. Campbell & Holland, 1972).

Die empirischen Untersuchungen von Holland stützen sich auf zwei groß angelegte Erhebungen des „US Departments of Health, Education & Welfare“, die jeweils im Frühjahr und Herbst des Jahres 1965 durchgeführt wurden.

Wie Holland selbst schreibt, konnten dabei zwei im strengen Sinne nicht repräsentative Stichproben von College-Studenten erhoben werden: „*Although these student samples are not precise representative samples, they appear to be reasonable approximations of the typical college freshman*“, (Holland, 1966, S. 279). Allerdings umfassten die beiden Stichproben immerhin $n = 12432$ und $n = 10646$ Studenten, welche an insgesamt 40 Hochschulen in verschiedene US-Bundesstaaten getestet wurden. Die Teilnahmeraten an den jeweiligen Hochschulen lagen zwischen 22% und 96% wobei allerdings der genaue Mechanismus des Datenausfalls nicht bekannt ist: „*The specific bias due to nonrespondents is not known*“ (Holland, 1966, S. 279). Das weitere Vorgehen beschreibt Holland damit, dass den Studenten Listen mit Berufsbezeichnungen (vgl. Holland, 1958) aus dem Vocational Preference Inventory (VPI – Holland, 1965, Fifth Revision), vorgelegt wurden, mit der Bitte diese auf einer zweistufigen Antwortskala „*gefällt mir*“ und „*gefällt mir nicht*“ einzuschätzen. Die Konstruktion des VPI beschreibt Holland (1958) in einem früheren Beitrag zunächst als einen Prozess der intensiven Literatur Recherche im Forschungsbereich der Interessen und Berufswahl mit einem Bezug zu Faktoren der Persönlichkeit. Als Ergebnis wurden daraus acht a priori Skalen konstruiert - „*Physical Activity*“, „*Intellectuality*“, „*Responsibility*“, „*Conformity*“, „*Verbal Activity*“, „*Emotionality*“, „*Reality*“ „*Orientation*“ sowie die Kontrollskala „*Acquiescence*“. Den inhaltlichen Skalen wurden dann verschiedene Berufsbezeichnungen auf Basis theoretischer Vorüberlegungen zugeordnet. Bei mehreren Revisionen des Verfahrens durch Analysen zur internen Konsistenz resultierten jeweils verschiedene Versionen des Inventars mit einer unterschiedlichen Anzahl von Skalen (Holland, 1958). Im Projektbericht zu den beiden US-amerikanischen Erhebungen von Abe, Holland, Lutz und Richards (1965, S.11) wurde dazu von den Autoren eine Interpretationshilfe sowie inhaltliche Definition zu den aus dem VPI extrahierten acht Skalen gegeben. Die Autoren des Projektberichts weisen dabei darauf hin, dass das VPI, im Hinblick auf den deskriptiven Charakter der Erhebungsdaten, lediglich als Inventar für berufliche Interessenorientierungen zu interpretieren ist. Die dabei gegebenen Definitionen lauteten im Englischen Original wie folgt:

Scale	Preference for:
Realistic	technical and skilled trades

Intellectual	scientific occupations
Social	teaching and helping occupations
Conventional	clerical occupations
Enterprising	supervisory and sales occupations
Artistic	artistic, musical, and literary occupations
Status	prestigeful occupations
	such as Lawyer, Doctor, Business Executive
Acquiescence	number of preferred occupations

Wie diese Liste mit den Definitionen der Skalen zunächst zeigt, sind bereits in dieser frühen Form des beruflichen Interessenmodells fünf Dimensionen mit ihrer auch später gewählten Bezeichnung enthalten. Auch die sechste im späteren Modell mit *Investigative* bezeichnete Skala lässt sich hier bereits, wenn auch unter der anderslautenden Bezeichnung *Intellectual*, so doch immerhin inhaltlich weitgehend konsistent finden. Interessanterweise wird die Interessenorientierung *Conventional* mit der Skalenbeschreibung „*clerical occupations*“ hier nur mit „Bürotätigkeit“ bzw. „Büroberuf“ beschrieben, was im Vergleich zu den später gegebenen Definitionen eher eine recht unscharfe Beschreibung dieser Interessendimension darstellt. Die in dieser Liste noch enthaltenen Skala *Status* ist in Hollands späterem Interessenmodell dagegen nicht mehr enthalten – ebenso wie die (eher methodisch orientierte) Skala *Acquiescence*. Obwohl J. L. Holland selbst von Beginn an die Berufswahl als Ausdruck der Persönlichkeit sah (Holland, 1973) und somit seine Theorie auch als Modell für individuelle *Persönlichkeitsorientierungen* angesehen werden kann (Holland, 1999), soll in dieser Arbeit, zur besseren Abgrenzung zum bereits dargestellten Big-Five-Modell, weiterhin vom Modell *beruflicher Interessenorientierungen* gesprochen werden.

J. L. Holland selbst ergänzte und erweiterte sein Modell im weiteren Verlauf um verschiedene zusätzliche Aspekte. So zum Beispiel im Hinblick auf die Zusammenhänge zwischen den postulierten sechs Dimensionen (Holland, 1971), den Möglichkeiten der beruflichen Entwicklung (Holland, 1973) und um Aspekte beruflicher Arbeitsumwelten (Holland, 1985, 1997). Insgesamt gesehen ist das Modell von Holland im Laufe seiner Entwicklung hinsichtlich seines Theoriegehaltes immer reichhaltiger geworden (Gottfredson, 1999). Einen aktuellen

Überblick über diese allmähliche „Evolution“ und die damit einhergehenden Forschungstätigkeiten rund um das Holland-Modell gibt Nauta (2010). In seiner letzten Fassung beinhaltet J. L. Hollands Theorie zunächst fünf Grundannahmen bzw. Theoreme (vgl. Holland, 1997), die sich auf die folgenden Punkte beziehen, welche auf der Ebene der Operationalisierung für den deutschsprachigen Raum im AIST–R Inventar von Bergmann und Eder (2005) umgesetzt ist:

1. **Die Personenorientierungen:**

   Die Grundlegende Annahme des Modells von Holland besteht darin, dass sich im westlichen Kulturkreis die folgenden sechs grundlegenden Interessenorientierungen unterscheiden lassen:

   Die *praktisch-technische* Orientierung (Realistic – **R**)

   die *intellektuell-forschende* Orientierung (Investigative – **I**)

   die *künstlerische* Orientierung (Artistic –**A**)

   die *soziale* Orientierung (Social – **S**)

   die *unternehmerische* Orientierung (Enterprising –**E**)

   und die *konventionelle* Orientierung (Conventional –**C**)

   Holland nimmt darüber hinaus an, dass die meisten Personen aus dem westlichen Kulturkreis einer dieser Personenorientierungen als Haupttyp zugeordnet werden können. Allerdings geht Holland davon aus, dass bei einer einzelnen Person, neben dem Haupttyp, auch noch weitere Personenorientierungen mehr oder weniger stark ausgeprägt vorliegen können. Im Zusammenhang mit der Typisierung einzelner Personen (und auch der beruflichen Umwelt, vgl. Punkt 2) wurden daher verschiedene Ansätze zur Klassifikation vorgeschlagen, welche in unterschiedlichem Ausmaß die Haupt- und Nebenorientierungen einer Person berücksichtigen. Holland selbst schlägt vor, Personen zunächst durch den Anfangsbuchstaben der am stärksten ausgeprägten Personenorientierung zu charakterisieren (der Holland *one-letter-code*). Für eine differenziertere Typisierung schlägt er die Verwendung der Anfangsbuchstaben der zwei, bzw. drei am stärksten

ausgeprägtesten Personenorientierungen in der entsprechenden Reihenfolge als so genanten drei-Buchstaben-Code bzw. *three-letter-code* vor. Bei einer Person mit dem Holland-Code „AIS“ liegt in diesem Sinne ihre höchste Ausprägung im Bereich der *künstlerischen*, die zweithöchste im Bereich der *intellektuell-forschenden* und die dritthöchste im Bereich der *sozialen* Personenorientierung. Aus der Kombination der sechs Personenorientierungen sind so bei einem einstelligen Klassifikationsansatz sechs, bei einem zweistelligen Klassifikationsansatz 30 und bei einem dreistelligen Klassifikationsansatz 120 verschiedene Personentypen möglich.

2. **Die Umweltorientierungen:**

   Bezüglich der beruflichen Umwelt nimmt Holland an, dass sich diese ebenfalls anhand derselben sechs Dimensionen wie bei den Personenorientierungen beschreiben lässt. Es bestehen demnach also auch *praktisch-technische* (Realistic – R), *intellektuell-forschende* (Investigative – I), *künstlerische* (Artistic – A), *soziale* (Social – S), *unternehmerische* (Enterprising – E) und *konventionelle* (Conventional – C) Umweltorientierungen. Eine dabei implizit getroffene Annahme besteht darin, dass sich alle Berufe anhand dieser Orientierungen – oder auch durch eine Kombination der sechs Umweltorientierungen – klassifizieren lassen.

   Jede nach den sechs Umweltorientierungen beschriebene berufliche Umwelt bietet den Raum für Aktivitäten, die mit der jeweiligen Personenorientierung korrespondieren und stellt auch entsprechende Anforderungen an die in ihr tätigen Personen. Eine *künstlerische* Umweltorientierung bietet demnach den Menschen, die in ihr arbeiten, offene, unstrukturierte Aktivitäten und ermöglicht das Schaffen kreativer Produkte. Sie erfordert Kreativität, Ideenreichtum und Ausdrucksfähigkeit und lässt unkonventionelles Handeln und Denken zu. In Verbindung mit dem Theorem der sechs spiegelbildlich konzeptualisierten Personenorientierungen können die Personen- und Umweltorientierungen damit direkt aufeinander bezogen werden. Die theoretische Grundlage für die den Personenorientierungen spiegelbildlich gegenüberstehenden Umweltorientierungen sieht Holland in deren historischen Entstehungsprozess verankert. Dabei sorgt ein systematischer Prozess der Auslese dafür, dass einerseits aus einzel-

nen Berufen diejenigen Tätigkeiten ausgelagert werden, die nicht zu den Personen passen die sie ausüben und andererseits diejenigen Personen berufliche Umwelten verlassen, deren Tätigkeiten oder Anforderungen nicht ihren Personenorientierungen entsprechen.

3. **Der strukturelle Zusammenhang der Personen- und Umweltorientierungen im RIASEC-Modell:**

   Für die oben beschriebenen sechs Personen- und Umweltorientierungen wird im Modell von Holland (1997) ein geometrisch, struktureller Zusammenhang angenommen. Die in der *Calculus-Hypothese* zum Ausdruck gebrachte hexagonale Repräsentation der beruflichen Interessen, bildet die Ähnlichkeitsbeziehung der sechs Dimensionen ab. Die zirkuläre Reihenfolge der sechs Dimensionen „RIASEC“ (wobei „C“ wiederum nahe bei „R“ steht) gab der Theorie beruflicher Interessenorientierungen auch den Namen „Hexagon-Modell der beruflichen Interessen“. Diese „räumliche“ Anordnung im Hexagon repräsentiert dabei die psychologische Nähe der einzelnen Interessenorientierungen (vgl. Abbildung 2.1). Nahe nebeneinanderliegende Interessenorientierungen haben einen ähnlichen und gegenüberliegende Interessenorientierungen haben einen gegensätzlichen Charakter.

4. **Die Zuordnung von Personen und Umwelt(en):**

   Personen streben nach dieser Annahme in eine berufliche Umwelt, die es ihnen erlaubt, die eigenen Fähigkeiten und Interessen zu realisieren. Die Personen können darin die Tätigkeiten auszuüben, welche ihrem Interessentyp (Persönlichkeitstyp) entsprechen. Personen, die z. B. dem künstlerischen Typ zuzuordnen sind, streben daher nach beruflichen Umwelten, in denen sie sich künstlerisch verwirklichen können. Dieses in der Person liegende Bestreben bildet die Kernannahme von Hollands Berufswahltheorie.

5. **Die Wechselwirkung zwischen Personen und Umwelt(en):**

   Aus der Wechselwirkung zwischen Personenorientierungen und Umweltorientierungen lassen sich Aussagen über das Verhalten der Person ableiten. Diese Verhaltensvorhersage bezieht sich auf unterschiedliche berufli-

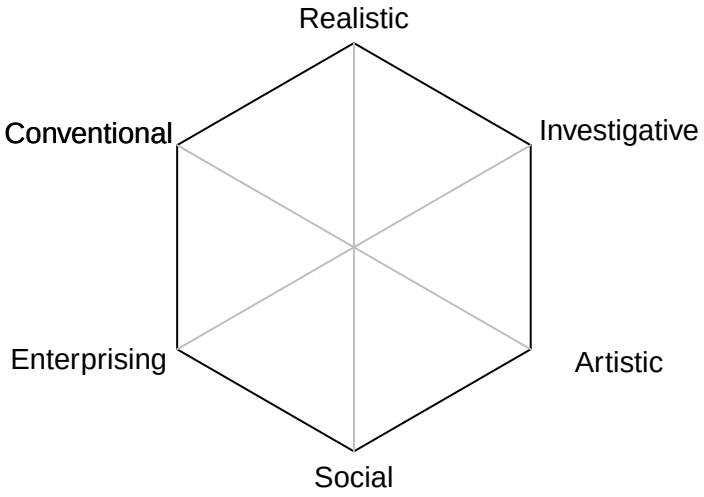


Abbildung 2.1 Schematische Darstellung zur räumlichen, hexagonalen Anordnung der beruflichen Interessenorientierungen im Modell von Holland.

che und persönliche Bereiche wie z. B. auf Leistung, Erfolg, Zufriedenheit und Stabilität im Beruf. Personen, deren berufliche Umweltorientierungen mit ihrer Personenorientierung übereinstimmt, haben eine höhere Wahrscheinlichkeit, in ihrem Beruf eine hohe Leistung zu erbringen, erfolgreich zu sein, zufrieden zu werden und langfristig in diesem Beruf zu bleiben.

Neben diesen fünf Grundannahmen bzw. Theoremen spezifiziert Holland nun vier Sekundärkonstrukte, die einerseits Aussagen bezüglich der Wechselwirkungen zwischen Personen und Umwelt(en) und über die Beschaffenheit der Interessenstruktur der Personen erlauben. Diese Sekundärkonstrukte ha-

ben somit vor allem praktische Implikationen in Bezug auf die Anwendung und die Vorhersage auf Basis des Models. Es sind die Sekundärkonstrukte *Kongruenz*, *Konsistenz*, *Differenziertheit* und die *(berufliche) Identität*, welche im Folgenden kurz beschrieben werden sollen:

- **Die Kongruenz**

  Das Sekundärkonstrukt *Kongruenz* bezieht sich auf das Ausmaß der Übereinstimmung (Passung) zwischen den Orientierungen der Person und den Orientierungen der (beruflichen) Umwelt. Eine hohe Kongruenz besteht beispielsweise für eine Person, die sich in einer exakt ihrem Persönlichkeitsmuster entsprechenden beruflichen Umwelt befindet. Zur Bestimmung der Person–Umwelt Passung kann die im RIASEC-Modell angenommene, hexagonale Struktur der sechs Dimensionen berücksichtigt werden. Befindet sich z. B. eine „R-Person“ in einer „R-Umwelt“, besteht eine hohe Kongruenz. Etwas geringer fällt die Kongruenz aus, wenn sich die „R-Person“ in einer Umwelt befindet, welche durch eine Umweltorientierung gekennzeichnet ist, die im hexagonalen Modell der „R-Orientierung“ benachbart ist – z. B. eine „I-Umwelt“. Demgegenüber würde die Kongruenz sehr niedrig ausfallen, wenn sich die „R-Person“ in der gegensätzlich orientierten „S-Umwelt“ befindet. Zur Operationalisierung der Kongruenz sind eine Reihe von Indizes entwickelt worden. Typischerweise stützt sich die Berechnung solcher Indices auf einen ein-, zwei- oder dreistelligen Personen- und Umwelt-Code, wobei häufig der dreistellige Code („3-letter-code“) verwendet wird. Häufig eingesetzte Indices sind der Zener-Schnuelle-Index (Zener & Schnuelle, 1976), der Iachan-Index (Iachan, 1984) und der N3-Index (Joerin Fux, 2003).

- **Die Konsistenz**

  Das Sekundärkonstrukt *Konsistenz* bezieht sich auf die innere Struktur einer individuellen Personen- oder Umweltorientierung. Ein konsistentes Orientierungsmuster besteht dann, wenn die am stärksten ausgeprägten Orientierungen im hexagonalen Modell nebeneinanderliegen (z. B. IAS). Demgegenüber besteht ein wenig konsistentes Orientierungsmuster, wenn die ausgeprägtesten Orientierungen in der hexagonalen Anordnung gegenüberstehen (z. B. IER). Das gleiche Prinzip kann auf berufli-

che Umwelten übertragen werden, die sich ebenfalls in ihrer Konsistenz unterscheiden können. Eine Operationalisierung, welche allerdings die Gültigkeit der Calculus-Hypothese voraussetzt, besteht darin, den Orientierungsmustern (den Holland-Codes) Zahlenwerte zuzuordnen, die den Distanzen in der hexagonalen Anordnung entsprechen. im Hexagon nebeneinanderliegende Orientierungsmuster (z. B. IA) erhalten dabei einen hohen Wert (hohe Konsistenz) und gegenüberliegende Muster (z. B. RS) einen niedrigen Wert.

- **Die Differenziertheit**

  Das Sekundärkonstrukt *Differenziertheit* steht für die Prägnanz und Eindeutigkeit des individuellen, als Profil gedachten, beruflichen Orientierungsmusters. Ein Profil der beruflichen Interessenorientierungen fällt dann differenziert aus, wenn es klare Höhen und Tiefen aufweist. Demgegenüber sind bei wenig differenzierten Personen die Interessenorientierungen etwa gleich schwach oder stark ausgeprägt. Ein einfacher Index für die Differenziertheit errechnet sich daher aus der Differenz zwischen dem Maximum und Minimum eines Interessenprofils. Andere Indices für die Differenziertheit stützen sich auf Maße der Streuung oder die Berechnung eines Differenziertheitsvektors (vgl. z. B. Bergmann & Eder, 2005).

- **Die berufliche Identität**

  Eine hohe berufliche Identität besteht aus einer klaren Vorstellung zu eigenen (beruflichen) Interessen, Fähigkeiten und Talenten (z. B. Holland, Gottfredson & Power, 1980). Das Ausmaß der beruflichen Identität kann mit Verfahren operationalisiert werden die nicht notwendigerweise am Modell der beruflichen Interessenorientierungen anknüpfen müssen. Die von Holland et al. (1980) vorgeschlagenen Skalen zur „Entscheidungsfindung und Persönlichkeit“ erfassen beispielsweise mit drei Dimensionen die berufliche Identität, den Bedarf für berufliche Informationen sowie persönliche Grenzen und Hindernisse bezüglich einer beruflichen Entscheidungsfindung (vgl. Holland et al., 1980; Holland, Johnston & Asama, 1993). Insgesamt weist das Identitätskonzept Bezüge zu den anderen beiden Sekundärkonstrukten Konsistenz und Differenziertheit auf.

Eng verknüpft mit der fortschreitenden Entwicklung der Theorie zum Modell der *beruflichen Interessenorientierungen*, verläuft auch die Entwicklung entsprechender Operationalisierungen zur Erfassung der beruflichen Interessenorientierungen und der beruflichen Umwelt. Im Hinblick auf eine individuelle Diagnostik sei hier z. B. auf das Inventar *„Self Directed Search“* – SDS (Holland, 1971, 1979), das *„Vocational Preference Inventory“* – VIP, (Holland, 1965, 1975; Taber, 2006), sowie speziell für den deutschsprachigen Raum auf das Inventar *EXPLORIX*, (Joerin Fux, Stoll, Bergmann & Eder, 2003) verwiesen. Weitere Informationen zum *EXPLORIX* sind in der Publikation von Singer, Decker und Glaesmer (2007) zu finden.

Ein sowohl im Bereich der Berufsberatung, als auch im Bereich der Forschung zu beruflichen Interessenorientierung weit verbreitetes Instrument ist der *Allgemeine-Interessen-Struktur-Test* in seiner revidierten Fassung (AIST–R – Bergmann & Eder, 2005). Dieses auch in der vorliegenden Arbeit eingesetzte Fragebogenverfahren erfasst mit insgesamt 60 Items die sechs Dimensionen beruflicher Interessenorientierungen nach Holland (1997). Eine detaillierte Darstellung des AIST–R von Bergmann und Eder (2005) wird in Abschnitt 5.1 in Kapitel 5 *Stichproben und Instrumente* gegeben.

2.3 Präferenzen des Musikgeschmacks

Die Untersuchung der Auswirkungen von Musik auf menschliches Erleben und Verhalten, sowie die Systematisierung unterschiedlicher Präferenzen zu verschiedenen Musikrichtungen hat in der psychologischen Forschung zu interindividuellen Unterschieden eine vergleichsweise lange Tradition (z. B. Bever, 1988; Cattell & Anderson, 1953; Cattell & Saunders, 1954b; Chamorro-Premuzic, Swami, Furnham & Maakip, 2009; Clark & Giacomantonio, 2013; Delsing, ter Bogt, Engels & Meeus, 2008; Langmeyer, Guglhör-Rudan & Tarnai, 2012; Levitin, Grahn & London, 2017; North & Hargreaves, 1996; Radocy & Boyle, 2003; Randall & Rickard, 2017; Rentfrow & Gosling, 2003; Rigg, 1937; Rohner, 1985; Schäfer, 2008; Schäfer & Mehlhorn, 2017; Seashore, 1938; Silvia, Fayn, Nusbaum & Beaty, 2015; J. Sloboda, 1986; Tan, Pfordresher & Harré, 2010; Tekman & Hortaçsu, 2002; Wing, 1941; M. Yamamoto, Naga & Shimizu, 2007; Zweigenhaft, 2008)

Bereits die frühen Arbeiten beziehen sich dabei auf die unterschiedlichsten Aspekte der *Psychologie der Musik* (Seashore, 1938). So wird beispielsweise der Frage nachgegangen in wie weit – im faktorenanalytischen Sinne – ein Generalfaktor Musikalischer Kompetenz [*musical ability*] besteht (Wing, 1941). Andererseits werden im Kontext von Fragestellungen im Bereich der Differentiellen Psychologie und Persönlichkeitspsychologie typischerweise die kognitiven (J. Sloboda, 1986), emotionalen (Rigg, 1937; Rohner, 1985; J. A. Sloboda, 1991; Zentner, Grandjean & Scherer, 2008) und affektiven (Clark & Giacomantonio, 2013; Litle & Zuckerman, 1986) Auswirkungen musikalischen Erlebens untersucht. So untersucht bereits Rigg (1937) den Einfluss von unterschiedlichen musikalischen Hörbeispielen auf Emotionen und Rohner (1985) den Zusammenhang zwischen kognitiver Komplexität, emotionaler Verarbeitung und unterschiedlichen musikalischen Stimuli. Zentner et al. (2008) analysieren in vier Studien die Struktur unterschiedlicher Emotionen in Abhängigkeit verschiedener musikalischer Stimuli. Als Ergebnis der Untersuchungen schlagen Zentner et al. (2008) ein neun Faktoren-Modell von durch Musik induzierten Emotionen vor. Recht prominent rezipiert sind beispielsweise auch die Befunde der Untersuchungen von Ivanov und Geake (2003); Rauscher und Shaw (2016); Rauscher, Shaw und Ky (1993, 1995) zum so genannten *Mozart-Effekt* (z. B.

D. Campbell, 2001). In einer experimentellen Untersuchung konnten Rauscher et al. (1993) zeigen, dass das Hören der Sonate für zwei Klaviere aus dem 23. Klavierkonzert in A-Dur, (KV 488) von Mozart, die Performanz bei einem nachfolgend vorgelegten Test zum räumlichen Schlussfolgern verbessert.

Ein Aspekt, der auch in der aktuelleren Literatur immer wieder untersucht wird, ist die Verbindung zwischen Musikpräferenzen und Dimensionen der Persönlichkeit (z. B. Langmeyer et al., 2012) – vgl. auch Schäfer und Mehlhorn (2017, für eine Übersicht). Diese Beziehung zwischen individuell unterschiedlich ausgeprägten Musikpräferenzen und Dimensionen der Persönlichkeit wurde dabei auch bereits in frühen empirischen Beiträgen untersucht (z. B. Cattell & Anderson, 1953; Cattell & Saunders, 1954b; Healey, 1973; Payne, 1967). Cattell und Anderson (1953); Cattell und Saunders (1954b) untersuchen dabei die Eignung der Erfassung musikalischer Präferenzen zum Zweck einer Diagnostik der Persönlichkeit in einem klinischen Kontext von Verhaltensstörungen.

Die frühen Untersuchungen zum Zusammenhang von Persönlichkeit und Musikpräferenzen operationalisierten die Musikpräferenzen mit dem IPAT Music Preferance Test (Cattell & Anderson, 1953; Cattell & Saunders, 1954b), bei dem die Probanden einzelne Musikstücke bewerten mussten, die sie zuvor gehört hatten. Die Ergebnisse solch einer Bewertung wurden dann als unbewusste Persönlichkeitsmerkmale interpretiert. Als Ergebnis einer faktorenanalytischen Betrachtung der erhobenen Daten verschiedener Gruppen von Individuen fanden (Cattell & Saunders, 1954b) zunächst 12 Faktoren. Die Ergebnisse zur faktoriellen Validität erwiesen sich jedoch insofern als inkonsistent, als dass diese in nachfolgenden Analysen nicht repliziert werden konnten (Healey, 1973). Ein vergleichsweise neues Verfahren zur Erfassung von Musikpräferenzen, welches ebenfalls Hörbeispiele einsetzt wurde von Rentfrow, Goldberg und Levitin (2011) vorgeschlagen. Rentfrow et al. (2011) schlagen dabei ein Modell musikalischer Präferenzen vor, das auf den affektiven Reaktionen der Probanden auf unterschiedlichste Hörbeispiele aus verschiedensten musikalischen Genres basiert. In drei unabhängigen Studien finden Rentfrow et al. (2011) dabei eine latente fünfdimensionale Struktur, die sich über die jeweils typischen affektiv–emotionalen Reaktionsmustern auf die vorgegebenen Musikausschnitte, beschreiben lässt.

Im Gegensatz zum Einsatz konkreter Hörbeispiele bei der Erfassung von Musikpräferenzen stützt sich ein anderer Ansatz zur Erfassung von Musikpräferenzen auf die Bewertung von Begriffen musikalischer Stilrichtungen durch Fragebogenverfahren. Litle und Zuckerman (1986) entwickelten beispielsweise die *Music Preference Scale* (MPS) und bezogen diese auf die Sensation Seeking Scale Form V von Zuckerman, Eysenck und Eysenck (1978). Die MPS enthält 60 Items, welche sich nach Litle und Zuckerman (1986) auf etablierte Kategorien von Musik und musikalischen Aktivitäten der kommerziellen Plattenindustrie in den USA [der damaligen Zeit] beziehen. Die 60 Bezeichnungen musikalischer Stilrichtungen wurden aus einem anfänglichen Pool von 150 Items unter Anwendung der Faktorenanalyse ausgewählt (Litle & Zuckerman, 1986). Nach den Befunden von Litle und Zuckerman (1986) war *Sensation Seeking* positiv mit allen Arten von Rockmusik und negativ mit eher farbloser Film- und Fernsehmusik assoziiert (vgl. auch McNamara & Ballard, 1999). Die Untersuchungen von (Dollinger, 1993) sowie Rawlings und Ciancarelli (1997) verwendeten eine leicht modifizierte Kurzform der MPS zur Erfassung der Musikpräferenz und das *NEO-Persönlichkeits-Inventar* (Costa & McCrae, 1985, 1992c) zur Erfassung der Persönlichkeit (vgl. auch Rawlings, Barrantes i Vidal & Furnham, 2000; Rawlings, Hodge, Sherr & Dempsey, 1995). Übergreifend wiesen die Befunde aus diesen Studien darauf hin, dass die Persönlichkeitsdimensionen *Extraversion* und *Offenheit* den stärksten Zusammenhang mit spezifischen Musikpräferenzen aufweisen (Dollinger, 1993; Rawlings & Ciancarelli, 1997).

Ein weiteres Fragebogenverfahren das musikalische Genrebezeichnungen zur Erfassung von Musikpräferenzen einsetzt, ist das *Musical Preference Questionnaire* (MPQ), welches von Sikkema (1999) entwickelt wurde. Das MPQ besteht aus einer Liste von 11 etablierten Kategorien von musikalischen Genrebezeichnungen. Diese 11 Items wurden teilweise auf der Grundlage von Interviews mit einer großen Anzahl von CD-Händlern und Jugendlichen an mehreren Sekundarschulen in den Niederlanden erstellt (Delsing et al., 2008). Die einzelnen Items des MPQ ähneln stark den Items des von Rentfrow und Gosling (2003) entwickelten *Short Test Of Music Preferences* (STOMP), welcher (in einer übersetzten Form) auch in der vorliegenden Arbeit eingesetzt wird. Der STOMP erfasst mit insgesamt 14 Items aus unterschiedlichen musikali-

schen Genrebezeichnungen vier Dimension der Musikpräferenz (Rentfrow & Gosling, 2003). Dabei besteht die Annahme, dass es sich hier, in Analogie zum Big-Five-Modell, um *orthogonale*, also voneinander unabhängige, unkorrelierte Dimensionen handelt (vgl. auch Rentfrow et al., 2012). Unterschiedliche Präferenzen für verschiedene Musikdimensionen bestehen nach dieser Modellvorstellung dabei unabhängig voneinander, sodass sich Präferenzen für einzelne oder mehrere Musikdimensionen nicht ausschließen.

Eine detaillierte Beschreibung des STOMP ist im Kapitel 5 *Stichproben und Instrumente* in Abschnitt 5.1.2 gegeben.

Kapitel 3

Theoretischer Hintergrund zu Antwortmustern

Werden Daten über Fragebogen zur Selbsteinschätzung erhoben, können sich unerwartete Antwortreaktionen ergeben. Solche unerwarteten Antwortreaktionen werden typischerweise entweder durch fehlende Motivation zur Bearbeitung der Fragen (z. B. Maniaci & Rogge, 2014), ein fehlendes Verständnis oder eine falsche Interpretation einzelner Fragen (z. B. Hardy & Ford, 2014), oder aber durch bewusste oder unbewusste Täuschungen seitens der antwortenden Personen begründet (z. B. Ziegler, MacCann & Roberts, 2012). Die unerwarteten Antwortreaktionen bei einzelnen Fragen resultieren dann in den zu analysierenden Daten in individuell unterschiedlichen, *idiosynkratischen* Antwortmustern, welche von den eigentlich (vom Testentwickler oder Anwender) erwarteten Antwortmustern abweichen.

3.1 Antwortverhalten, Antwortmuster, Antwortstile, Antwortverzerrung – ein Überblick

Die Erkenntnis und Feststellung, dass mit der Anwendung von Fragebogenverfahren zur psychologischen Diagnostik mit unbeabsichtigten Fehlerquellen zu rechnen ist, lässt sich bis zu recht frühen Beiträgen wie beispielsweise von Zubin (1937), Lorge (1937), Lentz (1938), Cronbach (1942) und Berg (1957) zurückverfolgen. Der Beitrag von Zubin (1937) ist dabei insofern bemerkenswert, als dass er als eine Hauptfehlerquelle bei der Interpretation von Fragebogenverfahren das methodische Vorgehen bei der (ausschließlich) summativen Auswertung benennt. Speziell diese rigorose Anwendung des Summenwertes (*Summenscores*) über alle Items als Index für die Merkmalsausprägung der Personen bei der Auswertung wird dabei bereits von Zubin (1937) kritisch in Frage gestellt.

Cronbach (1946) stellt in seiner Übersicht zur Validität bei Fragebogenverfahren allgemein fest, dass der resultierende Testwert durch andere Variablen als diejenigen, die eigentlich gemessen werden sollen, beeinflusst wird und diese damit einen negativen Effekt auf die Reliabilität und Validität des Testergebnisses nehmen. In einer empirischen Arbeit und zwei Übersichten unterschiedet Cronbach (1942, 1946, 1950) neben anderen im Wesentliche zwei Arten von Verzerrungen. Einerseits eine Tendenz zu mittleren oder extremen Polen der Antwortskala und andererseits mehr oder weniger mit dem Inhalt der Fragen begründete Verzerrungen wie eine Tendenz zur Zustimmung (*Akquieszenz*) oder auch unterschiedliche Sorgfalt bei der Beantwortung von Items. Eine der ersten allgemeinen Definitionen von Cronbach bezeichnet solche '*response sets*', also Antwortverzerrungen, durch idiosynkratische Antwortmuster wie folgt: „*A response set is defined as any tendency cousing a person consistently to give different responses to test items than he would when the same content is presented in a different form.*“ (Cronbach, 1946, S. 476); also als jegliche Tendenz einer (antwortenden) Person systematisch abweichende Reaktionen auf Testitems zu zeigen als wenn der gleiche Inhalt in einer anderen Form dargestellt würde. Wobei Cronbach (1946, S. 476) dies als eher theoretische Definition ansieht und einschränkend anmerkt, dass der Inhalt der Items wohl kaum vollständig von der Form (z. B. der Antwortskala) zu trennen sei. Ausgehend von dieser allge-

meinen Definition wurden im Rahmen einer Vielzahl von weiteren empirischen Arbeiten unterschiedliche Formen und Definitionen von Antwortverzerrungen identifiziert und definiert (vgl. Wiggins, 1964, S. 552, für eine erste Darstellung der unterschiedlichen Begriffe). Solche Verzerrungen der Antworten oder Abweichungen der Personenantworten von den eigentlich erwarteten werden in aktuellen Arbeiten unter Schlagworten wie *response bias, response sets, response styles, response distortion* und auch spezifischer im Hinblick auf eine vermutete Ursache, als *socially desirable responding*, *impression management*, *overclaiming* oder *malingering* und *faking* subsumiert (z. B. Helmes, Holden & Ziegler, 2015; Paulhus, 2012, für eine aktuellere Übersicht). In einer systematischen Übersichtsarbeit zur Literatur seit 2001 listet Rupp (2013) des Weiteren 18 englischsprachige Bezeichnungen für abweichendes Antwortverhalten auf.

Bei der Untersuchung von abweichenden Antwortmustern ist es sinnvoll, Theorien und Modelle zum Verständnis von Fragen (Gilbert, 1991; Knowles & Condon, 1999) und zum Antwortprozess allgemein (Tourangeau & Rasinski, 1988; Tourangeau, Rips & Rasinski, 2000) mit einzubeziehen, aus denen sich überprüfbare Hypothesen zum Antwortverhalten ableiten lassen (Ziegler, 2011). Ein prominentes Modell zum Antwortprozess wird von Tourangeau und Rasinski (1988) und Tourangeau et al. (2000) propagiert, wonach die Befragten vier kognitive Verarbeitungsprozesse durchlaufen (Tourangeau et al., 2000, vgl. dazu detaillierter auch Abschnitt 3.2.1). Jeder dieser vier Verarbeitungsprozesse kann zu Antwortverzerrungen führen. So können die Befragten beispielsweise die Frage falsch interpretieren, wichtige Informationen vergessen, falsche Schlussfolgerungen basierend auf ihrem Vorwissen ziehen, oder ihre Antworten auf eine unangemessene Antwortkategorie abbilden (Tourangeau et al., 2000).

Die Befundlage in der Literatur zu den Auswirkungen solcher Verzerrungen bei der Messung und die gegebenen Definitionen erscheint dabei insgesamt betrachtet recht heterogen. Bereits Rorer (1965) setzt sich im Rahmen eines kritischen Überblicks mit solchen unterschiedlichen Definitionen abweichenden Antwortverhaltens auseinander und stellt fest, dass „*In recent years the psychological literature dealing with response sets, response biases, or response styles has grown so large and been reviewed so many times that the reviews have themselves been reviewed.*“ (Rorer, 1965, S. 129). In diesem Sin-

ne fehlt auch aktuell, trotz umfangreicher Literatur zu den unterschiedlichen Aspekten abweichender Antwortmuster, eine allgemein akzeptierte, übergreifende Systematisierung der Begrifflichkeiten zu den unterschiedlichen Formen dieser Antwortverzerrungen.

Ein erster Rahmen für eine Strukturierung der verschiedenen Begriffe wurde schon relativ früh beispielsweise von Jackson und Messick (1958) eingeführt, welcher einerseits zwischen *response styles* (Antwortstilen) im Sinne einer (inhaltsunabhängigen) Konsistenz über Zeit und unterschiedliche Fragebögen, und andererseits *response sets* (Antwortverzerrung) im Hinblick auf eine spezifische Beurteilungssituation, unterscheidet. Innerhalb der response styles können nach dieser Taxonomie, die *Akquieszenz* (Zustimmungstendenz), die Tendenz zu extremen (*extrem response style* – ERS) oder mittleren (*midpoint response style* – MRS) Antwortkategorien einer mehrstufigen Antwortskala, unterschieden werden. Unter dem Oberbegriff *response sets* können im Wesentlichen drei weitere Begriffe und Phänomene unterschieden werden. Einerseits das *sozial erwünschte Antwortverhalten* (*social desirable responding* – SDR) und der inhaltlich verwandte, aber eher sozialpsychologisch fundierte Begriff der *Eindruckssteuerung* (*impression management* – IM). Ein weiterer Begriff, im Rahmen der *response sets* ist der, insbesondere in der neueren Literatur verbreitete, Begriff des *vorgetäuschten Antwortverhaltens* (faking).

Ein Kritikpunkt an dieser von Jackson und Messick (1958) eingeführten Einteilung kann darin gesehen werden, dass die Konnotation des Begriffs *Stil* [*style*] in Verbindung mit der entsprechenden Definition einer zeitlich und inhaltlich (über verschiedene Items und Fragebögen) übergreifenden Personeneigenschaft empirischen Befunden zu der Frage nach der Konsistenz abweichender *Antwortmuster* letztlich vorgreift. In dieser Hinsicht neutraler unterscheidet Häcker (2014) daher auf der obersten Stufe einer übergreifenden Taxonomie der Antwortverzerrungen zwischen inhaltsbezogenen Verzerrungen und solchen, welche sich eher auf die Form (z. B. die Antwortskala der Items) beziehen. Die Frage nach einer möglichen Konsistenz solcher Phänomene bleibt nach dieser Definition also zunächst offen und damit Gegenstand empirischer Untersuchungen. Ausgehend von dieser allgemeinen Definition von Häcker (2014) werden im folgenden Abschnitt einige Formen idiosynkratischen Antwortverhaltens nach einer erweiterten Taxonomie dargestellt.

3.2 Antwortmuster - eine erweiterte Taxonomie

In den folgenden Abschnitten sollen einige der unterschiedlichen, aus der Literatur vorliegenden theoretischen Betrachtungen und ausgewählte Befunde zu besonderem, idiosynkratischen Antwortverhalten bzw. den daraus resultierenden Antwortmustern in einer erweiterten Systematik dargestellt werden. Diese unterschiedlichen in der Literatur berichteten Formen des Antwortverhaltens werden hierbei in vier Bereiche unterteilt, welche sich aus den unterschiedlichen Aspekten der Beantwortung einzelner Items bzw. der Anwendung eines psychodiagnostischen Fragebogeninventars ableiten lassen. Der von Häcker (2014) gegebenen Einteilung folgend, können dabei bezogen auf die Items einerseits inhaltliche (vgl. Abschnitt 3.2.1) und andererseits formale (vgl. Abschnitt 3.2.4) Aspekte unterschieden werden. In Ergänzung dazu werden Aspekte der Operationalisierung in Abhängigkeit unterschiedlicher Konstruktdefinitionen, sowie die damit einhergehende Frage nach deren Polarität auf Ebene der psychometrischen Skalen und Antwortskalen der Items in Abschnitt 3.2.2 dargestellt. Eine spezielle Form des „Antwortverhaltens“, nämlich die nachlässige Bearbeitung oder das gänzliche Auslassen von einzelnen Items bzw. Itemgruppen [*careless responding*] wird in Abschnitt 3.2.3 diskutiert.

3.2.1 Antwortmuster in Abhängigkeit der Inhalte der Items

Sozial erwünschtes Antwortverhalten und Impression Management

Sozial erwünschtes Antwortverhalten (*socially desirable responding* – SDR) ist nach einer allgemeinen Definition durch eine „*Neigung, positive Selbstbeschreibungen zu geben* [tendency to give positive self-descriptions]“ (Paulhus, 2002, S. 49) gekennzeichnet. Soziale erwünschtes Antwortverhalten (SDR) wird von Edwards (1957) als ein wichtiger Aspekt bei der Beantwortung aller möglichen Aussagen zur Beschreibung von Personeneigenschaften angesehen. Die von Paulhus (2002) beschriebene Definition sozial erwünschten Antwortverhaltens wird von Tracey (2016) im Rahmen eines aktuellen Reviews der Literatur hinsichtlich dessen Funktion präzisiert und spezifischer definiert „*so-*

cially desirable responding (SDR) refers to the presentation of oneself in an overly favorable light“ (Tracey, 2016, S. 224); also als *Präsentation der eigenen Person in einer, im Vergleich zur Realität, vorteilhafteren Art und Weise*. Sozial erwünschtes Antwortverhalten wird dabei als unerwünschte Verzerrung bei der eigenschaftsbezogenen Selbstbeschreibung der eigenen Person im Vergleich zur (objektiven) Realität angesehen, welche potentiell systematische Fehlervarianz in die erzielten Testwerte einbringt und einen systematischen (verzerrenden) Bezug zu den zu messenden Inhalten aufweisen kann (Tracey, 2016). Im Sinne dieser Definition kann sozial erwünschtes Antwortverhalten aus der sozialpsychologischen Forschungsperspektive auch als potentiell dissonante Diskrepanz zwischen einer „öffentlich“ berichteten Einstellung zu einem bestimmten Einstellungsobjekt, im Vergleich zur eigentlichen, wahren Einstellung der befragten oder kommunizierenden Person angesehen werden. Erste empirische Evidenz zu den motivationalen und psychologischen Antezedenzien sozial erwünschten (Antwort-)Verhaltens resultierte dementsprechend aus der kritischen Überprüfung einer experimentellen Studie von Festinger und Carlsmith (1959) zur Theorie der kognitiven Dissonanz (vgl. Festinger, 1957) durch Tedeschi, Schlenker und Bonoma (1971). Nach der Theorie der kognitiven Dissonanz von Festinger (1957) ergibt sich eine tatsächliche Einstellungsänderung bei den untersuchten Personen allein durch die experimentelle Aufforderung, eine der ursprünglichen Einstellung entgegengesetzte Einstellung zu vertreten. Im Gegensatz dazu konnten Tedeschi et al. (1971) in ihren Untersuchungen zeigen, dass sich eine solche tatsächliche, intrapsychische Einstellungsänderung nur in Abhängigkeit der unterschiedlichen Beziehung zum Versuchsleiter als wahrgenommenen Adressaten der zu berichtenden Einstellung ergibt – oder eben ausbleibt. Die entscheidende experimentelle Variation der Bedingungen bei der Beantwortung von Fragen zur eigenen Einstellung (und damit der Selbstdarstellung) besteht dabei im Wesentlichen im Ausmaß der wahrgenommenen „Öffentlichkeit“ und damit einer potenziellen Überprüfbarkeit der bei der Beantwortung selbst berichteten Einstellung (vgl. Mummendey, 2015). Die Relevanz des Faktors „Öffentlichkeit“ beziehungsweise „Überprüfbarkeit“ setzt wiederum eine aktive, bewusste Wahrnehmung von Merkmalen der eigenen Person voraus, welche dann Gegenstand einer Selbstdarstellung werden. Die Relevanz der wahrgenommenen Öffentlichkeit im Hinblick auf das Ausmaß

und die Art der Selbstdarstellung konnte auch von Boeije (2004) im Rahmen einer qualitativen Untersuchung zu Interviewsituationen in Anwesenheit dritter Personen bestätigt werden.

Nach der allgemeinen Definition von sozial erwünschtem Antwortverhalten (Paulhus, 2002; Tracey, 2016) dient dieses in der Regel einer bewusst ausgeführten *Selbstdarstellung* der eigenen Person. Die allgemeine Intention einer Selbstdarstellung besteht darin, Selbstbilder an Interaktionspartner zu kommunizieren, um auf die eigene Person bezogene Attributionen und Eindrücke des Interaktionspartners zu beeinflussen (Schlenker, 2003; Tedeschi, 1981). Eine notwendige Voraussetzung für eine *bewusste* Selbstdarstellung ist die aktive Wahrnehmung von Eigenschaften der eigenen Person im Rahmen der Selbstwahrnehmungstheorie (Bem, 1967). Die auf der Selbstwahrnehmung aufbauende Selbstdarstellung ist allgemein der Einsatz jeglichen Verhaltens zur Kommunikation von Informationen über das eigene Selbst gegenüber anderen Personen im Rahmen sozialer Interaktionen auf der Basis unterschiedlicher Motive (Baumeister, 1982). Im Zusammenhang mit der Frage nach der Bewusstheit oder Unbewusstheit eigener Personenmerkmale bei der Beantwortung von Fragebögen schlägt Paulhus (1984, 2002) vor, zwischen zwei Komponenten sozial erwünschten Antwortverhaltens zu unterscheiden (vgl. auch Paulhus & Reid, 1991). Einerseits bewusst ausgeführte *Eindruckssteuerung* [impression management – IM] (Barrick & Mount, 1996; Davis, Thake & Weekes, 2012; Khorramdel, 2014) und andererseits, auf der nicht öffentlichen Ebene, unbewusste *Selbsttäuschung* [self-deception] (Gur & Sackeim, 1979; Sackeim & Gur, 1979), aber auch mehr oder weniger bewusste *Selbstverbesserung* [self-enhancement] (Baumeister, 1982; Kwan, John, Kenny, Bond & Robins, 2004; Kwan, John, Robins & Kuang, 2008). Das Zweikomponentenmodell für sozial erwünschtes Antwortverhalten wurde von Paulhus und Reid (1991) dahingehend erweitert, dass sich *Selbsttäuschung* in zwei Dimensionen bzw. Faktoren aufteilt. Einerseits *selbsverbessernde Täuschung* [self-deceptive enhancement] (SDE) welche aus positiv gepolten Items besteht und *selbsverleugnende Täuschung* [self-deceptive denial] (SDD), welche aus negativ gepolten Items besteht (vgl. auch Gignac, 2013) – zur Polarität von Items vgl. auch Abschnitt 3.2.2. Auch Mummendey (2015) betont im Zusammenhang mit Begriffen wie Selbstdarstellung und sozial erwünschtem Verhalten, dass deren Konnotation einer notwendi-

gerweise immer *bewussten* Täuschungsabsicht (vgl. auch dazu der Begriff des *faking* weiter unten) irreführend sei. Derartige Verzerrungen bei der Außendarstellung der eigenen Person oder grundsätzlich beobachtbarem Verhalten allgemein können nämlich sowohl bewusste als auch unbewusste Grundlagen haben (Baumeister, Masicampo & Vohs, 2011). Schlenker und Weigold (1992) betonen, dass sich bewusste und unbewusste Prozesse der Selbstdarstellung aus dem Bedürfnis von Menschen ergeben, ihre Umgebung zu strukturieren und zu beeinflussen, um diese für sie vorteilhafter und weniger bedrohlich zu gestalten.

Eine positive Selbsttäuschung bezieht sich dabei auf die Tendenz, weniger wünschenswerte Aspekte von sich selbst eher zu ignorieren oder zu verdrängen. Nach Befunden von Paulhus (1998a) hängt die Selbsttäuschung mit nachweisbaren Verzerrungen in Bezug auf bestimmte Aspekte der Selbstwahrnehmung zusammen. Selbsttäuschung als Teilaspekt von sozial erwünschtem Antwortverhalten scheint somit ein mangelndes Bewusstsein für einige weniger wünschenswerte Aspekte des Selbst zu erfassen. Personen, die in diesem Sinne einer Selbsttäuschung unterliegen, scheinen sich der sozial weniger wünschenswerten Aspekte der eigenen Person nicht bewusst zu sein, wodurch bezogen auf ihr Antwortverhalten bei der Selbsteinschätzung dann nicht unbedingt von einer aktiven oder bewussten Verstellung oder Täuschung gesprochen werden kann. Mit diesem kognitiven Phänomen der Selbsttäuschung wird auch der sogenannte *Barnum Effekt* in Verbindung gebracht (z. B. Meehl, 1956). Der Barnum Effekt beschreibt das Phänomen, dass vage, allgemein positive Aussagen über Persönlichkeitseigenschaften von Menschen (Forer, 1949), von diesen oft als einzigartig für die eigene Person angenommen werden, obwohl diese Aussagen für die meisten Menschen sehr wahrscheinlich zutreffend wären (Beins, 1994; Dickson & Kelly, 1985; Fichten & Sunerton, 1983; Forer, 1949; Furnham & Schofield, 1987).

Ein weiterer typischer Effekt bei solchen selbstreflexiven kognitiven Prozessen, der sich eher unbewusst und weniger aus einer spezifisch bewussten Täuschungsabsicht bei der Selbstdarstellung ergibt, ist der so genannte *über dem Durchschnitt Effekt* [above-average effect] (vgl. Chambers & Windschitl, 2004; Moore & Small, 2007). Dieser recht universell zu beobachtende Effekt zeigt sich insbesondere, wenn Personen eine größere Anzahl von selbstbezoge-

nen Merkmalsbewertungen in kurzer Zeit und unter kognitiver Belastung (z. B. zeitliche Limitationen) durchführen müssen (Alicke, 1985; Alicke, Klotz, Breitenbecher, Yurak & Vredenburg, 1995). Der above-average effect führt dazu, dass Personen dazu neigen bestimmte Merkmalsausprägungen der eigenen Person im Vergleich zu anderen Personen als eher überdurchschnittlich ausgeprägt einzustufen. Die unbewussten Anteile dieses Effekts ergeben sich aus der Art und Weise, in der typischerweise Informationen über das eigene Selbst im Vergleich zu Informationen über andere Personen verarbeitet werden. Dabei sind zunächst Gedanken über das Selbst und zu selbstrelevanten Informationen im Vergleich zu Gedanken über andere Personen und relevante Informationen üblicherweise größer und detaillierter mental repräsentiert. Die Konsequenz dieser Form eines *impliziten* Egozentrismus ist, dass selbstrelevante Informationen relativ zu anderen relevanten Informationen im Urteilsprozess ein unverhältnismäßiges Gewicht haben (Kruger, 1999). In Folge tendieren Personen dazu, personenbeschreibende Merkmalsdimensionen idiosynkratisch, bezogen auf ihre eigenen spezifisch wahrgenommenen Verhaltensweisen und Eigenschaften, zu definieren (Dunning & Cohen, 1992; Dunning & McElwee, 1995; Hayes & Dunning, 1997); vergleiche auch (Moore & Small, 2007). Ergänzend kann sich zusätzlich eine besondere Form der selektiven und selbstwertdienlichen Form der Informationsverarbeitung einstellen, welche dazu führt, dass negative Eigenschaften in geringerem Ausmaß repräsentiert sind als positive Eigenschaften (z. B. Kruger & Dunning, 1999; Moore & Small, 2007; Paulhus, 1998a; Regan, Snyder & Kassin, 1995; Williams, Dunning & Kruger, 2013). Als eine Folge solcher Formen der Informationsverarbeitung resultiert dann beispielsweise eine mehr oder weniger bewusste *Selbstverbesserung* [self-enhancement] (Kwan et al., 2004, 2008), welche sich in vergleichenden Untersuchungsdesigns mit Fremd- und Selbstbeschreibungen zu Persönlichkeitsmerkmalen empirisch finden lassen (z. B. Borkenau, Zaltauskas & Leising, 2009). Solche Prozesse der Selbstdarstellung können nach Mummendey (2015) bei jeglicher Art der sozialen Interaktion auftreten und sich auf jede Art von Adressat richten, also möglicherweise auch auf das eigene Selbst, im Rahmen selbstreflexiver kognitiver Prozesse, wie sie durch die Bearbeitung von Fragebögen ausgelöst werden können. Ferner betont Mummendey (2015), dass solche Prozesse der Selbstdarstellung als *Eindruckssteuerung* (IM) zwischen kommunizierenden Akteuren recht

universell auftreten und sich dabei sowohl auf sprachlicher als auch nichtverbaler Ebene beobachten lassen. Tedeschi (1981) definiert die *Eindruckssteuerung* [impression management] (IM) als jegliches Verhalten einer Person, das geeignet ist, die bei anderen entstehenden Attributionen und Eindrücke von der (eigenen) Person zu kontrollieren oder zu manipulieren – „*Impression management consists of any behavior by a person that has the purpose of controlling or manipulating the attributions and impressions formed of that person by others*“ (Tedeschi, 1981, S. 3). Im Sinne einer, gegenüber einer allgemeinen Selbstdarstellung differenzierter abgrenzenden Definition, definiert Schlenker (2003) IM als zielgerichtete Aktivität der Informationskontrolle gegenüber anderen Personen.

Ein Ausgangspunkt der systematisch, experimentellen Untersuchung zur (bewussten) *Eindruckssteuerung* (IM) im Rahmen der *Selbstdarstellungstheorie* in der Sozialpsychologie (vgl. Mummendey, 2015) kann in einer empirischen Untersuchung von E. E. Jones, Gergen und Jones (1963) zum (taktischen) Kommunikationsverhalten amerikanischer Marineangehöriger gesehen werden. E. E. Jones et al. (1963) beschreiben in ihrer Studie dabei das Ausmaß vom IM im Sinne einer bewussten und positiven Selbstdarstellung in Abhängigkeit von drei zentralen Variablen. Erstens, der wahrgenommenen Wichtigkeit und Relevanz der Frage bzw. der angesprochenen Thematik, zweitens, des Status der antwortenden Person (im Vergleich zum angenommenen Empfänger), sowie drittens, einer allgemeinen Tendenz zur Konformität, wobei der Einfluss letzterer im Sinne einer Interaktion durch die beiden ersten Variablen beeinflusst wird. Im Zusammenhang mit der wahrgenommenen Wichtigkeit des Tests bzw. der Wahrnehmung der daraus resultierenden Konsequenzen wird dementsprechend zwischen so genannten *high stake* und *low stake* Testsituationen unterschieden (Cole & Osterlind, 2008; Ziegler et al., 2012). Als *high stake* wird dabei eine Testsituation bezeichnet, aus deren Ergebnis sich vor dem Hintergrund konkreter diagnostischer Fragestellungen wichtige Konsequenzen für die teilnehmende Person ergeben (Einstellungstests im Arbeits- und Organisationsbereich, Abschlussprüfung in Schule und Studium sind hier typische Beispiele). Andererseits werden mit *low stake* so genannte Testsituationen bezeichnet, deren Ergebnis zwar für die antwortende Person interessant und relevant sein können, die jedoch keine wahrgenommenen gesellschaftlichen oder

beruflichen Konsequenzen haben. Diese unterschiedlichen Perspektiven auf die Testsituation bei der Bearbeitung von Fragebögen hat im Zusammenhang mit Antwortverzerrungen zur Konzeptualisierung des Phänomens *faking* beigetragen (vgl. folgender Abschnitt weiter unten). Bezogen auf SDR ist die Befundlage hinsichtlich des Einflusses der wahrgenommenen Testsituation uneinheitlich. So finden beispielsweise Smith und Ellingson (2002), dass Antwortverzerrungen einen geringen Effekt auf die Messung grundlegender Dimensionen der Persönlichkeit in Auswahlsituationen (*high stake testing*) haben. In ähnlicher Weise finden Ellingson, Sackett und Connelly (2007) im Rahmen einer experimentellen Untersuchung mit einem *within-subject-design* (high stake vs. low stake) einen geringen Einfluss auf SDR in Abhängigkeit der wahrgenommenen unterschiedlichen Testsituation. Demgegenüber konnte eine empirische Arbeit von Ziegler, Toomela und Bühner (2009) die Befunde des Einflusses der wahrgenommenen Wichtigkeit der Testsituation (vgl. E. E. Jones et al., 1963) bestätigen. Die aufgestellte Hypothese, dass sich eine situative Anforderung (*high stake* vs. *low stake* Testung), die sich aufgrund des militärischen Ranges einer Person ergab, Einfluss auf das Antwortverhalten nimmt konnte bestätigt werden (Ziegler et al., 2009).

Bezogen auf formalen „Ursachen“ für die verzerrte Selbstdarstellung bei Fragebogenverfahren untersucht dagegen Khorramdel (2014) die Auswirkungen unterschiedlicher Antwortskalen auf die Eindruckssteuerung (IM) und greift damit die bereits von Wiggins (1962) formulierte Idee auf, dass die Antwortskala den definierenden Rahmen möglicher Antwortverzerrungen darstellt. In der Untersuchung von Khorramdel (2014) wird dabei die Hypothese formuliert, dass eine sechsstufige Antwortskala im Vergleich zu einer dichotomen den Einfluss von IM (hier auch faking) verringert. Dabei wird argumentiert, dass es bei einer dichotomen Antwortskala möglicherweise schwieriger ist (weil potentiell leichter zu entdecken) die Antworten gemäß einer verzerrt positiven Eindruckssteuerung anzupassen. Die Befunde von Khorramdel (2014) weisen darauf hin, dass die mehrstufige Antwortskala im Vergleich zu der dichotomen tatsächlich den Einfluss von IM (hier auch faking) verringert, diesen aber nicht ganz verhindert.

Von den zwei Hauptdimensionen *Eindruckssteuerung* (IM – Barrick & Mount, 1996; Davis et al., 2012; Khorramdel, 2014) einerseits und *Selbsttäuschung* [self-

deception] (Gur & Sackeim, 1979; Sackeim & Gur, 1979) andererseits, welche nach Paulhus (1984, 2002) die soziale Erwünschtheit ausmachen, erscheint die Eindruckssteuerung (IM) zunächst einen erheblichen Einfluss auf die Validität von Fragebögen zur Selbstbeurteilung zu nehmen, da sie definitorisch das Ergebnis einer bewussten Verzerrung ist. Allerdings deuten Befunde von Barrick und Mount (1996); Ones, Viswesvaran und Reiss (1996); Smith und Ellingson (2002) darauf hin, dass sozial erwünschtes Antwortverhalten und spezifisch IM keinen substanziellen Einfluss auf die prädikative Validität von Fragebogenverfahren hat. Demgegenüber sprechen eine ganze Reihe von empirischen Befunden dafür, dass SDR potentiell negative Auswirkungen auf die Validität von Skalenwerten aus Fragebogenverfahren hat (Bäckström, 2007; Schmit & Ryan, 1993; Ziegler et al., 2012, 2009). Im Rahmen faktorenanalytischer Auswertungen von Persönlichkeitsinventaren nach dem Big-Five-Paradigma wird SDR dabei oft als übergeordneter Faktor gedeutet (z. B. Bäckström, Björklund & Larsson, 2009). Bereits Edwards (1957) formulierte in diesem Sinne im Zusammenhang mit sozial erwünschtem Antwortverhalten die Annahme, dass sich jegliche (personenbezogenen) Eigenschaftsbeschreibungen auf einer einzigen Dimension, nämlich dem Ausmaß der sozialen Erwünschtheit der jeweils erfassten Eigenschaft, anordnen lassen Edwards (1957, S. 3). Die übergeordnete Dimension stellt nach dieser Ansicht eine Eigenschaft der personenbeschreibenden Aussagen selbst dar, im Gegensatz zu den eigentlichen Inhalten der Aussagen, welche als Beschreibungen der Eigenschaften der Personen angesehen werden. Diese *Meta*eigenschaften der Aussagen bestehen nach Edwards (1957) unabhängig vom personenbezogenen Inhalt der jeweiligen Aussage, und damit deren Zuordnung zu unterschiedlichen Dimensionen des zu erfassenden Konstrukts. Solch eine enge Verbindung zwischen dem ursprünglich als Antwortverzerrung angesehenen *sozial erwünschten Antwortverhalten* (SDR) und personenbezogenen *Eigenschaftsbeschreibungen* wirft die Frage nach dessen Zusammenhang zu grundlegende Dimensionen der Persönlichkeit auf. Bereits recht früh wurden in dieser Hinsicht Überlegungen angestellt, dass derartiges Antwortverhalten wie SDR und IM nicht nur als Messfehler, sondern möglicherweise als Ausdruck von und Maß für grundlegende Persönlichkeitseigenschaften angesehen werden kann (Berg & Collier, 1953; Cronbach, 1950). Die Ergebnisse einer Untersuchung von McCrae und Costa (1983b) legen na-

he, dass Korrelationen zwischen Skalen zur Erfassung von Dimensionen der Persönlichkeit mit sozialer Erwünschtheit eher substanziell als artifiziell interpretiert werden sollten und dass die weit verbreitete Praxis der Korrektur von Scores durch SD-Skalen in Frage gestellt werden sollte. Ellingson, Sackett und Hough (1999) folgern daher, dass die Korrektur von sozial erwünschtem Antwortverhalten vergleichsweise ineffektiv ist und nicht dazu geeignet ist einen korrekten Messwert zu erzeugen. In Bezug auf die faktorielle Struktur des Konstruktes Persönlichkeit folgern Ellingson, Smith und Sackett (2001), dass sozial erwünschtes Antwortverhalten (SDR) einen geringen Einfluss auf die übergeordnete Faktorstruktur hat, welche den Zusammenhang zwischen grundlegenden Dimensionen der Persönlichkeit charakterisiert. In ähnlicher Weise interpretieren Lönnqvist, Paunonen, Tuulio-Henriksson, Lönnqvist und Verkasalo (2007) die Ergebnisse ihrer Untersuchungen, dass Skalen zur Erfassung von sozial erwünschtem Antwortverhalten in unterschiedlichem Ausmaß einerseits tatsächlich sozial erwünschtes Antwortverhalten erfassen, aber daneben auch inhaltlich bedeutsame, interindividuelle Unterschiede der Persönlichkeit. So zeigen beispielsweise Ones et al. (1996), dass Maße für soziale Erwünschtheit mit emotionaler Stabilität ($r = .37$), Gewissenhaftigkeit ($r = .20$) und Verträglichkeit ($r = .14$) korrelieren. Graziano und Tobin (2002) untersuchen im Hinblick auf die Erfassung von Persönlichkeitseigenschaften im Rahmen des Big-Five-Paradigmas den Zusammenhang zwischen der Dimension *Verträglichkeit* und *sozial erwünschtem Antwortverhalten*. In ihren theoretischen Betrachtungen betonen Graziano und Tobin (2002) dabei die bereits definitorisch bestehende Konfundierung zwischen der Persönlichkeitsdimension *Verträglichkeit* und *sozial erwünschtem Antwortverhalten*. In zwei empirischen Untersuchungen finden Graziano und Tobin (2002) mittlere Zusammenhänge zwischen der Persönlichkeitsdimension *Verträglichkeit* und *sozial erwünschtem Antwortverhalten* – jeweils operationalisiert über entsprechende Fragebogenskalen. In diesem Sinne untersuchen Sadler, Hunger und Miller (2010) den Zusammenhang zwischen 12 Taktiken der Eindruckssteuerung (IM) und verschiedenen Dimensionen der Persönlichkeit. Die Autoren finden dabei substanzielle Zusammenhänge zwischen einzelnen Dimensionen negativer Emotionalität und der Neigung zu bewusster Eindruckssteuerung (Sadler et al., 2010). Vor dem Hintergrund solcher Befunde argumentieren de Vries, Zettler und Hilbig (2014)

dass, bestimmte Anteile sozial erwünschten Antwortverhaltens (SDR) weniger einen Messfehler, sondern spezifische Persönlichkeitseigenschaften darstellen. In ihren Untersuchungen finden de Vries et al. (2014) bedeutsame Zusammenhänge zwischen der bewussten *Eindruckssteuerung* (IM) und den Big-Five-Persönlichkeitsdimensionen Gewissenhaftigkeit und Verträglichkeit. Ein vergleichsweise aktueller und umfassender Überblick zu SDR als Aspekt bei der Beantwortung von Fragebogenverfahren findet sich bei Paunonen und LeBel (2012). Dunlop, Telford und Morrison (2012) fassen einen zentralen Befund ihrer Untersuchungen zum Zusammenhang zwischen SDR und Persönlichkeitseigenschaften dahingehend zusammen, dass ein *nichtlinearer* Zusammenhang zwischen sozial erwünschtem Antwortverhalten (SDR) und Messwerten für das zu erfassende Persönlichkeitsmerkmal besteht. Dieser Befund erweist sich auch als konsistent mit vergleichbaren Befunden von Kuncel und Tellegen (2009) und Borkenau et al. (2009), sodass Dunlop et al. (2012) auf der Basis ihrer Untersuchungen schlussfolgern, dass idiosynkratisch interpretiert, sozial erwünschte Antworten ein Phänomen sein können, das am besten auf der Ebene der Items und nicht auf der Ebene der summativ verrechneten Skala untersucht werden sollte.

Vorgetäuschtes Antwortverhalten - Faking

Ebenfalls abhängig vom Inhalt der jeweiligen Fragestellung bzw. der einzelnen Items ist eine in der neueren Literatur diskutierte Form der Antwortverzerrung, die mit dem englischen Begriff *faking* bezeichnet wird. Das Konzept des *vorgetäuschten Antwortverhaltens* [faking] ist eng verknüpft mit dem Phänomen *sozial erwünschten Antwortverhaltens*, sodass hier ebenfalls die Untersuchung von entsprechenden Zusammenhängen mit Persönlichkeitsdimensionen angezeigt sind. In einer metaanalytischen Auswertung über insgesamt 51 Einzelstudien konnten Viswesvaran und Ones (1999) zeigen, dass das Ausmaß vorgetäuschten Antwortverhaltens, (*faking*) als spezifischer Aspekt *sozial erwünschten Antwortverhaltens* nicht innerhalb der Big-Five-Persönlichkeitsdimensionen variiert. Alle Big-Five-Faktoren waren dabei in gleicher Weise von *faking* beeinflusst. Demgegenüber zeigen Holden und Passey (2010), dass Maße für sozial erwünschtes Antwortverhalten zwar mit selbst berichteten Maßen der Persönlichkeit assoziiert sind, allerdings keine Zusammenhänge zu Fremdberichten

innerhalb desselben Konstruktes bestehen.

In einem von Ziegler et al. (2012) herausgegebenen Buch zur Thematik *faking* definieren die Herausgeber dabei *faking* wie folgt: *„Faking represents a response set aimed at providing a portrayal of the self that helps a person to achieve personal goals. Faking occurs when this response set is activated by situational demands and person characteristics to produce systematic differences in test scores that are not due to the attribute of interest“* Ziegler et al. (2012, p. 8). Im Gegensatz zu der von Mummendey (2015) allgemein gehaltenen Definition von inhaltsbezogenen Antwortverzerrungen, welche sowohl bewusst als auch unbewusst wirksam werden, handelt es sich bei *faking* also zunächst um eine Verzerrung, die *bewusst* und im Hinblick auf (unterschiedliche) persönliche Ziele ausgeführt wird. In Abgrenzung zur ebenfalls bewusst intendierten *Eindrucksteuerung* (IM) handelt es sich nach dieser Definition von Ziegler et al. (2012) bei *faking* nicht nur um eine bewusste, sondern auch im Hinblick auf *spezifische Adressaten* der Kommunikation, absichtsvoll ausgerichtete Täuschung einer anderen Person oder eines anderen Personenkreises, um dadurch einen persönlich empfundenen Vorteil zu erreichen. Während Mummendey (2015) im Rahmen der Definition der *Eindrucksteuerung* (IM) die prinzipielle Offenheit der Zielrichtung der bewussten Selbstdarstellung betont, ist die Definition des *faking* in dieser Hinsicht spezifischer in ihrer Zielrichtung der Selbstdarstellung bzw. Täuschung bezogen auf andere Personen in einer spezifischen Situation. Deutlich wird diese Unterscheidung auch in einer Definition von Kuncel und Borneman (2007), die betont, dass *faking* der bewusste Versuch einer Selbstdarstellung mit dem Ziel der Beeinflussung *Anderer* ist (Kuncel & Borneman, 2007, S. 221).

Ganz ähnlich wie bei SDR impliziert die Definition des *faking* den Einbezug bzw. die Relevanz der Inhalte der vorgegebenen Items und setzt so voraus, dass diese durch die antwortenden Personen im Hinblick auf deren individuellen Ziele „richtig“ interpretiert wurden. In Anlehnung an SDR definieren einige Arbeiten *faking* als (eigenständiges) Konstrukt bzw. Eigenschaft der Persönlichkeit (z. B. B. A. Martin, Bowen & Hunt, 2002). Allerdings betonen aktuelle Definitionen von *faking*, dass ein wichtiges Merkmal der Definitionen von *faking* darin besteht, dass dieses weniger als Eigenschaft definiert ist (vgl. Griffith, Lee, Peterson & Zickar, 2011; Kiefer & Benit, 2016), sondern als variierendes,

situationsbezogenes Verhalten, mit dem Ziel eine Selbstbeschreibung zu geben, welche dabei hilft ein bestimmtes Ziel zu erreichen. Die Situationsspezifität und Zielgerichtetheit von *faking* impliziert, dass die Funktion und Bedeutung eines Fragebogens für die antwortende Person im Hinblick auf wahrgenommene Konsequenzen in einer spezifischen Testsituation eine besondere Relevanz hat. Die beiden bereits oben erwähnten grundlegenden Testsituationen beim Einsatz von Fragebogenverfahren (*low stake* vs. *high stake*) erhalten hier beim *faking* insofern eine besondere Bedeutung (Ziegler et al., 2012).

Unter dem Oberbegriff *faking* unterscheiden verschiedene Autoren unterschiedliche Stile des *faking* (oder Klassen), welche sich beispielsweise in Abhängigkeit von der Fähigkeit zum *faking* (Levashina, Morgeson & Campion, 2009; Mersman & Shultz, 1998), der gewählten Strategie beim *faking* (Zickar & Robie, 1999), der Art der erfassten Merkmale (McFarland & Ryan, 2000), der Gelegenheit zum *faking* (Tett, Freund, Christiansen, Fox & Coaster, 2012) und auch dem Ausmaß der Selbsteinsicht und dem Missverstehen der experimentellen Instruktion in Studien zum *faking* (Zickar, Gibby & Robie, 2004), sowie der Motivation zum *faking* (Hauenstein, Bradley, O'Shea, Shah & Magill, 2017), ergeben können. Die in diesem Sinne mehrdimensionale Charakteristik von *faking* wird auch von Levin und Zickar (2002) betont, die im Rahmen einer konzeptuellen Betrachtung, sowohl hinsichtlich der Prozesse als auch der resultierenden Effekte, unterschiedliche Klassen von *faking* definieren. Diese theoretische Betrachtung wird durch Befunde von Zickar et al. (2004) gestützt, die bei ihren Analysen qualitativ verschiedene latente (*faking*-)Klassen finden, welche (inhaltlich) unterschiedlichen Strategien des *faking* zugeordnet werden können. Vor den Hintergrund solcher Befunde stellen Griffith et al. (2011) in einer Merkmalsklassifikationstheorie des *faking* eine Taxonomie von vier qualitativ unterschiedlichen Formen des *faking* vor (Selbstdarstellung, Übertreibung, reaktante Reaktion und betrügerische Reaktion). Die Merkmalsklassifikationstheorie des *faking* von Griffith et al. (2011) integriert dabei bestehende Theorien zur Interaktionen zwischen Persönlichkeitseigenschaften und Situationen (vgl. Cattell, 1980; Endler & Magnusson, 1976, sowie Abschnitt 2.1 in dieser Arbeit), um das Auftreten und Ausmaß bestimmter Formen des *faking* zu erklären.

Neben solchen qualitativ klassifikatorischen Ansätzen zum Phänomen *fa-*

king bestehen auch Ansätze, *faking* aus einer eher eigenschaftsorientierten, dimensionalen Perspektive zu untersuchen. Ausgehend von einer solchen Sichtweise auf *faking* untersuchen B. A. Martin et al. (2002), inwieweit eine individuell unterschiedliche ausgeprägte Fähigkeit zum *faking* in Abhängigkeit unterschiedlicher Einstellungen bezüglich erstrebenswerter Persönlichkeitseigenschaften bestehen. Tett et al. (2012) untersuchen die Zusammenhänge zwischen der Tendenz zum *faking* und kognitiver Fähigkeit und schlussfolgern, dass *faking* im Rahmen von Personalauswahlsituationen bei job-relevanten Eigenschaften eher bei intelligenteren Personen zu beobachten ist, wenn diese eine Gelegenheit dazu haben.

McFarland und Ryan (2006) erklären die Variabilität im *faking* unter Rückbezug auf die Theorie des geplanten Verhaltens nach Ajzen (1991). Ajzen (1991) schlägt in seiner Theorie des geplanten Verhaltens vor, dass die Absicht ein bestimmtes Verhalten durchzuführen, durch eine Kombination von *Einstellung*, *subjektiver Norm* und wahrgenommener *Kontrolle* in Bezug auf das Verhalten, vorhergesagt werden kann. *faking* resultiert danach als Ergebnis eines solchen Bewertungsprozesses. McFarland und Ryan (2006) zeigten in zwei Untersuchungen, dass 45% bis 57% der Varianz in der Tendenz zum *faking* durch eine Kombination aus Einstellung, wahrgenommener Kontrolle und subjektiver Norm bezüglich des *fakings* vorhergesagt werden können. Dieses theoretische Modell und die entsprechenden empirischen Befunde zur Erklärung von *faking* werden von Grieve und McSwiggan (2014) um (gesellschaftliche) moralische Normen und allgemein akzeptierte ethische Vorstellungen erweitert und erklären dabei zusätzliche 14% der Varianz in der Bereitschaft zum *faking*.

Bei der Untersuchung von *faking* bestehen, wie gezeigt, zwei Modellierungsansätze. Einerseits wird *faking* dabei als Manifestation von qualitativ unterschiedlichen Antwortmustern angesehen und andererseits als kontinuierliche und quantitative Variable. In einer aktuellen Arbeit von Ziegler, Maaß, Griffith und Gammon (2015) werden diese beiden Ansätze integriert. Die Befunde aus der Anwendung dieses integrierten Modellierungsansatzes stützen die Sichtweise, dass *faking* vor allem als kontinuierliche und quantitative Variable aufgefasst werden kann (Ziegler et al., 2015).

Wie bereits im Zusammenhang mit SDR erwähnt, bestehen in der Literatur widersprüchliche Befunde zu den Auswirkungen von Verzerrungen wie *faking*

und SDR auf die Validität von Fragebogenverfahren. In einer Metaanalyse von 51 Einzelstudien fanden Viswesvaran und Ones (1999), dass Personen die in einer experimentellen Bedingung zum *faking* aufgefordert wurden, ihre Testwerte in Big-Five-Inventaren im Durchschnitt um etwa 0,75 Standardabweichungen (within-subject-design) oder um 0,5 Standardabweichungen (between-subject-design) erhöhten konnten. Alle dabei untersuchten Big-Five-Faktoren waren in gleicher Weise von *faking* beeinflusst. Allerdings zeigt sich bei Untersuchungen mit kriteriumsbezogenem Design, dass *faking* in den meisten Studien die Kriteriumsvalidität kaum beeinflusst (Hough, Eaton, Dunnette, Kamp & McCloy, 1990; McCrae & Costa, 1983b). Solche Befunde legen nahe, dass die Personen ihre Testwerte durch *faking* beeinflussen können, wenn sie im Rahmen eines experimentellen Designs dazu aufgefordert werden. Gleichzeitig zeigt sich aber, dass *faking* in der praktischen Anwendung unter bestimmten Umständen einen geringen Einfluss auf die *relative Rangreihe* der getesteten Personen haben kann (Hough et al., 1990; B. A. Martin et al., 2002; Ones et al., 1996). Eine Erklärung für diese unterschiedlichen Ergebnisse könnte für einige Studien in einer fehlenden oder unzureichenden experimentellen Kontrolle individueller Unterschiede in der Fähigkeit zur Verfälschung begründete werden (Geiger, Sauter, Olderbak & Wilhelm, 2016; McFarland & Ryan, 2000; Mersman & Shultz, 1998; Tett et al., 2012). So zeigten McFarland und Ryan (2000) in ihrer Untersuchung, dass erhebliche Unterschiede zwischen einzelnen Personen hinsichtlich des Ausmaßes der Verfälschung bei drei verschiedenen Arten von Fragebogenverfahren bestehen. Nach den Befunden von McFarland und Ryan (2000) besteht demnach ein substanzieller Zusammenhang zwischen Gewissenhaftigkeit und Neurotizismus und einer Tendenz zur Verfälschung. Darüber hinaus legen die Befunde von Tett et al. (2012) nahe, dass *faking* in *high stake* Testsituationen mit höherer Intelligenz (vgl. auch Geiger et al., 2016) einhergeht, allerdings nicht in *low stake* Testsituationen.

Im Gegensatz zu SDR, welches zumindest implizit meist als unipolare Dimension aufgefasst wird (nicht vorhanden vs. in einem gewissen Ausmaß vorhanden), lassen sich beim *faking* bezogen auf das zu messende Merkmal zwei Pole theoretisch begründen. Einerseits ist dies das *faking good* im Sinne einer positiven Übertreibung, wie es beispielsweise in Persönlichkeitsinventaren im Rahmen von Bewerbungssituationen beobachtet wird (Birkeland, Manson,

Kisamore, Brannick & Smith, 2006; Lombardi & Pastore, 2015). Die Modellierung von *faking good* führt zu mehr gemeinsamer Varianz, die allerdings keinen Zusammenhang zu substanzieller Trait-Varianz hat (Zickar & Robie, 1999). Demgegenüber lässt sich insbesondere in klinischen Kontexten im Sinne einer Dissimulation das so bezeichnete *faking bad* in Persönlichkeitsinventaren und klinischen Skalen beobachten (Boon, Gozna & Hall, 2008; Ray et al., 2013), mit negativen Konsequenzen für die diagnostische Praxis (Franke, 2002). Hinsichtlich des Einsatzes von Persönlichkeitsinventaren in der Diagnostik folgert Kubinger (2002) als Resultat von drei experimentellen Untersuchungen, dass *faking* beim Einsatz dieser Inventare stattfindet, und dass dies in Abhängigkeit bestimmter Persönlichkeitsausprägungen mehr oder weniger stark auftritt, und dass die Personenauswahl auf Basis der so erzielten Testergebnisse unfair sein kann. Eine solche negative Auswirkung auf die Messung der eigentlich zu erfassenden Merkmale stehen in unmittelbarer Verbindung mit der Verletzung des angenommenen psychometrischen Antwort-, bzw. Messmodells (vgl. Kapitel 4 *Psychometrische Modellierung*). In einer Übersicht über vier einzelne Studien folgert Seiwald (2002) in Bezug auf Persönlichkeitsinventare in diesem Sinne, dass bei keinem der untersuchten Inventare in drei der vier Untersuchungen aufgrund von *faking* Rasch-Homogenität bezüglich der Skalen vorlag.

Trotz dieser empirisch belegten Relevanz von *faking* im Hinblick auf die Gültigkeit von Skalierungsmodellen besteht Uneinigkeit darüber, ob *faking* (in Abgrenzung zu SDR) als eigenständiges Konstrukt aufgefasst werden solle (Griffith & Peterson, 2011; Holden & Passey, 2010; Kuncel & Borneman, 2007; B. A. Martin et al., 2002). Holden und Passey (2010) schlussfolgern beispielsweise, dass bei Testbedingungen mit sozial und gesellschaftlich geringen Konsequenzen (low stakes testing) *sozial erwünschtes Antwortverhalten* als allgemeine Methodenvarianz auftritt, die nicht notwendigerweise als *faking* zu interpretieren ist. So diskutieren und untersuchen Bensch, Paulhus, Stankov und Ziegler (2017) den Zusammenhang und die Überlappung von *sozialer Erwünschtheit*, *overclaiming* und *faking* als potentiell unerwünschte Varianzquelle bei der Auswertung von Fragebogendaten. In Abgrenzung zu SDR stellen Kiefer und Benit (2016) als differentielle Definition von *faking* fest, dass dieses eher als Verhalten bezogen auf eine konkrete Situation, anstatt als überdauernde Persönlichkeitseigenschaft aufgefasst wird (C. MacCann, Ziegler & Roberts,

2012) und dass *faking* als intentionales Verhalten definiert wird (Griffith et al., 2011). Drittens besteht im Gegensatz zu SDR bezogen auf die Diskrepanz zwischen dem wahren Selbstbild und dem berichteten Selbstbild, bei *faking* keinerlei a priori Annahme bezüglich der Richtung dieser Diskrepanz. Die Definition des Konzepts *sozial erwünschten Antwortverhaltens* ist übergreifend mit dem Einsatz von Selbstberichten über Fragebogenverfahren verbunden, demgegenüber wird das Konzept des *faking* eher im Bereich der Personalauswahl und allgemein in Verbindung mit *high stake* Testsituationen diskutiert (Kiefer & Benit, 2016; Marcus, 2009; O'Connell, Kung & Tristan, 2011).

3.2.2 Akquieszenz und Polarität von Merkmalen und Items

Der Begriff *Akquieszenz* ist im Zusammenhang mit der Untersuchung von Antwortverhalten beim Einsatz von Fragebogenverfahren ein seit langem diskutiertes Phänomen. Im Sinne einer direkten Übersetzung des Fremdwortes *Akquieszenz* soll damit eine Tendenz zur Zustimmung zu Items in Fragebögen - mehr oder weniger unabhängig von deren Inhalt - bezeichnet werden (Cronbach, 1941, 1942; Lentz, 1938; Lorge, 1937). Wie von Rorer (1965) in einer ersten Übersicht zu den frühen Arbeiten zur Untersuchung von Antwortverzerrungen dargestellt, wurde das Phänomen der Akquieszenz erstmals von Lorge (1937) und ein Jahr später von Lentz (1938) jeweils als Konferenzbeitrag auf den Tagungen der American Psychological Association (APA) vorgestellt und diskutiert. Die beiden Beiträge von Lorge (1937) und Lentz (1938) beziehen sich dabei auf Skalen zur Erfassung von Persönlichkeitsmerkmalen. Die Beiträge von Lorge (1937) und Lentz (1938) wurden allerdings von den Autoren jeweils nicht als Aufsatz veröffentlicht, sodass hier lediglich die Zusammenfassungen der Tagungsbeiträge als Quelle zur Verfügung stehen. Nach der Definition von Lorge (1937) und Lentz (1938) besteht die grundlegende Idee des Phänomens der Akquieszenz darin, dass antwortende Personen aufgrund einer allgemeinen Neigung zur Zustimmung den einzelnen Items zuzustimmen, ohne notwendigerweise den Inhalt der Fragen explizit zu elaborieren. Bereits Lorge (1937) sieht dabei diese Neigung zu akquieszentem Antwortverhalten als Ausdruck eines speziellen Aspekts der Persönlichkeit an. Bereits Lentz (1938) betrachtet

Akquieszenz als eine potentiell schwerwiegende Verzerrung der Messung und schlägt als teilweise Lösung der Problematik die ausgewogene Verwendung von positiv und negativ formulierten Items vor, wobei er andererseits ein solches Vorgehen im Hinblick auf eine mögliche Konfundierung mit dem zu messenden Merkmal bereits sehr kritisch diskutiert. Die ersten systematischen und publizierten Untersuchungen zu akquieszentem Antwortverhalten wurden von Cronbach (1941, 1942) durchgeführt. Im Gegensatz zu den Beiträgen von Lorge (1937) und Lentz (1938) war Cronbach zunächst an der Verwendung von objektiven Tests zur Leistungsmessung interessiert und beschäftigte sich in diesem Zusammenhang mit dem Problem des Ratens. Personen, die beim Raten eher die Antwortkategorie „richtig“ bzw. die Zustimmung zu einer Frage wählen, werden von Cronbach dabei als „akquieszent“ bezeichnet (Cronbach, 1941, 1942). Als Ergebnis seiner Untersuchungen folgert Cronbach (1942, 1946), dass Akquieszenz negative Auswirkungen auf die Validität der Messung haben kann, sowie das negative und positiv formulierte Fragen möglicherweise unterschiedliche Aspekte des zu erfassenden Merkmals abbilden.

Ebenso wie sozial erwünschtes Antwortverhalten, so wurde auch Akquieszenz im Zusammenhang mit Merkmalen der Persönlichkeit diskutiert und als potentieller Ausdruck von ihren grundlegenden Dimensionen aufgefasst (vgl. z. B. Ferrando, Condon & Chico, 2004, für eine Übersicht). Dabei wird einerseits der Befund einer individuell unterschiedlich ausfallenden Tendenz zu akquieszentem Antwortverhalten als Persönlichkeitsvariable betont (z. B. Couch & Keniston, 1960; Morf & Jackson, 1972). Andererseits wird akquieszentes Antwortverhalten als (potentielle) universelle Fehlerquelle bei der Erfassung von Merkmalen und Einstellungen über Fragebogenverfahren angesehen (Ferrando & Lorenzo-Seva, 2010; Hofstee, Berge & Hendriks, 1998; Nunnally, 1978). Diese beiden Perspektiven auf akquieszentes Antwortverhalten korrespondieren auch mit unterschiedlichen Sichtweisen bezüglich der angenommenen Ursachen des Phänomens Akquieszenz (Mellenbergh, 2001). So kann akquieszentes Antwortverhalten einerseits als Spezifikum des eingesetzten Fragebogeninventars betrachtet werden, das bei allen Personen in gleicher Weise auftritt. Andererseits werden die interindividuell unterschiedlichen Effekte akquieszenten Antwortverhaltens betont (G. Johanson & Alsmadi, 2002; G. A. Johanson & Osborn, 2004). Im Zusammenhang mit der Erfassung von

Dimensionen der Persönlichkeit im Rahmen des Big-Five-Modells wird ein im Rahmen faktorenanalytischer Betrachtungen immer wieder gefundener übergeordneter Generalfaktor kontrovers diskutiert. Einerseits als substanzielle Eigenschaft oder andererseits als Artefakt, welcher sich aus akquieszentem Antwortverhalten herleitet (Davies, Connelly, Ones & Birkland, 2015).

Auch die enge konzeptuelle Überlappung und Konfundierung zwischen *Akquieszenz* und *sozial erwünschtem Antwortverhalten* wurde schon recht früh von Edwards (1961) in Bezug auf die Persönlichkeitsskalen des Minesota Multiphasic Personality Inventory (MMPI) untersucht und diskutiert. Diese und vergleichbare Arbeiten führten zu kontroversen Ansichten bezüglich des Phänomens der Akquieszenz und deren Bedeutung und Relevanz bei der Messung (vgl. Diers, 1964; Rorer, 1965; Rundquist, 1966). Einerseits wurde hier auf Basis empirischer Befunde argumentiert, dass es sich bei Akquieszenz letztlich (nur) um Effekte sozial erwünschten Antwortverhaltens (SDR) handelt (vgl. Abschnitt 3.2.1) und andererseits, dass eine allgemeine Tendenz zur Zustimmung als distinkte Eigenschaft der antwortenden Personen (neben SDR) nachgewiesen werden kann. So folgerte McGee (1962) aus seinen Untersuchungen anhand von Skalen zur Erfassung von Persönlichkeitsvariablen, dass es unabhängig von den Skalen- und Iteminhalten kein allgemeines Merkmal wie eine Neigung zu akquieszentem Antwortverhalten gibt (vgl. auch McGee, 1962; Rorer, 1965). Ausgehend von diesen beiden Hypothesen zu Antwortverzerrungen – sozial erwünschtes Antwortverhalten einerseits und Akquieszenz andererseits – untersucht Diers (1964) deren Unterscheidbarkeit als distinkte Phänomene an empirischen Daten aus dem MMPI Persönlichkeitsinventar. Aus den Untersuchungen folgert Diers (1964), dass sich Effekte von Akquieszenz lediglich in den Fällen nachweisen ließen, in denen Effekte der sozialen Erwünschtheit aufgrund von in dieser Hinsicht unauffälliger Iteminhalte nicht vorhanden waren. Nach Diers (1964) scheint daher sozial erwünschtes Antwortverhalten andere Antwortverzerrungen (wie Akquieszenz) zu überdecken, wobei dennoch bei einigen Items akquieszentes Antwortverhalten für einen Teil der Fehlervarianz verantwortlich ist. In diesem Sinne folgert Rundquist (1966), dass die (potentiell verzerrten) Antworten bei Items zur Erfassung von Dimensionen der Persönlichkeit immer das Ergebnis einer Interaktion zwischen sozial erwünschtem Antwortverhalten bezogen auf den Inhalt einer spezifischen Formulierung

der Items und einer gewissen akquieszenten Antworttendenz im Sinne der Definition von Cronbach (1946) ist.

Ein wichtiger Unterschied der frühen Untersuchungen zu Akquieszenz von Lorge (1937) und Lentz (1938) einerseits und derjenigen von Cronbach (1941, 1942) andererseits liegt in der unterschiedlichen empirischen Basis bzw. der verwendeten psychometrischen Skalen als Ausgangspunkt der Definition von akquieszentem Antwortverhalten. Während Cronbach (1941, 1942) seine Definition von Akquieszenz auf die Basis von Untersuchungen mit Leistungstests stützte, bezog sich die Definition von Lorge (1937) und Lentz (1938) auf Skalen zur Erfassung von Merkmalen der Persönlichkeit (vgl. auch Rorer, 1965). So folgert Cronbach (1942), dass akquieszentes Antwortverhalten insbesondere solche Fragen betrifft, welche für die antwortende Personen uneindeutig sind und daher das Raten (bei Leistungstests) wahrscheinlicher machen. Unter der Voraussetzung, dass ein Leistungstest mit insgesamt zehn Fragen beispielsweise fünf Fragen umfasst, die für eine richtige Antwort mit *ja* beantwortet werden müssen und fünf Fragen, die für eine richtige Antwort mit *nein* beantwortet werden müssen, berechnete Cronbach ein Maß für Akquieszenz, indem er die Anzahl der *nein-Antworten* von der Anzahl der *ja-Antworten* subtrahierte. Das relative Überwiegen von *ja-Antworten* (vor dem Hintergrund, dass bei vollständig richtiger Beantwortung aller Testaufgaben die Differenz aus *ja-Antworten* und *nein-Antworten* null ergibt), stellt dann ein Messwert für das Ausmaß der Antwortverzerrung durch Akquieszenz dar. Eine allgemeinere Definition von Cronbach (1946, S. 476) bezeichnet Antwortverzerrungen wie insbesondere Akquieszenz als jegliche Tendenz, die eine Person dazu veranlasst, unterschiedliche Antworten auf Testfragen zu geben, im Vergleich zu einer Darbietung desselben Inhalts in einer anderen Form. Diese (zweite) Definition von Cronbach (1946) erfordert insofern mindestens zwei Itemformulierungen desselben Inhalts in verschiedenen Formen. Für die Erfassung einer allgemeinen Tendenz zur Zustimmung zu Fragebogen-Items (Akquieszenz) werden daher Items in positiver und negativer Formulierung benötigt. Die Diskrepanzen in den Antworten, welche über diejenigen hinausgehen, die bei zwei Präsentationen desselben Inhalts in derselben Form zu erwarten wären, können dann unter bestimmten, inhaltlich zu begründenden Annahmen als Maß für die Antwortverzerrung im Sinne von Akquieszenz interpretiert werden.

Um zu dem Schluss zu kommen, dass eine Person nach dieser Operationalisierung akquieszent antwortet, ist es allerdings notwendig nachzuweisen, dass die Zustimmung sowohl zum positiv und negativ formulierten Item einen logisch und objektiv begründbaren Widerspruch darstellt. Bei Leistungstests ist ein solcher Widerspruch über objektiv richtige Antworten leicht begründbar, wohingegen das bei Einstellungs- und Persönlichkeitsfragebogen, wie sie in den Arbeiten von Lorge (1937) und Lentz (1938) eingesetzt wurden, nicht ohne weiteres der Fall ist. In diesem Sinne folgert auch Rorer (1965), dass bei Fragebogen zu Persönlichkeitsmerkmalen, Einstellungen und Interessen der Inhalt und die Polung der Fragen unweigerlich konfundiert sind, was allenfalls dadurch kontrolliert werden könnte, wenn die Fragen jeweils (paarweise) positiv und negativ formuliert werden. Wobei die paarweise Gruppierung der Fragen theoretisch oder inhaltlich mit Bezug zu dem erfassten Merkmal begründet werden muss.

Im Rahmen der Erfassung allgemeiner Einstellungen oder Dimensionen der Persönlichkeit werden Items als *positiv formuliert* bezeichnet, wenn sie in Richtung des dominanten Pols des zu erfassenden Merkmals formuliert sind. Als *negativ formuliert* werden dagegen Items bezeichnet, wenn sie in die entgegengesetzte Richtung des dominanten Pols des zu erfassenden Merkmals formuliert sind und daher deren beobachtete Antworten vor der (summativen) Berechnung eines Indexes für die Merkmalsausprägung umzukehren sind. Bereits Likert (1932) schlägt in diesem Sinne die Verwendung von unterschiedlich formulierten Items vor:

> To avoid any space error or any tendency to a stereoyped response it seems desirable to have the different statements so worded that about one-half of them have one end of the attitude continuum corresponding to the left or upper part of the reaction alternatives and the other half have the same end of the attitude continuum corresponding to the right or lower part of the reaction alternatives. (Likert, 1932, S. 46)

Bei der Entwicklung von negativ formulierten Items lassen sich zwei grundlegende Prinzipien unterschieden. Dies ist einerseits der Einsatz polarer, antagonistischer Gegensätze, welche inhaltlich begründet werden müssen (z. B. heiß vs. kalt). Andererseits können negierte Aussagen verwendet werden, welche sich aus regulären Items ableiten lassen (Schriesheim & Eisenbach, 1995;

Schriesheim et al., 1991). Gemäß dieser Prinzipien beschreiben van Sonderen et al. (2013) zwei allgemeine Strategien bei der negativen Formulierung von Items. Eine Strategie besteht darin, reguläre (positive) Aussagen entweder durch Wörter (z. B. „nicht“, „kein“, ...) oder um Affix-Morpheme („in-“, „un-“, „dis-“, „-los“, ...) zur Negation zu ergänzen. In diesem Fall soll die Richtung der Aussage des Items geändert werden, ohne deren Inhalt wesentlich zu verändern. Das so neu gebildete Item wird im Hinblick auf das zu erfassende Merkmal als umgekehrt orientiert betrachtet. Die zweite Strategie stützt sich, wie bereits oben angedeutet, auf die Auswahl von Begriffspaaren mit einer inhaltlich entgegengesetzten Bedeutung (van Sonderen et al., 2013). Beispielsweise könnte die (positive) Aussage bei einem Item zur Erfassung der Persönlichkeitsdimension *Gewissenhaftigkeit* zunächst wie folgt lauten „*Ich bin immer pünktlich*“. Zur negativen Formulierung könnte dann einerseits die polar, antagonistische (negative) Aussage „*Ich bin immer spät*“ ersetzt bzw. ergänzt werden. In diesem Fall wird die Richtung des neuen Items durch den Einsatz einer *inhaltlich antagonistischen* Aussage geändert. Andererseits würde die durch das Affix-Morphem („un-“) negierte Aussage aus dem regulären Item lauten „*Ich bin immer* ***un****pünktlich*“. Eine weitere (nicht zu empfehlende) Form der negativen Formulierung, welche durchaus in psychometrischen Skalen aufzufinden ist, besteht in der Ergänzung verneinender Worte wie „nicht“ oder „kein“ zu der ursprünglichen Item Formulierung – z. B. „*Ich bin* ***nicht*** *immer pünktlich*“. Derartige Formulieren, sowie auch doppelte Verneinungen führen in der Regel zu unbeabsichtigten Antworten (vgl. Abschnitt 3.2.3), die der eigentlichen Antwortabsicht entgegenstehen. In allen Fällen der Negierung müssen die beobachteten Antworten zu den Items umgekehrt bewertet werden (rekodiert werden), da die Zustimmung zu den negativ formulierten Items ein Indikator für geringe Gewissenhaftigkeit sein sollte.

Das Prinzip positiv und negativ formulierten Items zur Kontrolle von Akquieszenz bei der Beantwortung von Fragebögen einzusetzen, stützt sich auf die einfache Überlegung, dass (konsistent) antwortende Personen, die einem positiv formulierten Item zur Erfassung einer bestimmten Eigenschaft zustimmen, gleichzeitig ein negativ formuliertes Item zu derselben Eigenschaft dann auch ablehnen. Im Sinne dieser einfachen Überlegung existiert eine Reihe von Literatur, welche als Empfehlung bei der Skalenkonstruktion den Einsatz von

positiv und negativ formulierten (gepolten) Items empfiehlt. (z. B. Thorndike, Angoff & American Council on Education, 1971; Wiggins, 1973; Wright & Masters, 1982). Osterlind (2002, S. 146) empfiehlt beispielsweise als Daumenregel, dass ein Test etwas mehr als 5 bis 10% negativ formulierter Items enthalten sollte. Durch Variieren der Itempolung bei der Fragebogenkonstruktion soll nach Spector (1992, S. 24) eine Verzerrung der Messwerte, welche sich aus Antworttendenzen ergibt, minimiert werden. Spector (1992, S. 24) führt dazu detaillierter aus, dass bei gleicher Anzahl von positiv und negativ formulierten Items in einer Skala, Personen mit akquieszentem Antwortverhalten tendenziell eher mittlere Summenwerte für die erfasste Merkmalsdimension erhalten und deren (verzerrte) Schätzer für die Merkmalsausprägung auf diese Weise den Schätzungen der Mittelwerte für psychometrische Skalen und Tests weit weniger Schaden zufügen. Podsakoff, MacKenzie, Lee und Podsakoff (2003) liefern als mögliche Begründung einer Verwendung negativ (und positiv) formulierter Items die Erklärung, dass negativ formulierte Items (in Nachfolge auf positiv formulierte) als so genannte „kognitive Geschwindigkeitsbegrenzungen“ verstanden werden können, welche verhindern sollen, das die antwortenden Personen in eine automatische Antwortreaktion verfallen. Ergänzend lässt sich die von Podsakoff et al. (2003) gegebene Begründung zur Verwendung negativ (und positiv) formulierter Items durch eine Reihe von Befunden zum Antwortprozess rechtfertigen. So neigen manche der antwortenden Personen bei der Bearbeitung von Fragebogen-Items offenbar dazu, eine automatische Antworttendenz zur Zustimmung zu den Fragen zu zeigen, um die kognitiven Belastung (Lenzner, Kaczmirek & Lenzner, 2010) bei der Fragebogenbearbeitung zu reduzieren (vgl. auch Krosnick, 1991). Diese Art des Antwortverhaltens bezeichnen Krosnick und Alwin (1987) und Krosnick (1999) als *satisficing* [etwa: „Zufriendenstellung“] und definieren damit das Phänomen, ohne größere Anstrengung die erste akzeptable Antwort zu geben, anstatt nach einer optimalen Antwort zu suchen (vgl. auch Roßmann, 2017, sowie auch Abschnitt 3.2.3). In diesem Zusammenhang weisen die Befunde von Knowles und Nathan (1997) darauf hin, dass akquieszentes Antwortverhalten assoziiert ist mit bestimmten Persönlichkeitseigenschaften wie *kognitive Simplizität*, *Intoleranz* und *rigiden*, vorgefassten Denkmustern. Knowles und Condon (1999) schlagen in diesem Zusammenhang, aufbauend auf Gilberts (1991) *Zweistufenmodell des Verste-*

hens, ein Prozessmodell zum Antwortverhalten vor, welches zur Erklärung der Tendenz zur Zustimmung zu Items herangezogen werden kann. Nach Gilbert (1991) erfordert das Verständnis der Frage auf der ersten Stufe zunächst deren automatische Akzeptanz, welche in der zweiten Stufe in einem tiefer gehenden und damit längeren Evaluationsprozess neu überdacht wird. Die Befunde von Knowles und Condon (1999) zeigen, dass Personen mit einer Tendenz zur Zustimmung die entsprechenden Fragen in kürzerer Zeit beantworten, als Fragen, welche abgelehnt werden – wobei sich die Antwortzeiten bei den drei Gruppen der Antworttendenz (Ja-, Nein-Sager und neutral) nicht bedeutsam unterschieden (Knowles & Condon, 1999). Diese Befunde stützen die Modellannahme von Gilbert (1991) in dem Sinne, dass Personen mit einer Tendenz zur Zustimmung die zweite Evaluationsphase des Prozesses zur Beantwortung der Fragen möglicherweise zu einer Effizienzsteigerung (vgl. Krosnick, 1999) bei der Bearbeitung abkürzen (Knowles & Condon, 1999). Das Zweistufenmodell des Verständnisses von Fragen integriert sich in ein allgemeineres und übergreifendes Modell von Tourangeau und Rasinski (1988); Tourangeau et al. (2000) zum Antwortprozess bei der Beantwortung von Fragen im Rahmen sozialwissenschaftlicher Untersuchungen. Danach erfordert die Beantwortung von Fragen einen vierstufigen Prozess: Auf der ersten Stufe steht dabei das *Verständnis* der Frage [*comprehension*], auf der zweiten die *Aktivierung* und das *Wiederabrufen* relevanter Informationen bzw. Vorwissens aus dem Gedächtnis [*retrieval*], auf der dritten die *Integration* der Gedächtnisinhalte in die Fragestellung und die *Beurteilung* im Hinblick auf eine Antwortfindung [*judgement*] und schließlich auf der vierten und letzten Stufe die *Anpassung* der Antwort auf eine Antwortkategorie des vorgegebenen Antwortformats [*response*] (Tourangeau et al., 2000).

Im Zusammenhang mit dem Einsatz positiv und negativ kodierter Items innerhalb einer Skala zur Kontrolle von akquieszentem Antwortverhalten schlagen Schmitt und Stults (1985) die Inspektion der Personenantworten auf der Ebene einzelner Items vor, um die Plausibilität der gegebenen Antworten zu überprüfen. In diesem Sinne argumentieren auch Swain, Weathers und Niedrich (2008), dass die Überprüfung von inkonsistenten Antworten auf negativ (und positiv) formulierten Items dazu verwendet werden kann, um Personen zu identifizieren, welche die einzelnen Items nach einem automatischen Antwort-

muster beantwortet haben. Nach Ziegler (2015) ist der solcherart gemischte Einsatz von positiv und negativ formulierten Items bei der Skalenkonstruktion auch aktuell noch ein häufig propagierter Ansatz zur Kontrolle von Akquieszenz bei der Beantwortung von Fragebögen. Der Ansatz die Verwendung von negativ (und positiv) formulierten Items mit dem Ziel zu begründen, damit eine Verzerrung durch Antworttendenzen zu mindern, welche sich unabhängig vom Skaleninhalt ergeben, ist insofern erstaunlich, als dass es bereits seit etlichen Jahren empirische Hinweise dafür gibt, dass mit einem solchen Vorgehen eine Reihe von methodischen Problemen einhergehen. So erfordert beispielsweise die Bearbeitung (durch Negation gebildeter) negativ formulierter Items in einem höheren Maße verbale Fähigkeiten (Marsh, 1996). Darüber hinaus haben explorative und konfirmatorische Faktorenanalysen vieler Autoren gezeigt, dass negativ formulierte Items in einer gemischten Skala einen zusätzlichen Faktor bilden. So konnten Schmitt und Stults (1985) zeigen, dass bereits eine relativ kleine Personengruppe innerhalb einer befragten Stichprobe mit nachlässigem Antwortverhalten für das Auftreten eines (zusätzlichen) Faktors verantwortlich sein kann, welcher ausschließlich auf negativ gepolte Items zurückgeht (vgl. auch Arias & Arias, 2017; DiStefano & Motl, 2009; Horan et al., 2003; Lam & Stevens, 1994; Matschinger & Krebs, 1998; McPherson & Mohr, 2005; Schriesheim & Eisenbach, 1995; Spector, Van Katwyk, Brannick & Chen, 1997).

Allerdings kann die gemischte Verwendung negativ und positiv formulierter Items auch auf der Basis theoretischer Definitionen des zu erfassenden Konstruktes *inhaltlich* begründet werden. In diesem Zusammenhang stellt sich dann die Frage nach der Polarität der zu entwickelnden psychometrischen Skala bzw. der damit zu erfassenden psychologischen Merkmalsdimension (z. B. Reise & Waller, 2009; Russell & Carroll, 1999a, 1999b; Watson & Tellegen, 1999). So wird der Einsatz negativ und positiv formulierter Items bei der Erfassung grundlegender Dimensionen der Persönlichkeit auch mit der bipolaren Natur des zu erfassenden Merkmals begründet. Als ein Beispiel sei hier die Merkmalsdimension *Extraversion* (bzw. *Introversion*) genannt, welche bereits von Jung (1921) als grundlegendes, bipolar antagonistisches Merkmal der Persönlichkeit eingeführt wurde (vgl. auch Abschnitt 2.1). Für die Merkmalsdimension *Extraversion* wird dabei eine bipolare Skala bei der Entwicklung der einzelnen

Items angenommen, was sich aus der theoretischen Konzeption des Merkmals als bipolares Kontinuum mit den antagonistisch gegenüberliegenden Extremen *extravertiert* bis *introvertiert* begründet. In diesem Sinne argumentieren beispielsweise Russell und Carroll (1999a), dass sich durch die Annahme bipolarer Konstruktdimensionen als inhaltlich zusammenhängende Pole ein und desselben Merkmals manche empirische Daten besser erklären lassen.

Die Annahme einer unipolaren Skala bietet sich dagegen immer dann an, wenn sich für das zu erfassende Merkmal nach inhaltlich theoretischen Überlegungen ein natürlicher Nullpunkt definieren lässt, wie es zum Beispiel bei Fragen nach *Häufigkeiten*, *Intensitäten*, *Wahrscheinlichkeiten* und *Prozentangaben* der Fall wäre, nicht aber notwendigerweise bei allgemeinen *Bewertungen* wie zum Beispiel dem Kontinuum zwischen den beiden Valenzpolen *totale Ablehnung* und *absolute Zustimmung* (vgl. dazu Kaplan, 1972). In seinem Aufsatz zum Problem der Ambivalenz von bipolaren Merkmalsdimensionen argumentiert Kaplan (1972), dass Situationen denkbar sind, bei denen die antwortende Person gleichzeitig eine positive als auch ablehnende Haltung als Ausdruck von (wahrgenommener) Ambivalenz zum Ausdruck bringen möchte. In solchen Fällen wäre die angenommene Bipolarität der eindimensionalen Skala zugunsten der Annahme eines eher zweidimensionalen Merkmalsraumes kritisch zu hinterfragen (vgl. Spector et al., 1997). Demgegenüber weisen manche empirische Untersuchungen darauf hin, dass allgemeine Einstellungen und Bewertungen in ihrer kognitiven Repräsentation eine bipolare Struktur aufweisen (Judd & Kulik, 1980). In diesem Sinne konzeptionalisieren Pratkanis (1989) und Pratkanis und Greenwald (1989) die kognitive Repräsentation allgemeiner Einstellungen als bipolares Kontinuum. Diese individuellen *Einstellungen* stellen nach Ajzen und Fishbein (1972) eine Prädisposition dar, um allgemein in einer entweder günstigen (zustimmenden) oder ungünstigen (ablehnenden) Weise in Bezug auf den Gegenstand der Beurteilung zu antworten. Im Rahmen der Theorie zu individuellen Einstellungen lassen sich hierbei übergreifend affekt- und kognitionsbasierte Einstellungen unterschieden (Ajzen, 2001). Affektbasierte Einstellungen sind mit einer starken und unmittelbaren (zustimmenden oder ablehnenden) affektiven Reaktion auf ein Einstellungsobjekt assoziiert, während kognitionsbasierte Einstellungen auf evaluativen Überzeugungen oder bestehenden Schemata über ein Einstellungsobjekt beruhen (Willson, Dunn, Kraft & Lisle, 1989).

Vor dem Hintergrund solcher inhaltlich begründeter „negativer“ (und „positiver“) Itemformulierungen innerhalb einer psychometrischen Skala muss dann allerdings die Praxis der Kontrolle oder die Erfassung von akquieszentem Antwortverhalten über gemischt gepolte Items im Sinne einer Entdeckung von inkonsistentem Antwortverhalten, kritisch betrachtet werden, da hier eine Konfundierung mit dem zu erfassenden Merkmal zu befürchten ist (vgl. z. B. Tellegen, 1988). Die Untersuchung solch einer möglichen Konfundierung inhaltlicher Aspekte mit akquieszentem Antwortverhalten war bereits relativ früh Gegenstand einer Studie von Bass (1955), welcher die bekannte Autoritarismus *F*–Skala von Adorno (1950) zum Untersuchungsgegenstand hatte. Bass (1955) folgerte aus seinen Untersuchungen, dass insgesamt Akquieszenz in den Antworten zur der *F*–Skala zu finden ist (vgl. auch Chapman & Campbell, 1957) und Korrelationen zwischen der *F*–Skala und anderen Konstrukten hauptsächlich auf akquieszente Antworttendenzen [response sets] zurückgehen und diese mit steigender Ambivalenz der Items bzw. deren Inhalten ansteigt. Der ursprüngliche Befund, dass drei Viertel der Varianz in der *F*–Skala mit akquieszenten Antworttendenzen assoziiert sind (vgl. Bass, 1955, S. 623), wurde dagegen später bei einer Überprüfung der Analysen dahingehend relativiert, dass aufgrund eines Vorzeichenfehlers in den ursprünglichen Analysen drei Viertel der Varianz in der *F*–Skala tatsächlich auf das Merkmal Autoritarismus zurückzuführen ist (vgl. Messick & Jackson, 1957; Zuckerman & Norton, 1961).

In einem klassischen Aufsatz zeigt Bentler (1969), dass akquieszentes Antwortverhalten die tatsächlich vorliegende Bipolarität des zu erfassenden Merkmals maskieren kann, indem die Korrelation zwischen den unterschiedlich gepolten Items in eine positive Richtung (anstatt eine negative Richtung) verschoben wird. In ähnlicher Weise legen Green, Goldman und Salovey (1993) in drei Untersuchungen nahe, dass Befunde, welche angeblich die Unabhängigkeit scheinbar gegensätzlicher Affektzustände zeigen – im Sinne von niedrigen Korrelationen zwischen positiven und negativen Affekten – das Ergebnis von Versäumnissen sein können, Verzerrungen aufgrund von zufälligen und nichtzufälligen Antwortfehlern angemessen zu berücksichtigen. Die Berücksichtigung dieser Fehlerquellen, welche sich durch den Einsatz von Fragebogen zur Selbstbeurteilung ergeben können, durch den Einsatz alternativer Methoden zur Erfassung der affektiven Zustände, zeigt dann demgegenüber eine weit-

gehend bipolare Affektstruktur (Green et al., 1993; Green, Salovey & Truax, 1999). Vor dem Hintergrund derartiger Befunde schlägt Diener (1999) neben dem Einsatz alternativer Methoden zur Erfassung von affektiven Zuständen für die zukünftige Forschung den Einsatz statistischer Verfahren zur Entdeckung von Antwortverzerrungen und die systematische Variation der formalen Struktur der Antwortskalen für die Items vor. Segura und González-Romá (2003) untersuchen dazu beispielsweise die möglicherweise unterschiedliche Interpretation der Antwortskala als entweder uni- oder bipolar bei Fragen zu affektbasierten Einstellungen. Die Befunde von Segura und González-Romá (2003) stützen zunächst die Hypothese einer bipolaren Repräsentation affektbasierter Einstellungen. Allerdings ist dabei die Polarität der mentalen Repräsentation des zu erfassenden Konstruktes eng verknüpft mit der Polarität der Formulierung der einzelnen Fragen und fällt individuell unterschiedlich aus. Segura und González-Romá (2003) merken daher weiter an, dass sich für einige der befragten Personen unerwartete Antwortmuster nicht etwa ergeben, weil deren Antworten falsch wären, sondern weil diese nicht mit den Vorstellungen des Testentwicklers (zur Polarität) übereinstimmen.

In Anlehnung an konzeptionelle Überlegungen von Russell und Carroll (1999a) können bezüglich der Antwortskalen für Items zunächst grundlegend, eindeutig *unipolare* [strictly unipolar] und eindeutig *bipolare* [strictly bipolar] Antwortskalen unterschieden werden. Die Polarität der Antwortskalen ist dabei zwar einerseits zu unterscheiden von der Polarität des zu messenden Merkmals, steht aber andererseits auch in Verbindung mit der (wahrgenommenen) Uni- oder Bipolarität des zu messenden Merkmals. Zur Verdeutlichung dieser Kategorisierung der Antwortskalen für Items sei hier auf die untenstehende Darstellung in Abbildung 3.1 verwiesen. Typischerweise umfassen bipolare Antwortskalen antagonistische Begriffspaare, welche jeweils die beiden Endpunkte der Antwortskala spezifizieren. Beispielsweise könnte die Antwortskala eines Items zur Erfassung des „Gemütszustandes“ die beiden „Endpunkte“ *traurig* und *froh* beinhalten. Als eindeutig *bipolar*, kann eine Antwortskala definiert werden, welche (theoretisch) das gesamte Kontinuum der als ebenfalls antagonistisch, bipolar angenommenen Merkmalsdimension umfasst. (vgl. Abbildung 3.1 unten). Als eindeutig *unipolar* ist dagegen eine Antwortskala zu bezeichnen, wenn diese das (eindimensionale) Merkmalskontinuum, ausgehend von einem Null-

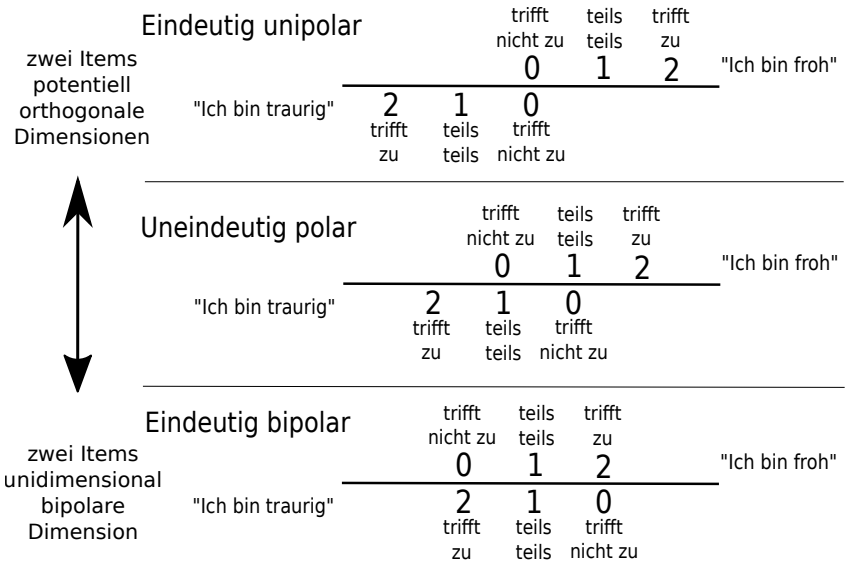


Abbildung 3.1 Schematische Darstellung zu Interpretationsmöglichkeiten der Polarität der Antwortskala von zwei (potentiell) antagonistisch konzipierten Items

punkt als Ausmaß der Intensität des Merkmals im Sinne einer Angabe von Prävalenzen zu bestimmten Aussagen, abdeckt (vgl. Abbildung 3.1 oben). Im Sinne einer unipolaren Interpretation der Antwortskalen des hypothetischen Konstruktes „Gemütszustand“ mit den potentiell unabhängig, orthogonalen Dimensionen *Traurigkeit* und *Fröhlichkeit* sind dabei in der schematischen Darstellung in Abbildung 3.1 oben zwei unabhängige Items (unterschiedlicher Polarität) abgebildet. In Analogie zu der Argumentation von Kaplan (1972), dass die antwortende Person gleichzeitig weder *traurig* noch *fröhlich* ist, also beiden Items gegenüber eine ablehnende Haltung als Ausdruck von Ambivalenz zum Ausdruck bringen möchte, wäre die angenommene Bipolarität der Antwortskala zugunsten der Annahme eines eher zweidimensionalen Merkmalsraumes aufzugeben (vgl. Spector et al., 1997). Zwischen diesen beiden Extremen in Bezug auf die (Eindeutigkeit) der Polarität (unipolar vs. bipolar) können sich unterschiedliche Abstufungen der wahrgenommenen Polarität (in Abhängigkeit der Anzahl der Antwortstufen) ergeben, was in der Abbildung 3.1 in der Mitte exemplarisch dargestellt ist.

Der hier exemplarisch dargestellte Komplex aus Polarität des zu erfassen-

den Merkmals einerseits und der Antwortskalen der korrespondierenden Items andererseits, ist Gegenstand kontroverser Sichtweisen auf kognitive Repräsentationen von Gefühlen, Affekten und Einstellungen – vgl. z. B. Russell und Carroll (1999a, 1999b) einerseits und Watson und Tellegen (1999) andererseits. Unabhängig von dieser Debatte argumentieren Segura und González-Romá (2003), dass die Interpretation und Zuordnung der Antwortskalen als uni- bzw. bipolar letztlich nur mehr oder weniger eindeutig objektiv gegeben ist (vgl. auch Rey, Abad, Barrada, Garrido & Ponsoda, 2014). Vielmehr kann davon ausgegangen werden, dass möglicherweise interindividuelle Unterschiede in der Wahrnehmung der Polarität der Antwortskalen der Items und der dahinterliegenden Merkmalsdimension bestehen können. Die Abbildung 3.1 zeigt am Beispiel der oben bereits erwähnten hypothetischen Merkmalsdimension des „Gemütszustandes“ schematisch die unterschiedlichen Interpretationsmöglichkeiten von zwei potentiell antagonistisch interpretierbaren Items, welche sich aus der unterschiedlichen Wahrnehmung der Polarität der Antwortskalen ergeben können. Vor dem Hintergrund solcher theoretischen Überlegungen lassen sich dann Befunde zu den hinsichtlich der Messgenauigkeit negativen Auswirkungen einer Verwendung von negativ und positiv formulierten Items in gemischten Skalen erklären. So zeigen sich bei empirischen Untersuchungen (je nach Stichprobenzusammensetzung) oft unterschiedliche Befunde zur Dimensionalität (z. B. Rey et al., 2014) der Konstrukte oder der Polarität der gemessenen Merkmalsdimension. Bereits Schriesheim und Hill (1981) finden beispielsweise, dass die Verwendung positiv und negativ formulierter Items zur Kontrolle von Akquieszenz die Messgenauigkeit negativ beeinflusst. Ferner stellen Matschinger und Krebs (1998) dar, dass eine gemischte Verwendung gegensätzlicher Itempolungen innerhalb einer Skala oft eine Bipolarität bzw. eine zusätzliche Dimension im Rahmen faktorenanalytischer Auswertungen erzeugt, auch wenn sich diese theoriebasiert für das gemessene Konstrukt nicht begründen lässt (vgl. auch DiStefano & Motl, 2009; Merritt, 2012). Ferner weisen Kulas, Klahr und Knights (2018) auf eine mögliche Konfundierung hin, welche zwischen dem jeweiligen Ausmaß der sozialen Erwünschtheit des Inhaltes eines Items und dessen Polarität im Hinblick auf das zu erfassende Merkmal, bestehen kann.

Aus eher inhaltlicher Perspektive folgern Rodebaugh, Woods und Heimberg (2007) auf der Basis ihrer Untersuchungen mit Skalen zur Erfassung von Ängstlichkeit, dass negative formulierte Items und das Gegenteil von Ängstlichkeit nicht dasselbe Merkmal darstellen. Ferner finden Rodebaugh et al. (2004), dass der Einsatz positiv formulierter Items in einer höheren Validität der betreffenden Skalen resultiert. G. A. Johanson und Osborn (2004) modellieren die Effekte akquieszenten Antwortverhaltens (im Sinne der unterschiedlichen Rezeption der Polarität der Merkmale) als differentielle Personenpassung auf das implizit zugrunde gelegte (kumulative) Antwortmodell. Als Ergebnis einer aktuelleren Analyse im Rahmen der Item-Response-Theory folgern Sliter und Zickar (2014), dass positiv und negativ formulierte Items nicht als (im inhaltlichen Sinne) austauschbar zu verstehen sind und negativ formulierte Items bei der Kontrolle akquieszenten Antwortverhaltens eine begrenzte Nützlichkeit haben. Dennoch ist die Praxis einer gemischten Verwendung von negativ und positiv formulierten Items bei der Konstruktion von psychometrischen Skalen so weit verbreitet, dass viele Praktiker ihre Skalen entsprechend entwickeln, ohne dies weiter zu begründen. Es erscheint, dass diese „gängige Praxis“ so gut etabliert ist, dass eine weitere Begründung nicht notwendig ist (Dalal & Carter, 2015b). In einer Übersicht zu Aufsätzen seit dem Jahr 2000, welche die Entwicklung von psychometrischen Skalen im Bereich der Arbeits- und Organisationspsychologie thematisieren, zeigen Dalal und Carter (2015b), dass 19 der 50 in den Aufsätzen berichteten Skalen (38%) negativ (und positiv) formulierte Items verwenden, dabei aber nur 10 Aufsätze (20%) diese Verwendung explizit begründen. Von denjenigen Aufsätzen, welche eine Begründung enthalten, sind die zwei am häufigsten genannten Gründe für die Aufnahme von Items mit negativem Wortlaut: (1) dass die Items auf früheren Skalen basieren, die negativ formulierte Items enthalten (40%) und (2) dieses Prinzip der Skalenkonstruktion Antwortverzerrungen reduzieren solle (30%). Die Konstruktion von Skalen mit gemischt formulierten Items basiert also einerseits auf der simplen Übernahme etablierter Konstruktionsprinzipien und andererseits auf der zumindest kritisch infrage zu stellenden Annahme, dadurch verzerrenden Antworttendenzen vorzubeugen. Insofern bezeichnen Dalal und Carter (2015b) die Verwendung gemischt formulierter Items in psychometrischen Skalen als „urbane Legende“, welche sich in der Praxis der Fragebogenkonstruktion eta-

blieren konnte, obwohl es in erheblichem Umfang empirische Hinweise dafür gibt, dass mit einem solchen Vorgehen eine ganze Reihe von methodischen Problemen einhergehen (z. B. Arias & Arias, 2017; Barnette, 2000; Bentler, Jackson & Messick, 1971; Cronbach, 1942, 1946; DiStefano & Motl, 2009; Gu, Wen & Fan, 2015; Idaszak & Drasgow, 1987; Kulas et al., 2018; Lam & Stevens, 1994; Marsh, 1996; McPherson & Mohr, 2005; Merritt, 2012; Michaelides et al., 2016; Pilotte & Gable, 1990; Roszkowski & Soven, 2010; Schmitt & Stults, 1985; Schriesheim & Eisenbach, 1995; Schriesheim et al., 1991; Schriesheim & Hill, 1981; Spector et al., 1997; Stewart & Frye, 2004; van Sonderen et al., 2013; Wang, Chen & Jin, 2015).

3.2.3 Fehlantworten und unaufmerksames Antwortverhalten

Die erhöhten kognitiven Anforderungen bei der Beantwortung von negativ formulierten Antworten (Marsh, 1996) kann, neben intendierten Verzerrungen wie SDR und *faking*, bei manchen Personen auch dazu führen, versehentliche *Fehlantworten* [*missresponse*] aufgrund einer unaufmerksamen Bearbeitung bei der Beantwortung zu erzeugen. Solche Fehlantworten [missresponse] werden von Swain et al. (2008) insbesondere in Verbindung mit dem Einsatz von positiv und negativ formulierten Items dadurch definiert, dass Personen aufgrund der negativen Codierung (versehentlich) die ihrer eigentlichen Überzeugung gegenteilig gegenüberliegende Antwortkategorie wählen (vgl. auch Weijters & Baumgartner, 2012). Bereits Schmitt und Stults (1985) zeigen in Ihren Untersuchungen, dass bereits ein kleiner Anteil von unaufmerksam antwortenden Personen bei einer faktoranalytischen Auswertung der resultierenden Daten dazu führt, dass sich für negativ formulierte Items ein gesonderter Faktor ergibt. Insofern kann ein solcher, bereits weiter oben dargestellter *Methodenfaktor* (vgl. auch Arias & Arias, 2017; DiStefano & Motl, 2009; Horan et al., 2003; Kam & Meyer, 2015; Lam & Stevens, 1994; Maraun & Rossi, 2001; Matschinger & Krebs, 1998; McPherson & Mohr, 2005; Schriesheim & Eisenbach, 1995; Spector et al., 1997), nicht nur aus der Perspektive der Inhalte oder der Form der jeweiligen psychometrischen Skala, sondern auch aus der Perspektive der antwortenden Personen und deren motivationaler Situation interpretiert werden.

Maniaci und Rogge (2014) untersuchten die negativen Auswirkungen von unaufmerksamem Antwortverhalten [*Careless Responding*] auf Befunde zu Kon-

struktzusammenhängen. Die Ergebnisse von Maniaci und Rogge (2014) deuten zunächst darauf hin, dass 3-9% der Befragten eher ein unaufmerksames Antwortverhalten zeigen und bei einer personenorientierten Auswertung der Daten eine *latente Klasse* bilden (vgl. Abschnitt 4.6 in Kapitel 4 *Psychometrische Modellierung* zum Begriff der latenten Klasse). Für diese Personengruppe ergeben sich Daten von deutlich schlechterer Qualität, die geeignet sind, Ergebnisse zu inhaltlichen Fragestellungen aus Regressionsanalysen oder Effekte aus experimentellen Untersuchungsdesigns zu überdecken (Maniaci & Rogge, 2014). Insgesamt scheint aber der Anteil unaufmerksam antwortender Personen bei unterschiedlichen Untersuchungen allerdings erheblich zu variieren. Über verschiedene Untersuchungen zu unterschiedlichen Inhaltsbereichen hinweg reichen die berichteten Anteile unaufmerksam antwortender Personen über einen weiten Bereich von 3% bis 46% (z. B. Berry et al., 1992; J. A. Johnson, 2005; Maniaci & Rogge, 2014; Meade & Craig, 2012; Oppenheimer, Meyvis & Davidenko, 2009).

Unaufmerksames Antwortverhalten unterscheidet sich konzeptionell von anderen Formen der Antwortverzerrung wie sozial erwünschtem Antwortverhalten (SDR), *faking* oder *Eindruckssteuerung* (IM). So implizieren SDR, *faking* und IM gerade durch die Motivation, sich in einer bestimmten Art und Weise zu präsentieren, eine *aufmerksame* Bearbeitung der Fragen, welche erst die Voraussetzung für eine verzerrte Antwort bildet. Insofern können SDR, *faking* und IM negative Zusammenhänge mit unaufmerksamen Antwortverhalten aufweisen (z. B. Meade & Craig, 2012). In diesem Sinne kann unaufmerksames Antwortverhalten als ein Mangel an Motivation sich in einer bestimmten Art und Weise zu präsentieren, definiert werden (Maniaci & Rogge, 2014). Extreme Formen eines nach solch einer Definition konzeptualisierten unaufmerksamen Antwortverhaltens steht in Einklang mit Befunden zu einer latenten Personenklasse *intrinsisch unskalierbarer* Personen (Dayton & Macready, 1980), für die das zugrunde gelegte (kumulative) Antwortmodell nicht sinnvoll anwendbar ist (vgl. auch Bem, 1977; Bem & Allen, 1974; Bem & Funder, 1978; Dayton & Macready, 1980; Formann, 2002; Ponocny & Klauer, 2002; Rost et al., 1999; Rost & Georg, 1991).

Unaufmerksames und nachlässiges Antwortverhalten korrespondiert auch mit dem von Nichols, Greene und Schmolck (1989) als *Inhalts-Nichtresponsivität*

bezeichnetem Phänomen eines schematisch, stereotypen Antwortverhaltens. Die Ursache solch einer *Nichtresponsivität* bezüglich der zu messenden latenten Variablen kann auch in Verbindung mit dem im Zusammenhang von SDR weiter oben beschriebenen Konzept des *satisficing* (Krosnick, 1991; Lenzner et al., 2010; Roßmann, 2017) beschrieben werden. Nach Krosnick (1991) lassen sich demnach zwei grundlegend verschiedenen Strategien bei der Beantwortung von Fragebogen-Items unterscheiden. Einerseits das *optimizing*, wobei die antwortenden Personen eine Antwortstrategie wählen, die auf eine möglichst genaue Beantwortung der vorgegebenen Fragen abzielt, und andererseits das *satisficing*, wobei die antwortenden Personen eine Strategie anwenden, welche eine Reduzierung des Aufwands bei der Beantwortung der Fragen zum Ziel hat (Roßmann, 2017). Im Verbindung mit dem Konzept der Metaeigenschaften (Baumeister & Tice, 1988; Britt, 1993; Tellegen, 1988), welche im Hinblick auf die Erfassung inhaltlich motivierter latenter Eigenschaften dazu führen, dass bei bestimmten Personen bestimmte Eigenschaften nicht immer messbar sind (Bem, 1977; Bem & Funder, 1978; Tracey, 2003), wird die individuell unterschiedlich gezeigte Strategie bei der Beantwortung von Fragebogen dabei im Sinne einer konsistenten Personeneigenschaft interpretiert. Allerdings weisen bereits Knowles und Byers (1996) im Zusammenhang mit der Frage nach der Konsistenz von Antwortverhalten und Reliabilität der Messung im Verlauf der Beantwortung einzelner Fragen auf die mögliche *Reaktivität* der Messung hin. Demnach lernen die antwortenden Personen auch durch die vorhergegangenen Fragen das Konstrukt und auch die Messintention während der Fragebogenbearbeitung immer besser kennen (Knowles, 1988), wodurch sich die Messgenauigkeit und auch die Antwortstrategie im Laufe der Bearbeitung verändern kann. Bei verpflichtend vollständigen Antworten, wie sie sich beispielsweise im Rahmen von Online–Erhebungen mit Überprüfung der Vollständigkeit der gegebenen Antworten zu den einzelnen Items realisieren lässt, zeigt sich dabei eine vergleichsweise neue Art der Antwortverweigerung in Form von konstanten Antwortmustern in der resultierenden Datenmatrix. Dieses Antwortmuster wird beispielsweise von Menictas, Wang und Fine (2011) auch als *flat-lining* bezeichnet. Die befragten Personen „beantworten“ die einzelnen Items einer Skala dabei über die unreflektierte Wahl immer derselben Antwortkategorie. Meade und Craig (2012) identifizieren in ihrer Untersuchung dementsprechend

zwei unterscheidbare Typen von Antwortmustern als Ergebnis unaufmerksamen Antwortverhaltens. Einerseits ein zufälliges Antwortmuster, welches sich aus der spontanen Wahl einer beliebigen Antwortkategorie der vorgegebenen Antwortskala ergibt. Andererseits ein nichtzufälliges Antwortmuster, welches mit dem von Menictas et al. (2011) als *flat-lining* bezeichnetem Antwortmuster korrespondiert. Ebenso wie das Phänomen Akquieszenz wird auch unaufmerksames Antwortverhalten im Zusammenhang mit der unterschiedlichen Polarität von Items in entsprechend konstruierten, gemischten Skalen diskutiert (Weijters & Baumgartner, 2012). Der Mechanismus des unaufmerksamen Antwortverhaltens wird dabei nach Weijters und Baumgartner (2012) im Rahmen des vierstufigen Prozessmodells zur Beantwortung von Fragebogen (Tourangeau & Rasinski, 1988; Tourangeau et al., 2000) auf dessen erster Stufe, dem *Verständnis* der Frage [*comprehension*], zugeordnet. Zur Entdeckung (zufälliger) abweichender Antwortmuster aufgrund einer nachlässigen Bearbeitung der vorgelegten Skalen schlagen T. Liu, Lan und Xin (2016) in einer aktuellen Arbeit so genannte Personen-Fit-Statistiken (z. B. Tatsuoka, 1985; Tatsuoka & Linn, 1983) vor, wie sie aus entsprechenden Antwortmodellen im Rahmen der Item-Response-Theory (IRT) abgeleitet werden können (vgl. dazu Abschnitt 4.4.2).

3.2.4 Antwortmuster bei unterschiedlichen Antwortskalen

Neben den im vorangegangenen Abschnitt dargestellten Formen der Antwortverzerrung, welche sich im Wesentlichen auf inhaltliche Aspekte, entweder der einzelnen Items oder auch der zu messenden Eigenschaften, beziehen lassen, kann bei der Auswertung von Daten aus der Anwendung von Fragebogenverfahren auch eine Form der Verzerrung gefunden werden, welche sich scheinbar unabhängig vom Inhalt der einzelnen Items ergibt. Eine solche Form der Antwortverzerrungen [*response bias*] definiert Paulhus (1991) allgemein als „*... a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content ...*“ (Paulhus, 1991, S. 17) – also als eine systematische Tendenz auf eine Reihe von Fragen in bestimmter Weise zu antworten, unabhängig von deren Inhalt. Diese Definition von Paulhus

(1991) weist eine starke Analogie zu der bereits von (Cronbach, 1946, S. 476) gegebenen allgemeinen Definition von „*response sets*“, also Antwortverzerrungen durch idiosynkratische Antwortmuster auf. Während allerdings Cronbach (1946) in seiner Definition den Schwerpunkt auf die *unterschiedliche Form* des ansonsten gleichen Iteminhalts legt „*A response set is defined as any tendency causing a person consistently to give different responses to test items than he would when the same content is presented in a different form.*“ (Cronbach, 1946, S. 476), bezieht sich die Definition von Paulhus (1991) demgegenüber eher auf *antwortenden Personen* und deren (Antwort-)Tendenz: „*...a systematic tendency to respond to a range of questionnaire ...*“ (Paulhus, 1991, S. 17). So lässt sich in den Daten aus der Anwendung von Fragebögen mit mehrstufigen Antwortskalen oft eine unterschiedliche Tendenz entweder zu mittleren (*middle response style* – MRS) oder extremen (*extreme response style* – ERS) Antwortkategorien feststellen, die von einigen Autoren mit einem differentiellen Gebrauch des mehrstufigen Antwortformats in Verbindung gebracht wird (z. B. Adams-Webber & Benjafield, 1973; E. J. Austin, Deary & Egan, 2006; J. A. Bond, 1987; González-Romá & Espejo, 2003; Hernández, Drasgow & González-Romá, 2004; Hernández, Espejo & González-Romá, 2006; Hui & Triandis, 1989; Khorramdel, 2014; Kulas & Stachowski, 2009; Kulas, Stachowski & Haynes, 2008; Moors, 2008; Nunnally, 1978; Wiggins, 1962).

Einige Untersuchungen bezieht sich dabei auf den Bereich der Erfassung interindividuell unterschiedlich ausgeprägter Merkmale wie zum Beispiel Persönlichkeitseigenschaften (Arce-Ferrer, 2006; E. Austin, Deary, Gibson, McGregor & Dent, 1998; E. J. Austin et al., 2006; Berg & Collier, 1953; Damarin & Messick, 1965; Eid & Zickar, 2007; Gollwitzer, Eid & Jürgensen, 2005; Heine, 2010; Hurley, 1998; Meisenberg & Williams, 2008; Minkov, 2017; Naemi, Beal & Payne, 2009; Rost, 2002; Rost, Carstensen & von Davier, 1997; Rost et al., 1999; van der Kloot, Kroonenberg & Bakker, 1985; Warr & Coffman, 1970).

Andere Untersuchungen befassen sich außerdem mit den Auswirkungen von extremen und mittleren Antworttendenzen im Zusammenhang mit kulturvergleichenden Studien (z. B. Bertea & Zait, 2014; Hamamura, Heine & Paulhus, 2008; Roster, Albaum & Rogers, 2006). Ein weiteres Feld ist die Untersuchung von extremen und mittleren Antworttendenzen im Zusammenhang mit Problemstellungen aus dem Bereich Marketing (Dolnicar & Grün, 2007; Rocereto,

Puzakova, Anderson & Kwak, 2011) sowie Fragestellungen aus dem Bereich der Arbeits- und Organisationspsychologie (z. B. Carter, Dalal, Lake, Lin & Zickar, 2011) in Verbindung mit Fragen der Personalauswahl (Zickar et al., 2004; Zickar & Robie, 1999; Ziegler, 2011). Im Sinne der Definitionen von Paulhus (1991) beschreibt bereits Nunnally (1978, S. 612) das oft zu beobachtende Phänomen von ERS beziehungsweise MRS als potentiell zeitlich stabile (Weijters, Geuens & Schillewaert, 2010b; Wetzel, Lüdtke, Zettler & Böhnke, 2016) und konstruktübergreifend, individuell unterschiedlich konsistent (Weijters, Geuens & Schillewaert, 2010a; Wetzel, Carstensen & Böhnke, 2013) ausgeprägte Eigenschaft der antwortenden Personen. Auch Jackson und Messick (1958) führten in diesem Sinne im Rahmen der Betrachtung unterschiedlicher Formen von Messfehlern bei Fragebögen bereits die Unterscheidung zwischen *Antwortverzerrungen* [*response sets*] und *Antwortstilen* [*response styles*] ein. Nunnally (1978, S. 595) führt zu der Unterscheidung zwischen *response styles* und *response bias* aus, dass erstere auf Artefakte der Messung durch individuelle Unterschiede zurückzuführen seien, wohingegen es sich bei *response sets* nicht notwendigerweise um stabile und damit reliabel zu erfassende individuelle Unterschiede handeln muss. Bei einem *response bias* handele es sich demnach eher um Artefakt(e) der Messung, die sich bei den durchschnittlichen Antworten einer Gruppe von Personen zeigen. Als weiteres Beispiel für einen *response bias* führt Nunnally (1978) hier zum Beispiel auch die Verzerrung des Messwertes durch Rate-Effekte bei mehrfach Wahlaufgaben an.

Es kann aber grundsätzlich die Frage gestellt werden, inwieweit sich beispielsweise die von Jackson und Messick (1958) unter dem Oberbegriff *response styles* subsumierten Phänomene ERS und MRS nur als übergeordnete Eigenschaft der antwortenden Personen, oder aber als besondere Eigenschaft des Messinstrumentes interpretieren lassen. So argumentiert beispielsweise Wiggins (1962), dass stilistische Konsistenzen (der Personen) durch die Form des Messinstrumentes mediiert werden. Die formale Struktur des Messinstruments mit in einer bestimmten Form fest vorgegebenen Antwortalternativen sieht Wiggins (1962) dabei als definierenden Rahmen aller möglichen, zu beobachtenden Antwortstile an. Insofern müssen solche stilistischen Tendenzen dann nicht nur im Hinblick auf die antwortenden Personen, sondern auch im Hinblick auf das eingesetzte Messinstrument interpretiert werden (Wiggins, 1962,

S. 224). Im Gensatz zu der von Nunnally (1978) formulierten Hypothese, dass sich Menschen unabhängig vom Inhalt [der Items] in der Tendenz unterscheiden, die Extreme der Ratingskalen auszuwählen und nicht etwa die Mitte der Skala [„*The Hypothesis is that, regardless of the content, peaople differ in the Tendency to mark the extremes of rating scales rather than points near the middle of the scale.*“ (Nunnally, 1978, S. 612)], deutet eine neuere Arbeit von Kulas und Stachowski (2013) darauf hin, dass die Ursache für diese beiden Antwortmuster eher in der Beschaffenheit der Items begründet ist. Im Gegensatz zur Antwortverzerrung der *Sozialen Erwünschtheit* hängt es demnach eher von der Beschaffenheit der Antwortskala als vom unterschiedlichen Inhalt der einzelnen Items einer Skala ab, ob solch unterschiedlich ausgeprägtes Antwortverhalten bei den untersuchten Personen gefunden werden kann. Eine Übersicht über häufig zu beobachtende Antwortverzerrungen auf der Basis unterschiedlicher Antwortformate geben Baumgartner und Steenkamp (2001). Greenleaf (1992b) stellt in einer Überblicksarbeit verschiedene Möglichkeiten zur Operationalisierung und Erfassung von extremem Antwortverhalten außerhalb der Item-Response-Theory dar. Ferner zeigt Greenleaf (1992a), wie solche Informationen dazu herangezogen werden können, den Fehleranteil bei der Erfassung von Merkmalen über Fragebogenverfahren zu minimieren.

Unabhängig von der Frage nach der Ursache der beiden Antwortmuster konnten einige Untersuchungen zur Rasch-Skalierbarkeit von psychodiagnostischen Inventaren zeigen, dass bereits das Vorhandensein bzw. Fehlen einer mittleren Antwortkategorie in einer Antwortskala dazu führt, dass sich die betreffende Skala als nicht eindimensional erweist (z. B. Hernández et al., 2004, 2006; Rost, 2002; Rost et al., 1999). Geht man davon aus, dass mit jedem einzelnen Item einer Skala tatsächlich (nur) eine latente Eigenschaft gemessen wird, welche für alle Personen die Gemeinsamkeit aller Items in einer Einstellungsskala bildet, so können unterschiedlich ausgeprägte Antwortmuster im Sinne eines systematischen Messfehlers dazu führen, dass sich die betreffende Skala dennoch als nicht eindimensional erweist. Im Rahmen der Skalierung solcher Skalen mit probabilistischen Testmodellen zeigt sich dann meist, dass zur Modellierung des Antwortverhaltens so genannte *mixed-Rasch-Modelle* (vgl. Abschnitt 4.2.4) mit zwei *latenten Klassen* (vgl. Abschnitt 4.6.2) eher geeignet sind als das „einfache„ Rasch-Modell (Rost, 2002; Rost et al., 1999). Vergleicht

man die als Schwierigkeitsprofil der Itemkategorien abgetragenen Schwellenparameter der Items einer Skala (Schwellenparameterprofile) beider Klassen, so lassen sich bei polytomen Antwortformaten für beide Personengruppen (latente Klassen) unterschiedliche Abstände zwischen den Profillinien finden. Ein geringer Abstand dieser Schwellenparameterprofile weist darauf hin, dass für diese Personengruppe die Wahrscheinlichkeit für die Wahl der äußeren bzw. extremen Antwortkategorien einer mehrstufigen Antwortskala steigt (*Extremkreuzer* – ERS). Große Abstände der Schwellenparameterprofile weisen dagegen auf eine Tendenz zur Wahl von mittleren Antwortkategorien (*Mittelkreuzer* – MRS) hin (Rost et al., 1997, 1999). Der individuell unterschiedliche Antwortstil wird dabei über die Zuordnung jeder Person zu einer der beiden latenten Klassen als kategoriale Merkmalsvariable, im Sinne von qualitativen Unterschieden, zwischen den beiden Personengruppen, operationalisiert und interpretiert (z. B. E. J. Austin et al., 2006; Eid & Rauber, 2000; Gollwitzer et al., 2005; Maij-de Meij, Kelderman & van der Flier, 2008; Rost et al., 1997, 1999; Rost & Langeheine, 1997).

3.3 Übergreifende Betrachtung von Antwortmustern

In der recht umfangreichen Literatur zu den unterschiedlichsten Arten von Antwortverzerrungen werden neben den hier bereits erwähnten noch weitere spezifische Formen und Ursachen diskutiert. Dabei werden jeweils detaillierte und sich teilweise einander ausschließende Definitionen gegeben, wobei diese dennoch letztlich einen gemeinsamen Kern teilen. Im Hinblick auf die jeweils zu messende Eigenschaft oder das Personenmerkmal wird vermutet, dass es eine oder mehrere Variationsquellen außerhalb der eigentlichen Messintention gibt, welche sich unter Umständen reliabilitäts- und validitätsmindernd auf die intendierten Testergebnisse auswirken (vgl. Cronbach, 1946). Je nach theoretischem Modell wird dabei die Ursache der eigentlich unerwünschten Variation entweder in den Testpersonen (z. B. Fiske & Rice, 1955; Ziegler et al., 2012), der Methode der Erhebung – z. B. durch unterschiedliche Antwortskalen der einzelnen Items (z. B. J. A. Bond, 1987; Hernández et al., 2006; Kulas & Stachowski, 2009; Moors, 2008) oder auch in den Inhalten der einzelnen Items (z. B. Higgins, Zumbo & Hay, 1999) gesehen. Insgesamt steht also, wenn auch mit unterschiedlicher Schwerpunktsetzung, bei der Betrachtung von abweichendem Antwortverhalten und den daraus resultierenden Antwortmustern immer die Reaktion von einzelnen Personen und / oder die Charakteristik bestimmter Items in Fokus des Interesses. Die Untersuchung solcher systematischen oder unsystematischen Reaktionen von Personen auf Fragebögen zur Selbstauskunft zu unterschiedlichsten Inhalten und Konstrukten hat, wie hier gezeigt, in der psychologischen Forschung eine lange Historie. Neben dem starken Interesse an der Entdeckung etwaiger Verzerrungen der zu erzielenden Messwerte im Sinne einer möglichst reliablen und validen Erfassung der jeweiligen Konstrukte bestand auch ein Interesse an der Entwicklung von spezifischen Antwort– und Messmodellen, welche die Abweichungen der beobachteten von den erwarteten Antwortmustern der Personen erklären können.

Die Fokussierung auf die Personen und deren Eigenschaften beim Antwortprozess ist assoziiert mit der Frage nach der situationsübergreifenden (über unterschiedliche Itemfragestellungen) Konsistenz von Eigenschaften und dem damit verbundenen Verhalten der antwortenden Personen. Diese Frage bezieht

sich im Zusammenhang mit der Erfassung von Persönlichkeit auf die *Person–Situation Debatte* (vgl. Mischel, 1968, 2009) in der Persönlichkeitspsychologie (vgl. Abschnitt 2.1 im Kapitel 2 *Theorie zu den untersuchten Konstrukten*). Vor dem Hintergrund der Person–Situation Debatte und der Beobachtung, dass nicht immer alle Eigenschaften bei allen Personen messbar sind (vgl. Bem & Allen, 1974; Bem & Funder, 1978) entwickelten Baumeister und Tice (1988) das Konzept der Metaeigenschaft (*metatrait*). Baumeister und Tice (1988) definieren dabei die Metaeigenschaft [metatrait] als die übergeordnete Eigenschaft, eine bestimmte (Persönlichkeits-)Eigenschaft zu haben oder nicht zu haben - „*A metatrait is the trait of having versus not having a particular trait*“ (Baumeister & Tice, 1988, S. 573). Im Rahmen der Erfassung von Dimensionen der Persönlichkeit über Fragebogenverfahren untersuchte Tellegen (1988) inkonsistentes Verhalten bei der Itembeantwortung und greift dabei zur Erklärung der gefundenen Inkonsistenzen im Sinne idiosynkratischer Antwortmuster das Konzept der *traitedness* wieder auf. Die Metaeigenschaft hat nach diesem Konzept einen Einfluss auf die Konsistenz bzw. Inkonsistenz und damit die Skalierbarkeit der Antworten einzelner Personen (Britt, 1993; Britt & Shepperd, 1999; Dwight, Porter Wolf & Golden, 2002; Reise & Waller, 1993; Tellegen, 1988, vgl. auch Abschnitt 4.4.2 in dieser Arbeit). Auch wenn das Konzept der Metaeigenschaft Gegenstand kontroverser Diskussionen ist (z. B. Britt, 1993; Dwight et al., 2002; Funder & Colvin, 1991; Haaga et al., 1995; Siem, 1998), bleibt die Untersuchung individuell unterschiedlicher Verhaltensweisen bei der Beantwortung von Fragebogen im Hinblick auf die Reliabilität und Validität der Messung relevant (Viswanathan, 2005). Die Diskussion von Antwortverzerrungen aus der Perspektive eines Messfehlers umfasst dabei unterschiedliche Aspekte. So wird auf die künstliche Erhöhung der Testwerte [*score inflation*] (Koretz, 2005; Nye, Do, Drasgow & Fine, 2008) und eine künstliche Erhöhung der Reliabilität aufgrund einer Tendenz der antwortenden Personen, ihre Antworten über unterschiedliche Items hinweg anzugleichen (Peer & Gamliel, 2011), hingewiesen. Ferner wird eine Varianzeinschränkung oder auch Varianzüberschätzung, welche zu einer Über- oder Unterschätzung der beobachteten korrelativen Zusammenhänge der gemessenen Konstrukte führt, diskutiert (Spector, Rosen, Richardson, Williams & Johnson, 2017) und auf schiefe Verteilungen und Deckeneffekte bezogen auf die gemessenen Merkmale

(Finney & DiStefano, 2006), oder auf eine unterschiedliche individuelle Varianz in der Beantwortung von einzelnen Fragen hingewiesen (Churchyard, Pine, Sharma & Fletcher, 2014). Solche potentiell störenden Messfehler werden in Folge oft unter dem Begriff der gemeinsamen Methoden (bedingten) Varianz (D. T. Campbell & Fiske, 1959), zum Beispiel bei der Erfassung von Merkmalen der Persönlichkeit diskutiert (Arias & Arias, 2017; Biderman, Nguyen, Cunningham & Ghorbani, 2011; Dicken, 1967; Wiggins, 1962). Allerdings argumentiert Spector (2006), dass der Begriff der *gemeinsamen Methodenvarianz* zugunsten eines Fokus auf den *Messfehler*, welcher das Produkt des Zusammenspiels der erfassten Merkmale und der Methoden, mit denen die Merkmale erfasst werden, aufgegeben werden sollte (Spector, 2006; Spector et al., 2017).

Sozial erwünschtes Antwortverhalten (SDR) und das Konzept der Akquieszenz haben sich zu zentralen Konzepten in der Forschung zur Erfassung von Eigenschaften mit Fragebogenverfahren entwickelt. Allerdings liefern empirische Untersuchungen, die auf diesen Konzepten basieren, oft uneinheitliche Befunde dazu, ob beispielsweise SDR den Nutzen der Erfassung von interindividuellen Unterschieden beeinträchtigt. Die Konzepte SDR und Akquieszenz stellen jeweils recht unscharfen Oberbegriffe dar, deren Nutzen daher in den letzten Jahren mehr und mehr in Frage gestellt werden. Demgegenüber werden andere, spezifischere Konzepte wie beispielsweise das *faking*, eingeführt (Ziegler, 2011; Ziegler et al., 2012). Zusätzlich ist eine zunehmend differenziertere Betrachtung von Phänomenen wie SDR und Akquieszenz in neueren Arbeiten zu beobachten, welche beispielsweise die Konfundierung von SDR mit der Polarität der Fragen und Konstrukte behandeln (z. B. Kulas et al., 2018; Segura & González-Romá, 2003). Demgegenüber bestehen aber auch Ansätze einer eher „simplifizierenden“ Betrachtung der einzelnen Formen von Antwortverzerrungen wie sie in diesem Kapitel behandelt wurden, im Sinne einer Integration zu einem übergeordneten Response-Style Faktor (z. B. He & van de Vijver, 2015; He, van de Vijver, Espinosa & Mui, 2014). Allerdings sollten solche Ansätze kritisch hinterfragt werden, da gezeigt werden kann, dass die unterschiedlichen Formen der Antwortverzerrungen jeweils mit inhaltlich konzeptuell unterschiedlichen, kognitiven Prozessen und motivationalen Affektlagen assoziiert sind (Bensch et al., 2017; Helmes et al., 2015; Ziegler, 2015).

Zur Entdeckung von Antwortverhalten, welches von dem erwarteten abweicht, werden unterschiedliche methodischen Ansätze diskutiert. Historisch relativ alt sind dabei Ansätze zu sehen, die beispielsweise zur Entdeckung von SDR oder auch *faking* so genannte „*Lügenskalen*“ einsetzen (z. B., Konstabel, Aavik & Allik, 2006; McCrae & Costa, 1983b). So wurden zur Kontrolle von SDR bei Fragebogenverfahren Skalen entwickelt, welche das Ausmaß von SDR erfassen sollen (z. B. Crowne & Marlowe, 1960; Paulhus, 1998b), um diese zur Korrektur der gemessenen Merkmalsausprägung einzusetzen. Der Einsatz solcher Skalen stellt auch aktuell eine häufig verwendete Methode dar, um SDR zu operationalisieren (S. 104 Kuncel, Borneman & Kiger, 2012). Die Messwerte aus den Skalen werden dabei entweder als Proxy oder als direktes Maß für die sozial erwünschte Antworttendenz verwendet. Skalen zur Erfassung von SDR bestehen aus Items, die sich entweder auf Verhaltensweisen beziehen, die im Allgemeinen für gut und wünschenswert angesehen werden und gleichzeitig aber sehr unwahrscheinlich sind (*ich lüge nie*), oder aber Items welche unerwünschte Verhaltensweisen ausdrücken, die im Allgemeinen aber häufig vorkommen. Dieses Prinzip weist Parallelen mit der beispielsweise von Paulhus (2012) beschriebenen *overclaiming-Technik* auf, wonach die antwortenden Personen nach ihrer Vertrautheit mit in der Realität nichtexistierenden Konzepten befragt werden. Bing, Kluemper, Kristl Davison, Taylor und Novicevic (2011) setzen in diesem Sinne die overclaiming-Technik als Maß für die Erfassung von *faking* ein. Eine der bekanntesten Skalen zur Erfassung von SDR ist die *Marlowe-Crowne Social Desirability Scale* (Crowne & Marlowe, 1960). Eine weitere weit verbreitet Skala zur Erfassung von SDR ist die Skala von Paulhus (1998b) und deren neuere, überarbeitet Version das *Balanced Inventory of Desirable Responding – Short Form* (BIDR-16 – Hart, Ritchie, Hepper & Gebauer, 2015). In der neueren Literatur werden derartige Ansätze hinsichtlich ihrer Fähigkeit zur Entdeckung von *faking* oder SDR allerdings eher kritisch diskutiert (z. B. Ellingson et al., 1999; Piedmont, McCrae, Riemann & Angleitner, 2000; Smith & Ellingson, 2002). Insofern wird die Validität und Praktikabilität solcher Skalen auch aus der Perspektive spezifischerer Konzepte von Antwortverzerrungen wie der *Eindruckssteuerung* [*impression management*] (IM) und des *faking* kritisiert (Baer, Wetter & Berry, 1992; de Jong, Pieters & Fox, 2010; de Vries et al., 2014; Heilbrun, 1962; McCrae & Costa, 1983b; Nederhof, 1985;

Pauls & Crost, 2004; Uziel, 2010). So weisen beispielsweise Pauls und Crost (2004) darauf hin, dass die Skalen zur Erfassung von SDR selbst durch Effekte von *faking* und *Eindruckssteuerung* (IM) konfundiert sind. Einen guten Überblick zu dieser kritischen Diskussion mit einer Schwerpunktesetzung auf *faking* geben beispielsweise Ziegler et al. (2012). Neueren Arbeiten setzen dementsprechend zur Modellierung von einerseits inhaltlichen und andererseits durch sozial erwünschtes Antwortverhalten bedingten Varianzanteilen in psychometrischen Skalen oft Strukturgleichungsmodelle ein (z. B. Jo, 2000; Jo, Nelson & Kiecker, 1997; McIntyre, 2011; Ziegler & Bühner, 2009) oder greifen auf Modelle aus der Item-Response-Theory zurück. So propagieren beispielsweise Holden und Book (2009) den Einsatz von HYBRID- oder mixed-Rasch-Modellen (vgl. Abschnitte 4.2.4 und 4.6.2) zur Identifikation von *faking*. Aus inhaltlicher Perspektive beleibt letztlich auch offen, in wie weit sozial erwünschtes Antwortverhalten, Akquieszenz und auch extreme oder mittlere Antworttendenzen bei polytomen Antwortformaten entweder als Messfehler oder andererseits als substantieller Ausdruck spezifischer Persönlichkeitsanteile angesehen werden muss (z. B. Arias & Arias, 2017; Holden & Passey, 2010).

Im Rahmen der psychometrischen Modellierung (vgl. Kapitel 4 *Psychometrische Modellierung*) wird in der Literatur unter dem Begriff *appropriateness measurement* (Birenbaum, 1985; Drasgow, Levine & Williams, 1985; M. V. Levine & Rubin, 1979) eine Methode diskutiert, welche die individuellen Antwortmuster der Personen in der Stichprobe analysiert. Bei diesem, erstmalig von M. V. Levine und Rubin (1979) propagierten Ansatz, handelt es sich um eine Methode zur Bestimmung der Passung von Antwortmustern einzelner Personen – der „*person-fit*“ – unter der Annahme eines bestimmten psychometrischen Antwortmodells. Dabei wird das Ausmaß der Abweichung des beobachteten Antwortmusters vom erwarteten Antwortmuster, gegeben die Parameter des jeweiligen Modells, operationalisiert.

Eine Gemeinsamkeit der in diesem Kapitel dargestellten Literatur zur Thematik „Antwortverhalten“, „Antwortmuster“ oder „Antwortverzerrungen“ kann, wenn auch nicht in allen Publikationen explizit thematisiert, darin gesehen werden, dass sich die jeweilige Definition einer Antwortverzerrung oder eines abweichenden Antwortverhaltens stets auf ein bestimmtes *Antwortmodell* zum *Scoring* und zur *Skalierung* der Personenantworten beziehen (vgl. Abschnitt

1.3). Die jeweils angenommenen Antwortmodelle für ein spezifisches Fragebogeninventar bilden dabei die definitorische Grundlage für die Klassifikation der Personenantworten. Die unterschiedlichen Antwortmodelle zur Skalierung lassen sich wiederum über unterschiedliche psychometrische Modelle zur Modellierung des Antwortverhaltens abbilden. In Kapitel 4 *Psychometrische Modellierung* der vorliegenden Arbeit wird daher ein Überblick und eine Einführung in die verschiedenen psychometrischen Antwortmodelle, welche in der Literatur vorherrschen, gegeben.

Kapitel 4

Psychometrische Modellierung

Im vorliegenden Kapitel werden die Grundprinzipien von einigen psychometrischen Modellen dargestellt, die sich einerseits aus den in Abschnitt 1.3 im Kapitel 1 *Einleitung, Überblick und Einführung in die Dissertation* dargestellten Techniken zur Skalierung von Fragebogendaten ableiten lassen, und andererseits die explorative Untersuchung von typischen Antwortmustern in Fragebogendaten ermöglichen. In der wissenschaftlichen Forschung ist die Modellbildung als hinreichend genaue und näherungsweise Abbildung einer empirischen Realität ein übliches Prinzip (z. B. Stachowiak, 1983). Nach Stachowiak (1983) ist das allgemeine Prinzip der Modellbildung bzw. das Modelldenken und der Modellbegriff als solcher so alt wie die Menschheit selbst. So lässt sich beispielsweise das Denken in Modellen und Abbildern einer wahrgenommenen (empirischen) Realität bis zu den steinzeitlichen Höhlenmalereien ca. 20000 Jahre vor unserer Zeit zurückverfolgen (Stachowiak, 1983, S.18). Die Modellbildung hat in der Wissenschaft eine wichtige Erkenntnisfunktion und dient der Veranschaulichung abstrakter und komplexer Aspekte und Zusammenhänge aus der realen Welt. Die Erkenntnisfunktion des Modellkonzepts folgt in diesem Sinne nach Stachowiak (1973, S. 58) in vielen Aspekten den Einsichten der „Forschungslogik“ von Karl Popper (1935).

4.1 Modellbildung zum Antwortverhalten

Auch in der wissenschaftlichen Psychologie ist die Modellbildung zur Generierung von Erklärungsansätzen und Ableitung von Hypothesen bezüglich menschlicher Phänomene ein verbreitetes Prinzip (z. B. Gigerenzer, 1981). Im spezifischen Kontext der Erfassung von individuellen Merkmalen über Fragebogenverfahren werden dabei *psychometrische Antwortmodelle* entwickelt, um diese anhand empirischer Daten zu testen. Solche psychometrischen Antwortmodelle können zur Beschreibung, Erklärung und auch zur Vorhersage menschlichen Antwortverhaltens auf psychodiagnostische Fragebogeninventare herangezogen werden. In diesem Sinne deckt sich die Begründung für die psychometrische Modellbildung zur Fragebogenbeantwortung mit einigen grundlegenden Zielsetzungen der Psychologie als Wissenschaftsdisziplin.

In seiner *Allgemeinen Modelltheorie* definiert Stachowiak (1973) drei Hauptmerkmale des allgemeinen Modellbegriffs. Danach sind Modelle durch ihr *Abbildungsmerkmal*, das *Verkürzungsmerkmal* und durch ihr *pragmatisches Merkmal* charakterisiert (Stachowiak, 1973, S. 131-133). Aus diesen allgemeinen Charakteristika lassen sich, bezogen auf die im Folgenden dargestellten psychometrischen Antwortmodelle, einige wichtige Aspekte vorab betonen. Insbesondere das *Verkürzungsmerkmal* erscheint hierbei besonders beachtenswert. So korrespondiert dieses Merkmal mit einer in der psychometrischen Literatur immer wieder, teils vehement, geführten Debatte um den angemessenen Auflösungsgrad psychometrischer Antwortmodelle zur Beschreibung von Datenstrukturen (z. B. die Frage nach der Anzahl der Modellparameter, vgl. Abschnitt 4.2.4 und 4.7). Dabei ist allerdings stets zu berücksichtigen, dass gemäß seinem *Abbildungsmerkmal* ein wahres Modell, unabhängig vom Auflösungsgrad und seiner Komplexität, zumindest in den Sozialwissenschaften wohl kaum existieren dürfte. Als Analogie könnte man hier auf den Abbildungscharakter der digitalen Fotografie verweisen: Ganz unabhängig von der Anzahl der Pixel mit der die Umwelt oder Personen in einem Foto aufgenommen werden, handelt es sich bei dem Foto nie um die Realität selbst, sondern stets nur um ein Abbild derselben (vgl. auch Stachowiak, 1973, S. 160 ff.). Diese Realität wird dabei, je nach Bedarf, aus unterschiedlichen Perspektiven abgebildet. Etwas pointierter drückte diesen Umstand der Statistiker George

Box (1979) aus: „*All models are wrong but some are usefull*“ (Box, 1979, S. 202) [„*Alle Modelle sind falsch, aber manche sind nützlich*“]. Gerade der letzte Aspekt dieses Zitates von George Box – die *Nützlichkeit* von Modellen – weist einen Bezug zu dem dritten von Stachowiak (1973) definierten Charakteristikum von Modellen auf – dem *pragmatischen Merkmal*. So impliziert die Frage nach der Nützlichkeit stets auch immer die Frage: Nützlich *wofür*? Diese Frage nach dem Zweck und dem Geltungsbereich der Modellierung ist Teil der Definition des *pragmatischen Merkmals* eines Modells. So schreibt Stachowiak (1973) zum *pragmatischen Merkmal* von Modellen:

> Modelle sind nicht nur Modelle von etwas. Sie sind auch Modelle für jemanden, einen Menschen oder einen künstlichen Modellbenutzer. Sie erfüllen dabei ihre Funktionen in der Zeit, innerhalb eines Zeitintervalls. Und sie sind Modelle zu einem bestimmten Zweck. (Stachowiak, 1973, S. 133)

Die Frage nach dem unterschiedlichen Zweck und dem Geltungsbereich stellt sich auch bei psychometrischen Antwortmodellen. Zur vergleichenden Beurteilung der Angemessenheit verschiedener und unterschiedlich komplexer Antwortmodelle können daher durchaus verschiedene Maßstäbe und Kriterien angelegt werden (vgl. Buckland, Burnham & Augustin, 1997; Burnham, Anderson & Burnham, 2002). Zur (vergleichenden) Bewertung der *relativen*, globalen Modellpassung können dazu unterschiedliche informationstheoretische Kriterien wie zum Beispiel das *Akaike Information Criteria* (AIC – Akaike, 1974) oder das *Bayes Information Criteria* (BIC – Schwarz, 1978) herangezogen werden (vgl. auch Rost, 2004, S.329, sowie Abschnitt 4.4). Demgegenüber können auch sogenannte lokale Modellverletzungen analysiert werden. Die Idee besteht dabei darin, entweder einzelne Personen (Zeilen) oder Items (Spalten) innerhalb einer Datenmatrix zu identifizieren, deren Antwortmuster bzw. psychometrische Eigenschaften nur schlecht mit der getroffenen Modellannahme zu vereinbaren sind. Derartige identifizierte lokale Modellabweichen können als Indiz für einen eingeschränkten Geltungsbereich des entsprechenden Antwortmodells sein. Ein Überblick zu den Möglichkeiten der Testung der Angemessenheit der unterschiedlichen psychometrischen Antwortmodelle wird in Abschnitt 4.4 gegeben.

Die Kernidee eines (parametrischen) psychometrischen Antwortmodells be-

steht darin, das Zustandekommen der empirisch vorgefundenen Datenmatrix über eine mathematisch formale Verknüpfung von verschiedenen Modellparametern zu erklären. Durch die Anwendung unterschiedlicher Modelle auf empirische Datensätze können dann über die Evaluation der jeweiligen Modellpassung unterschiedliche Hypothesen bezüglich des Antwortverhaltens der Personen auf die einzelnen Fragen eines Fragebogens überprüft werden (z. B. Chernyshenko, Stark, Chan, Drasgow & Williams, 2001). Dieses Prinzip der Analyse von Fragebögen und der Modellierung des Antwortverhaltens wird allgemein unter dem Begriff *Item-Response-Theory* (IRT) subsumiert. Den Darstellungen zur Skalierung in Abschnitt 1.3 in Kapitel 1 folgend, werden die psychometrischen Antwortmodelle zunächst zwei grundlegend verschiedenen Annahmen in Bezug auf den Prozess der Beantwortung einer Reihe von Fragebogen-Items zugeordnet. Demnach werden prinzipiell psychometrische Antwortmodelle für kumulative *Dominanz*-Antwortprozesse (z.B. *Guttman-Modell*, *Rasch-Modell*) und solche für Präferenzen mit *Nähe–Distanz*-Antwortprozessen (Unfolding) z. B. das *Hyperbelcosinus-Modell* (HCM) und das *generalisierte Unfoldingmodel* (GUMM) unterscheiden (vgl. auch van Schuur, 2011, sowie Abschnitt 1.3, zur Darstellung der fundamentalen Unterschiede im Antwortprozess und den daraus folgenden Modellunterschieden).

Die folgenden Abschnitte dieses Kapitels sind jeweils einigen wichtigen Modellen oder prominenten Vertretern einer Modellfamilie gewidmet.

4.2 Modelle für Dominanz-Antwortprozesse

4.2.1 Das Guttman-Modell

Ein Vertreter der Familie derjenigen Modelle, welche einen Dominanz-Antwortprozess postulieren, ist das Guttman-Modell (Guttman, 1950). Es kann aufgrund seiner Einfachheit und der historisch recht frühen Formulierung als Vorläufer aller weiteren Modelle, welche explizit einen Dominanz-Antwortprozess postulieren, betrachtet werden und wird daher hier als erstes dargestellt. Das hier beschriebene Testmodell wurde von Guttman (1944, 1947) zunächst unter dem Begriff *Skalogrammanalyse* als Prozedur zu Testung der Hypothese einer kumulativ, summativen Skalierbarkeit einer Reihe von Items für eine bestimmte Population, eingeführt (vgl. auch Mokken, 1971, S.70). Dabei wird bei dichotomen Itemantworten (z. B. $1 \equiv$ *richtig* bzw. *Zustimmung*; $0 \equiv$ *falsch* bzw. *Ablehnung*) durch systematische Umsortierung der Zeilen und Spalten einer Datenmatrix überprüft, ob sich diese so als *Skalogramm* anordnen lässt, sodass in jeder Zeile jeweils links die gelösten Items (1) und rechts die nicht gelösten Items (0) stehen (vgl. Tabelle 4.1; rechte Abbildung). Beide Darstellungen in Tabelle 4.1 beziehen sich auf dieselbe Datenmatrix mit $n = 5$ Personen die für $k = 4$ Items (A bis D), lediglich modellkonforme Antwortmuster produziert haben. Die beiden unterschiedlichen Darstellungen bezeichnet Zysno (1993) einerseits als *Reaktions-Skalogramm* und andererseits als *Code-Skalogramm*.

Tabelle 4.1 Beispiel für zwei Skalogramm Darstellungen einer umsortierten (perfekten) Datenmatrix für dichotome Items nach dem Guttman-Modell.

	Reaktions-Skalogramm									Code-Skalogramm				
	$A_{\{0\}}$	$B_{\{0\}}$	$C_{\{0\}}$	$D_{\{0\}}$	$A_{\{1\}}$	$B_{\{1\}}$	$C_{\{1\}}$	$D_{\{1\}}$		$A_{\{0,1\}}$	$B_{\{0,1\}}$	$C_{\{0,1\}}$	$D_{\{0,1\}}$
$P1$	1	1	1	1	0	0	0	0	$P1$	0	0	0	0
$P2$	0	1	1	1	1	0	0	0	$P2$	1	0	0	0
$P3$	0	0	1	1	1	1	0	0	$P3$	1	1	0	0
$P4$	0	0	0	1	1	1	1	0	$P4$	1	1	1	0
$P5$	0	0	0	0	1	1	1	1	$P5$	1	1	1	1

Anmerkungen: Darstellungen des *Reaktions-Skalogramm* (links) und des *Code-Skalogramm* (rechts) einer umsortierten (perfekten) Datenmatrix ($n = 5$) für $k = 4$ dichotome Items (A bis D) nach dem Guttman-Modell.

Im *Reaktions-Skalogramm* (linke Darstellung in Tabelle 4.1), sind die Items (A bis D) jeweils für **beide** Antwortkategorien nach ihrer Schwierigkeit in auf-

steigender Reihenfolge angeordnet. Die ersten Gruppe von vier Spalten ($A_{\{0\}}$ bis $D_{\{0\}}$) bezieht sich auf die Beobachtung des Auftretens der Antwortkategorie '0' ($0 \equiv$ *falsch* bzw. *Ablehnung*) und die zweite Gruppe von vier Spalten ($A_{\{1\}}$ bis $D_{\{1\}}$) bezieht sich auf die Beobachtung des Auftretens der Antwortkategorie „1“ ($1 \equiv$ *richtig* bzw. *Zustimmung*). In beiden Spaltengruppen ist die Kodierung dabei jeweils so vorgenommen, dass der Wert „1“ für die Beobachtung der entsprechenden Kategorie steht ($1 \equiv$ *beobachtet*) und der Wert „0“ für das Fehlen einer Beobachtung der entsprechenden Kategorie ($0 \equiv$ *nicht beobachtet*). Auch im *Code-Skalogramm* (rechte Darstellung in Tabelle 4.1) sind die Items nach ihrer Schwierigkeit in aufsteigender Reihenfolge von links nach rechts und die Personen nach ihrer Merkmalsausprägung in aufsteigender Reihenfolge von oben nach unten angeordnet. Für das *Code-Skalogramm* (rechte Darstellung in Tabelle 4.1) ergibt sich bei einer perfekten Guttman-Skala das typische, dreieckige Muster aus den Kodierungen für die Kategorien „1“ und „0“. Demgegenüber ergibt sich in der Darstellung der Daten im Reaktions-Skalogramm eine typische Struktur in Form eines Parallelogramms.

Obwohl von Guttman (1947) zunächst als reine Technik zur „Datensortierung“ beschrieben, impliziert das Modell letztlich zur psychometrischen Modellierung der Personenantworten auf die Items einer Skala zwei *Modellparameter*. Diese beiden impliziten Parameter beziehen sich einerseits auf das Ausmaß der Merkmalsausprägung der Personen und andererseits auf die Schwierigkeit der Items. Beim Guttman-Modell ergibt sich das Ausmaß der Merkmalsausprägung einer Person einfach aus der Anzahl der gelösten Items und die Itemschwierigkeit der aus der Anzahl der Personen welche das betreffende Item gelöst haben. Für die folgenden Darstellungen soll das Ausmaß der Merkmalsausprägung der Person v als *Personenparameter* mit θ_v und die *Itemschwierigkeit* des Items i mit σ_i bezeichnet werden. Die beiden Parameter θ_v und σ_i repräsentieren dabei die Position der Personen und Items auf dem latenten Kontinuum der Merkmalsdimension. Diese Parameter werden im Guttman-Modell auf *deterministische* Weise miteinander verknüpft. Das Auftreten einer der beiden Antwortkategorien auf der Antwortvariable X_{vi}, nämlich $x_{vi} = 1$ (richtige Antwort) oder $x_{vi} = 0$ (falsche Antwort) ergibt sich im Guttman-Modell eindeutig, deterministisch aus der *Relation* der beiden Parameter θ_v und σ_i. Immer dann, wenn die Merkmalsausprägung θ_v einer Person die Schwie-

rigkeit eines Items σ_i übersteigt beträgt die „Wahrscheinlichkeit“ [1] einer richtigen Antwort $p(x_{vi} = 1) = 1$. Im umgekehrten Fall, wenn die Schwierigkeit eines Items σ_i die Merkmalsausprägung θ_v einer Person übersteigt, beträgt die „Wahrscheinlichkeit“ einer richtigen Antwort $p(x_{vi} = 1) = 0$. Die beiden, aus der deterministischen Modellformulierung resultierenden, zulässigen Antwortwahrscheinlichkeiten $p = 0$ und $p = 1$ lassen sich daher in Abhängigkeit der beiden Parameter θ_v und σ_i formal wie folgt darstellen (vgl. Gleichungen 4.1 und 4.2).

$$p\left(x_{vi} = 1 | \sigma_i < \theta_v\right) = 1 \tag{4.1}$$

$$p\left(x_{vi} = 0 | \sigma_i > \theta_v\right) = 0 \tag{4.2}$$

Die in den beiden Gleichungen 4.1 und 4.2 definierte Beziehung drückt hierbei eine deterministisch formulierte *Ordnungsrelation* zwischen den Positionen der Personen und Items auf dem latenten Merkmalskontinuum aus. Trägt man die beiden „Antwortwahrscheinlichkeiten“ an der y-Achse, sowie die Differenz aus Itemschwierigkeit und Merkmalsausprägung der Personen an der x-Achse ab, ergibt sich als sogenannte *Item Characteristic Curve* (ICC), die für das Guttman-Modell typische Sprungfunktion (vgl. Abbildung 4.1).

Die Sprungstelle auf der x-Achse, bei der die Antwortwahrscheinlichkeit von $p = 0$ auf $p = 1$ springt, definiert dabei die Itemschwierigkeit auf dem gemeinsamen latenten Kontinuum von Personenfähigkeit und Itemschwierigkeit. Dabei muss angemerkt werden, dass auf Basis der formalen Modelldefinition durch die beiden Gleichungen 4.1 und 4.2 keine numerisch, auf einer gemeinsamen Skala vergleichbaren Schätzungen für die beiden Parameter θ_v und σ_i zu erreichen sind. Dies resultiert schlicht und ergreifend daraus, dass es sich bei den in den Gleichungen 4.1 und 4.2 gegebenen Modelldefinitionen lediglich um *Ordnungsrelationen* handelt, und nicht um Funktionsgleichungen, welche die beiden Parameter θ_v und σ_i mit den Antwortwahrscheinlichkeiten der Itemkategorien kontinuierlich verknüpfen und damit den Verlauf des Graphen der Antwortwahrscheinlichkeiten definieren (vgl. dazu auch Sijtsma & Molenaar, 2002, S. 15).

[1] Der Begriff „Wahrscheinlichkeit“ wir hier insofern in Anführungszeichen gesetzt, da es sich bei dem Modell von Guttman um ein deterministisches Modell handelt, dass letztlich keine „Wahrscheinlichkeiten“, die von $p = 1$ oder $p = 0$ abweichen, kennt.

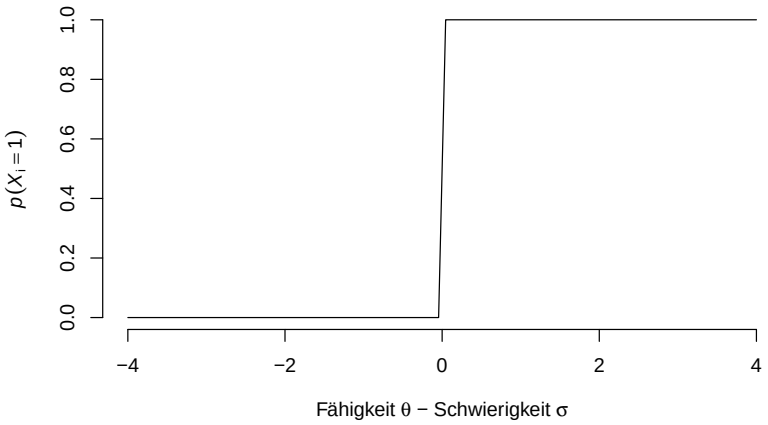


Abbildung 4.1 Darstellung der *Item Characteristic Curve* (ICC) des Guttman-Modells für ein Item mit der Schwierigkeit $\sigma = 0$.

Wird die zugrunde liegende Datenmatrix spaltenweise vom psychometrisch schwierigsten (links) bis zum psychometrisch leichtesten (rechts) und zeilenweise von der Person mit der geringsten Merkmalsausprägung (oben) bis zur Person mit der höchsten Merkmalsausprägung (unten) angeordnet, so ergibt sich bei perfekter Modellpassung das in Abbildung 4.1 (Darstellung auf der rechten Seite), beispielhaft gezeigte Skalogramm, also ein dreieckiges Muster aus Einsen und Nullen. Für eine perfekte Guttman-Skala wird daher die vollständige Transitivität in den Antwortmustern vorausgesetzt. In Bezug zu den in Abschnitt 1.4 nur kurz erwähnten Bedingungen zur Verrechnung der einzelnen Antworten der Personen zu einem Summenwert muss angemerkt werden, dass bei nachgewiesener Geltung des Guttman-Modells für eine Skala (und Population) nicht nur die notwendige Bedingung eines kumulativen Antwortmodells, sondern auch eine weitere (hinreichende) Bedingung, als erfüllt angesehen werden kann. Diese Bedingung bezieht sich dann auf das Vorliegen einer stichproben-, beziehungsweise personeninvarianten (aufsteigenden) Ordnung der Items nach ihrer Schwierigkeit. Unter dieser Voraussetzung der Modellgeltung lassen sich die Randsummen der Datenmatrix als suffiziente (erschöpfende) Statistiken der beiden Parameter θ_v für die Personenfähigkeit

und σ_i für die Itemschwierigkeit, interpretieren, welche beim Guttman-Modell ein ordinales Skalenniveau aufweisen.

Ebenso wie die weiter unten beschriebenen Modelle kann auch das Guttman-Modell auf polytome Items mit mehr als zwei Antwortkategorien verallgemeinert werden. Nach Borg und Staufenbiel (2007, S. 134) kann dabei ein mehrkategorielles Item mit m Kategorien als eine „Batterie“ von $m-1$ dichotomen Items aufgefasst werden (vgl. Abbildung 4.2). Bei einem Item I_i mit beispielsweise fünf aufsteigend geordneten Kategorien m (von $m = 0$ bis $m = 4$) wird das Überschreiten der vier jeweiligen Kategoriegrenzen (0|1, 1|2, 2|3 und 3|4) durch jede Person in der Datenmatrix in virtuelle Dummyitems aufgelöst, welche die jeweils übersprungenen Schnittpunkte der Kategoriegrenzen repräsentieren (vgl. Abbildung 4.2). Eine auf diese Art und Weise rekodierte

	I_i	$I_{i0\|1}$	$I_{i1\|2}$	$I_{i2\|3}$	$I_{i3\|4}$
P_1	0	0	0	0	0
P_2	1	1	0	0	0
P_3	2	1	1	0	0
P_4	3	1	1	1	0
P_5	4	1	1	1	1

Abbildung 4.2 Beispiel für die Rekodierung eines polytomen Items I_i mit fünf aufsteigend geordneten Kategorien (von $m = 0$ bis $m = 4$) in vier virtuelle Items ($I_{i0|1}$ bis $I_{i3|4}$) für ($n = 5$) Personen nach dem (polytomen) Guttman-Modell.

Datenmatrix mit (ursprünglich) mehrstufigen Items kann dann in Folge, wie oben bereits beschreiben, durch Umsortieren beziehungsweise durch Bestimmung der Itemkategorieschwierigkeiten $\sigma_{I_{i=m}}$ und der Personen Ausprägung θ_v analysiert werden.

In der praktischen Anwendung des Guttman-Modells auf reale empirische Datensätze zeigt sich in der Regel, dass die deterministische Modellannahme und die damit verbundene Forderung nach einer zum perfekten Skalogramm umsortierbaren, empirischen Datenmatrix meist unrealistisch ist. Zur Quantifizierung des Ausmaßes der Abweichung der realen Daten von dieser idealistischen Modellvorstellung wurden daher bereits von Guttman (1950, S. 64)

der *Reproduzierbarkeitskoeffizient* eingeführt. Dieser Koeffizient relativiert die Anzahl der Abweichungen (Fehler) in den Antwortmustern aller Personen von dem unter der Modellgeltung angenommenen, perfekten Antwortmuster mit der jeweils gleichen (Zeilen) Randsumme an der Größe des Skalogramms (vgl. z. B. Rost, 2004, S. 104, für eine detaillierte Darstellung), sowie Borg und Staufenbiel (2007, S. 127 ff.). Die nachträgliche Bestimmung des Ausmaßes der Modellpassung realer Daten auf das deterministische Guttman-Modell über einen letztlich probabilistisch motivierten Reproduzierbarkeitskoeffizienten birgt einen gewissen Widerspruch in sich. So weist Rost (2004, S. 103) darauf hin, dass nach logisch, strenger Auslegung der deterministischen Modelldefinition, schon ein einziges abweichendes Antwortmuster bereits zur Ablehnung der Modellgültigkeit führen müsste. Ein weiteres Problem ergibt sich aus der in den beiden Gleichungen 4.1 und 4.2 über eine Ordnungsrelation gegebene Modelldefinition, welche für die ICCs der einzelnen Items in einer Sprungfunktion resultiert (vgl. Abbildung 4.1). Paradoxerweise lässt gerade die deterministische Modelldefinition über eine Ordnungsrelation und die daraus resultierende Sprungfunktion den Spezialfall der (exakten) Gleichheit der beiden impliziten Parameter, also den Fall $\theta_v = \sigma_i$, ungenügend definiert. Liegt ein solcher, wenn auch wohl eher selten anzunehmender, Fall vor, so bleibt theoretisch unklar, ob die Lösungswahrscheinlichkeit für die betreffende Person auf dem entsprechenden Item nun $p = 1$ oder $p = 0$ beträgt. Die auf der Basis von Plausibilitätsüberlegungen naheliegende Antwortwahrscheinlichkeit von $p = .5$ (für jeweils beide Antwortkategorien) ist durch die Sprungfunktion, welche an der Sprungstelle eine unendliche Steigung aufweist, nicht definiert. Eine Möglichkeit zur Auflösung des Widerspruchs zwischen einem post hoc anzuwendenden, probabilistisch motivierten Reproduzierbarkeitskoeffizienten und der deterministischen Modellcharakteristik besteht darin, den deterministischen Zusammenhang zwischen den beiden implizit angenommenen Modellparametern θ_v und σ_i durch eine probabilistische Beziehung zu ersetzen. Bei solchen Modellen kann die jeweilige Antwortwahrscheinlichkeit dann sämtliche Werte im Bereich zwischen $p = 0$ und $p = 1$ annehmen. Dabei wird auch gleichzeitig das Problem des unzureichend definierten Spezialfalles $\theta_v = \sigma_i$ gelöst.

Abschließend muss im Hinblick auf die hier gewählte Darstellung des Guttman-Modells betont werden, dass das Guttman-Modell letztlich als *nichtparametrisches* Skalierungsverfahren bezeichnet werden muss. So lassen sich die beiden hier hilfsweise eingeführten impliziten Modellparameter θ_v und σ_i im Rahmen des Guttman-Modells nicht explizit auf einer gemeinsamen Metrik bestimmen. Dies resultiert aus der Tatsache, dass die beiden zulässigen Antwortwahrscheinlichkeiten lediglich als bedingte Zuweisungen (vgl. Gleichungen 4.1 und 4.2) definiert sind und keine parametrisch, funktionale Spezifikation im Sinne eines spezifischen, kontinuierlich, stetigen Funktionsverlaufs erfolgt.

4.2.2 Die Mokken-Analyse, einfache und doppelte Monotonie

Der gegenüber dem Guttman-Modell angeführten Kritik, dass die deterministische Modellannahme und die damit verbundene Forderung nach einer perfekten Datenmatrix in der Praxis meist unrealistisch sind, kann durch die Einführung einer probabilistischen Modellannahme begegnet werden. Dieses Prinzip ist in der Skalenanalyse nach Mokken (1971) realisiert. Die nach ihrem Begründer benannte Mokken-Analyse wurde als nichtparametrische Methode der Skalierung für dichotome Itemantworten entwickelt (Mokken, 1971; Mokken & Lewis, 1982; Mokken, Lewis & Sijtsma, 1986), und ist auch unter den Begriffen *Monotone Homogeneity Model* (MHM) und *Double Monotonicity Model* (DMM) bekannt geworden. Beide Modelle (MHM und DMM) wurden ausgehend von deren Anwendung für dichotome Antwortformate später für Items mit polytomen Antwortformaten erweitert (vgl. Sijtsma & Molenaar, 2002). Im Gegensatz zum Guttman-Modell (vgl. Abschnitt 4.2.1) kann die Skalierung nach Mokken (1971) als ein nichtparametrischer, aber dennoch probabilistisch motivierter Ansatz zur Item- und Skalenanalyse bezeichnet werden. Dieser besteht aus einer Reihe von exploratorischen Schritten zur Konstruktion von Skalen mit nützlichen Messeigenschaften wie Monotonie in der latenten Variablen und der invarianten Anordnung von Personen und Items. Konkret basiert Mokkens Skalierungsmodell auf Forderungen, die einerseits im *Monotone Homogeneity Model* (MHM) und andererseits im *Double Monotonicity Model* (DMM – Sijtsma & Molenaar, 2002, S. 23) formalisiert sind. Das MHM basiert

zunächst auf der Annahme einer eindimensionalen latenten Merkmalsdimension, lokaler stochastischer Unabhängigkeit der Personenantworten, sowie eines monotonen Zusammenhanges zwischen der latenten Variable und den manifest zu beobachtenden Antworten auf die Items – bzw. deren Lösungswahrscheinlichkeiten. Die lokale stochastische Unabhängigkeit der Personenantworten bedeutet, dass die Antwort einer Person auf einem Item nicht durch die Antwort derselben Person auf ein anderes Item beeinflusst wird (Sijtsma & Molenaar, 2002). Die Wahrscheinlichkeit für eine korrekte Antwort bei einem Item mit dichotomer Antwortskala ist dabei eine im Hinblick auf ihre Parameter unspezifizierte, aber monoton ansteigende Funktion der latenten Variablen. Eine derartige Funktion impliziert eine Ordnungsrelation für die Personen und ihre Merkmalsausprägung, sodass z. B. für zwei Personen A und B mit $\theta_A > \theta_B$ gleichzeitig die folgende Bedingung für die Wahrscheinlichkeit $P(X_i = 1)$, d. h. für eine korrekte Antwort auf einem Item i gilt:

$$P(X_i = 1|\theta_A) \quad \geq \quad P(X_i = 1|\theta_B) \tag{4.3}$$

Der hier in Ungleichung 4.3 dargestellte Ausdruck besagt, dass die Lösungswahrscheinlichkeit für das Item i monoton mit dem Grad der Merkmalsausprägung ansteigen muss.

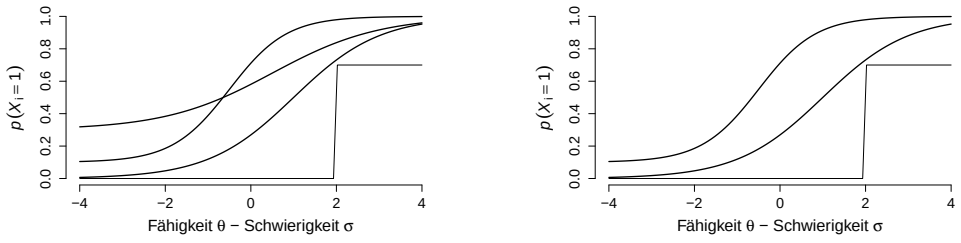


Abbildung 4.3 Darstellung von vier monoton ansteigenden *Item Characteristic Curves* (ICCs) des MHM (links), sowie drei überschneidungsfrei, monoton ansteigenden *Item Characteristic Curves* (ICCs) des DMM (rechts).

Die Lösungswahrscheinlichkeit $P(X_i = 1)$ muss für die Person A mit einer größeren Merkmalsausprägung höher ausfallen als für die Person B, welche eine

geringere Merkmalsausprägung θ aufweist. Als Funktionsgraphen der Lösungswahrscheinlichkeit eines Items – die *Item Characteristic Curve* – kommen hier beliebige, monoton ansteigende Kurvenverläufe in Frage, wie sie exemplarisch in Abbildung 4.3 (links) dargestellt sind.

Beim *Double Monotonicity Model* (DMM) werden die drei Modellannahmen des MHM (*Eindimensionalität*, *lokale stochastische Unabhängigkeit* und *Monotonie*) um eine vierte Eigenschaft ergänzt. Diese bezieht sich auf die *universelle* Ordnungsrelation der Items anhand deren Lösungswahrscheinlichkeiten. Bei dieser Annahme werden die Relationen der Lösungswahrscheinlichkeiten mehrerer Items i mit $i = 1 \ldots k$ gleichzeitig betrachtet. Während sich also die Annahmen im MHM jeweils (nur) auf die ICC eines Items in Abhängigkeit der unterschiedlichen Merkmalsausprägung (verschiedener Personen) beziehen, wird im DMM die Relation der ICCs aller Items in Abhängigkeit unterschiedlicher Merkmalsausprägungen betrachtet. Die doppelte Monotonie der Personen und Items im DMM impliziert die Forderung nach überschneidungsfreien ICCs aller Items einer Skala (vgl. Abbildung 4.3, rechts). Das Prinzip der doppelten Monotonie und damit der Überschneidungsfreiheit der ICCs ist eng verknüpft mit einer der Eigenschaften des parametrisch, probabilistischen Modells von Georg Rasch (1960), nämlich der *spezifischen Objektivität* der Messung. Dieses Modell und einige seiner Erweiterungen werden im Folgenden dargestellt.

4.2.3 Das Modell von Georg Rasch

Das logistische Testmodell, welches von Georg Rasch (1960) ursprünglich für die Auswertung von Intelligenztests eingeführt wurde, formalisiert die Antwortwahrscheinlichkeiten einer Person für jeweils eine von zwei vorgegebenen Antwortkategorien (z. B. $richtig \equiv 1$ und $falsch \equiv 0$). In dem psychometrischen Antwortmodell von Rasch (1960) werden dazu explizit zwei Modellparameter eingeführt analog zu den beiden impliziten Modellparametern im zuvor beschriebenen Guttman-Modell. Diese sind der *Personenparameter* θ_v, welcher, am Beispiel der Intelligenztestung, die latente *Fähigkeit* einer Person v repräsentiert und der *Itemparameter* σ_i, welcher die *Schwierigkeit* eines Items i repräsentiert. Der zentrale Unterschied zum Guttman-Modell und eine Gemeinsamkeit mit der Mokken-Analyse besteht nun darin, dass diese beiden Modellparameter beim Rasch-Modell durch eine *stetige, monoton steigen-*

de Funktion mit den Antwortwahrscheinlichkeiten der beiden Itemkategorien verknüpft werden. Im Unterschied zur Mokken-Analyse wird allerdings beim Rasch-Modell eine *spezifische* durch die beiden Parameter σ_i und θ_v in ihrem Funktionsverlauf zu beschreibende Funktion für die Antwortwahrscheinlichkeiten eingeführt. Die Antwortwahrscheinlichkeiten sind dadurch (im Gegensatz zum Guttman-Modell) nicht nur auf die beiden Werte $x_{vi} = 1$ und $x_{vi} = 0$ beschränkt.

Die gegebene Antwort einer Person v auf ein Item i wird dabei als Zufallsvariable X_{vi} aufgefasst deren Realisationen $x_{vi} = 1$ (richtige Antwort) und $x_{vi} = 0$ (falsche Antwort) in Abhängigkeit einer durch die beiden Parameter zu beschreibenden Funktion dargestellt werden sollen. Die Wahrscheinlichkeit $p\left(X_{vi} = x_{vi}\right); x_{vi} \in \{0, 1\}$ des Auftretens einer der beiden Antwortkategorien (0 und 1) ist dabei, im Gegensatz zum Guttman-Modell und der Mokken-Analyse, als (probabilistische) logistische Funktion der *Differenz* der beiden Parameter θ_v und σ_i definiert. Zur Herleitung der Modellgleichung des logistischen Testmodells setzt Rasch (1960) die Differenz der beiden Parameter θ_v und σ_i im ersten Schritt mit dem logarithmierten Wettquotient (der Chance) zur Lösung eines Items gleich. Dieser Wettquotient entspricht dem logarithmierten Quotienten aus den beiden Kategoriewahrscheinlichkeiten $p\left(X_{vi} = 1\right)$ und $p\left(X_{vi} = 0\right)$ zur „Wahl“ einer der beiden vorgegebenen Antwortkategorien; also $richtig = 1$ und $falsch = 0$ (vgl. Gleichung 4.4).

$$ln\left(\frac{p\left(X_{vi}=1\right)}{p\left(X_{vi}=0\right)}\right) = (\theta_v - \sigma_i); \quad \text{oder umgeformt:} \quad \frac{p\left(X_{vi}=1\right)}{p\left(X_{vi}=0\right)} = e^{(\theta_v - \sigma_i)} \tag{4.4}$$

Die im Ausdruck 4.4 auf der rechten Seite bereits umgeformte Gleichung lässt sich nun jeweils nach $p\left(X_{vi} = 1\right)$ und $p\left(X_{vi} = 0\right)$ auflösen. Es ergeben sich dabei zwei Gleichungen, welche jeweils die Wahrscheinlichkeit des Auftretens einer der beiden Kategorien, $richtig \equiv 1$ und $falsch \equiv 0$, in Abhängigkeit der Differenz aus den beiden Parametern θ_v und σ_i formal beschreiben (vgl. Gleichungen 4.5).

$$p\left(X_{vi} = 1|\theta_v; \sigma_i\right) = \frac{e^{(\theta_v - \sigma_i)}}{1 + e^{(\theta_v - \sigma_i)}}; \qquad p\left(X_{vi} = 0|\theta_v; \sigma_i\right) = \frac{1}{1 + e^{(\theta_v - \sigma_i)}} \tag{4.5}$$

Die beiden im Ausdruck 4.5 gegebenen Gleichungen lassen sich zu einer

Modellgleichung (vgl. Gleichung 4.6) kombinieren. Dabei nutzt man die Kodierung der Antworten der Personen mit „0“ und „1“ in der Datenmatrix aus. Der Ausdruck $\theta - \sigma$ im Exponenten des Zählers wird dazu mit der Kodierung der Antworten der Personen, also der jeweiligen Realisation x_{iv} der Zufallsvariable X_{vi}; (mit $x_{vi} \in \{0, 1\}$), multiplikativ verknüpft. Für den Fall $p\,(X_{vi} = 1)$ bleibt durch die Multiplikation des Exponenten im Zähler mit $X_{vi} = 1$ der Ausdruck $e^{(\theta_v - \sigma_i)}$ unverändert erhalten. Für den Fall $p\,(X_{vi} = 0)$ nimmt der Exponent den Wert null an, sodass insgesamt im Zähler der Wert eins stehen bleibt. Die resultierende formale Darstellung der Modellgleichung ist in Gleichung 4.6 wiedergegeben.

$$p\,(X_{vi} = x_{vi}|\theta_v; \sigma_i) = \frac{e^{(x_{vi} \cdot (\theta_v - \sigma_i))}}{1 + e^{(\theta_v - \sigma_i)}}; x_{vi} \in \{0, 1\} \tag{4.6}$$

Die durch diese Modellgleichung definierten Kategoriewahrscheinlichkeiten eines Items können daher jeden Wert innerhalb des Bereiches $p > 0$ und $p < 1$ annehmen. Sie sind an jeder Stelle des latenten Kontinuums in Abhängigkeit der Differenz aus θ und σ eindeutig über diese beiden Parameter definiert.

Es ist zu betonen, dass die Wahrscheinlichkeit des Auftretens einer der beiden Kategorien „0“ oder „1“ nach dieser Gleichung aus der *Differenz* der beiden Parameter θ_v und σ_i resultiert, was den *kumulativen Charakter* des Modells zur Abbildung einer *Dominanz-Relation* zwischen Items und Personen konstituiert. Ergibt sich als Differenz der Wert $\theta_v - \sigma_i = 0$, so nimmt, in den beiden Fällen $x_{vi} = 1$ und $x_{vi} = 0$ der Zähler in der Gleichung 4.6 den Wert $e^{(x_{vi}(\theta_v - \sigma_i))} = 1$ und der Nenner den Wert $1 + e^{(\theta_v - \sigma_i)} = 2$ an. Die Wahrscheinlichkeit für jede der beiden Antwortkategorien beträgt so jeweils $p = .5$ (vgl. Leunbach, 1961).

Mit steigender **positiver** Differenz der beiden auf der x-Achse abgetragenen Parameter nähert sich die Wahrscheinlichkeit für eine richtige Antwort $(x_{vi} = 1)$ dem Wert $p = 1$ asymptotisch an. Wohingegen sich mit steigender **negativer** Differenz die Wahrscheinlichkeit für eine richtige Antwort dem Wert $p = 0$ asymptotisch annähert. Die sich aus dieser formalen Darstellung ergebenden ICC ist in Abbildung 4.4 dargestellt. Konventionsgemäß wird bei der Darstellung der ICC eines Items jeweils Bezug auf den Verlauf der Wahrscheinlichkeit für eine richtige Antwort $(x_i = 1)$ in Abhängigkeit der Schwierigkeit σ_i eines Items und unterschiedlicher Merkmalsausprägungen der Personen θ_v dargestellt (schwarze Kurve in Abbildung 4.4).

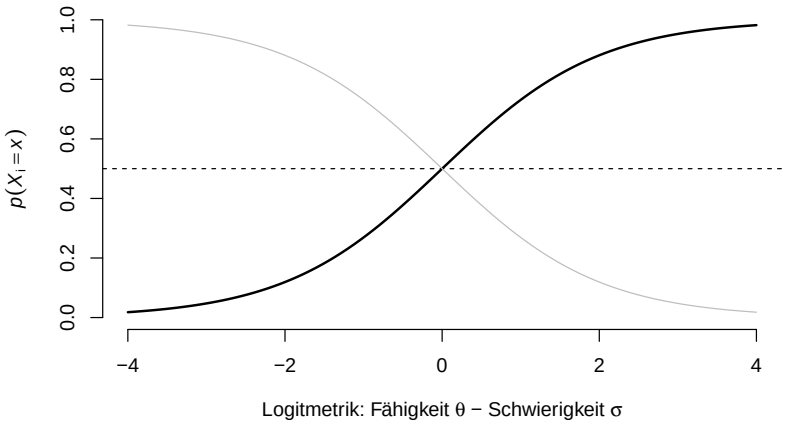


Abbildung 4.4 Darstellung der *Item Characteristic Curve* (ICC) des Rasch-Modells für ein Item i mit der Schwierigkeit $\sigma_i = 0$.

Das in Abbildung 4.4 als Beispiel dargestellte Item weist eine Schwierigkeit von $\sigma_i = 0$ auf, sodass für Personen mit der Fähigkeit $\theta_v = 0$ eine Lösungswahrscheinlichkeit von $p = .5$ (gestrichelte Linie in Abbildung 4.4) für das Item besteht. Die graue Kurve in Abbildung 4.4 verdeutlicht den Wahrscheinlichkeitsverlauf einer falschen Antwort ($x_{vi} = 0$) in Abhängigkeit der von links nach rechts auf der x-Achse größer werdenden (positiven) Differenz zwischen den beiden Parametern θ_v und σ_i. Die Wahrscheinlichkeiten der beiden Kurven der (zwei) Antwortkategorien addieren sich an jeder Stelle des latenten Merkmalskontinuums so zu einem Wert von $p = 1$. Die monoton ansteigende Form der ICC des Rasch-Modells (schwarze Kurve in Abbildung 4.4) impliziert dabei den summativen Charakter dieses Skalierungsmodells für die beiden Antwortkategorien der einzelnen Fragen einer psychometrischen Skala.

4.2.4 Erweiterungen des Modells von Georg Rasch

Eine prominente Modellerweiterung des logistischen Testmodells von Georg Rasch geht auf Masters (1982) zurück. Das sogenannte *Partial Credit Model* (PCM) generalisiert das Modell von Rasch (1960) für Items mit mehrstufi-

gen Antwortformaten. Seine Bezeichnung leitet sich daraus ab, dass durch das mehrstufig, *polytome* Antwortformat zum Beispiel bei Leistungstests für teilrichtige Lösungen eben auch Teilpunkte [engl. partial credit] vergeben werden können.

Polytomen Items sind zunächst ebenso kategoriale Items wie dichotome Items, mit dem einzigen Unterschied, dass einfach mehr als zwei mögliche Antwortkategorien vorliegen. Der naheliegende Grund für die Entwicklung und Anwendung von polytomen IRT-Modellen ist die Tatsache, dass polytome Antwortformate bei der psychometrischen Messung recht weit verbreitet sind. Insbesondere bei der Erfassung von individuellen Einstellungen und Dimensionen der Persönlichkeit werden oft polytome Antwortformate eingesetzt (vgl. Abschnitte 2.1.3 und 5.1 in dieser Arbeit). Dabei wird häufig argumentiert, dass dichotome Items (im Vergleich zu polytomen) oft weniger gut dazu geeignet sind, subtile Abstufungen der jeweils zu erfassenden Einstellungen oder Eigenschaften abzubilden (z. B. Kamakura & Balasubramanian, 1989). Prinzipiell besteht natürlich die Möglichkeit, mehrstufig kodierte Antworten nachträglich zur Auswertung mit dem oben beschrieben Rasch-Modell anhand bestimmter Kriterien zu dichotomisieren. Allerdings zeigt J. Cohen (1983), dass ein solches Vorgehen zu einem erheblichen und systematischen Verlust von Messinformationen führt. In diesem Sinne argumentiert auch Cox (1980), dass Items mit lediglich zwei oder drei Kategorien nur wenig Information enthalten und darüber hinaus geeignet sind, eine differenzierte Beantwortung der Items zu unterdrücken, und so die antwortenden Personen zu frustrieren. In diesem Sinne legt Andrich (1996) in seinen vergleichenden Darstellungen zur psychometrischen Skalenbildung nach Likert (1932) und Thurstone (1927a, 1928) dar, dass Likert (1932) die abgestuften Antwortskalen möglicherweise als Ausgleich für das Vorherrschen von Items mit eher extremen oder eindeutigen Aussagen (im Vergleich zu eher ambivalenten Aussagen), wie sie für die summative Indexbildung notwendig sind, einführte.

Die Frage nach der optimalen Anzahl von alternativen Antwortkategorien bei mehrstufigen Antwortskalen ist ein kontrovers (und lange) diskutierter Punkt (z. B. Boyce, 1915; Cox, 1980; Finn, Ben-Porath & Tellegen, 2015; Haladyna & Downing, 1993; Lord, 1944; Matell & Jacoby, 1971). Als Ergebnis einer Literaturübersicht zu dieser (damals bereits) 80 Jahre geführten Debat-

te folgert Cox (1980) zunächst, dass es offenbar keine universelle Antwort auf die Frage nach der optimalen Anzahl von alternativen Antwortkategorien gebe. Für die Praxis gibt Cox (1980) allerdings einige ungefähre Leitpunkte zur Orientierung. So resultiert nach Cox (1980) aus der Verwendung von mehr als neun Antwortalternativen kaum eine Verbesserung des Messinstruments; und eine ungerade statt einer geraden Anzahl von Antwortalternativen ist unter Umständen vorzuziehen, sodass die befragte Person eine neutrale Position einnehmen kann. Eine übermäßige Verwendung der neutralen Kategorie durch die Befragten soll dadurch vermieden werden, dass eine angemessene Anzahl plausibler Antwortalternativen zur Verfügung gestellt wird, welche mit nachvollziehbaren Anweisungen zur Beantwortung und einer eindeutigen Kennzeichnung der Antwortalternativen kombiniert werden (Cox, 1980). Demgegenüber vertreten Kulas et al. (2008) auf Basis ihrer Untersuchungen zur Funktion von mittleren Antwortkategorien, dass diese eher selten eine mittlere Trait-Ausprägung repräsentieren, sondern oft als „Mülleimer-Kategorie“ oder „Restkategorie“ im Fall einer unsicheren, oder auch ängstlichen (Hurley, 1998) Antwortentscheidung verwendet werden. In diesem Sinn zeigen Kulas und Stachowski (2009), dass die Tendenz zur Wahl einer mittleren Antwortkategorie oft mit einer uneindeutigen Itemformulierung einhergeht.

Zur Frage nach der Anzahl der einzusetzenden Antwortkategorien bei mehrstufigen Antwortskalen folgert Rodriguez (2005) auf Basis einer Metaanalyse über 27 Einzeluntersuchungen, dass sich drei Antwortkategorien in den meisten Anwendungsfällen als optimal erweisen. Danach zeigt sich, dass durch die Reduktion der fünf (klassischen) auf drei Antwortkategorien die Itemschwierigkeit eher geringfügig sinkt, sowie die Itemtrennschärfe und die Reliabilität unbeeinflusst bleiben (Rodriguez, 2005).

Neben solchen aus empirischen Befunden herrührenden Argumenten für eine bestimmte, „optimale“ Anzahl von Antwortkategorien werden für polytome Items allgemein oft auch psychometrische, messtheoretische Argumente für mehrstufige Antwortformate angeführt (z. B. Bejar, 1977; Donoghue, 1994; Lord, 1944; Lozano, García-Cueto & Muñiz, 2008; R. G. MacCann, 2004; Masters, 1988; Matell & Jacoby, 1971). Allgemein wird dabei argumentierte, dass sich mit polytomen Antwortformaten (mit weniger Items) ein größerer Bereich des latenten Kontinuums abdecken lässt (z. B. Matell & Jacoby, 1971; Thissen,

Reeve, Bjorner & Chang, 2007). Es wird ferner argumentiert, dass sich durch mehrstufige Antwortformate der Grad der Merkmalsausprägung einer Person auf dem latenten Kontinuum präziser abschätzen lässt (Bejar, 1977; R. G. MacCann, 2004; Masters, 1988). Damit übereinstimmend finden Maydeu-Olivares, Kramp, Garcia-Forero, Gallardo-Pujol und Coffman (2009) zunächst positive Effekte von mehrstufigen Antwortformaten auf die interne Konsistenz als Schätzer für die Reliabilität von psychometrischen Skalen, allerdings nur unbedeutende Verbesserungen der konvergenten Validität – operationalisiert als Korrelation mit externen Kriterien. Gleichzeitig finden Maydeu-Olivares et al. (2009) dagegen, dass sich die Modellpassung für ein eindimensionales Skalierungsmodell bei der Verwendung polytomer, im Vergleich zu dichotomen Antwortskalen, verschlechtert. Trotzdem lässt sich in der Regel zeigen, dass die Informationsfunktionen für gut diskriminierende, polytome Items über einen weiten Bereich des latenten Kontinuums relativ hoch ist (Thissen et al., 2007). Donoghue (1994) zeigt empirisch, dass polytome Antwortformate die meisten Informationen bei Personen mit mittlerer bis hoher Merkmalsausprägung liefern und die Testinformationsfunktion im Vergleich zum dichotomen Antwortformat ein Maximum erreichte. Die optimale Anzahl von Antwortkategorien liegt dabei nach Lozano et al. (2008) zwischen vier und sieben Kategorien.

Ausgehend von der Modellgleichung des RM (vgl. Gleichung 4.6; für dichotome Items) wird der Itemparameter eines Items mit m Kategorien x; mit $x \in \{1, 2, \ldots, m\}$, bei der polytomen Erweiterung des Modells in sogenannte einzelne *kumulierte Schwellenparameter* σ_{ix} zerlegt. Diese kumulierten Schwellenparameter ergeben sich aus den *Schwellenwahrscheinlichkeiten* der einzelnen Antwortkategorien. Die Schwellenwahrscheinlichkeiten sind dabei der relative Anteil der jeweils höheren Kategoriewahrscheinlichkeiten $p_{x_{vi}}$ und $p_{x-1_{vi}}$ der beiden benachbarten Kategorien $x - 1_{vi}$ und x_{vi}, deren Verhältnis sich gemäß Gleichung 4.7 darstellen lässt.

$$q_{(x_{vi} \equiv s)} = \frac{p_{x_{vi}}}{p_{x-1_{vi}} + p_{x_{vi}}}; \quad mit \quad s \in \{1, 2, \ldots, m\} \tag{4.7}$$

Die Schwellenwahrscheinlichkeiten $q_{x_{vi}}$ zweier benachbarter Kategorien eines mehrstufigen Items werden dabei jeweils durch die logistische Funktion des Rasch-Modells (vgl. Gleichung 4.6) definiert. Danach ist die Wahrscheinlichkeit des Auftretens einer Kategorie $x_{vi} \equiv s$ (mit $s \in \{1, 2 \ldots, m\}$) bei der

Antwort von Person v auf Item i, also $q_{x_{vi}}$, nach der allgemeinen Gleichung 4.8 durch die Lage der Kategorieschwellen auf dem latenten Kontinuum – die (Kategorien-)*Schwellenparameter* τ_{is} (gegeben θ_v) bestimmt.

$$q_{(x_{vi} \equiv s|\theta_v;\tau_{is})} = \frac{e^{(x_{vi} \cdot (\theta_v - \tau_{is}))}}{1 + e^{(\theta_v - \tau_{is})}}; \quad x_{vi} \in \{0, 1\}; \quad mit \quad s \in \{1, 2, \ldots, m\} \tag{4.8}$$

Bei dieser Darstellung in Gleichung 4.8 ist zu beachten, dass der Ausdruck x_{vi} hier, gewissermaßen dummykodiert, für das Auftreten einer bestimmten Kategorie s steht. Ist die betreffende Kategorie s gewählt, so erhält x_{vi} den Wert $x_{vi} = 1$, ist die Kategorie nicht gewählt, so ist $x_{vi} = 0$. Die Gleichung 4.8 lässt sich für die Wahrscheinlichkeit des Auftretens einer bestimmten Kategorie $x \equiv s$ und der entsprechende Gegenwahrscheinlichkeit auch wie folgt in vereinfachter Darstellung (ohne Berücksichtigung der Laufindizes für Personen und Items) in zwei Gleichungen auflösen.

$$q_s = \frac{e^{(\theta - \tau_s)}}{1 + e^{(\theta - \tau_s)}} \tag{4.9}$$

$$1 - q_s = \frac{1}{1 + e^{(\theta - \tau_s)}} \tag{4.10}$$

Da bezogen auf das Auftreten einer bestimmten Antwortkategorie der Ausdruck x_{vi} aus Gleichung 4.8 lediglich die Werte „0“ (Kategorie tritt nicht auf) und „1“ (Kategorie tritt auf) annehmen kann, vereinfacht sich die allgemeine Gleichung 4.8 für die Wahrscheinlichkeit des Auftretens einer bestimmten Kategorie gemäß Gleichung 4.9; und deren entsprechenden Gegenwahrscheinlichkeit gemäß Gleichung 4.10. Es lässt sich zeigen, dass sich die *kumulierten Schwellenparameter* σ_{ix} durch einsetzen der Gleichungen 4.9 und 4.10 in die umgeformte Gleichung 4.7 aus den (Kategorie-)*Schwellenparametern* τ_{is}, nach Gleichung 4.11 ergeben (für eine detaillierte Darstellung der Ableitung vgl. Rost, 2004, S. 207 – 209).

$$\sigma_{ix} = \sum_{s=1}^{x} \tau_{is} \tag{4.11}$$

mit den so parametrisierten, *kumulierten Schwellenparametern* σ_{ix} (vgl. Gleichung 4.11) ergibt sich die Modellgleichung des ordinalen „Rasch-Modells“,

das *Partial Credit Model* (PCM – Masters, 1982) wie in Gleichung 4.12 dargestellt (vgl. Rost, 2004, S. 209, Gleichung 8') sowie Masters (1982, 1988).

$$p\left(X_{vi} = x_{vi} | \theta_v; \sigma_i\right) = \frac{e^{(x_{vi} \cdot \theta_v - \sigma_{ix})}}{1 + \sum_{s=1}^{m} e^{(s \cdot \theta_v - \sigma_{is})}}; \quad s \in \{1, 2, \ldots, m\} \tag{4.12}$$

Die bei dieser Parametrisierung eingeführten (Kategorie-)*Schwellenparameter* τ_{is} repräsentieren diejenigen Punkte auf dem gemeinsamen latenten Kontinuum der Itemschwierigkeit σ_i und Personenfähigkeit θ_v, an denen sich jeweils die Antwortwahrscheinlichkeiten von nebeneinander liegenden Antwortkategorien gleichen, und bilden so die mehrstufig, ordinale Antwortskala der Items ab. Grafisch verdeutlichen lassen sich diese Schwellenparameter als Schnitt-

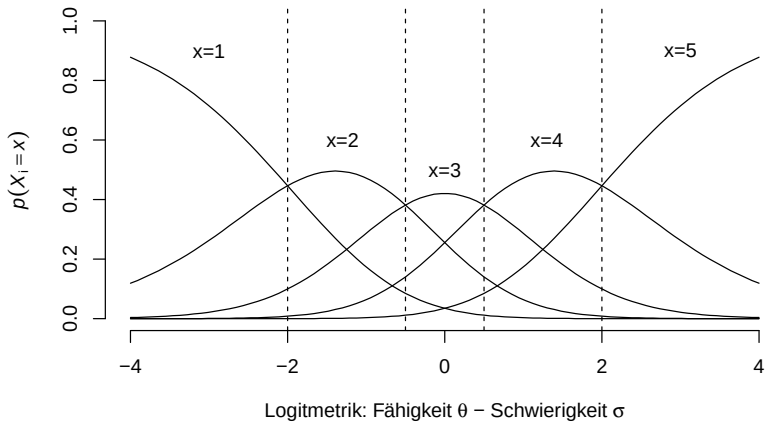


Abbildung 4.5 Darstellung der *Item Category Response Functions* (ICRF) des Partial Credit Modells für ein Item i mit den Schwellenparametern $\tau_{i,1} = -2$, $\tau_{i,2} = -.5$, $\tau_{i,3} = .5$ und $\tau_{i,4} = 2$.

punkte zweier benachbarter Itemkategoriefunktionen und deren horizontaler Position auf der x-Achse, wie sie in Abbildung 4.5 exemplarisch über vertikale, gestrichelte Linien dargestellt sind.

Neben dieser Erweiterung für mehrstufige Antwortformate hat das Modell von G. Rasch einige andere Erweiterungen erfahren. Diese nehmen Bezug auf

die, trotz seiner im Vergleich zum Guttman-Modell (vgl Abschnitt 4.2.1) probabilistischen Formulierung, immer noch vergleichsweise restriktiven Annahmen bezüglich der Qualität der Items einer Skala einerseits und dem Antwortverhalten der Personen andererseits. Da im RM (und im PCM) die Itemcharakteristik jeweils nur über einen Parameter beschrieben wird (die Schwierigkeit σ_i), müssen die ICCs der Items einer Skala alle die gleiche Steigung (vgl. Abbildungen 4.4 und 4.6) und damit die gleichen Trennschärfen aufweisen. Bei der Anwendung des Rasch-Modells auf empirische Datensätze zeigt sich allerdings, dass diese Restriktion nicht selten dazu führt, dass das Modell zur Erklärung der Daten verworfen werden muss. Anders betrachtet, kann man in solchen Fällen der Modellablehnung auch davon sprechen, dass die unter der Modellannahme vorhergesagten Antwortkategoriewahrscheinlichkeiten nicht gut mit den in den empirischen Daten vorgefundenen Wahrscheinlichkeiten übereinstimmen. Zur besseren Erklärung solcher Diskrepanzen oder zur besseren Anpassung des theoretischen Modells an die empirisch vorgefundenen Daten werden daher in verschiedenen Erweiterungen des RM (und des PCM) zusätzliche Parameter eingeführt mit dem Ziel den Verlauf der ICC genauer zu beschreiben. Um eine bessere Passung zwischen dem aus der theoretischen Modellannahme vorhergesagten Verlauf der ICC und dem sich aus den empirischen Daten ergebenden Verlauf zu erreichen, kann zunächst die Annahme gleicher Trennschärfen für alle Items einer Skala, also grafisch betrachtet die Parallelität der ICCs, aufgegeben werden. Mathematisch formal entspricht diese Modellerweiterung der Einführung eines itemspezifischen Steigungsparameters (Diskriminationsparameters) α_i für die jeweilige ICC des Items i. Die Gleichung 4.6 des RMs (für dichotome Items) erweitert sich so zu Gleichung 4.13

$$p\left(X_{vi} = x_{vi} | \theta_v; \sigma_i; \alpha_i\right) = \frac{e^{(x_{vi} \cdot \alpha_i \cdot (\theta_v - \sigma_i))}}{1 + e^{(\alpha_i \cdot (\theta_v - \sigma_i))}}; x \in \{0, 1\} \tag{4.13}$$

Diese Erweiterung des Rasch-Modells wurde als Spezialfall des logistischen Testmodells von Birnbaum (1968) diskutiert und ist seitdem auch unter dem Namen „*Birnbaum-Modell*“ (z. B. Bühner, 2011, S.503), oder auch als *2–PL–Modell* bekannt geworden. Wie man am Vergleich der beiden Gleichungen 4.6 und 4.13 leicht erkennen kann, reduziert sich das Birnbaum-Modell zum Rasch-Modell, wenn man den itemspezifisch variierenden Steigungsparameter α_i für

alle Items $i = 1, 2, ..., k$ auf den Wert $\alpha_i = 1$ setzt. Das *Birnbaum-Modell* lässt sich auch für die mehrstufig, polytomen Itemantwortskalen des PCMs verallgemeinern. Diese Verallgemeinerung des PCMs von Masters (1982) auf ein Modell mit itemspezifisch variierenden Steigungsparametern (Trennschärfen) wurde von Muraki (1992) entwickelt und wird als *Generalized Partial Credit Model* (GPCM) bezeichnet (vgl. Gleichung 4.14).

$$p\left(X_{vi} = x_{vi}|\theta_v; \sigma_i; \alpha_i\right) = \frac{e^{(\alpha_i \cdot x_{vi} \cdot (\theta_v - \sigma_{ix}))}}{\sum_{s=0}^{m} e^{(\alpha_i \cdot (s \cdot \theta - \sigma_{ix}))}}; \quad x \in \{0, 1, \ldots, m\} \tag{4.14}$$

wobei der Parameter σ_{ix} der Modellgleichung in 4.14 in der Publikation von Muraki (1992) nochmal reparametrisiert wird und dabei über eine additive Verknüpfung in zwei Parameter zerlegt wird (vgl. Gleichung 4.15)

$$\sigma_{ix} = \sigma_i - \delta_{is} \tag{4.15}$$

der Parameter σ_i stellt nach dieser Parametrisierung die mittlere Itemschwierigkeit dar (Lage des Items auf dem latenten Merkmalskontinuum), und entspricht dem Mittelwert der (kumulierten) Schwellenparameter aus dem PCM (vgl. Gleichung 4.12). Der Parameter δ_{is} kennzeichnet die *relative* Lage (relativ zur mittleren Itemschwierigkeit) der Itemkategoriegrenzen.

Die Flexibilität des *Partial Credit* Modells von Masters (1982) lässt sich auf unterschiedliche Art und Weisen restringieren. Einerseits können die Abstände zwischen den einzelnen Antwortkategorien über alle Items jeweils gleichgesetzt werden (wobei aber die Abstände z. B. zwischen der ersten und zweiten Kategorie nicht gleich groß sein müssen), was zum *Rating Scale Model – RSM* (Andrich, 1978a, 1978b, 1978c) führt. Andererseits können die Abstände zwischen allen Antwortkategorien jeweils *eines* Items gleichgesetzt werden was zum *Äquidistanzmodell* führt Andrich (1982). Eine weitere Erweiterung bzw. Restriktion des *Partial Credit* Modells von Masters (1982) stellt das *Dispersionsmodell* (Rost, 1988) dar, dass die Eigenschaften des Äquidistanzmodells und des RSM kombiniert (Rost, 2004, S. 218 und 225)

Das GPCM von Muraki (1992) reduziert sich wiederum auf das *Partial Credit* Modell von Masters (1982) wenn der itemspezifische Steigungsparameter α_i für alle Items auf einen Wert von $\alpha_i = 1$ gesetzt wird. Diese, in Bezug auf die variierenden Trennschärfen der Items, allgemeine Modellformulierung kann

nun wiederum dahingehend eingeschränkt werden, indem die Variabilität des Steigungsparameters im Hinblick auf die Menge der Items eingeschränkt wird. Sowohl für das 2–PL–Modell als auch das GPCM lässt sich der itemspezifisch variierende Steigungsparameter dahingehend restringieren, als dass dieser einerseits nicht wie im RM und PCM auf den konkreten Wert $\alpha = 1$ gesetzt wird, aber andererseits auch nicht itemspezifisch variiert. Vielmehr kann er auf einen beliebigen, allerdings für alle Items gleichen Wert restringiert werden. Diese Restriktion, welche einige positiven Eigenschaften des RM mit der Flexibilität des 2–PL–Modells kombiniert, wird auch als *1–PL Modell*, oder *One Parameter Logistic Model* (OPLM) bezeichnet (Verhelst & Glas, 1995).

Im Rasch-Modell (und auch im 1–PL Modell) sowie in dessen polytomer Verallgemeinerung impliziert die Identität der Itemtrennschärfe über alle Items eine wichtige Modelleigenschaft im Hinblick auf die Vergleichbarkeit von Personen und Items. Diese wichtige Eigenschaft eines Tests, welche bei nachgewiesener Geltung des Rasch-Modells gegeben ist, ist die so genannte *spezifische Objektivität*. Diese Eigenschaft besagt, dass die Rangfolge der Items (und Personen) entsprechend ihrer Schwierigkeit (bzw. Merkmalsausprägung) immer gleich ist – unabhängig von der jeweils betrachteten (Teil-)Stichprobe einer Population. So ist z.B. für jede Person das Item $i = 2$ schwieriger als das Item $i = 1$, ganz unabhängig vom Ausmaß der Merkmalsausprägung der jeweiligen Personen. Gleichermaßen ist (nach diesem Beispiel) die Lösungswahrscheinlichkeit eines Items $i = 2$ demzufolge für alle Personen immer geringer als die Lösungswahrscheinlichkeit des Items $i = 1$. Die mit zunehmender Merkmalsausprägung θ_v der Personen ansteigenden Lösungswahrscheinlichkeiten, der nach ihrer Schwierigkeit σ_i geordneten Items, sind bei Modellgeltung für alle Personen gültig. Diese, im Hinblick auf eine objektive Vergleichbarkeit von Personen und Items vorteilhafte Eigenschaft, ergibt sich aus der Parallelität der ICC's der Items, bzw. dem Umstand, dass diese sich nicht überschneiden. Der relative Schwierigkeitsunterschied zwischen zwei Items ist, unabhängig vom Ausmaß der Merkmalsausprägung der jeweiligen Person, immer gleich (vgl. Abbildung 4.6).

So stehen alle Items in einer einheitlichen Beziehung zum Fähigkeitsparameter θ. Eine Folge aus diesem Umstand besteht darin, dass unter Geltung des RM allein die Betrachtung der Anzahl der „gelösten“ Items ausreichend

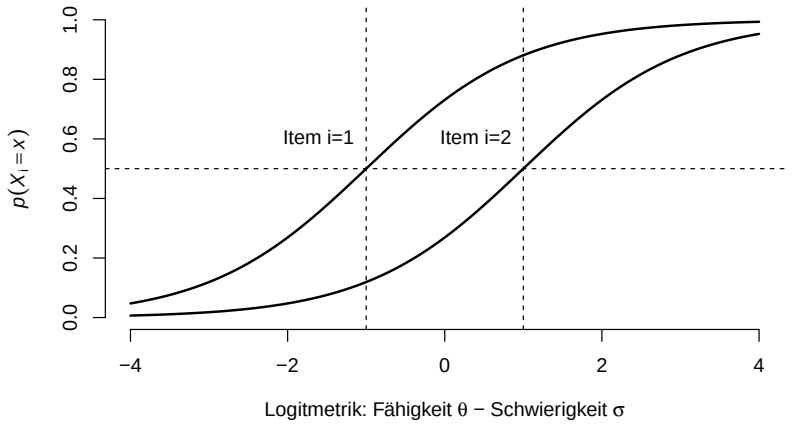


Abbildung 4.6 Darstellung der Item Characteristic Curves (ICC) für zwei Items unterschiedlicher Schwierigkeit nach dem Rasch-Modell; $\sigma_{i=1} = -1$, $\sigma_{i=2} = 1$.

ist, um die Personen im Hinblick auf ihre Merkmalsausprägung (*spezifisch objektiv*) zu vergleichen – unabhängig davon, welche Items die Personen beantwortet haben. Gleiches gilt analog dazu auch für den Vergleich der Items. Bei Geltung des RM stellen die *Randsummen* der Datenmatrix *suffiziente* Statistiken dar, welche also erschöpfende Aussagen über die Verhältnisse zwischen den Items und Personen untereinander erlauben (z. B. Andrich, 2010). Fischer (1974, S. 184) diskutiert eine detaillierte Definition des Begriffes der *suffizienten*, *erschöpfenden* Statistiken. Die Eigenschaft der *spezifischen Objektivität* im dichotomen RM kann bei einem über alle Items konstanten Diskriminationsparameter α_i auch bei der Verallgemeinerung des Modells für Items mit mehrstufigem Antwortformat gelten. So zeigt Mellenbergh (1995), dass es sich bei der *spezifischen Objektivität* des RM um einen Spezialfall handelt, welcher auch bei mehrstufigen Antwortformaten gilt, wenn dabei die Invarianz der Differenz der logarithmierten Wahrscheinlichkeiten der einzelnen Antwortstufen – den *log-odds* – gegeben ist. Die von Masters (1982) vorgeschlagene Modellerweiterung auf mehrstufige Antwortformate *kann* diese Eigenschaft aufweisen, da hier bei der Modellierung der Schwellenwahrscheinlichkeiten (vgl. Gleichung

4.8) durch den Verzicht auf einen Diskriminationsparameter α_i die Trennschärfen aller Items implizit auf den gleichen Wert $\alpha_i = 1$ gesetzt werden. Ergibt sich darüber hinaus bei der Schätzung der Modellparameter für einen empirischen Datensatz eine über alle Items hinweg konsistente aufsteigende ordinale Abfolge der Itemkategorieschwellenparameter, so erfüllt sich damit eine notwendige Voraussetzung für eine spezifisch objektive Messung.

Eine eigentlich unvorteilhafte Folge der Einführung eines itemspezifisch variierenden Steigungsparameters im *2–PL–Modell* besteht darin, dass die Items nicht mehr in einer einheitlichen Beziehung zum Fähigkeitsparameter θ stehen. Die oben beschriebene Eigenschaft der *spezifischen Objektivität* wird dabei zugunsten der flexibleren Modellanpassung aufgegeben. Je nach Trennschärfe, also der Steigung der ICC des einzelnen Items, trägt dieses mit unterschiedlichem Anteil zur Differenzierung der Personen auf dem Kontinuum der erfassten Merkmals θ bei. Im 2–PL–Modell zeigt sich, dass ein gewissermaßen gewichteter Summenscore der einzelnen Antworten auf die Items als Schätzer für die Merkmalsausprägung der Person dient (Sijtsma & Hemker, 2000). Die Gewichte stellen dabei nicht die Schwierigkeiten, sondern die unterschiedlichen Trennschärfen der Items dar (Rost, 2004).

Es lässt sich leicht zeigen, dass durch die Einführung des zusätzlichen Diskriminationsparmeters im Birnbaum-Modell die Invarianz der Itemordnung nach ihren Lösungswahrscheinlichkeiten, in Bezug zu einer unterschiedlichen Merkmalsausprägung der Personen, verloren geht. In Abbildung 4.7 sind zur Illustration dieses Umstandes die ICCs von zwei Items unterschiedlicher Schwierigkeit und Trennschärfe abgetragen. Die Schwierigkeiten der beiden Items unterscheiden sich um eine Logiteinheit ($\sigma_{i=1} = -.5$ und $\sigma_{i=2} = .5$) und die Trennschärfen weichen jeweils von dem im RM (und PCM) fixierten Wert von $\alpha = 1$ ab ($\alpha_{i=1} = .5$ und $\alpha_{i=2} = 1.5$ – vgl. Abbildung 4.7). Aufgrund der unterschiedlichen Trennschärfen überschneiden sich die beiden Kurven der Lösungswahrscheinlichkeiten. Dadurch entsteht die paradoxe Situation, dass sich die Relation der Lösungswahrscheinlichkeiten der beiden Items (abgetragen auf der y-Achse) für bestimmte Bereiche der Merkmalsausprägung auf dem latenten Kontinuum θ (abgetragen als Differenz $\theta - \sigma$ auf der x-Achse) umkehrt. So ergibt sich für den grau schraffierten Bereich in Abbildung 4.7 für das schwierigere Item $i = 2$ ($\sigma_{i=2} = .5$) eine höhere Lösungswahrscheinlich-

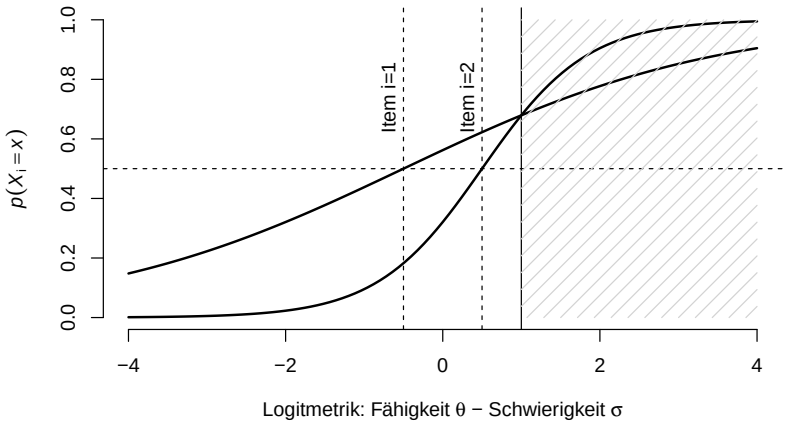


Abbildung 4.7 Darstellung der Item Characteristic Curves (ICC) für zwei Items unterschiedlicher Schwierigkeit und Trennschärfe nach dem Birnbaum-Modell; $\sigma_{i=1} = -.5$, $\alpha_{i=1} = .5$; $\sigma_{i=2} = .5$, $\alpha_{i=2} = 1.5$.

keit als für das eigentlich leichtere Item $i = 1$ ($\sigma_{i=1} = -.5$); Die Kurve für den Verlauf der Lösungswahrscheinlichkeit des Items $i = 2$ liegt hier über derjenigen für Item $i = 1$. Demgegenüber entspricht die Relation der Lösungswahrscheinlichkeiten der beiden Items $i = 1$ und $i = 2$ im nicht schraffierten Bereich von Abbildung 4.7 den Erwartungen, wie sie sich aus der Relation der Schwierigkeiten ($\sigma_{i=1} = -.5$ und $\sigma_{i=2} = .5$) der beiden Items ergeben. Das hier freilich konstruierte Beispiel ist allerdings im Hinblick auf die praktische Anwendung insofern nicht unrealistisch, als dass sich bei empirischen Daten nicht selten die Situation einstellt, dass eine höhere Itemschwierigkeit (σ) mit einer größeren Trennschärfe (α) einhergeht. Dies wiederum resultiert aus dem einfachen Umstand, dass demgegenüber leichtere Items eher von vielen Personen der Stichprobe gelöst werden, und damit weniger gut zwischen Personen mit unterschiedlichen Merkmalsausprägungen diskriminieren.

Ebenfalls von Birnbaum (1968, S. 404) vorgeschlagen, wurde eine weitere Modellergänzung um einen dritten Parameter zur Modellierung der ICCs der Items. Aufgrund des dritten zusätzlichen Modellparameters wird dieses Modell in der Literatur meist als *drei–Parameter–logistisches–Modell* (3–PL–

Modell) bezeichnet. Der dritte Modellparameter berücksichtigt dabei den insbesondere bei mehrfach Wahlaufgaben [multiple choice] im Bereich der Leistungsmessung bestehenden Umstand, dass hier aus inhaltlicher Perspektive, unabhängig von der Aufgabenschwierigkeit und der Personenfähigkeit, immer eine gewisse Wahrscheinlichkeit zum richtigen Raten der Aufgabenlösung besteht (vgl. Puchhammer, 1988). Der dritte Parameter repräsentiert dabei den y-Achsenabschnitt der unteren Asymptote der ICC der Items, welche beim RM, 1–PL–Modell und 2–PL–Modell immer bei einem Wert von $p = 0$ liegt (vgl. z.B. Ayala, 2009, S. 123 ff.). Diese insbesondere für den Bereich der Leistungsmessung mit mehrfach Wahlaufgaben plausible Modellerweiterung, kann aber auch in anderen Bereichen der psychometrischen Messung von individuellen Eigenschaften eingesetzt werden. So zeigen z. B. Reise und Waller (2003), dass sich durch die Modellierung eines zusätzlichen dritten Modellparameters für verschiedenen Skalen eines Persönlichkeitsinventars die Modellpassung auf die Daten – wenn auch geringfügig – verbessern lässt. Inhaltlich führen Reise und Waller (2003) den Parameter für die untere Asymptote der ICC dabei auf die Uneindeutigkeit der Inhalte einzelner Items zurück. Abbildung 4.8 zeigt zur Illustration zwei Items $i = 1$ und $i = 2$ mit unterschiedlichen Werten für die Parameter σ, α und den Parameter χ für die untere Asymptote.

Wie die Abbildung 4.8 allerdings zeigt, können sich die drei Parameter im Hinblick auf die tatsächliche Lösungswahrscheinlichkeit eines Items gegenseitig beeinflussen. So liegt in diesem Beispiel die mittlere Lösungswahrscheinlichkeit $p = .5$ für beide Items bei $\theta = -.625$ (gepunktete Linien in Abb. 4.8), obwohl beide Items unterschiedliche Schwierigkeiten σ aufweisen, welche definiert ist als Lage des *Wendepunktes* der ICC. In diesem Beispiel unterscheiden sich die Schwierigkeiten (Lage der Wendepunkte der Kurven) um eine Einheit auf der Logitskala ($\sigma_{i=1} = -.5$, durchgezogene vertikale Linie und $\sigma_{i=2} = .5$, gestrichelte vertikale Linie). Das sich dennoch für beide Items eine Lösungswahrscheinlichkeit von $p = .5$ für Personen mit einer Merkmalsausprägung $\theta = -.625$ ergibt, hängt mit den unterschiedlichen Werten ($\chi_{i=1} = .1$ und $\chi_{i=2} = .3$) für die untere Asymptote zusammen.

Der Idee die untere Asymptote der ICC der Items über einen zusätzlichen Modellparameter zu modellieren, lässt sich genauso auch für die obere Asymptote der ICC anwenden. Analog zu den Bezeichnungen der anderen Modelle

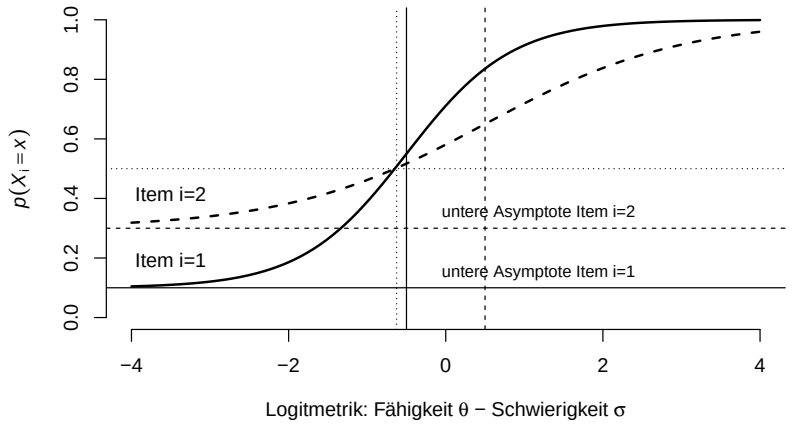


Abbildung 4.8 Darstellung der Item Characteristic Curves (ICC) für zwei Items unterschiedlicher Schwierigkeit, Trennschärfe und unterer Asymptote nach dem 3–PL–Modell; $\sigma_{i=1} = -.5$, $\alpha_{i=1} = 1.5$, $\chi_{i=1} = .1$; sowie $\sigma_{i=2} = .5$, $\alpha_{i=2} = .8$, $\chi_{i=2} = .3$.

mit mehreren Parametern zur Modellierung der ICCs der Items wird dieses Modell als *4–PL–Modell* bezeichnet und wurde von Barton und Lord (1981) eingeführt. Wie bereits dargestellt, kann eine untere Asymptote mit einem Wert für den y–Achsenabschnitt $p > 0$ inhaltlich zumindest im Bereich der Leistungsmessung, vergleichsweise eindeutig als Wahrscheinlichkeit zum richtigen Raten einer Aufgabenlösung interpretiert werden. Weicht die Differenz zum Wert $p = 1$ für die obere Asymptote zu dem im 1–, 2– und 3–PL–Modell fixierten Wert von $1 - p_{max} = 0$ ab, so stellt sich auch hier zunächst die Frage nach der inhaltlichen Interpretation dieses Parameters. Bei der Erfassung von kognitiven Fähigkeiten im Bereich der Leistungsmessung bietet sich eine zumindest vorstellbare Erklärung im Rahmen des erfassten Konstrukts an. So wird dabei der vierte Parameter oft mit der unsorgfältigen Bearbeitung der vorgegebenen Testaufgaben in Verbindung gebracht (z. B. Linacre, 2004a). Allerdings bleibt bei der Erfassung von individuellen Eigenschaften oder Einstellung die inhaltliche Interpretation eines Parameters für die untere und obere Asymptote weitgehend offen (z. B. Glas, 2009; G. Maris & Bechger,

2009).

Eine Gemeinsamkeit aller bisher dargestellten Erweiterungen des Rasch-Modells besteht darin, dass die getroffenen Modellannahmen und die auf dieser Basis bestimmten Modellparameter zur Modellierung des Antwortverhaltens der Personen jeweils für die gesamte Stichprobe bzw. Population gelten. Diese Annahme impliziert, dass alle Personen innerhalb einer Stichprobe die einzelnen Items und deren Antwortkategorien hinsichtlich ihrer Inhalte und Schwierigkeitsrangfolge in gleicher Weise interpretieren. Die angenommene Eindimensionalität einer psychometrischen Skala bildet sich im Modell dadurch ab, dass die Personenparameter θ_v bei der Bestimmung der Antwortwahrscheinlichkeiten als Konstanten[2] über alle Items und deren Schwellenparameter hinweg in die Modellgleichung (vgl. 4.12) eingehen (Rost, 2002). Diese strenge Annahme wird auch als *Personenhomogenität* bezeichnet und wird im *mixed-Rasch-Modell* gelockert (Rost, 1990). Das mixed-Rasch-Modell für dichotome Antwortdaten wurde von Rost (1990) als Erweiterung des RM und für polytome Daten von Rost (1991) als Erweiterung des PCMs von Masters (1982) entwickelt. Die grundlegende Idee bei dieser Erweiterung besteht darin, unterschiedliche Personenklassen (vgl. Abschnitt 4.6.2) anzunehmen, zwischen denen unterschiedliche Item- und Personenparameter gelten. Es ist so zum Beispiel möglich, die Annahme zu überprüfen, ob zwei oder mehrere Subpopulationen innerhalb einer Stichprobe entweder die Items oder deren Antwortskala oder auch beides unterschiedlich interpretieren (Rost et al., 1997). Das mixed-Rasch-Modell kombiniert also den klassifikatorischen Ansatz der *Latent Class Analysis* (LCA – vgl. Abschnitt 4.6.2) mit dem Skalierungsansatz nach einem Dominanz-Antwortprozess des Rasch-Modells. Das mixed-Rasch-Modell skaliert und klassifiziert die empirischen Daten damit gleichzeitig (Rost, 2004). Eine über einen solchen Ansatz hinausgehende Erweiterung stellte das so genannte *HYBRID-Modell* (K. Yamamoto, 1989) dar, bei dem in den unterschiedlichen latenten Personenklassen jeweils unterschiedliche Modelle angenommen werden können.

[2]Die Personenparameter variieren hierbei nur aufgrund der unterschiedlichen tatsächlichen Merkmalsausprägung der Personen – Personen gleicher Merkmalsausprägung erhalten den gleichen Personenparameter

4.2.5 Zusammenfassung und Übersicht zu Modellen für Dominanz-Antwortprozesse

Das Guttman-Modell und insbesondere das probabilistisch formulierte Rasch-Modell (RM) und seine im diesem Abschnitt beschriebene Erweiterung für mehrstufig, ordinale Antwortformate das *Partial Credit Model* (PCM) haben einen Bezug zu den in Abschnitt 1.3.1 in Kapitel 1 angesprochenen Voraussetzungen für die Anwendung der summierten Rating-Skalierung von Likert. Während bei der Fragebogenauswertung nach Likert (1932); Likert et al. (1934), wie auch bei der (anschließenden) Anwendung der KTT, die Gültigkeit der summativen Verrechnung der einzelnen Werte der gescorten Antwortkategorien einfach angenommen wird, lassen sich diese Annahmen, wie im Abschnitt 1.3.2 bereits erwähnt, im Hinblick auf einen empirischen Datensatz durch die Anwendung des Rasch-Modells überprüfen. Die unterschiedlichen Prinzipien und Methoden zur Überprüfung der Passung bzw. Angemessenheit eines psychometrischen Antwortmodells werden in Abschnitt 4.4 behandelt. Eine Übersicht zu verschiedenen Modellen zur Modellierung eines *Dominanz*-Antwortprozesses, die *nichtparametrisch* oder *parametrisch*, *deterministisch* oder *probabilistisch*, sowie für *dichotome* oder *polytome* Antwortformate formuliert sein können, ist in Tabelle 4.2 gegeben.

Trotz ihrer Unterschiedlichkeit sind alle in Tabelle 4.2 dargestellten Modelle mit der in Abschnitt 1.3.1 beschriebenen Skalierung nach (Likert, 1932; Likert et al., 1934) und dem Konzept zur Erfassung psychologischer Merkmale nach dem Prinzip der *summierten Ratingskalierung* (Borg & Staufenbiel, 2007; Spector, 1992) verbunden. Die zentrale Idee der in Tabelle 4.2 dargestellten Modelle besteht darin, dass Personen diejenigen Items lösen oder den Items zustimmen, deren Schwierigkeit *kleiner* als die der Merkmalsausprägung der Person ist. Parametrisch und probabilistisch aufgefasst muss für die Zustimmungswahrscheinlichkeiten p_i zu einem Item i (im dichotomen Fall) gegeben die unterschiedliche Merkmalsausprägung $(\theta_a) < (\theta_b)$ von zwei Personen a und b gelten: $p_i(\theta_a) < p_i(\theta_b)$. Diese Formulierung des summativen Skalierungsprinzips entspricht dem *Monotone Homogeneity Model* (vgl. Ungleichung 4.3 in Abschnitt 4.2.2). Die Verbreiterung dieses Prinzips auf mehrere Items ist im *Double Monotonicity Model* formalisiert. Das dabei formulierte Prinzip

Tabelle 4.2 Übersicht zur Klassifikation von eindimensionalen IRT-Modellen für den *Dominanz*-Antwortprozess.

	deterministisch	*probabilistisch*
nichtparametrisch	dichotom	
	Guttman Model – GM, (Guttman, 1944, 1947)	*Monotone Homogeneity Model* –MHM und *Double Monotonicity Model* – DMM, (Mokken, 1971; Mokken & Lewis, 1982; Mokken et al., 1986)
	polytom	
	Polytomous Scalogram Analysis – PSA, (Zysno, 1993); (Borg & Staufenbiel, 2007, S.134)	*Nonparametric Partial Credit Model* – NP-PCM, (Hemker, Sijtsma, Molenaar & Junker, 1997); *Nonparametric Graded Response Model* – NP-GRM, (Hemker, Sijtsma, Molenaar & Junker, 1996; Molenaar, 1997b)
parametrisch	probabilistisch dichotom	
	Rasch Model – RM, (Rasch, 1960); *One Parameter Logistic Model* – OPLM, (Verhelst & Glas, 1995); *Two Parameter Logistic Model* – 2–PL–Model, (Birnbaum, 1968); *Three Parameter Logistic Model* – 3–PL–Modell, (Birnbaum, 1968, S. 404); *Four Parameter Logistic Model* – 4–PL–Modell, (Barton & Lord, 1981).	
	probabilistisch polytom	
	Partial Credit Model –PCM (Masters, 1982); *Rating Scale Model* – RSM (Andrich, 1978a, 1978b, 1978c); *Equidistance Model*, (Andrich, 1982); *Dispersion Model*, (Rost, 1988); *Generalized Partial Credit Model* – GPCM (Muraki, 1992); *Graded Response Model* – GRM (Samejima, 1969); *General Graded Response Model* – GGRM (Samejima, 1999); *Sequential Rating Scale Model* – SRSM (Tutz, 1990)	

Anmerkungen: In dieser tabellarischen Übersicht werden die englischsprachigen Bezeichnungen für die Modelle verwendet.

einer doppelten Monotonie, welches darin besteht, dass für *alle* Personen die Lösungswahrscheinlichkeiten *mehrerer* Items analog zu deren Schwierigkeiten ansteigen, ist verknüpft mit dem Konzept der *spezifischen Objektivität* im parametrisch, probabilistischen Modell von Georg Rasch (1960). Dieses Prinzip wird aber – wie in Abschnitt 4.2.4 *Erweiterungen des Modells von Georg Rasch* dargestellt – bei manchen Modellerweiterungen des Rasch-Modells unter bestimmten Bedingungen (lokal) verletzt.

4.3 Modelle für Nähe–Distanz-Antwortprozesse

Die im vorangegangenen Abschnitt vorgestellten Modelle bilden trotz mancher Unterschiedlichkeit in deren Formulierung letztlich alle die im Abschnitt 1.3 vorgestellte Idee einer summativen Verrechnung der einzelnen *Item Scores* ab, und implizieren damit eine monoton steigende Item Characteristic Curve (ICC). Wie bereits dargelegt liegt solchen Modellen die Annahme einer *Dominanz-Relation* zwischen der latenten Merkmalsausprägung der Personen einerseits und der Itemschwierigkeit andererseits zugrunde. In diesem Abschnitt sollen nun demgegenüber psychometrische Modelle vorgestellt werden, deren Basis eine *Nähe–Distanz-Relation* zwischen den Personen und den Items ist (vgl. Abschnitt 1.3.3 in Kapitel 1). Obwohl beispielsweise das von Coombs (1950) konzeptuell vorgestellte Prinzip des *Unfolding* zur Abbildung einer *Nähe–Distanz-Relation* historisch bereits vor dem Rasch-Modell vorgestellt wurde, haben solche Modelle in der psychometrischen Literatur allerdings wenig Aufmerksamkeit erfahren (vgl. z. B. M. S. Johnson, 2006). In gleicher Weise wie bei den Modellen für Dominanz-Antwortprozesse lassen sich die Modelle für Nähe–Distanz-Antwortprozesse zunächst in deterministische, probabilistische, nichtparametrische und parametrische Modellformulierungen einteilen. Weiterhin lassen sich auch die Modelle für Nähe–Distanz Antwortprozesse ausgehend von Modellformulierungen für dichotome Daten als Generalisierung für polytome Daten erweitern. So wie das in Abschnitt 4.2.1 beschriebene Guttman-Modell lässt sich auch das von Coombs (1950) vorgestellte Modell als deterministische Grundlage vieler weiterer Modelle für Nähe–Distanz-Antwortprozesse ansehen. Es wird daher im folgenden Abschnitt als erstes – und etwas ausführlicher – vorgestellt.

4.3.1 Das Unfoldingmodell nach Coombs

Das von Coombs (1950, 1967) erstmals unter dem Begriff *Unfolding* konzeptionell dargestellte Modell formalisiert die bereits von Thurstone und Chave (1929) beschriebene Skalierungsmethode zur Messung von individuellen Einstellungen (vgl. Abschnitt 1.3.3). Dem Modell von Coombs (1950) liegt dabei weniger eine Dominanz-Relation zwischen den Personen und Items zugrunde, sondern vielmehr die Annahme einer *Nähe–Distanz-Relation* zwischen den

Personen und den Items. Coombs (1950) geht zunächst ebenfalls davon aus, dass sich die einzelnen Items entlang eines eindimensionalen Kontinuums einer Einstellungsdimension anordnen lassen. Im Gegensatz zu dem von Thurstone und Chave (1929) vorgeschlagenen Vorgehen zur Bestimmung der Itemschwierigkeiten ergibt sich die (relative) Anordnung der Items auf dem Kontinuum der Einstellungsdimension im Rahmen des Modells allerdings nicht durch die a priori Auswertung von Expertenurteilen, sondern direkt im Rahmen der Messung der Personen. Durch den Messvorgang, welcher letztlich auch aus einer Itembewertung durch die zu testenden Personen besteht, werden dabei die Personen und Items gleichzeitig auf einer gemeinsamen Skala der Merkmalsdimension positioniert. Diese gemeinsame Skala für Items und Personen bezeichnet Coombs (1950) als *joint scale* oder auch J–Skala.

In der aus Coombs (1950, S. 147) übernommenen Abbildung 4.9 sind hier diese Positionen (relativen Schwierigkeiten) von vier Items (A, B, C und D) sowie die Positionen von zwei Personen (X und Y) auf einer J–Skala dargestellt.

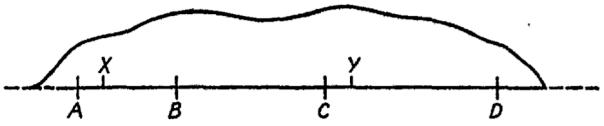


Abbildung 4.9 Darstellung der gemeinsamen J–Skala und Verteilung von Items und Personen, Beispiel entnommen aus Coombs (1950, S. 147).

Die Position der beiden Personen X und Y in obiger Abbildung 4.9 lassen sich dabei inhaltlich als deren Ausprägung auf dem latenten Kontinuum, also deren zu messender Eigenschaftsausprägung interpretieren. Ordnet man die Items (A, B, C und D) nun für jede einzelne Person ausgehend von ihrem *Idealpunkt* auf dem latenten Kontinuum (in Abbildung 4.9 die Punkte X oder Y) nach ihrer Distanz zur Position der Person in aufsteigender Reihenfolge, so ergibt sich für jede Person zunächst eine unterschiedliche Rangfolge der vier Items. Für die Person „X“ lautet diese Rangfolge A, B, C, D, wohingegen für Person „Y“ die Rangfolge C, D, B, A lautet. Diese individuellen (Rang-) Skalen bezeichnet Coombs (1950) als *individual scale* oder auch I–Skala (vgl. auch Coombs, 1967, S. 335). Nach der von Coombs (1967, S. 27) aufgestellten Klassifikation von Daten aus Verhaltensbeobachtungen, welche vier grundle-

gende Datentypen unterscheidet, resultieren solche I–Skalen aus *preferential choice data*, also Daten aus Präferenzurteilen. Die individuellen Rangreihen (I–Skalen) ergeben sich dabei entweder direkt durch die Aufforderung an die Personen, für die vorgelegten Stimuli oder Items eine Präferenzreihenfolge anzugeben – zum Beispiel: „*Ordnen Sie bitte die folgenden Aussagen in der Reihenfolge ihrer persönlichen Zustimmung*“ oder aber durch die Angabe derjenigen Aussage, der die jeweilige Person zustimmt. In letzterem Fall handelt es sich um ein unvollständiges, dichotomes Präferenz–Rating, da hierbei lediglich das (am meisten) präferierte Item aus den insgesamt vorgegebenen Items ausgewählt werden muss. Die individuelle I–Skala einer Person ergibt sich dann unter den Voraussetzungen, dass erstens die Position der Person auf dem latenten Kontinuum (dem Idealpunkt) durch diejenigen Items definiert ist, welchen die Person zugestimmt hat (vgl. LCJ–Skalierung in Abschnitt 1.3.3), und zweitens, dass die relativen Schwierigkeiten der Items auf dem latenten Kontinuum der Einstellungsdimension, durch die Bewertungen der Gesamtheit der zu testenden Personen definiert sind. Klappt bzw. *faltet* man die beiden Enden der gemeinsamen J–Skala mit den so definierten Positionen der Items am Idealpunkt einer Person jeweils um 90 Grad nach unten, so ergibt sich als Senkrechte die individuelle I–Skala (vgl. Abbildung 4.10).

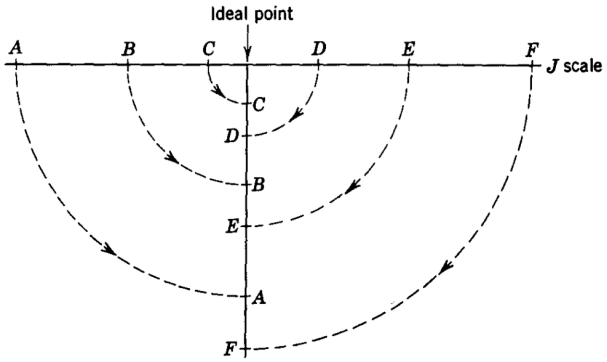


Abbildung 4.10 Darstellung der Entfaltung der gemeinsamen J–Skala (Waagerechte – mit Items A, B, C, D, E und F) und der sich durch Faltung am Idealpunkt ergebenden I–Skala (Senkrechte), Abbildung entnommen aus Coombs (1967, S. 80).

Aus dieser grafischen Veranschaulichung der Überführung der J–Skala in

individuelle I–Skalen (und umgekehrt) leitet sich der von Coombs (1950) eingeführte Begriff *Unfolding* ab. Der Begriff Entfaltung [engl.: Unfolding] wird hier deswegen gewählt, weil die als individuelle Präferenzen erhobenen Daten vor deren Auswertung ja zunächst für jede Person (nur) als „gefaltete“ I–Skalen in den erhobenen Daten vorliegen. Auf Basis dieser Daten wird dann im Rahmen der Skalierung und Auswertung durch deren „Entfaltung“ auf die gemeinsame J–Skala geschlossen.

Es lässt sich an diesem Beispiel mit vier Items zeigen, dass sich durch Auswertung der Antwortmuster eines Datensatzes, welcher sich aus den unterschiedlichen Präferenzurteilen einer Reihe von Personen ergibt, auf die relativen (Rang–) Positionen der Items und Personen auf einer gemeinsamen J–Skala schließen lässt. Nimmt man zunächst hypothetisch an, dass die individuellen Rangreihen bezüglich der Items (A, B, C, und D) der befragten Personen unsystematisch und zufällig entstanden sind, so ergäbe sich die Anzahl der möglichen Antwortmuster (*pattern*) $\Sigma_{Pmax.}$ beziehungsweise die Anzahl aller möglichen individuellen Rangreihen (I–Skalen) hier als $\Sigma_{Pmax.} = 4! = 24$, allgemein also für k Items als $k!$. Wären in einem Datensatz tatsächlich alle möglichen individuellen Rangreihen beobachtbar, so würden sich die einzelnen Rangurteile, betrachtet man sie als Ganzes, allerdings unter der Annahme einer gemeinsamen latenten Eigenschaftsdimension (J–Skala), gegenseitig gewissermaßen widersprechen. Unter der Annahme einer eindimensionalen Eigenschaftsdimension würden also inkonsistente, *nicht transitive* Rangurteile (vgl. *Ordinalskalen* in Abschnitt 1.2) bezüglich der relativen Itempositionen auf der J–Skala vorliegen (für eine praktische Darstellung von intransitiven Präferenzurteilen vgl. auch Laux, 2005, S. 33). Coombs (1950) konnte nun zeigen, dass unter der Bedingung der *Transitivität* der individuellen I–Skalen in einer Datenmatrix für eine gemeinsame J–Skala für k Items nur $\binom{k}{2}+1$ verschiedene I–Skalen zulässig sind. Für vier Items existieren in diesem Beispiel zwei Sets mit jeweils $\Sigma_{Ptrans.} = \binom{4}{2} + 1 = 7$ zulässigen *transitiven* Präferenzrangfolgen (vgl. Abbildung 4.11).

Das überraschende Ergebnis, dass für diese vier Items gleich **zwei** Sets von transitiven I–Skalen vorliegen, resultiert aus der Tatsache, dass, obwohl zunächst nur ordinale Information in den I–Skalen enthalten ist, die Existenz der beiden Sets unterschiedliche Rückschlüsse bezüglich der Abstände der vier

A	*B*	*C*	*D*
B	**A**	*C*	*D*
B	*C*	**A**	*D*
B	*C*	*D*	**A**
C	*B*	*D*	**A**
C	*D*	*B*	**A**
D	*C*	*B*	**A**

A	*B*	*C*	*D*
B	**A**	*C*	*D*
B	*C*	**A**	*D*
C	*B*	**A**	*D*
C	*B*	*D*	**A**
C	*D*	*B*	**A**
D	*C*	*B*	**A**

Abbildung 4.11 Zulässige transitive Präferenzrangfolgen für $k = 4$ Items auf einer gemeinsame J–Skala, Beispiel entnommen aus Coombs (1950, S. 151).

Items auf der gemeinsamen J–Skala zulassen. So korrespondiert das I–Skalen Set in Abbildung 4.11 (links) mit der (quantitativen, *quasi metrischen*) Information, dass der Abstand zwischen den beiden Items A und B größer als der Abstand der Items C und D auf der gemeinsamen J–Skala ist. Wohingegen das I–Skalen Set in Abbildung 4.11 (rechts) mit der Information korrespondiert, dass der Abstand zwischen den Items A und B kleiner als der Abstand der Items C und D ist (Coombs, 1950). Aufgrund dieser Erkenntnis, dass sich aus der Art der Zusammensetzung der individuellen I–Skalen in den Daten (quasi-) metrische Information bezüglich der J–Skala ableiten lässt, welche über die rein ordinale Information hinausgeht, bezeichnet Coombs (1950) das Skalenniveau der J–Skala als *ordered metric scale*, deren Skalenniveau zwischen den von Stevens (1946) definierten Skalenniveaus *ordinal* und *intevall* liegt. Es liegt zwar, wie beim Intervallskalenniveau gefordert, keine absolute Information bezüglich der Abstände der Skalenpunkte vor (Gleichheit der Abstände bei Intervallskalen), aber dennoch relative Information bezüglich der Verhältnisse der Abstände. Die aus diesen Analysen interessierende relative Anordnung der Items auf dem Merkmalskontinuum ergibt sich direkt aus denjenigen individuellen Rangreihen (I–Skalen) eines Sets, für die eine gespiegelte Präferenz–Rangreihe vorliegt. Für die beiden in Abbildung 4.11 dargestellten Sets von I–Skalen sind dies die Rangreihen A, B, C, D sowie D, C, B, A. Beide Rangreihen repräsentieren jeweils ein extremes Ende des latenten Merkmalskontinuums. Die Position einer einzelnen Person auf dieser so definierten

J–Skala lässt sich wiederum aus ihrer jeweils individuellen I–Skala bestimmen.

	A	B	C	D
$P1$	1	0	0	0
$P2$	1	0	0	0
$P3$	0	1	0	0
$P4$	0	1	0	0
$P5$	0	1	0	0
$P6$	0	0	1	0
$P7$	0	0	1	0
$P8$	0	0	0	1
$P9$	0	0	0	1
$P10$	0	0	0	1

Abbildung 4.12 Beispiel für eine umsortierte (perfekte) Datenmatrix ($n = 10$) in Parallelogrammstruktur für dichotome Präferenzurteile ($k = 4$ Items) nach dem Unfolding-Antwortprozess.

Betrachtet man zum Beispiel für Item A die Positionen der einzelnen Personen in den jeweiligen Präferenz–Rangreihen (I–Skalen) in Abbildung 4.11, so fällt auf, dass sich diese entlang der Diagonalen der Matrix der Präferenzrangreihen von der ersten Position bis zur letzten Position anordnen. Diese diagonale Anordnung der Zustimmung zu einem Item (erste Position auf der vollständigen Präferenz–Rangreihe) ergibt sich beim Unfoldingmodell nach Coombs (1950) für den Antwortprozess immer dann, wenn die resultierende Datenmatrix mit den I–Skalen bezüglich der Spalten (Items) nach der Itemschwierigkeit aufsteigen sortiert wird und gleichzeitig bezüglich der Zeilen (Personen) nach dem Ausmaß der Merkmalsausprägung der Personen sortiert wird. Liegen nun Daten mit unvollständigen (dichotomen) individuellen Präferenz–Ratings vor, wie sie z.B. aus der von Thurstone und Chave (1929) vorgeschlagenen Methode resultieren (vgl. Abschnitt 1.3.3) und werden diese jeweils mit $1 \equiv \textit{Zustimmung}$ und $0 \equiv \textit{Ablehnung}$ kodiert, ergibt sich für perfekt passende Daten nach deren Umsortierung ein typisches *Paralellogrammmuster*, wie es als Besipiel in Abbildung 4.12 für vier Items (A, B, C und D) und zehn Personen (P1 bis P10) dargestellt ist.

Das hier dargestellte, von Coombs (1950) formalisierte Unfoldingprinzip wurde von verschiedenen Autoren aufgegriffen. Die Unfoldingmodelle von Bechtel (1968) und Sixtl (1973) stellen Varianten probabilistischer Erweiterungen des Modells von Coombs (1950) dar. Bechtel (1968) zeigt zum Beispiel, wie sich durch die Einführung bestimmter parametrischer Annahmen bezüglich der (Normal-)Verteilung der Positionen der Personen auf dem latenten Kontinuum, das sich das aus der Anwendung des Modells von Coombs (1950) resultierende ordinale Skalenniveau auf ein Intervallskalenniveau „anheben“ lässt. Sixtl (1973) wiederum erweitert diese parametrische Annahme von Bechtel (1968) dahingehend, dass die Verteilung der Idealpunkte der Personen (deren Position) auf dem latenten Kontinuum empirisch geschätzt wird und damit die Notwendigkeit einer bestimmten (Normal-)Verteilungsannahme entfällt. Einen Überblick über weitere Entwicklungen auf Basis des von Coombs (1950) vorgestellten Unfoldingprinzips wird von Bossuyt (1990) gegeben.

4.3.2 Parametrische Unfoldingmodelle

Ebenso wie bei den psychometrischen Modellen für Dominanz-Antwortprozesse lassen sich auch für Nähe–Distanz-Antwortprozesse Modelle aufstellen, welche die Antwortwahrscheinlichkeiten in den einzelnen Kategorien der Items in Abhängigkeit von bestimmten Modellparametern beschreiben, und dabei probabilistische Aussagen zum Auftreten bestimmter Antwortkategorien der einzelnen Items machen. Ein solches Modell für dichotome Daten ist das von Andrich (1988, 1989) vorgeschlagene (einfache) quadratische logistische Testmodell [*Squared Simple Logistic Model* – SSLM]. Zur Herleitung des SSLM greift Andrich (1988) die Idee der logittransformierten Wahrscheinlichkeiten zur Lösung bzw. Zustimmung zu einem Item i – also $p\left(X_{vi}=1\right)$ aus dem Rasch-Modell auf, und setzt diese gleich mit der negativen **quadrierten** Differenz zwischen der Merkmalsausprägung der Person θ_v und der Schwierigkeit des Items σ_i (vgl. Gleichung 4.16).

$$log\frac{p\left(X_{vi}=1\right)}{1-p\left(X_{vi}=1\right)}=-\left(\theta_v-\sigma_i\right)^2 \tag{4.16}$$

Die resultierende Modellgleichung (vgl. Gleichung 4.17) weist Ähnlichkeit

mit derjenigen das Rasch-Modells auf (vgl. Gleichung 4.6 in Abschnitt 4.2.3).

$$p\left(X_{vi} = x_{vi} | \theta_v; \sigma_i\right) = \frac{e^{\left(-x_{vi} \cdot (\theta_v - \sigma_i)^2\right)}}{1 + e^{\left(-(\theta_v - \sigma_i)^2\right)}}; x_{vi} \in \{0, 1\} \tag{4.17}$$

Der einzige Unterschied zwischen der Modellgleichung in 4.17 und derjenigen des Rasch-Modells (vgl. Gleichung 4.6) besteht darin, dass der Ausdruck $\theta_v - \sigma_i$ quadriert wird und innerhalb der Modellgleichung mit negativen Vorzeichen eingefügt wird. Dies führt dazu, dass die Wahrscheinlichkeiten für die Zustimmung zu einem Item mit zunehmender *Distanz* des Personenparameters θ_v von der Schwierigkeit des Items σ_i sinken – und zwar unabhängig davon, ob der Personenparameter oberhalb oder unterhalb der Itemschwierigkeit auf dem latenten Kontinuum liegt, also die einfache Differenz aus θ und σ positiv oder negativ ausfällt (vgl. Abbildung 4.13).

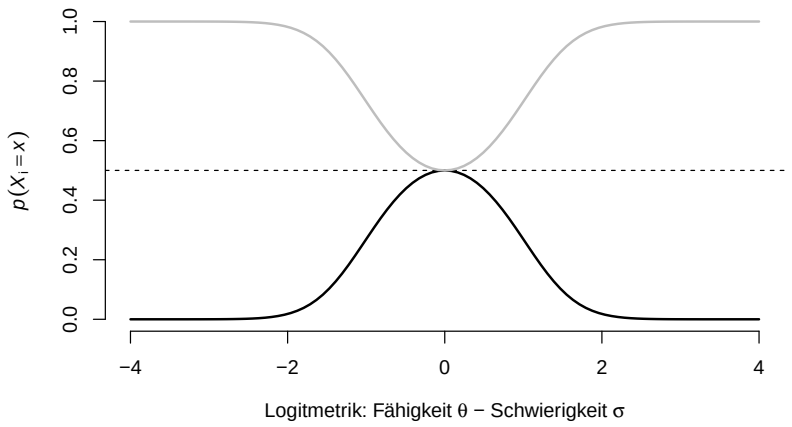


Abbildung 4.13 Darstellung der *Item Characteristic Curve* (ICC) des (einfachen) quadratisch logistischen Modells (Andrich, 1988) für ein Item i mit der Schwierigkeit $\sigma_i = 0$.

In der Abbildung 4.13 gibt die schwarze Kurve den Verlauf der Zustimmungswahrscheinlichkeit zu einem Item mit der Schwierigkeit $\sigma = 0$ in Abhängigkeit der Differenz zwischen θ_v und $\sigma_i = 0$ an und die graue Kurve die entsprechende Gegenwahrscheinlichkeit (Ablehnung des Items). Dabei zeigt

sich ziemlich deutlich eine „inhaltliche“ Schwäche dieses psychometrischen Antwortmodells. Trotz seiner recht eleganten Ableitung erscheint es dennoch psychologisch unplausibel zu sein, dass im Falle einer perfekten Übereinstimmung zwischen der Merkmalsausprägung der Person und der Position des Items auf dem latenten Kontinuum die maximale Zustimmungswahrscheinlichkeit lediglich den Wert von $p = .5$ – also letztlich die Ratewahrscheinlichkeit – erreicht. Rost (2004) merkt im Zusammenhang mit seiner Darstellung dieses Modells von Andrich (1988) an, dass *„dieser Sachverhalt die Brauchbarkeit [dieses Modells] einschränken“* (Rost, 2004, S. 143 - Ergänzungen in eckigen Klammern). Allerdings muss als Exkurs angemerkt werden, dass dieser Nachteil, zumindest bei der Anwendung des Modells auf Daten aus mehrfach Wahlaufgaben mit fest vorgegebenen Antwortalternativen, relativ leicht umgangen werden könnte. So könnte eine Möglichkeit zur Lösung der Problematik zum Beispiel darin bestehen, die Ratewahrscheinlichkeit einfach als festen, nicht zu schätzenden Modellparameter (hier beispielsweise $\chi_i = .5$) zusätzlich in die Modellgleichung aufzunehmen. Ein vergleichbares Prinzip schlagen beispielsweise Kubinger und Draxler (2007a) vor, um beim (kumulativen) Rasch-Modell bzw. dessen Erweiterung für mehrstufig Antwortformate (Masters, 1982) den Rate-Effekt bei *multiple-choice* Aufgaben zu kontrollieren, ohne einen weiteren zu *schätzenden* Modellparameter einzuführen (vgl. auch Kubinger & Draxler, 2007b).

Dieser feste, nicht zu schätzende Parameter wird dann zu der Zustimmungswahrscheinlichkeit, die sich in Abhängigkeit der beiden (zu schätzenden) Parameter θ_v und σ_i ergibt, hinzuaddiert (vgl. Gleichung 4.18) sowie von der Gegenwahrscheinlichkeit (Ablehnung des Items) abgezogen (vgl. Gleichung 4.19)

$$p\left(X_{vi} = 1|\theta_v;\sigma_i\right) = \frac{e^{\left(-x_{vi}\cdot(\theta_v-\sigma_i)^2\right)}}{1+e^{\left(-(\theta_v-\sigma_i)^2\right)}} + \chi_i \;;\chi_i = .5 \tag{4.18}$$

$$p\left(X_{vi} = 0|\theta_v;\sigma_i\right) = \frac{e^{\left(-x_{vi}\cdot(\theta_v-\sigma_i)^2\right)}}{1+e^{\left(-(\theta_v-\sigma_i)^2\right)}} - \chi_i \;;\chi_i = .5 \tag{4.19}$$

Der Umstand das dieser Parameter dabei ausgerechnet auf einen Wert von $\chi_i = .5$ fixiert wird, ergibt sich aus dem Kehrwert der Anzahl der fest vorgegebenen Antwortkategorien $m = 2$. Für die Anwendung dieses Prinzips auf Modelle für den Dominanz-Antwortprozess bezeichnen Kubinger und Draxler (2007a, 2007b) die so erweiterten Modelle als „Difficulty plus Guessing

PL Modell“ (Kubinger & Draxler, 2007b, S. 138) – vgl. auch Keats (1974). Nimmt man beispielsweise für den Rateprozess bei einem dichotomen Antwortformat ein Laplace Zufallsexperiment an, muss dann für den Wert von χ_i gelten: $\chi_i = \frac{1}{m} = 1/2$. Die in Abbildung 4.13 dargestellt schwarze Kurve der Zustimmungswahrscheinlichkeit verschiebt sich dabei um den Wert des festen Parameters $\chi_i = .5$ entlang der y-Achse nach oben und weist dadurch eine untere Asymptote auf dem Niveau von $p = .5$ auf, welche der „Ratewahrscheinlichkeit“ entspricht. Gleichzeitig erreicht die Zustimmungswahrscheinlichkeit an der Stelle $\theta - \sigma = 0$ mit $p = 1$ ihr Maximum. Die graue Kurve der Gegenwahrscheinlichkeit weist dementsprechend ebenfalls eine obere Asymptote bei $p = .5$ auf. Während bei Modellen für den Dominanz-Antwortprozess der nach diesem Prinzip eingeführte Parameter χ_i für die untere Asymptote inhaltlich als *Rateparameter* zu interpretieren ist, dürfte sich dessen inhaltliche Interpretation bei einem Nähe–Distanz-Antwortprozess aber eher in Richtung „baseline Parameter“ für zufälliges oder nachlässiges Antwortverhalten (careless responding) verschieben (vgl. Abschnitt 3.2.3 in Kapitel 3 *Theoretischer Hintergrund zu Antwortmustern*). Diese hier nur kurz als Exkurs skizzierte, mögliche Erweiterung durch eine einfache und plausible Annahme zum Rate- bzw. Antwortverhalten bei einer fest vorgegebenen Anzahl von Antwortkategorien, ist allerdings, nach Kenntnis des Autors der vorliegenden Arbeit, in der psychometrischen Literatur für Unfoldingmodelle bislang nicht diskutiert worden.

Ein ähnliches Modell, das ebenfalls von einem Modell für den Dominanz-Antwortprozess abgeleitet wurde, ist das von Andrich und Luo (1993) und Verhelst und Verstralen (1993) parallel entwickelte Hyperbelcosinus-Modell [*Hyperbel Cosinus Model* – HCM] (vgl. auch Andrich, 1995). Im Gegensatz zum quadratisch logistischen Testmodell von Andrich (1988) leitet sich das Hyperbelcosinus-Modell nicht vom Rasch-Modell (für dichotome Antwortformate), sondern von dessen polytomer Erweiterung, dem *Partial Credit Model* (PCM – Masters, 1982), ab. Speziell von der Betrachtung des Spezialfalles von Items mit einer Antwortskala mit drei Kategorien. Die grundlegende Idee beruht zunächst auf der „Beobachtung“, dass die mittlere Kategorie bei dreikategoriellen Items bereits eine unimodale ICC aufweist. Zur Herleitung des Hyperbelcosinus-Modells aus dem PCM werden dann die beiden Randkategori-

en mit einer jeweils monoton sinkenden (unterste Kategorie „0“) und monoton steigenden (oberste Kategorie „2“) ICC, zu einer verbundenen ICC für eine gemeinsame Ablehnungskategorie – bestehend aus zwei Ablehnungsgründen – kombiniert (z. B. Rost, 2004, S. 144 ff., für eine Darstellung der Herleitung) – vgl. auch Andrich (1995) sowie Andrich und Luo (1993). Bei dieser Herleitung führen Andrich und Luo (1993) zunächst einen zusätzlichen, von ihnen als „*unit Parameter*“ bezeichneten Modellparameter δ [1] ein. Inhaltlich lässt sich dieser Parameter als Dispersions- oder auch Trennschärfeparameter interpretieren, welcher zunächst im Wesentlichen die Breite der ICC bestimmt. Je größer der Parameter ausfällt, desto breiter fällt die ICC des Items aus und desto geringer seine Trennschärfe (S. 272 Andrich, 1995). Dies folgt aus der Definition des „Unit“- oder Dispersionsparameters aus den beiden Schwellenparametern (zwischen den Kategorien „0“ und „1“, sowie „1“ und „2“) des PCM: $\delta = \frac{(\tau_1 - \tau_2)}{2}$. Weiter auseinanderliegende Schwellen τ_1 und τ_2 entsprechen daher einer breiteren ICC des abgeleiteten Hyperbelcosinus-Modells. Gleichung 4.20 gibt die Modellgleichung des noch allgemein formulierten Hyperbelcosinus-Modells wieder, in welcher der Dispersionsparameter δ noch enthalten ist.

$$p\left(X_{vi} = x_{vi} | \theta_v; \sigma_i; \delta\right) = \frac{\left(e^{\delta}\right)^{x_{vi}} \cdot \left(2 \cdot \cosh\left(\theta_v - \sigma_i\right)\right)^{1 - x_{vi}}}{\left(e^{\delta}\right) + \left(2 \cdot \cosh\left(\theta_v - \sigma_i\right)\right)}; x_{vi} \in \{0, 1\} \quad (4.20)$$

Durch eine mathematisch günstige Wahl eines festen Wertes für den Parameter δ (vgl. Andrich & Luo, 1993, S. 260-261) lässt sich die in 4.20 gegebene Modellgleichung erheblich vereinfachen. Wird für den Parameter δ der Wert $\delta = \ln 2$ gewählt reduziert sich die in 4.20 gegeben Modellgleichung zu der in 4.21 gegeben Gleichung.

$$p\left(X_{vi} = x_{vi} | \theta_v; \sigma_i\right) = \frac{\left(\cosh\left(\theta_v - \sigma_i\right)\right)^{1 - x_{vi}}}{1 + \cosh\left(\theta_v - \sigma_i\right)}; x_{vi} \in \{0, 1\} \quad (4.21)$$

Bereits Andrich und Luo (1993) weisen darauf hin, dass dieser Spezialfall des Hyperbelcosinus-Modells wie er in Gleichung 4.21 wiedergegeben ist,

[1] in der orignal Publikation von Andrich und Luo (1993) wird dieser Parameter mit θ bezeichnet. Um hier in dieser Arbeit aber den Buchstaben θ konsistent als Merkmalsausprägung der Personen beizubehalten, wurde hier auf den Buchstaben δ ausgewichen. Dies auch deshalb, weil es sich inhaltlich letztlich um einen *D*ispersionsparameter handelt (daher δ), welcher die Breite der ICC bestimmt.

äquivalent zum (einfachen) quadratisch logistischen Testmodell von Andrich (1988) ist (vgl. Abbildung 4.14 - durchgezogene Kurven und Abbildung 4.13). Rost (2004) weist darauf hin, dass die Gleichung 4.21 für den Spezialfall des Hyperbelcosinus-Modells (mit $\delta = \ln 2$) der Gleichung des Rasch-Modells (vgl. Gleichung 4.6 in Abschnitt 4.2.3) sehr ähnlich ist: „*In beiden Fällen hängt die Antwortwahrscheinlichkeit nur von der Differenz von Personenfähigkeit und Itemschwierigkeit ab, jedoch einmal mittels der Exponentialfunktion und einmal mittels des Hyperbelcosinus.*“ (Rost, 2004, S. 147). Der Unterschied des Verlaufs der Graphen dieser beiden Funktionen (vgl. Abbildung 66. in Rost, 2004, S. 147) spiegelt in gewisser Weise den unterschiedlichen Antwortprozess wieder, welcher mit dem Rasch-Modell einerseits (monoton steigende ICC) und dem Hyperbelcosinus-Modell (unimodale ICC) modelliert werden.

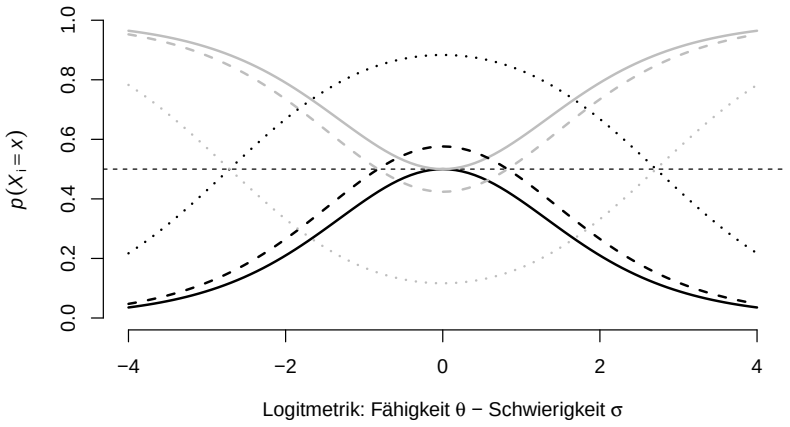


Abbildung 4.14 Darstellung der *Item Characteristic Curve* (ICC) des Hyperbelcosinus-Modells Andrich und Luo (1993) für ein Item i mit der Schwierigkeit $\sigma_i = 0$ bei unterschiedlichen Werten für den „*unit Parameter*“ δ; durchgezogene Kurven: $\delta = \ln 2$, vgl. (4.21)

Als zentraler Unterschied zwischen dem Hyperbelcosinus-Modell und dem (einfachen) quadratisch logistischen Testmodell von Andrich (1988) mag die psychologisch (noch) plausiblere Herleitung des Hyperbelcosinus-Modells aus

dem PCM (für drei Antwortkategorien) angesehen werden. Dabei wird die Idee modelliert, dass ein Item mit allgemein formulierter Aussage letztlich von „zwei Seiten“ her abgelehnt werden kann. Einerseits weil die Merkmalsausprägung der antwortenden Person *über* und andererseits auch *unter* der durch das jeweilige Item repräsentierten Merkmalsintensität liegen kann (vgl. Abschnitt 1.3.1, sowie auch van Schuur, 2011).

Dennoch teilt der in Gleichung 4.21 dargestellte Spezialfall des Hyperbelcosinus-Modells das gleiche „Problem“ einer maximalen Zustimmungswahrscheinlichkeit von lediglich $p = .5$ für den Fall einer perfekten Übereinstimmung der Merkmalsausprägung der Person und der Lage (Schwierigkeit) des Items (Andrich & Luo, 1993, S. 261). Allerdings eröffnet die generalisierte Form des Hyperbelcosinus-Modells (vgl. Gleichung 4.20) immerhin die Möglichkeit, die Zustimmungswahrscheinlichkeit (ohne zusätzliche Parameter) über einen Wert von $p = .5$ zu heben, wenn für den „*unit Parameter*“ δ (Dispersionsparameter) ein größerer Wert als $\delta = \ln 2$ gewählt wird. In Abbildung 4.14 werden in Anlehnung an die Darstellung von (Andrich & Luo, 1993, S. 260) verschiedene Verläufe der ICC in Abhängigkeit unterschiedlicher Werte für den Parameter δ gegeben. Wie dabei aus der Abbildung 4.14 recht deutlich hervorgeht, lässt sich zwar der Wert für die maximale Zustimmungswahrscheinlichkeit, je nach gewähltem Wert für den Parameter δ, nahezu beliebig steigern, jedoch geht dies mit einer zunehmend geringer werdenden Trennschärfe (die Breite der ICC nimmt zu), einher.

Diese Problematik wird mit einem weiteren Modell für dichotome Daten, die nach einem Nähe–Distanz-Antwortprozess modelliert werden sollen, umgangen. Dieses als *PARELLA-Modell* bezeichnete Testmodell mit unimodaler ICC wurde bereits vor dem Hyperbelcosinus-Modell von Hoijtink (1990) vorgeschlagen. Das PARELLA-Modell von Hoijtink (1990) verfügt im Gegensatz zum quadratisch logistischen Testmodell von Andrich (1988) über einen plausiblen Verlauf der Itemfunktion, welche bei einer perfekten Übereinstimmung von Item- und Personeneigenschaft auch den Wert $p = 1$ erreicht. Hoijtink (1990) gibt formal die in Gleichung 4.22 dargestellte Definition zur Bestimmung der Zustimmungswahrscheinlichkeit $p\left(X_{vi} = 1\right)$ in Abhängigkeit der Parameter θ für die Merkmalsausprägung der Person, σ für die Schwierigkeit des Items sowie einem zusätzlichen, von Hoijtink (1990) als *power parameter*

bezeichneten Parameter γ.

$$p\left(X_{vi}=1|\theta_v;\sigma_i;\gamma\right)=\frac{1}{1+\left(\left(\theta_v-\sigma_i\right)^2\right)^{\gamma}}=1-p\left(X_{vi}=0|\theta_v;\sigma_i;\gamma\right) \quad (4.22)$$

Der *power parameter* γ kann dabei anlog zum weiter oben dargestellten 2–PL–Modell (vgl. Abschnitt 4.2.4) als Trennschärfeparameter aufgefasst werden, der einen theoretischen Wertebereich von $-\infty$ bis $+\infty$ aufweist. Mit steigendem positiven Werten für γ nähert sich die probabilistische Charakteristik des Modells einem deterministischen Modell an (vgl. Abbildung 1 in Hoijtink, 1990, S. 643; sowie Abbildung 4.15). Die entsprechende Gegenwahrscheinlichkeit $p\left(X_{vi}=0\right)$ zur „Wahl“ der Ablehnungskategorie lässt sich gemäß Gleichung 4.23 schreiben.

$$p\left(X_{vi}=0|\theta_v;\sigma_i;\gamma\right)=\frac{\left(\left(\theta_v-\sigma_i\right)^2\right)^{\gamma}}{1+\left(\left(\theta_v-\sigma_i\right)^2\right)^{\gamma}} \quad (4.23)$$

Werden die Daten, wie beim dichotomen Rasch-Modell, mit „0“ für die Ablehnung und „1“ für die Zustimmung zu einem Item kodiert, lassen sich die beiden Gleichungen 4.22 und 4.23 zu der in Gleichung 4.24 gegebenen Modellgleichung kombinieren.

$$p\left(X_{vi}=x_{vi}|\theta_v;\sigma_i;\gamma\right)=\frac{\left(\left(\theta_v-\sigma_i\right)^{2\cdot\left(1-x_{vi}\right)}\right)^{\gamma}}{1+\left(\left(\theta_v-\sigma_i\right)^2\right)^{\gamma}};x_{vi}\in\{0,1\} \quad (4.24)$$

Trotz seines plausiblen Verlaufes der ICC mit einem Maximum von $p=1$ kann an dem PARELLA Modell bemängelt werden, dass „*hier jedoch keine nachvollziehbare Ableitung der Modellgleichung aus einfachen Annahmen über das Antwortverhalten*“ (Rost, 2004, S. 144) besteht, also nicht eine logistische Funktion wie beim Rasch-Modell zugrunde gelegt wird. Betrachtet man die Modellgleichung in 4.24 – und dort insbesondere den Nenner – genauer, fällt darüber hinaus auf, dass die maximale Zustimmungswahrscheinlichkeit ($p=1$) nur unter bestimmten mathematischen Konventionen bezüglich des Ergebnisses des Ausdruckes 0^0 erreicht wird. Für den Fall einer perfekten Übereinstimmung der Merkmalsausprägung der Person (θ) und der Schwierigkeit des Items (σ) ergibt sich als Differenz der beiden Parameter zunächst der Wert

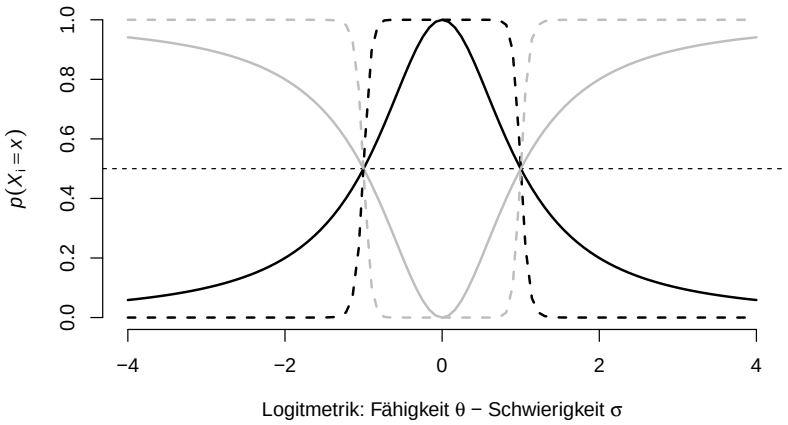


Abbildung 4.15 Darstellung der *Item Characteristic Curve* (ICC) des PARELLA-Modells (Hoijtink, 1990) für zwei Items $i = 1$ und $i = 2$ mit der Schwierigkeit $\sigma_{i=1} = \sigma_{i=2} = 0$; durchgezogene Linien der ICC:$\gamma_{i=1} = 1$, gestrichelte Linien der ICC:$\gamma_{i=2} = 10$.

$\theta - \sigma = 0$. Für die Kategoriewahrscheinlichkeit der Ablehnungskategorie (kodiert mit „0“) ergibt sich für den Ausdruck $2 \cdot (1 - x_{vi})$ im Exponenten des Nenners der Wert zwei, sodass der Nenner insgesamt den Wert null ($0^2 = 0$) erreicht. Zusammen mit dem Zähler erreicht damit die Ablehnungswahrscheinlichkeit den (plausiblen) Wert $p = 0$; ($0/1 = 0$). Für die Kategoriewahrscheinlichkeit der Zustimmungskategorie (kodiert mit „1“) im selben Fall, muss sich daher der Wert $p = 1$ ergeben. Allerdings ergibt sich hier für den Ausdruck $2 \cdot (1 - x_{vi})$ im Exponenten des Nenners der Wert null sowie für die Differenz aus θ und σ ebenfalls der Wert null, sodass hier der Ausdruck 0^0 stehen bleibt. Unter der Voraussetzung einer Definition von $0^0 = 1$ ergibt sich gemeinsam mit dem Zähler der Wert $p = 1$ ($1/1 = 1$) für die Zustimmungswahrscheinlichkeit. Diese Definition des Ergebnisses des Ausdrucks $0^0 = 1$ wird allerdings in der Mathematik durchaus umstritten diskutiert (vgl. Knuth, 1992, S. 407-408). Das PARELLA Model von Hoijtink (1990) formalisiert somit zwar einen psychologisch durchaus plausiblen Verlauf der ICCs, kann aber hinsichtlich seiner fehlenden plausiblen Ableitung aus einfachen Annahmen aus dem Antwort-

prozess (vgl. Rost, 2004, S. 144) sowie aufgrund der hier kurz dargestellten, aus mathematischer Perspektive, formalen Schwächen, kritisiert werden.

Für Items mit polytomen Antwortformaten wurde das Hyperbelcosinus-Modell (Andrich & Luo, 1993; Verhelst & Verstralen, 1993) von Andrich (1996) für mehrstufige Antwortformate zum generalisierten Hyperbelcosinus-Modell [*Generalized Hyperbel Cosinus Model* – GHCM] erweitert (vgl. Rost & Luo, 1997, für ein praktisches Anwendungsbeispiel) – Luo (1998a, 1998b) gibt hierzu eine Übersicht zu den Zusammenhängen zu vergleichbaren Modellen. Von Roberts und Laughlin (1996) wurde das abgestufte Entfaltungsmodell [*Graded Unfolding Model* – GUM] vorgeschlagen, welches von Roberts et al. (2000) als generalisiertes abgestuftes Entfaltungsmodell [*Generalized Graded Unfolding Model* – GGUM] durch Hinzunahme eines Trennschärfeparameters verallgemeinert wurde.

Das von Andrich (1996) und die von Roberts und Laughlin (1996) sowie Roberts et al. (2000) vorgeschlagenen Modelle weisen einige Parallelen auf (vgl. Luo, 1998a, 1998b). Das GGUM Modell von Roberts et al. (2000) wurde ausgehend vom GPCM von Muraki (1992) abgeleitet und im Vergleich zum GUM von Roberts und Laughlin (1996) um einen Trennschärfeparameter erweitert. Während Roberts und Laughlin (1996) und Roberts et al. (2000) das GUM bzw. GGUM für polytome Antwortformate mit einer geraden Anzahl von Antwortkategorien entwickeln, bezieht sich die polytome Erweiterung des Hyperbelcosinus-Modells, das GHCM (Andrich, 1996), auf eine ungerade Anzahl von Antwortkategorien - „*Roberts (1995) [Dissertation: publiziert als (Roberts & Laughlin, 1996)] has operationalized the same model but with an even number of categories always rather than the odd number in the construction in this paper*“ (Andrich, 1996, S. 361; Anmerkungen in eckigen Klammern).

Luo (2001) formuliert eine allgemeine Klasse von parametrischen Unfoldingmodellen für polytome Antwortformate, innerhalb derer sich die Modelle von Andrich (1996), Roberts und Laughlin (1996) und Rost und Luo (1997) jeweils als Spezialfälle einordnen lassen. Aus Kapazitäts- und Platzgründen sollen die unterschiedlichen formalen Ableitungen und Parametrisierungen dieser einzelnen Modelle hier nicht im Einzelnen dargestellt werden. Stattdessen sei an dieser Stelle auf die hier jeweils zitierten Originalpublikationen verwiesen. Allerdings soll zum besseren Verständnis das den hier erwähnten parametrischen

Unfoldingmodellen für polytome Antwortformate zugrunde liegende Prinzip, bei deren Herleitung aus den jeweiligen kumulativen Modellen, grafisch veranschaulicht werden.

Das Prinzip besteht zunächst darin, für jede beobachtete Antwortkategorie (BAK), wie sie in den Daten kodiert ist (also z. B. $1, 2, 3...m$ für m Antwortkategorien), jeweils zwei subjektive Antwortkategorien (SAK) zu postulieren. Diese beiden subjektiven Antwortkategorien repräsentieren die beiden Ablehnungsgründe der jeweiligen vorgegebenen Antwortkategorien in den empirischen Daten (vgl. Beispiel in Abschnitt 1.3.1), welche beim einem Nähe–Distanz-Antwortprozess bestehen.

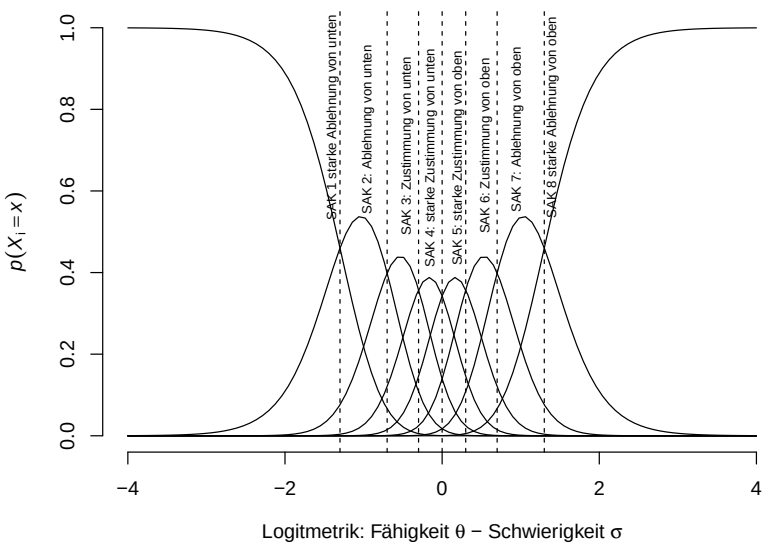


Abbildung 4.16 Darstellung der *Item Characteristic Curve* (ICC) der subjektiven Antwortkategorien (SAK) des GGUM (Roberts et al., 2000)

Die Abbildung 4.16 zeigt zunächst den Verlauf der Wahrscheinlichkeiten dieser sogenannten „subjektiven Antwortkategorien“ (SAK). Diese repräsentieren jeweils die *beiden* Gründe der Ablehnung der „beobachteten Antwortkategorie“ (BAK). Jede einzelne Antwortkategorie einer ordinalen polytomen Antwortskala kann beim Nähe–Distanz-Antwortprozess aus genau zwei Gründen nicht

ausgewählt werden. Entweder, weil die Merkmalsausprägung der antwortenden Person oberhalb, oder weil sie unterhalb der Kategorieschwierigkeit bzw. deren Lage auf dem latenten Kontinuum liegt. Die beiden jeweiligen Ablehnungskategorien werden nun bei polytomen Unfoldingmodellen (z. B. beim GGUM) zu einer gemeinsamen Ablehnungskategorie addiert. Die subjektiven Antwortkategorien (SAK) werden sozusagen gemäß dem bereits von Coombs (1950) beschriebenen Prinzip der Entfaltung und Faltung der individuellen I-Skala und der gemeinsamen J-Skala, an ihrer Symmetrieachse gefaltet. Damit werden die Antwortwahrscheinlichkeiten der SAK jeweils addiert und beschreiben die Antwortkategoriewahrscheinlichkeiten der BAK in den beobachteten Daten. Die Abbildung 4.17 gibt den Verlauf der Wahrscheinlichkeiten für die

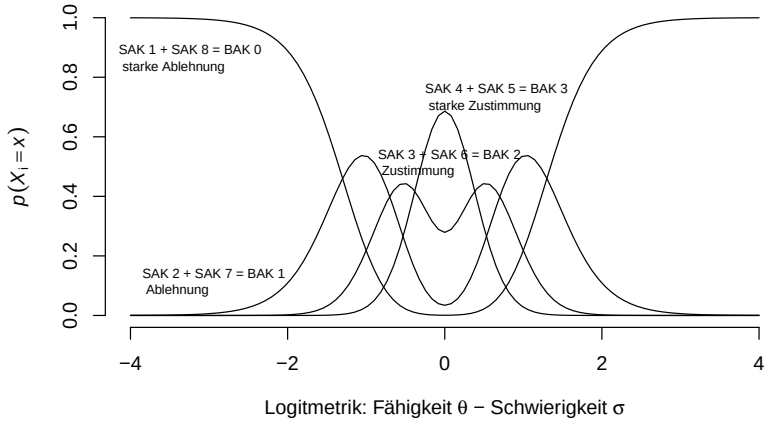


Abbildung 4.17 Darstellung der *Item Characteristic Curve* (ICC) der beobachteten Antwortkategorien (BAK) des GGUM (Roberts et al., 2000)

beobachteten Antwortkategorien in den Daten (BAK) in Abhängigkeit der Differenz aus θ und σ wieder. Bei der Betrachtung der Kurvenverläufe in Abbildung 4.17 wird deutlich, dass lediglich die oberste Antwortkategorie zur maximalen Zustimmung zu einem Item einen unimodalen Verlauf nimmt. Alle anderen ICCs weisen eine bimodale Verteilung der Antwortwahrscheinlichkeiten auf, welche sich symmetrisch um den Nullpunkt bewegen. Der Nullpunkt ist dabei derjenige Punkt auf dem latenten Kontinuum, bei dem die Distanz

zwischen der Merkmalsausprägung der Person und der Lage des Items den Wert null annimmt. Dieser Punkt weist somit eine maximale Wahrscheinlichkeit zur Wahl der höchsten beobachteten Antwortkategorie (BAK) auf. Obwohl auch der höchsten Antwortkategorie aus zwei Gründen zugestimmt werden kann (Zustimmung von Oben und Unten – vgl. Abbildung 4.16), weist die Wahrscheinlichkeitsverteilung zur Zustimmung zu der höchsten beobachteten Antwortkategorie (BAK) deswegen einen unimodalen Verlauf auf, weil keine weitere, höhere Antwortkategorie zwischen diesen beiden Gründen der Zustimmung zu dem Item liegt. Das hier grafisch aufgezeigte Prinzip der Unterscheidung latenter, subjektiver Antwortkategorien, aus denen sich in Folge dessen dann beobachtete Antwortkategorien ableiten, welche in den Daten eine Unfoldingstruktur erzeugen, nennt E. Maris (1995) die *collapsing condensation rule* (dt. etwa: kollabierende Kondensationsregel). Danach „kondensieren" gewissermaßen die nicht beobachtbaren subjektiven Antwortkategorien (SAK), welche durch eine *Nähe–Distanz*-Relation gekennzeichnet sind, zu den beobachteten Antwortkategorien als Ergebnis eines internen kognitiven Prozesses bei der Beantwortung von Fragebogen-Items. Die beobachteten Reaktionen der Personen werden so mit deren latenten Antworten über eine parametrische Modellformulierung eines Nähe–Distanz-Antwortprozesses in Beziehung gesetzt (M. S. Johnson, 2006).

4.3.3 Weitere Modelle zur Abbildung des Unfoldingprozesses

Neben der parametrischen Entwicklungsrichtung zur Abbildung eines Nähe–Distanz-Antwortprozesses sind auch nichtparametrische Modelle nach dem Unfoldingprinzip von Coombs (1950) zur Bestimmung der relativen Skalenpositionen von Personen und Items entwickelt und eingesetzt worden (z. B. M. S. Johnson, 2006; Post, 1992; van Schuur, 1984, 1992, 1997). Analog zu dem weiter oben dargestellten Prinzip der Skalierung nach Mokken (1971) entwickelt van Schuur (1984) eine nichtparametrische Methode zur eindimensionalen Unfoldingskalierung dichotomer Items (vgl. auch Post & Snijders, 1993). Diese nichtparametrische Methode zur Unfoldingskalierung wurde von van Schuur (1992) auf polytome Items erweitert. Eine Übersicht zu nichtparametrischen

Modellen zur Unfoldingskalierung findet sich bei van Schuur (1997) sowie eine praktische Anwendung solcher Modelle auf empirische Daten bei van Schuur (1995). Bossuyt und Roskam (1989) stellen eine probabilistische Modellformulierung eines Nähe–Distanz-Antwortprozesses ohne parametrische Annahmen dar.

Ein weiteres parametrisches und probabilistisches Model stellt Klinkenberg (2001) vor, das mit einem Vorzeichen versehene ein-Parameter-logistische-Modell [*Signed One-Parameter Logistic Model* – Signed OPLM]. Das Signed OPLM wird von Klinkenberg (2001) vom *One Parameter Logistic Model* (OPLM) von Verhelst und Glas (1995) abgeleitet. Dieses stellt einen Spezialfall des Rasch-Modells dar, wobei die Trennschärfen der Items im OPLM zwar als Parameter geschätzt, aber über eine Modellrestriktion alle auf den gleichen Wert gesetzt werden – was die Überschneidungsfreiheit der ICCs gewährleistet (vgl. Abschnitt 4.2.3) – wohingegen im Rasch-Modell diese Trennschärfen (über das Fehlen eines solchen Parameters) implizit mit einem Wert von $\alpha = 1$ für alle Items angenommen werden. Die Erweiterung von Klinkenberg (2001) besteht nun darin, den Trennschärfeparameter des OPLM als a priori zu setzenden *pseudo* Parameter (Klinkenberg, 2001, S. 176) zu implementieren. Entweder auf Basis von theoretischen (Vor-)Überlegungen oder anhand empirischer (Vor-)Untersuchungen basierend auf den Kategoriehäufigkeiten in den Daten (vgl. Verhelst & Glas, 1995), muss der Trennschärfeparameter ein positives oder negatives Vorzeichen tragen. Dies erlaubt die Modellierung sowohl ansteigender als auch absteigender ICCs innerhalb ein und derselben Skala für die Zustimmungskategorie der Items einer Skala. Bei entsprechender Polung oder Formulierung der Items besteht damit die Möglichkeit, eine Ablehnung entsprechender Items aufgrund der eigenen Position oberhalb oder unterhalb auf dem latenten Kontinuum der gemessenen Eigenschaft zum Ausdruck zu bringen. Einschränkend muss zu dem Signed OPLM von Klinkenberg (2001) angemerkt werden, dass es sich im Gegensatz zu den bisher in diesem Abschnitt vorgestellten Modellen nicht um ein „echtes“ Unfoldingmodell mit eingipfliger ICC zur Abbildung einer *Nähe–Distanz*-Relation zwischen Personen und *jedem einzelnen* Item handelt. Allerdings erlaubt die Anwendung des Modells die Modellierung einer empirischen Situation, in der die ICCs einiger Items abnehmen und andere zunehmen. Eine solche Situation kann sich als

Ergebnis einer bipolaren latenten Merkmalsdimension ergeben, für deren Erfassung einerseits extrem negativ und andererseits extrem positiv formulierte Items eingesetzt werden (vgl. auch Abschnitt 3.2.2). Zumindest auf Skalenebene erlaubt das Modell dabei die Modellierung zweier Ablehnungsgründe zu den Items – entweder aufgrund der Position der antwortenden Person oberhalb oder unterhalb auf dem latenten Kontinuum. Ein weiteres, dem Modell von Klinkenberg (2001) ähnlichen Modell, stellen G. Maris und Maris (2002) vor, das allerdings (für jedes Item) zunächst eine eingipflige ICC vorsieht, welche als Spezialfall auch einen monoton ansteigenden (oder absteigenden) Verlauf annehmen kann.

4.3.4 Zusammenfassung und Übersicht zu Modellen für *Nähe–Distanz*-Antwortprozesse

Coombs (1950) legt mit seiner nichtparametrischen Formulierung der *Unfolding*-Technik die Grundlage für weitere Modelle zur Modellierung eines *Nähe–Distanz*-Antwortprozesses. Ähnlich wie das Guttman-Modell (Guttman, 1947), hat das Skalierungsmodell von Coombs (1950) einen deterministischen Charakter. Allerdings müssen (Antwort-)Entscheidungen, welche auf *Nähe–Distanz*-Prozessen basieren, bei den antwortenden Personen nicht notwendiger Weise konsequent und konsistent ausfallen – im Sinne von daraus resultierenden vollständig *transitiven* Rangreihen. Aus diesem Grund wurden viele probabilistische Erweiterungen des Unfoldingprinzips in Form von parametrischen (Andrich, 1988; Andrich & Luo, 1993; Bechtel, 1968; Hoijtink, 1990; Klinkenberg, 2001; Luo, 2001; Roberts et al., 2000; Roberts & Laughlin, 1996; Sixtl, 1973; Verhelst & Verstralen, 1993) und nichtparametrischen (Bossuyt & Roskam, 1989; van Schuur, 1984, 1992) Modellen entwickelt (vgl. Abschnitt 4.3). Diese Modelle führen die Wahrscheinlichkeit zur Wahl eines Items, oder einer Antwortkategorie, auf die *Distanz* zwischen der Position der Person und dem jeweiligen Item auf der gemeinsamen latenten Dimension, zurück. Alle diese probabilistischen Unfoldingmodelle sind geometrisch. Das bedeutet, die Items oder deren Antwortkategorien und die Idealpunkte der Personen werden als Punkte in einem metrischen Raum dargestellt. Die (Un-)Ähnlichkeit zwischen Items und Idealpunkten wird dabei durch eine metrische Abstands-

funktion bestimmt, welche den Abstand zwischen den beiden entsprechenden Punkten in diesem Raum modelliert. In der Praxis ergeben sich bei der Parameterbestimmung für Unfoldingmodelle nicht selten Probleme im Hinblick auf deren Identifikation und Genauigkeit (Carter & Zickar, 2011; de la Torre, 2006; Heiser & Meulman, 1983; M. S. Johnson & Junker, 2003, vgl. auch Abschnitt 4.5). Zur Umgehung dieser Probleme kann das *Nähe–Distanz*-Prinzip der parametrischen Unfoldingmodelle durch die Multidimensionale Skalierung (MDS) beschrieben werden (Meulman, Hubert & Heiser, 1998; Warrens & Heiser, 2006, vgl. auch Abschnitt 4.5.4). Eine Übersicht zu verschiedenen Modellen zur Modellierung eines *Nähe–Distanz*-Antwortprozesses, die *deterministisch* oder *probabilistisch*, *nichtparametrisch* oder *parametrisch*, sowie für *dichotome* oder *polytome* Antwortformate formuliert sein können, ist in Tabelle 4.3 gegeben.

Die in der Tabelle 4.3 systematisierten IRT-Modelle für den *Nähe–Distanz*-Antwortprozess modellieren, trotz ihrer Unterschiedlichkeit, alle den psychologischen Mechanismus einer *Nähe–Distanz-Relation* zwischen Items und Personen. Das gemeinsame zugrunde liegende Prinzip besteht darin zu postulieren, dass die Zustimmungswahrscheinlichkeit, z. B. bei einem dichotomen Item, mit geringer werdender (*absoluter*) Differenz zwischen der Merkmalsausprägung der Person θ_v und der Itemschwierigkeit σ_i ansteigt, woraus (im dichotomen Fall) eine eingipflige ICC resultiert. Formal kann, parametrisch und probabilistisch ausgedrückt, die Zustimmungswahrscheinlichkeit p_i zu einem Item i (im dichotomen Fall), trotz unterschiedlicher Merkmalsausprägung $\theta_a \neq \theta_b$ von zwei Personen a und b, gleich ausfallen – also $p_i(\theta_a) = p_i(\theta_b)$ – solange gilt $|\sigma_i - \theta_a| = |\theta_b - \sigma_i|$. Beim (einfachen) quadratisch logistischen Modell von Andrich (1988) wird diese Gleichheit der (absoluten) Differenz über die Exponentialfunktion mit quadratischem Exponenten (vgl. Gleichung 4.17), und beim Hyperbelcosinus-Modell von Andrich und Luo (1993) und Verhelst und Verstralen (1993) über die Hyperbelcosinusfunktion (vgl. Gleichung 4.21) erreicht.

Dieses gemeinsame Prinzip der Modelle in Tabelle 4.3 ist verbunden mit der Eigenschaftsmessung nach Thurstone (1927c) und Thurstone und Chave (1929), welches in der konzeptionellen Darstellung durch Coombs (1950) mit dem Begriff *Unfolding* bezeichnet wird. Die grundlegende Idee besteht darin, dass Personen nur denjenigen Items zustimmen, deren Aussagen sie ihrer

Tabelle 4.3 Übersicht zur Klassifikation von eindimensionalen IRT-Modellen für den *Nähe–Distanz*-Antwortprozess.

	deterministisch	*probabilistisch*
nichtparametrisch		dichotom
	Coombs (1950)	*Multiple Stochastic Unidimensional Unfolding* – MUDFOLD, (van Schuur, 1984; van Schuur & Molenaar, 1982); *Nonparametric Unfolding Models*, (Post & Snijders, 1993)
		polytom
		Probabilistic Midpoint Unfolding Theory – PMUT, (Bossuyt & Roskam, 1989); *Nonparametric Unidimensional Unfolding for Multicategory Data*, (van Schuur, 1992)
parametrisch		probabilistisch dichotom
	Stochastic Folding Model – SFM, (Bechtel, 1968); *Probabilistic Unfolding Model* – PUM, (Sixtl, 1973); *Squared Simple Logistic Model* – SSLM, (Andrich, 1988, 1989); *PARELLA Model*, (Hoijtink, 1990); *Hyperbel Cosinus Model* – HCM, (Andrich, 1995; Andrich & Luo, 1993; Verhelst & Verstralen, 1993); *Signed OPLM* (Klinkenberg, 2001)	
		probabilistisch polytom
	Generalized Hyperbel Cosinus Model – GHCM, (Andrich, 1996); *Graded Unfolding Model* – GUM, Roberts und Laughlin (1996); *Generalized Graded Unfolding Model* – GGUM, Roberts et al. (2000)	

Anmerkungen: In dieser tabellarischen Übersicht werden die englischsprachigen Bezeichnungen für die Modelle verwendet.

Einstellung nach *psychisch nahe* sind. Das Item mit der höchsten Zustimmungswahrscheinlichkeit repräsentiert daher den Punkt auf dem (latenten) Merkmalskontinuum, welcher die Merkmalsausprägung einer Person am besten, oder auch *ideal* beschreibt, weshalb solche psychometrischen Modelle auch als *Idealpunktmodelle* bezeichnet werden (vgl. Brady, 1985, 1989, 1990; Gediga, 1998, sowie Abschnitt 1.4).

4.4 Überprüfung der Passung von psychometrischen Antwortmodellen

Wie bereits in der Einleitung zu diesem Kapitel kurz dargelegt, zeichnen sich Modelle im allgemeinen und psychometrische Antwortmodelle für Fragebogendaten im speziellen dadurch aus, dass sie eine Vorstellung bzw. eine Erklärung (ein Modell) darüber entwickeln, wie die empirischen Daten zustande kommen. Im Rahmen der Testung dieser Erklärungen muss daher überprüft werden, in welchem Ausmaß die einzelnen Modelle mit der empirischen Realität (den Daten) in Einklang stehen. Das grundlegende Prinzip bei solch einer Überprüfung anhand eines parametrischen Modells besteht darin, nachzuweisen, dass sich die empirisch vorgefundenen Antworten der Personen auf eine Reihe von Items mehr oder weniger vollständig, ausschließlich durch die im Modell eingeführten Parameter (Item- und Personenparameter) erklären lassen. Diesem Prinzip einer parametrischen Modelltestung, z. B. des probabilistischen Rasch-Modells, geht, wie der Name bereits andeutet, die Schätzung oder Bestimmung der Modellparameter voraus (Rost, 2004), wobei hierzu unterschiedliche Verfahren eingesetzt werden können, wie sie in Abschnitt 4.5 noch dargestellt werden. Das Prinzip der Überprüfung der empirischen Daten mit bestimmten Modellannahmen im Rahmen der Modelltestung kann letztlich auch für deterministische Modelle, wie beispielsweise für das oben dargestellte Guttman-Modell angewendet werden. Dabei werden aus den deterministisch formulierten Modellen jeweils entsprechende Forderungen an die Struktur der empirischen Daten abgeleitet, welche im Rahmen einer axiomatischen Modelltestung überprüft werden (z. B. Adams & Messick, 1958; Karabatsos, 2001, 2006; Orth, 1989).

Der Vorteil einer probabilistischen Modellformulierung, wie zum Beispiel im Rasch-Modell, liegt allerdings darin, dass diese bis zu einem gewissen Umfang auch nicht modellkonforme Antworten zulässt, ohne gleich die Gültigkeit des Modells als Ganzes in Frage zu stellen. Die hier am Beispiel der beiden Modelle (Rasch und Guttman) für den Dominanz-Antwortprozess skizzierten Prinzipien der Modelltestung sind weitgehend universell und lassen sich analog auch für die entsprechenden Modelle für den Nähe–Distanz-Antwortprozess formulieren. Ebenfalls weitgehend unabhängig von der Beschaffenheit des jeweiligen psychometrischen Antwortmodells und dessen unterschiedlicher Annahmen lassen

sich dabei zunächst *globale Tests* zur Überprüfung der Modellgeltung und Tests zur Untersuchung von *lokalen Modellverletzungen* unterscheiden (z. B. Rost, 2004, S. 331). Bei Tests zur Untersuchung von lokalen Modellverletzungen besteht das vorrangige Ziel in der Identifikation von einzelnen Personen (z. B. M. V. Levine & Rubin, 1979) oder Items (z. B. Gulliksen, 1950; Wright & Panchapakesan, 1969, Kap. 21) – oder auch beides gleichzeitig – innerhalb eines Datensatzes, welche den jeweiligen Modellannahmen widersprechen. Globale Tests zur Modellpassung beziehen sich dagegen auf die Frage nach der Passung des psychometrischen Antwortmodells auf die gesamte Datenmatrix.

4.4.1 Globale Maße zur Modellpassung

Zur Feststellung der Angemessenheit von Modellen zur Beschreibung empirischer Phänomene sind also Maße gesucht, welche die Abweichung der durch das jeweilige Modell vorhergesagten Antworten auf die Fragen eines Fragebogeninventars und den tatsächlich vorliegenden Antworten in den empirischen Daten quantifizieren. Dabei können innerhalb der globalen Modelltestung zwei Prinzipien unterschieden werden. Einerseits die Bewertung der *absoluten Modellpassung* und andererseits die Bewertung einer *relativen Modellpassung* (Bühner, 2006). Indizes zur Beurteilung der absoluten Modellpassung (bezogen auf eine Stichprobe) im Rahmen der IRT beziehen sich entweder auf die Summe der Abweichungen der empirischen von den erwarteten Antworten der Personen unter der jeweiligen Modellspezifikation (z. B. Baghaei, Yanagida & Heene, 2017; Christensen, Makransky & Horton, 2017; Cressie & Read, 1984; Heine et al., 2016; Ranger & Kuhn, 2012; Wright & Panchapakesan, 1969; Yen, 1984) oder aber auf den Vergleich der potentiell unterschiedlichen Modellpassung, bezogen auf unterschiedlichen Teilstichproben aus der Gesamtstichprobe von Personen oder Items (z. B. Andersen, 1973b; Fischer & Scheiblechner, 1970a; Martin-Löf, 1973; van den Wollenberg, 1982). Einen Überblick und eine Evaluation der unterschiedlichen Ansätze und Prinzipien der globalen Modelltestung geben beispielsweise Glas und Verhelst (1995), Gustafsson (1980b), Suárez-Falcón und Glas (2003), Thissen (2013) und Maydeu-Olivares (2013, 2015). Das allgemeine Prinzip bei der Bewertung der absoluten Passung eines Modells besteht darin, zu evaluieren, ob das Modell die beobachteten Daten generiert haben könnte (Maydeu-Olivares, 2015). Wie Maydeu-Olivares (2015)

weiter anmerkt, sind allerdings im Anwendungsbereich der IRT die Freiheitsgrade bei der Modelltestung meistens so groß, dass von keinem Modell erwartet werden kann, dass eine exakte Passung zu den Daten besteht. So umfasst ein eindimensionales IRT-Modell für zehn Items mit jeweils fünf Antwortkategorien wie beispielsweise für eine einzelne Skala des AIST–R (Bergmann & Eder, 2005) bereits $m^k = 5^{10} = 9765625$ Antwortmuster mit einer entsprechen hohen Anzahl an Freiheitsgraden. Bei solchen Modellen ist daher die vergleichende Bewertung der relativen Modellpassung oft ein angemessener Weg (z. B. Henson, Reise & Kim, 2007; Maydeu-Olivares, 2015).

Eine mögliche Grundlage zur Beurteilung der *relativen Modellpassung* kann in der unterschiedlichen „Plausibilität“ der empirisch beobachteten Daten unter der Annahme der verschiedenen, zu vergleichenden Modelle bestehen (vgl. auch Geisser, 1992). Zur Bestimmung dieser „Plausibilität“ der Daten wird auf das Prinzip der *Kullback-Leibler Information* zurückgegriffen (Kullback & Leibler, 1951). Allgemein formuliert kann die die Kullback-Leibler Information (K-L) als Maß für die Differenz zwischen der empirischen Realität und einem, diese Realität beschreibenden, Modell angesehen werden (z. B. Burnham & Anderson, 2001) und hat Bezüge zu dem von Boltzmann (1877) eingeführten Konzept der *Entropie*. Die K-L Information quantifiziert die von R. A. Fisher als zentrale Zielsetzung der Statistik propagierte Reduzierung der empirischen Daten auf *relevante Information*, welche sich in *suffizienten Statistiken* (Parameter eines Modells) zur Beschreibung der Daten ausdrückt. Fisher führte als Begriff für die „Plausibilität“ dieser Statistiken (gegeben die Daten und ein Modell zu deren Beschreibung) den Begriff der *Likelihood* (L) ein (Fisher, 1922, 1925; Geisser, 1992). Die K-L Information quantifiziert nun dieses von Fisher propagierte Konzept als „Differenz“ zwischen der (beobachteten) Realität f und einem diese Realität approximierenden Modell g (vgl. Gleichung 4.25).

$$I\left(f,g\right) = \int f\left(x\right) \log\left(\frac{f\left(x\right)}{g\left(x|\Theta\right)}\right) dx \tag{4.25}$$

Die K-L Information $I(f,g)$ definiert den Informationsverlust, der entsteht, wenn die Realität f durch das Modell g beschrieben wird (vgl. auch Burnham & Anderson, 2004). Formal ist dieser Informationsverlust $I(f,g)$ dabei als Integral über die Differenz (bzw. den logarithmierten Quotienten) zwischen f und g (mit den Parametern Θ) durch die Gleichung 4.25 definiert, wobei f und g

als kontinuierliche n–dimensionale Wahrscheinlichkeitsverteilungen angesehen werden.

Bei der Modellierung der vorgefundenen empirischen Realität wird nun jenes Modell gesucht, bei dem der Informationsverlust $I(f, g)$ am geringsten ausfällt. Bei diesem theoretischen Ansatz ist zunächst in Bezug auf die Praxis zu beachten, dass die Realität f und auch die Parameter Θ des Modells g unbekannt sind. Zweitens wird durch diese Darstellung schnell ersichtlich, dass die Abweichung zwischen der empirischen Realität und einem diese beschreibenden Modells desto geringer ausfällt, je mehr Parameter das betreffende Modell zur Beschreibung einsetzt. Gesucht ist also eine sinnvolle Balancierung bzw. Gewichtung zwischen der Anzahl der eingesetzten Parameter in einem Modell und der dabei erzielten Genauigkeit in der Beschreibung der empirischen Realität. Das einer solchen Gewichtung zugrunde liegende Prinzip basiert letztlich auf einer bereits von Wilhelm von Ockham (1288–1347) formulierten Forderung nach sparsamen Modellen zur Erklärung beobachteter Phänomene: „*Pluralitas non est ponenda sine necessitate*“ [*Die Mehrzahl sollte nicht ohne Notwendigkeit behauptet werden*] (Wilhelm von Ockham, 1967, S. 74; Liber I., Prologus, Quaestio I) – sinngemäß bezogen auf die Modellierung von Daten aus Fragebogenverfahren also, dass die Modellkomplexität zur Erklärung nur dann erhöht werden sollte, wenn sich dabei eine substanzielle Verbesserung ergibt. Dieses Prinzip korrespondiert mit einem der von Stachowiak (1973) definierten drei Hauptmerkmalen [3] des allgemeinen Modellbegriffs - dem *Verkürzungsmerkmal* (Stachowiak, 1973, S. 131-133), wonach Modelle immer eine auf das wesentliche eines Phänomes reduzierte Beschreibung der empirischen Realität darstellen (vgl. auch Abschnitt 4.1).

Eine solche Gewichtung nehmen sogenannte *informationstheoretische Kriterien* vor, wie beispielsweise das **A**kaike **I**nformation theoretic **C**riterion (AIC – Akaike, 1974) und das **B**ayes **I**nformation theoretic **C**riterion (BIC – Schwarz, 1978).

Akaike (1973, 1974) stellte dabei eine formale Beziehung zwischen der K-L Information und dem zur Schätzung der Modellparameter einsetzbaren Like-

[3] In seiner *Allgemeinen Modelltheorie* definiert Stachowiak (1973) drei Hauptmerkmale des allgemeinen Modellbegriffs: Das *Abbildungsmerkmal*, das *Verkürzungsmerkmal* und das *pragmatische Merkmal*, (vgl. Abschnitt 4.1).

lihood Prinzip von Fisher (1925) her und konnte damit ein Maß für den *relativen* Modellvergleich schaffen. Akaikes informationstheoretisches Kriterium (AIC) verbindet so die Schätzung der Modellparameter nach dem *maximum Likelihood* oder auch *kleinste Quadrate* Prinzip (vgl. Abschnitt 4.5) mit dem Vergleich der relativen Passung unterschiedlicher Modelle. Der AIC definiert sich nach Akaike (1974) gemäß Gleichung 4.26

$$AIC = -2\log\left(L\left(\Theta|daten\right)\right) + 2K \tag{4.26}$$

Nach Gleichung 4.26 setzt sich der AIC dabei aus zwei Termen zusammen, die addiert werden und damit ein Maß für die relative Modellpassung liefern. Der Term links vom Zeichen für die Addition ist die mit -2 multiplizierte logarithmierte Likelihood der Daten nach dem Prinzip der Kullback-Leibler-Divergenz. Diese wird ergänzt um den Strafterm $2K$ für die Modellkomplexität, wobei K für die Anzahl der geschätzten Modellparameter steht. Als Erweiterung schlägt Sugiura (1978) einen Korrekturfaktor für den AIC vor, dessen Anwendung Hurvich und Tsai (1989) allgemein empfehlen, wenn der Quotient zwischen der Anzahl der zu schätzenden Modellparameter K und der Stichprobengröße n (also K/n) ansteigt (vgl. auch Anderson, Burnham & White, 1998). Auch der bereits erwähnte BIC (Schwarz, 1978) weist einen solchen Strafterm für die Modellkomplexität (die Anzahl der Parameter) auf (Kuha, 2004).

Der BIC ergibt sich nach Gleichung 4.27

$$BIC = -2\log\left(L\left(daten|\Theta\right)\right) + \left(\log n\right) \times K \tag{4.27}$$

Der BIC gewichtet dabei das Kriterium der Modellsparsamkeit in stärkerem Ausmaß als der AIC (Rost, 2004, S. 90). Während der AIC gewissermaßen eine „eins zu eins“ Gewichtung der Anzahl der Modellparameter und der logarithmierten Likelihood vornimmt (vgl. Gleichung 4.26), wird der Strafterm beim BIC mit einem variablen Wert bestimmt, der sich als die logarithmierte Stichprobengröße n definiert (vgl. Gleichung 4.27).

In einer vergleichenden Überblicksarbeit zu den beiden informationstheoretischen Kriterien AIC und BIC weisen Burnham und Anderson (2004) darauf hin, dass diese einer unterschiedlichen wissenschaftstheoretischen Fundierung entspringen. Während beim AIC dasjenige Modell favorisiert wird, dessen theoretische Häufigkeitsverteilung nach dem Prinzip der Kullback-Leibler-

Divergenz (Kullback & Leibler, 1951) den geringsten „Abstand“ zur (realen) Verteilung der empirischen Daten aufweist, wird beim BIC das Modell favorisiert, bei dem die *bedingte* Wahrscheinlichkeit der empirischen vorgefundenen Daten unter der Annahme eines bestimmten (zu testenden) Modells am wahrscheinlichsten wird. Diese beiden Ansätze lassen sich somit zwei unterschiedlichen erkenntnistheoretischen Grundhaltungen zuordnen. Während beim BIC die Annahme besteht, dass (unter den zu vergleichenden Modellen) tatsächlich ein „wahres“ Modell zur Beschreibung der Daten existiert, besteht beim AIC die Annahme, dass zur Erklärung der Realität im Prinzip unendlich komplexe Modelle erforderlich wären und daher mit dem Kriterium AIC lediglich das Modell, welches die beste Annäherung an diese Komplexität liefert, gefunden werden kann. Neben diesen beiden unterschiedlichen erkenntnistheoretischen Grundhaltungen, welche die Entscheidung zur Verwendung des einen oder anderen informationstheoretischen Kriteriums motivieren können, besteht die Gemeinsamkeit bei beiden informationstheoretischen Kriterien darin, dass bei deren Berechnung die Komplexität der zu schätzenden Modelle, in jeweils unterschiedlicher Gewichtung als „strafender“ Faktor mit einfließt (Kuha, 2004).

Die Bewertung der Güte der Modellpassung auf einen vorliegenden Datensatz kann mit Hilfe der oben beschriebenen Likelihood-basierten Kriterien nun vergleichend vorgenommen werden. Das grundlegende Prinzip dieser Strategie zur Modelltestung besteht darin, zunächst verschiedene Modellspezifikationen an die vorliegenden Daten anzupassen um diese anschließend zu vergleichen (z. B. Henson et al., 2007). Zur Überprüfung einer universellen Modellgeltung im Rasch-Modell oder PCM (im Sinne einer Personenhomogenität) wird beispielsweise der klassifizierende Ansatz der Latent-Class-Analysis (LCA – Formann, 1984; Lazarsfeld, 1950, 1959, vgl. auch Abschnitt 4.6.2) mit dem dimensionalen Skalierungsansatz des Rasch-Modells (vgl. Abschnitt 4.2.3) kombiniert (Rost, 1990). Das *mixed-Rasch-Modell* (Rost, 1990) klassifiziert und skaliert die Personen somit gleichzeitig (Rost, 2004). Zur Testung der Hypothese einer vorliegenden Personenhomogenität (eine latente Klasse) werden dann weitere Modellspezifikationen mit einer unterschiedlichen Anzahl latenter Personenklassen aufgestellt (vgl. auch Abschnitt 4.6.2). Innerhalb dieser latenten Klassen werden dann entweder, wie beim *HYBRID-Modell*, unterschiedliche

psychometrische Antwortmodelle (K. Yamamoto, 1989; K. Yamamoto & Everson, 1995) angenommen, oder aber die jeweils gleichen Modelle (*mixed-Rasch-Modell*), allerdings mit unterschiedlichen Werten für die Parameterschätzung (z. B. Rost, 1991; Rost et al., 1997). Über die hier beschriebenen informationstheoretischen Kriterien können die unterschiedlichen Modellspezifikationen dann verglichen werden.

Die Anwendung dieses Prinzips der Testung auf das Vorliegen einer eindimensional, nach dem Rasch-Modell zu skalierenden, homogenen Personenstichprobe über die vergleichende Beurteilung der Modellpassung, hat in der psychometrischen Praxis eine weite Verbreitung gefunden (z. B. E. J. Austin et al., 2006; Eid & Zickar, 2007; Herzberg, 2002; Hong & Min, 2007; Keller & Kempf, 1997; Preinerstorfer & Formann, 2012; Rost et al., 1997, 1999; Vittersø, Biswas-Diener & Diener, 2005; Wetzel, Böhnke, Carstensen, Ziegler & Ostendorf, 2013; Wetzel, Böhnke & Rose, 2016; Ziegler et al., 2015). Nach vergleichenden Untersuchungen von Preinerstorfer und Formann (2012) führt die Anwendung des BIC bei der Wahl des am besten passenden Modells bei mixed-Rasch-Modellen zu annähernd perfekten Ergebnissen und führt im Vergleich zur Anwendung des AIC zu zuverlässigeren Ergebnissen bei der Modellwahl. Nach Rost (2004, S. 329) kann als grobes Auswahlkriterium ferner gelten, dass der AIC bei einer kleineren Anzahl von Items mit großen Häufigkeiten der Antwortmuster und der BIC bei einer großen Anzahl von Items bei kleinen Häufigkeiten der Antwortmuster vorzuziehen ist. In den einzelnen Analysen der in dieser Arbeit berichteten Untersuchungen wird daher in der Regel der BIC zur Modellwahl herangezogen.

4.4.2 Lokale Maße zur Modellpassung - und Antwortmuster

Aus der allgemeinen Frage nach einer angemessenen Methodik zur Auswertung von Fragebogen und speziell der Frage nach einer gültigen Verrechnungsvorschrift für die einzelnen Fragen zu einem Index, wie sie bereits von Zubin (1934) als zweite Hautfehlerquelle bei der Indexbildung identifiziert wurde (vgl. auch Abschnitte 1.2 und 3.1), lässt sich die Frage nach dem Geltungsbereich der bei dieser Skalierung (implizit oder explizit) aufgestellten *Antwortmodelle* ab-

leiten. Die in den vorangegangenen Kapiteln vorgestellten unterschiedlichen Methoden der Skalierung und psychometrischen Modellbildung beziehen sich dabei auf die Interaktion von Personen ($v; v = 1...n$) und Items ($i; i = 1...k$). Das Ergebnis dieser Interaktion wird bei der Messung aufgezeichnet und resultiert dann in einer zweidimensionalen $n \times k$ Datenmatrix, welche die Wechselbeziehung von zwei Gruppen von Objekten abbildet – also einer *2-way, 2-mode* Datenstruktur (vgl. Carroll & Arabie, 1980; Jacoby, 1991; Young, 1984, zur Klassifikation von Datenstrukturen).

Auf Grundlage dieses Entstehungsprozesses von psychologischen Messdaten lässt sich die Frage nach dem Geltungsbereich eines bestimmten Antwortmodells daher im Prinzip aus zwei Perspektiven betrachten. Der eine Aspekt bezieht sich dabei auf die Personen und deren aus ihren Eigenschaften resultierendem Antwortverhalten. Der andere Aspekt bezieht sich dagegen auf die Beschaffenheit der Items und ihrer Antwortskala und auf die daraus resultierende Eigenschaft der gesamten psychometrischen Skala. Diese beiden Perspektiven korrespondieren mit einem variablen- oder personenorientierten Ansatz bei der Betrachtung von abweichenden Antwortmustern (Gaier & Lee, 1953; Gaier, Lee & McQuitty, 1953). Der personenorientierte Ansatz kann beispielsweise mit dem Einsatz von klassifizierenden Modellen und Methoden zur Analyse von Datenstrukturen verbunden werden (vgl. Abschnitt 4.6), wobei dabei im Sinne eines explorativen Vorgehens nicht unbedingt ein bestimmtes Antwortmodell angenommen werden muss. Der variablenorientierte Ansatz ist dagegen vor dem Hintergrund der Forderung eines universellen Geltungsbereiches eines psychometrischen Antwortmodells eher mit den Prinzipien der Testkonstruktion assoziiert (z. B. Bühner, 2011), wie sie bei der Skalierung und Indexbildung mit dem Ziel einer eindimensionalen Erfassung und Messung individueller Unterschiede, angewendet werden.

Die Forderung eines universellen Geltungsbereiches und diagnostischer Eindeutigkeit psychometrischer Skalen und Antwortmodelle ist sowohl aus der variablen- als auch personenorientierten Perspektive verbunden mit der Forderung nach *Eindimensionalität* einer psychometrischen Skala. Eindimensionalität bedeutet dabei, dass sich das Ergebnis der Interaktion von Personen und Items – für alle Personen und Items – auf eine einzige (latente) Eigenschaftsdimension als Personen- und Itemmerkmal zurückführen lässt. Die Erfassung

dieser Eigenschaftsdimension ist bezogen auf die Differenzierung der Personen in der Regel das Ziel der Messung mit Fragebogenverfahren. Im Hinblick auf die beiden Aspekte (Personen und Items) des Geltungsbereiches psychometrischer Messungen lassen sich dementsprechend unterschiedliche Gründe für das Fehlen von Eindimensionalität und damit der universellen Geltung eines Antwortmodells identifizieren. Am Beispiel der Erfassung von Persönlichkeitseigenschaften nennt Rost (2002) dementsprechend auch einerseits die Items und andererseits die Personen als mögliche Quellen fehlender Eindimensionalität einer psychometrischen Skala, wobei diese beiden Aspekte nochmals differenzierter betrachtet werden können.

Auf Seiten der Items identifiziert Rost (2002) dabei erstens die spezifische Form ihres Inhalts (vgl. Abschnitt 3.2.2) und zweitens (übergreifend), die Beschaffenheit der Antwortskala (vgl. Abschnitt 3.2.4) als mögliche Ursachen für das Fehlen von Eindimensionalität. Die typische Herangehensweise bei der Entwicklung von psychometrischen Skalen basiert nun auf der zunächst naheliegend erscheinenden Annahme, dass die Ursache für die Verletzung der Eindimensionalität und damit der Einschränkung des universellen Geltungsbereichs einer psychometrischen Skala in den Eigenschaften einzelner Items begründet ist. Die dahinterliegende Vorstellung besteht beispielsweise darin, dass die Inhalte einzelner Items dafür verantwortlich sind, dass sich die resultierende Skala als nicht eindimensional erweist (Horan et al., 2003; Porst, 2000; Schmitt & Stults, 1985; Schriesheim & Eisenbach, 1995; Wang et al., 2015). Im Hinblick auf das Ziel eines universellen Geltungsbereichs und damit speziell der Itemhomogenität einer Skala gilt es dementsprechend durch die Anwendung geeigneter Methoden und der Interpretation entsprechender statistischer Kennwerte diejenigen Items auszusortieren, welche zur Verletzung der Eindimensionalität einer Skala führen (Ferguson, 1942; Khalid & Glas, 2014; H. Lee & Geisinger, 2015; Spector, 1992). Ein solches Vorgehen erfolgt entweder im Rahmen der Klassischen Testtheorie (vgl. Abschnitt 1.3.2) mit faktorenanalytischen Verfahren (z. B. Ferrando, 2007) oder aber (sinnvollerweise) durch eine Skalierung der Items mit unterschiedlichen Modellen der Item-Response-Theory (z. B. Reise, 1990) – vgl. auch Abschnitte 4.2 und 4.3.

Auf Seiten der Personen identifiziert Rost (2002) als mögliche Ursachen für die Verletzung der Eindimensionalität zunächst deren differentielle Ausprä-

gung auf der eigentlich zu messenden Merkmalsdimension, deren unterschiedliche Einstellungen zu spezifischen Inhalten einzelner Items (vgl. Abschnitt 3.2.1) und andererseits, im Hinblick auf die Beschaffenheit der Antwortskala der Items, die möglicherweise unterschiedlich ausgeprägte Reaktionstendenz (vgl. Abschnitt 3.2.4) in Bezug auf bestimmte Antwortkategorien.

Auf der inhaltlichen Ebene hat die Identifikation von nach einem bestimmten Antwort- und Skalierungsmodell nicht passenden Antwortmustern einzelner Personen(-Gruppen) Verbindungen zu unterschiedlichen Konzepten der Differentiellen Psychologie. Dies ist der Aspekt der (Personen-)*Konsistenz* bei Persönlichkeitseigenschaften (Bem, 1983; Lanning, 1991; Tellegen, 1988), das Konzept der *Metaeigenschaft* (Britt, 1993; Britt & Shepperd, 1999; Dwight et al., 2002) und der Begriff der *traitedness* (Reise & Waller, 1993), welche auch im Rahmen der *Person–Situation* Debatte (Bem & Allen, 1974; Bem & Funder, 1978) diskutiert werden. Die *Metaeigenschaft* hat danach beispielsweise einen Einfluss auf die Skalierbarkeit der Personen und damit auf die Passung der jeweils gegebenen Antworten zu der Struktur des angenommenen Antwort- und Skalierungsmodells – vgl. auch Abschnitt 3.3 in Kapitel 3 *Theoretischer Hintergrund zu Antwortmustern* in dieser Arbeit. So propagieren beispielsweise Formann (2002) und Ponocny und Klauer (2002) die Identifikation *nicht-skalierbarer* Personen nach einem angenommenen Dominanz-Antwortmodell. Ferrando (2004) diskutiert das Konzept abweichender Antwortmuster zu personenbeschreibenden Merkmalen und Einstellungen im Hinblick auf die individuell unterschiedliche Variabilität solcher Merkmale und Konsistenz bei der Selbstbeschreibung – mit der Folge einer unterschiedlich ausfallenden Personen Reliabilität. Im Hinblick auf die resultierenden Daten bezieht sich das Konzept *nicht passender Personen* und deren idiosynkratische Antwortmuster zunächst auf sämtliche im Kapitel 3 *Theoretischer Hintergrund zu Antwortmustern* im Abschnitt 3.2 dargestellten personenbezogenen Antwortverzerrungen (vgl. auch Berg, 1957; Emons, 2009; Meijer, Niessen & Tendeiro, 2016).

Die Identifikation solcher lokalen Modellabweichung auf der Ebene der Antwortmuster einzelner Personen oder Personengruppen kann als ergänzend, komplementäre Strategie zu der oben skizzierten Itemselektionsstrategie angesehen werden (Andrich, 1985; Tarnai & Rost, 1990). Für die Analyse psychometrischer Skalen und deren Testung im Hinblick auf eine universelle Gül-

tigkeit einer bestimmten Verrechnungsvorschrift zur Indexbildung und Skalierung (vgl. 1.3), kann diese Strategie als ebenso legitim angesehen werden wie die klassische Strategie der Item-Selektion (Tarnai & Rost, 1990). Darüber hinaus kann die Identifizierung von abweichenden Antwortmustern über entsprechende Indizes diagnostische Relevanz besitzen, weil sie diejenigen Personen identifiziert, bei denen die Inspektion des individuellen Antwortmusters mehr Information liefert als der einfache Summenwert aus den kodierten Items (Rost, 2004; Tarnai & Rost, 1990).

Übergreifend können unterschiedliche Prinzipien bei der Entwicklung von lokalen Indizes für die Personenpassung unterschieden werden. Zwei dieser Ansätze basieren auf Modellen aus der IRT. Daneben bestehen weitere Analyseverfahren und Indizes, welche nicht auf Modelle aus der IRT aufbauen (vgl. Harnisch & Linn, 1981, für einen Überblick). Dabei wird beispielsweise auf dem Ausmaß der Übereinstimmung eines individuellen Antwortmusters mit den Lösungswahrscheinlichkeiten der Items, wie sie anhand einer Stichprobe ermittelt wurden, betrachtet (Donlon & Fischer, 1968), oder der Zusammenhang zwischen der Wahl falscher Antwortkategorien (M. V. Levine & Drasgow, 1983), oder allgemeiner gefasst, die Analyse der Korrespondenz spezifischer Antwortmuster und dem jeweils erzielten Test-(Summen)-Wert (Emons, 2009).

Im Rahmen der IRT lassen sich einerseits Fit-Indizes zusammenfassen, welche auf den Annahmen *parametrischer* IRT-Modelle basieren (Drasgow et al., 1985; Molenaar & Hoijtink, 1990, 1996; Wright & Masters, 1990). Andererseits lassen sich Indizes zusammenfassen, welche die Passung individueller Antwortmuster unter den Annahmen eines *nichtparametrischen* IRT-Modells und dessen axiomatischer Prüfung, operationalisieren (Emons, 2008; Meijer, Molenaar & Sijtsma, 1994; Meijer, Muijtjens & Vleuten, 1996; Sijtsma, 1986; Tendeiro & Meijer, 2014) In der psychometrischen Literatur wird die Identifikation der Passung einzelner Personen und deren Antwortmuster oft unter Begriffen wie *appropriateness measurement* (M. V. Levine & Drasgow, 1988; M. V. Levine & Rubin, 1979) und *person-fit analysis* (Meijer, 1996) diskutiert. Die entsprechend entwickelten Indizes werden unter Begriffen wie *caution index* [dt. etwa: Vorsicht-Index] (Tatsuoka, 1984), *appropriateness-indices* [dt. etwa: Passungs-Indizes] (M. Levine & Drasgow, 1979; M. V. Levine & Drasgow, 1982) oder *person-fit-Indizes* (Klauer, 1995; Meijer, 1996; Wright, 1977) diskutiert.

Derartige Indizes stellen insofern *lokale Maße* für die Modellpassung dar, als das es darum geht, einzelne Personen oder Personengruppen innerhalb einer Stichprobe zu identifizieren, für die das jeweils angenommen psychometrische Antwortmodell nicht angemessen gilt (Meijer & Sijtsma, 2001a, 2001b). Das Konzept der Identifikation von idiosynkratischen, das jeweilige Antwortmodell (lokal) strukturverletzende Antwortmuster, hat so allgemein eine starke Verbindung zur IRT. Übersichten zu dieser Thematik findet sich beispielsweise bei M. V. Levine und Drasgow (1982, 1988); Meijer et al. (2016); Meijer und Sijtsma (1995, 2001b).

Bei einer *person-fit Analysis* [dt. etwa: Personenpassung Analyse] im Rahmen der IRT wird die Passung von Antwortmustern einzelner Personen – der „person-fit“ – unter der Annahme eines bestimmten psychometrischen Antwortmodells untersucht. Dabei kann das Ausmaß der individuellen Passung als Abweichung des beobachteten Antwortmusters vom unter der Annahme eines bestimmten psychometrischen Antwortmodells erwarteten Antwortmuster operationalisiert werden (z. B. Haberman, Sinharay & Chon, 2013; van den Wittenboer, Hox & De Leeuw, 1997). Ein daraus folgender Ansatz zur Entwicklung von Indizes für lokale Modellabweichungen basiert daher auf der Analyse von Antwort-Residuen (Andersen, 1995; Haberman et al., 2013; M. Wu & Adams, 2013), welche sich aus der Differenz zwischen beobachteten und erwarteten Item-Antworten ergibt. Die Residuen liegen dabei zunächst in einer (Daten-)Matrix vor, welche die gleichen Ausmaße wie die Antwortmatrix aufweist. Zur Ableitung eines Index für eine lokale Modellpassung können die Residuen dann entweder über die Personen (spaltenweise) oder die Items (zeilenweise) aggregiert werden. Die Residuen können ferner standardisiert werden, indem das Residuum durch die Standardabweichung (über Personen oder Items) des beobachteten Scores dividiert wird (Linacre, 2002; Wright & Masters, 1982, 1990). Im Allgemeinen erfolgt die Aggregation (über Personen oder Items) in Form von quadrierten standardisierten Residuen, welche dann durch die Gesamtzahl der Items (für Indizes zur Personenpassung) oder durch die Gesamtzahl der Personen (für Indizes zur Item-Passung), dividiert werden, woraus eine mittlere quadratische Statistik resultiert. Diese Statistik wird als ungewichtetes mittleres Quadrat oder auch als *OUTFIT*-Statistik bezeichnet (T. G. Bond & Fox, 2015; Linacre, 2002; Wright & Masters, 1982). Bei einer

gewichteten Version dieser Statistik werden die quadrierten standardisierten Residuen mit der beobachteten Varianz der Antworten multipliziert und dann durch die Summe der Varianzen – entweder über Personen oder Items – dividiert. Die resultierende Fit-Statistik – für Personen oder Items, wird als informationsgewichtetes mittleres Quadrat oder auch als *INFIT*-Statistik bezeichnet (T. G. Bond & Fox, 2015; Linacre, 2002; Wright & Masters, 1982). Im Vergleich zu den *INFIT*-Statistiken, erweisen sich die *OUTFIT*-Statistiken eher gegenüber Ausreißern sensitiv (Linacre, 2002). Demgegenüber berücksichtigen die *INFIT*-Statistiken bei der Beurteilung von lokalen Modellabweichungen aufgrund deren Gewichtung eher Personen und Items, deren Merkmalsausprägung sich einander annähern. Für eine praktische Anwendung dieser beiden Varianten der Fit-Statistiken (in Bezug auf die Items), weisen T. G. Bond und Fox (2015) darauf hin, dass die *INFIT*-Statistiken eher berücksichtigt werden sollten. Das angeführte Argument besteht darin, dass die *INFIT*-Statistiken diejenigen Personen, deren Eigenschaftsausprägung etwa der des betreffenden Items nahekommt, stärker berücksichtigt werden und damit besser geeignet sind, die Passung eines Items zu bewerten (T. G. Bond & Fox, 2015). Nach T. G. Bond und Fox (2015) können nach den strengsten Kriterien Items mit mehrstufigen Likert-Skalen, deren (root-mean-square) Item-Fit-Statistiken innerhalb eines Bereiches von $FIT_{MSQ} \geq .8$ bis $FIT_{MSQ} \leq 1.2$ liegen, noch als akzeptabel gewertet werden (vgl. Tabelle 12.7 in T. G. Bond & Fox, 2015, S. 282).

Allgemein kann die Verwendung von residuumbasierten Fit-Statistiken, insbesondere in deren z-standardisierter Form zur Signifikanztestung aber auch kritisiert werden. So weisen (z-standardisierte) Fit-Statistiken, welche auf Residuen nach erfolgter Rasch-Skalierung basieren, letztlich eine Verteilung mit unbekannten Eigenschaften auf (Masters & Wright, 1997), wobei die (übliche) Annahme einer Normalverteilung (zur Signifikanztestung) zweifelhaft scheint (H. J. Rogers & Hattie, 1987). Auch Karabatsos (2000) weist, bezogen auf die Testung des probabilistisch formulierten Rasch-Modells darauf hin, dass die Identifikation lokaler Modellabweichungen auf der Ebene der Personen nicht so einfach und unkompliziert ist, wie es scheint. So kann die Verwendung eines einzelnen festen Schwellwertes (kritischer z-Wert) für eine Fit-Statistik bei unterschiedlichen Analysesituationen, bei denen sich die empirischen Ge-

gebenheiten im Hinblick auf die Stichprobe und die Testeigenschaften unterscheiden, entweder zu einer Überdetektion oder zu geringen Erkennung von abweichenden Antwortmustern führen (Karabatsos, 2000). In einer Simulations-Studie untersucht Karabatsos (2003) insgesamt 36 unterschiedliche Personen-Fit-Statistiken. Dabei zeigt sich, in Übereinstimmung mit der bereits von Karabatsos (2000) formulierten Forderung zu Anwendung von nichtparametrischen, Guttman-Fehler basierten Fit-Statistiken (z. B. Meijer, 1994), dass die beiden nichtparametrischen Statistiken H^T (Sijtsma, 1986; Sijtsma & Meijer, 1992) sowie die U^3-Statistik von van der Flier (1982) am besten abschneiden. Diese Befunde stehen im Einklang mit einer von Tendeiro, Meijer und Niessen (2016) formulierten Forderung nach einer Anwendung von einfachen nichtparametrischen Fit-Statistiken zur Entdeckung abweichender Antwortmuster. In diesem Sinne formulieren Niessen, Meijer und Tendeiro (2016) eine praxisorientierte Anleitung zur Identifikation von abweichenden Antwortmustern mit dem *R*-Paket `PerFit` (Tendeiro et al., 2016). Der in der vorliegenden Arbeit eingesetzte, und im folgenden Abschnitt vorgestellte *Personen-Q-Index* (Tarnai & Rost, 1990) stützt sich einerseits auch auf die Prüfung einfacher Annahmen zum Dominanz-Antwortprozess in kumulativen Skalierungsmodellen und basiert andererseits auf dem parametrischen Prinzip des Rasch-Modells.

4.4.3 Der Personen-Q-Index und dessen polytome Verallgemeinerung

Der Q-Index wurde von Tarnai und Rost (1990) zur Identifikation abweichender Antwortmuster für das dichotome Rasch-Modell vorgestellt. Die Identifikation von abweichenden Antwortmustern kann einerseits als komplementäre Strategie zur klassischen Item-Selektion bei der Modellanpassung angesehen werden (Andrich, 1985) und weist andererseits diagnostische Relevanz auf (Tarnai & Rost, 1990). Im Gegensatz zu seinem symmetrischen Pendant – dem Item-Q-Index, welcher unter Bezug auf Tarnai und Rost (1990) von Rost und von Davier (1994) vorgestellt wurde, war für den Personen-Q-Index bislang keine rechnerische Implementierung in aktuellen Software-Paketen verfügbar. Während der Item-Q-Index im Rahmen einer rechnerischen Umsetzung in dem Programm *WinMira* (von Davier, 2001) für dichotome und po-

lytome Item-Antwortformate verfügbar ist, existierte darüber hinaus für den Personen-Q-Index von Tarnai und Rost (1990) auch keine polytome Generalisierung. Diese Lücken sollen in diesem Abschnitt hier in zweierlei Hinsicht geschlossen werden. Zum einen wird eine leicht verfügbare Softwareimplementierung des Personen-Q-Index in der aktuellen Version des *R*-Pakets `pairwise` (Heine, 2019) vorgestellt. Darüber hinaus wird der ursprünglich nur für das (dichotome) Rasch-Modell verfügbare Personen-Q-Index im Paket `pairwise` auch für polytome Item-Antwortformate erweitert. Ein Vorteil des Personen-Q-Index im Vergleich zu anderen Personen-Fit-Indizes (vgl. z. B. Meijer et al., 2016; Rupp, 2013; Tendeiro & Meijer, 2014, für eine Übersicht und praktische Anwendungen), besteht in dessen *quasi nichtparametrischer* Fundierung. Der Personen-Q-Index basiert auf der grundlegenden algebraischen Struktur des zugrunde gelegten (kumulativen) Dominanz-Antwortmodells (Rasch-Modell). Allerdings ist dabei eine Klassifikation der Antwortmuster möglich, ohne dass eine vorausgehende Bestimmung der Personenparameter notwendig ist. Die Beurteilung der Passung von Antwortmustern anhand des Personen-Q-Index stützt sich dabei vor allem auf die relativen Item-Kategorie-Schwierigkeiten im Sinne einer Item-übergreifenden Rangreihe. Verbunden wird dies mit einer axiomatischen Prüfung der kumulativen Skalierbarkeit der einzelnen Itemkategorien nach den Prinzipien des Guttman-Modells. Der Personen-Q-Index greift somit in gewisser Weise das von Meijer (1994) propagierte Prinzip des Zählens der Anzahl von Guttman-Fehlern als nützliche und einfache Alternative zu komplexeren Personen-Fit-Statistiken auf.

Für die Ableitung des Personen-Q-Index wird die bedingte Wahrscheinlichkeit des beobachteten Antwortmusters dabei formal ins Verhältnis zur bedingten Wahrscheinlichkeit des entsprechenden Guttman- und Anti-Guttman-Antwortmusters, bei gegebener gleicher Randsumme (Test-Summenwert), gesetzt. Der Personen-Q-Index lässt sich in Folge als vergleichsweise einfache Funktion der Itemparameter ableiten, ohne dass die Personenparameter θ_v benötigt werden. Rost (2004, S. 363) merkt dazu an, dass die Möglichkeit des Verzichts auf Personenparameter bei der Berechnung von Personen-Fit-Statistiken ein entscheidender Vorteil ist. Im Folgenden soll zunächst die formale Ableitung des Personen-Q-Index nach Tarnai und Rost (1990) dargestellt werden. Darauf folgend wird eine Erweiterung für polytome Antwortformate entwickelt

und vorgestellt. Die formale Ableitung des Personen-Q-Index beginnt mit der Darstellung der Wahrscheinlichkeiten einzelner Antwortpattern bei konstanter Randsumme (Summenscore). Zur Ableitung des Personen-Q-Index wird die algebraische Struktur des Rasch-Modells ausgenutzt (Tarnai & Rost, 1990). Angewendet wird dabei ein konditionales Prinzip. Dabei werden drei Gruppen von Antwortmustern (*pattern* – Zeilenvektoren der Datenmatrix) betrachtet, welche jeweils die gleiche Randsumme (Summenwert der Items) aufweisen. Durch die Kürzbarkeit der Personenparameter im Rasch-Modell lassen sich die bedingten Patternwahrscheinlichkeiten bei konstantem Summenscore r (z. B. $r = 3$) als Funktion der Itemparameter darstellen, ohne dass die Personenparameter θ_v benötigt werden, (vgl. Gleichungen 4.28 bis 4.30). Die Ableitung dieser bedingten Patternwahrscheinlichkeiten aus den unbedingten Patternwahrscheinlichkeiten und der Modellgleichung des Rasch-Modells ist bei Rost (2004, S. 126–127) ausführlich dargestellt und wird daher hier nicht wiederholt. Betrachtet werden die bedingten Wahrscheinlichkeiten von drei pattern, gegeben die Itemparameter σ_i und einem konstanten Summenscore r: Erstens die Wahrscheinlichkeiten des beobachteten pattern $P_{obs}\left(\underline{x}_v|r_v,\sigma_i\right)$, vgl. Gleichung 4.28, (z. B. „011001“), zweitens die Wahrscheinlichkeiten des Guttmanpattern $P_{Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)$, vgl. Gleichung 4.29, (z. B. „111000“) und drittens die Wahrscheinlichkeiten des Anti-Guttman-pattern $P_{Anti-Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)$, vgl. Gleichung 4.30, (z. B. „000111“)[4].

$$P_{obs}\left(\underline{x}_v|r_v,\sigma_i\right) = \frac{exp\left(-\sum_{i=1}^{k} x_{vi}\sigma_i\right)}{\gamma_r\left(\sigma\right)} \tag{4.28}$$

$$P_{Guttman}\left(\underline{x}_v|r_v,\sigma_i\right) = \frac{exp\left(-\sum_{i=1}^{r} \sigma_i\right)}{\gamma_r\left(\sigma\right)} \tag{4.29}$$

$$P_{Anti-Guttman}\left(\underline{x}_v|r_v,\sigma_i\right) = \frac{exp\left(-\sum_{i=k-r+1}^{k} \sigma_i\right)}{\gamma_r\left(\sigma\right)} \tag{4.30}$$

Der Ausdruck $\gamma_r\left(\sigma\right)$ in den Gleichungen 4.28 bis 4.30 stellt die jeweilige symmetrische Grundfunktion der Ordnung r aller Itemparameter σ dar. Die

[4]Bei dieser Darstellung der Antwortpattern und deren Benennung als Guttman- bzw. Anti-Guttman-pattern wird davon ausgegangen, dass die Items nach ihrer Schwierigkeit in aufsteigender Reihenfolge von links nach rechts geordnet sind.

formale Darstellung dieser Funktion nach Rost (2004, S. 127) ist in Gleichung 4.31 gegeben.

$$\gamma_r(\sigma) = \sum_{\underline{x}|r} \prod_{i=1}^{k} x_i exp(-\sigma_i) \tag{4.31}$$

Diese symmetrische Grundfunktion ist (für dichotome Items) eine Summe einzelner Produkte mit r Faktoren. Die einzelnen Summanden aus r Faktoren stellen alle möglichen Kombinationen der Exponenten der Itemparameter für jeweils einen Summenwert dar. Die *exakte* Berechnung der Koeffizienten der symmetrischen Grundfunktion höherer Ordnung r (bei Skalen mit vielen Items), insbesondere bei polytomen Antwortformaten, stellt, unabhängig von der Leistungsfähigkeit des eingesetzten Computersystems, ein nicht triviales Problem dar (z. B. Baker & Harwell, 1996; Formann, 1986; Gustafsson, 1980a). Dies resultiert aus der kombinatorischen Charakteristik der Funktion, welche dazu führt, dass die symmetrische Grundfunktion mit steigender Anzahl von Items und Antwortkategorien immer komplexer ausfällt (vgl. Rost, 2004, S. 214). Allerdings existieren Algorithmen zur näherungsweisen Bestimmung der symmetrischen Grundfunktion (vgl. Baker & Harwell, 1996, für eine Übersicht), die auf heutigen Computersystemen für eine übliche Anzahl von Items und Antwortkategorien mit in der Regel hinreichender Genauigkeit arbeiten. Dennoch kann es günstig sein die Berechnung der symmetrischen Grundfunktion durch eine geeignete Ableitung von darauf aufbauenden Koeffizienten (wie der Personen-Q-Index), zu umgehen.

Zur Ableitung des Personen-Q-Indexes werden die bedingten Wahrscheinlichkeiten der *beobachteten-* (vgl. Gleichung 4.28) und der *Anti-Guttman-* (vgl. Gleichung 4.30) pattern an den bedingten Wahrscheinlichkeiten der *Guttman*-pattern (vgl. Gleichung 4.29) standardisiert. Es resultieren die Gleichungen 4.32 und 4.33.

$$\frac{P_{obs}(\underline{x}_v|r_v,\sigma_i)}{P_{Guttman}(\underline{x}_v|r_v,\sigma_i)} = \frac{\frac{exp\left(-\sum_{i=1}^{k} x_{iv}\cdot\sigma_i\right)}{\gamma_r(\sigma)}}{\frac{exp\left(-\sum_{i=1}^{r}\sigma_i\right)}{\gamma_r(\sigma)}} \tag{4.32}$$

$$\frac{P_{Anti-Guttman}(\underline{x}_v|r_v,\sigma_i)}{P_{Guttman}(\underline{x}_v|r_v,\sigma_i)} = \frac{\frac{exp\left(-\sum_{i=k-r+1}^{k}\sigma_i\right)}{\gamma_r(\sigma)}}{\frac{exp\left(-\sum_{i=1}^{r}\sigma_i\right)}{\gamma_r(\sigma)}} \tag{4.33}$$

Die Doppelbrüche in den beiden Gleichungen 4.32 und 4.33 werden zunächst als Multiplikationen mit dem Kehrwert dargestellt (vgl. 4.34 und 4.37). In Folge können diese jeweils durch Kürzen der symmetrischen Grundfunktion (4.35 und 4.38) und Logarithmieren (4.36 und 4.39) zu jeweils zwei einfachen Differenzen von zwei Summen vereinfacht werden.

$$\frac{P_{obs}\left(\underline{x}_v|r_v,\sigma_i\right)}{P_{Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)} = \frac{exp\left(-\sum_{i=1}^{k} x_{iv}\cdot\sigma_i\right)}{\gamma_r\left(\sigma\right)} \times \frac{\gamma_r\left(\sigma\right)}{exp\left(-\sum_{i=1}^{r}\sigma_i\right)} \tag{4.34}$$

$$\frac{P_{obs}\left(\underline{x}_v|r_v,\sigma_i\right)}{P_{Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)} = \frac{exp\left(-\sum_{i=1}^{k} x_{iv}\cdot\sigma_i\right)}{exp\left(-\sum_{i=1}^{r}\sigma_i\right)} \tag{4.35}$$

$$\frac{P_{obs}\left(\underline{x}_v|r_v,\sigma_i\right)}{P_{Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)} = \sum_{i=1}^{k} x_{iv}\cdot\sigma_i - \sum_{i=1}^{r}\sigma_i \tag{4.36}$$

$$\frac{P_{Anti-Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)}{P_{Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)} = \frac{exp\left(-\sum_{i=k-r+1}^{k}\sigma_i\right)}{\gamma_r\left(\sigma\right)} \times \frac{\gamma_r\left(\sigma\right)}{exp\left(-\sum_{i=1}^{r}\sigma_i\right)} \tag{4.37}$$

$$\frac{P_{Anti-Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)}{P_{Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)} = \frac{exp\left(-\sum_{i=k-r+1}^{k}\sigma_i\right)}{exp\left(-\sum_{i=1}^{r}\sigma_i\right)} \tag{4.38}$$

$$\frac{P_{Anti-Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)}{P_{Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)} = \sum_{i=k-r+1}^{k}\sigma_i - \sum_{i=1}^{r}\sigma_i \tag{4.39}$$

Der Koeffizient Q des Personen-Q-Index ist definiert als Quotient der beiden standardisierten Patternwahrscheinlichkeiten gemäß Gleichung 4.40

$$Q = \frac{log\left(P_{obs}\left(\underline{x}_v|r_v,\sigma_i\right)\right) - log\left(P_{Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)\right)}{log\left(P_{Anti-Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)\right) - log\left(P_{Guttman}\left(\underline{x}_v|r_v,\sigma_i\right)\right)} \tag{4.40}$$

Durch Einsetzen von Gleichung 4.36 in den Zähler und 4.39 in den Nenner aus Gleichung 4.40, ergibt sich Gleichung 4.41,

$$Q|\underline{x}_v,\sigma = \frac{\sum_{i=1}^{k} x_{iv}\cdot\sigma_i - \sum_{i=1}^{r}\sigma_i}{\sum_{i=k-r+1}^{k}\sigma_i - \sum_{i=1}^{r}\sigma_i} \tag{4.41}$$

sodass sich der Personen-Q-Index bei dichotomen Antwortskalen für einen Antwortvektor $\underline{x}_v$ einer Person v als einfacher Quotient aus der Differenz zweier Summen ergibt, welche lediglich die Itemparameter enthalten. Durch die durchgeführten Schritte der Standardisierung variiert der Personen-Q-Index

inerhalb eines Bereiches von $Q = 0$ für ein perfekt passendes Guttman-pattern und $Q = 1$ für ein perfektes (unpassendes) Anti-Guttman-pattern.

Die Berechnung von Q stützt sich auf die, bei dichotomen Daten gegebene, eindeutige Definition des jeweiligen Guttman- und Anti-Guttman-patterns bei gleichem Summenwert. Für polytome Antwortformate lässt sich dagegen für einen konstanten Summenwert das jeweilige Guttman- und Anti-Guttman-pattern zunächst nicht eindeutig bestimmen. Diese Problematik ist beispielhaft für einen Summenwert von $r = 5$ bei 3 Items mit einer jeweils vierstufigen Antwortskala (mit der Kodierung $0, 1, 2, 3$) in Tabelle 4.4 dargestellt.

Tabelle 4.4 Beispiel zur Darstellung der Problematik einer eindeutigen Definition von Guttman- und Anti-Guttman-pattern für polytome Antwortformate.

pattern	$i = 1$	$i = 2$	$i = 3$	r
Beobachtet	**3**	**0**	**2**	**5**
Guttman (1?)	2	2	1	5
Guttman (2?)	3	1	1	5
Guttman (3?)	3	2	0	5
Anti-Guttman (1?)	1	2	2	5
Anti-Guttman (2?)	1	1	3	5
Anti-Guttman (3?)	0	2	3	5

Anmerkungen: 3 Items; jeweils vierstufige Antwortskala mit der Kodierung $0, 1, 2, 3$; Summenwert $r = 5$; Items nach Schwierigkeit aufsteigend sortiert (vlnr.).

Eine minimale Voraussetzung für ein Guttman-pattern besteht darin, dass bei aufsteigender Sortierung der Items nach deren Schwierigkeit, der numerische Wert der aufsteigend kodierten ordinalen Antwortkategorien bei einem leichteren Item stets größer oder gleich dem numerischen Wert der kodierten Antwortkategorie des jeweils nächst schwierigeren Items ist (vgl. Abschnitt 4.2.1). Basierend auf dieser für polytome Antwortskalen generalisierten axiomatischen Definition des Guttman-pattern, ergibt sich für mehrstufige Antwortskalen die Problematik einer nicht eindeutigen Identifikation eines Guttman-pattern für einen bestimmten, konstanten Summenwert. So zeigt die

Tabelle 4.4 für ein Beispiel mit drei Items ($i = 1$ bis $i = 3$) mit jeweils vierstufiger Antwortskala, dass für einen Summenwert von beispielsweise $r = 5$ jeweils drei pattern existieren, welche als Guttman- (Zeilen 2-4 in Tabelle 4.4) bzw. Anti-Guttman-pattern (Zeilen 5-7 in Tabelle 4.4) klassifiziert werden könnten.

Die Lösung dieser in Tabelle 4.4 dargestellten Problemantik besteht darin, dass bei polytomen Antwortformaten die Daten von der Code-Darstellung in die Reaktions-Darstellung überführt werden müssen (vgl. Zysno, 1993, sowie auch Abschnitt 4.2.1 in dieser Arbeit). In Abbildung 4.18 ist exemplarisch für ein Item i mit vier Antwortkategorien dargestellt, wie die gewählte Kategorie jeweils über drei dichotome Indikatorvariablen repräsentiert wird. Dieses Prinzip ist analog zu dem in Abschnitt 4.2.1 beschriebenen Prinzip der Generalisierung des Guttman-Modells für polytome Items zu sehen (vgl. auch Borg & Staufenbiel, 2007, S. 134).

	i_m	$i_{m0\|1}$	$i_{m1\|2}$	$i_{m2\|3}$
K_1	0	0	0	0
K_2	1	1	0	0
K_3	2	1	1	0
K_4	3	1	1	1

Abbildung 4.18 Beispiel für die Rekodierung eines polytomen Items i mit vier aufsteigend geordneten Kategorien (von $m = 0$ bis $m = 3$) in drei Indikatoren für die übersprungenen Kategoriegrenzen ($i_{m0|1}$ bis $i_{m2|3}$).

Kodiert wird dabei, welche Kategoriegrenzen durch die Wahl einer bestimmten Antwortkategorie „übersprungen“ wurden (z. B. für die dritte Kategorie $i_m = 2$ die Kategoriegrenzen $i_{m0|1}$ und $i_{m1|2}$). In den auf diese Weise rekodierten Daten repräsentieren die Indikatorvariablen die Kategoriegrenzen der einzelnen polytomen Items. Diese Kategoriegrenzen können dann wiederum nach deren psychometrischen Schwierigkeiten, wie sie sich im Rahmen der Bestimmung der Schwellenparameter im ordinalen Rasch-Modell (Masters, 1982) ergeben haben, aufsteigend sortiert werden (vgl. Beispiel in Abbildung 4.19).

Für eine weitere Illustration des Prinzips nehmen wir an, dass für das Beispiel mit drei Items bei vierstufiger Antwortskala (0, 1, 2, 3) die folgenden Schwellenparameter in Abbildung 4.19 ermittelt wurden.

	τ_1	τ_2	τ_3
$i = 1$	-3.00	-0.75	1.50
$i = 2$	-2.25	0.00	2.25
$i = 3$	-1.50	0.75	3.00

Abbildung 4.19 Beispiel für Schwellenparameter τ_1 bis τ_3 für drei Items $i = 1$ bis $i = 3$ mit jeweils vier Antwortkategorien.

In den rekodierten und geordneten Daten in der Reaktionsdarstellung ergibt sich dann für das pattern „302“ mit Summenwert $r = 5$ bei nach Schwierigkeit aufsteigender Sortierung (vlnr.) die in Tabelle 4.5 gegebene Darstellung:

Tabelle 4.5 Beispiel für nach Schwierigkeit aufsteigend sortierte und rekodierte Daten zur eindeutigen Definition von Guttman- und Anti-Guttmanpattern für polytome Antwortformate.

	Schwellenparameter	-3.00	-2.25	-1.50	-0.75	0.00	0.75	1.50	2.25	3.00	0
	pattern	$i = 1_{m0\|1}$	$i = 2_{m0\|1}$	$i = 3_{m0\|1}$	$i = 1_{m1\|2}$	$i = 2_{m1\|2}$	$i = 3_{m1\|2}$	$i = 1_{m2\|3}$	$i = 2_{m2\|3}$	$i = 3_{m2\|3}$	r
Beobachtet	’302’	1	0	1	1	0	1	1	0	0	5
Guttman	’221’	1	1	1	1	1	0	0	0	0	5
Anti-Guttman	’122’	0	0	0	0	1	1	1	1	1	5

Anmerkungen: 3 Items; jeweils vierstufige Antwortskala mit der Kodierung $0, 1, 2, 3$; Summenwert $r = 5$; Items nach Schwierigkeit aufsteigend sortiert (vlnr.).

Wie aus Tabelle 4.5 deutlich wird, lässt sich aus den auf diese Weise kodierten Daten für einen gegeben Summenwert (hier im Beispiel $r = 5$) jeweils eindeutig ein Guttman- und Anti-Guttman-pattern ableiten. Der entscheidende Punkt bei diesem Vorgehen der Datenrekodierung besteht darin, dass die einzelnen Antwortkategorien, definiert über deren Kategoriegrenzen, über die Sortierung nach den Schwellenparametern gewissermaßen aus ihrem „Itemverbund“ herausgelöst werden. Dieses Prinzip erlaubt dann die eindeutige Definition eines Guttman- und Anti-Guttman-pattern.

Der hier anhand eines Datenbeispiels dargestellte Personen-Q-Index mit seiner hier entwickelten Generalisierung für polytome, mehrstufig ordinale Ant-

wortskalen, ist in dem *R*-Paket `pairwise` (Heine, 2019) implementiert. Der Person-Q-Index wird in dieser Form für die im empirischen Teil der vorliegenden Arbeit durchgeführten Analysen eingesetzt (vgl. Abschnitt 7.1). Identifiziert werden dabei diejenigen Personen, deren Antwortmuster zu einem kumulativen Dominanz-Antwortmodell, dem *Partial Credit Model* (PCM – Masters, 1982), passen.

4.5 Methoden zur Bestimmung der Modellparameter

Im Zusammenhang mit der Entwicklung der unterschiedlichen Modelle in der Item Response Theory (IRT) wurden verschiedene Verfahren zur Bestimmung der jeweiligen Modellparameter entwickelt und vorgeschlagen. Einige dieser Methoden sind dabei spezifisch auf die jeweilige Modellformulierung angepasst und andere lassen sich, ausgehend von der jeweiligen Modellgleichung, eher nach einem universellen Prinzip anwenden. Die existierenden Verfahren lassen sich nach Linacre (1999) zunächst in *iterative* und *nichtiterative* Verfahren klassifizieren. Die iterativen Verfahren stützen sich dabei auf das Maximum Likelihood Prinzip (ML). Das gemeinsame Prinzip dieser ML-basierten Methoden ist, dass sie die Modellparameter als Randsummen der empirischen Daten *schätzen*, indem sie deren Wahrscheinlichkeit in einem iterativen Prozess maximieren – in der Regel nach dem Newton-Raphson-Algorithmus (Linacre, 2004). Das in der vorliegenden Arbeit für Antwortmodelle nach einem Dominanz-Antwortprozess ergänzend verwendete Verfahren der Parameter*bestimmung* nach dem *PAIR*-Algorithmus (vgl. Anhang A und B für eine Ableitung und detaillierte Darstellung), kann in die zweite (nichtinterative) Klasse von Techniken zur Parameterbestimmung eingeordnet werden. Da die iterativen, ML-basierten Verfahren auf einer recht verbreiteten und universellen Methodik basieren, welche nach ihrem grundlegenden Prinzip sowohl für Modelle für einen Dominanz- als auch einen Nähe–Distanz-Antwortprozess eingesetzt werden kann, wird deren Prinzip daher im folgenden Abschnitt zuerst dargestellt.

4.5.1 Iterative, Likelihood-basierte Schätzverfahren

Iterative Schätzmethoden gehen zunächst von prinzipiell beliebigen Startwerten für die zu schätzenden Parameter aus, z. B. dem Wert null, welcher als erste Schätzung für alle Modellparameter eingesetzt wird. Dies Schätzung wird nun verwendet, um (gemäß einer entsprechend gewählten Modellgleichung) erwartete Werte für alle Datenpunkte zu bestimmen. Die so bestimmten erwarteten Werte werden mit den empirischen Werten in der Datenmatrix verglichen, wor-

aufhin basierend auf den jeweils beobachteten Diskrepanzen bessere Schätzungen für die Modellparameter erzeugt werden. Diese beiden Vorgänge werden wiederholt, also iteriert, bis die Diskrepanzen zwischen den erwarteten und beobachteten Werten als klein genug angesehen werden. Die bis dahin geschätzten Werte werden dann als beste Schätzer der Modellparameter angesehen. Die Schätzmethode nach diesem Prinzip basiert somit auf der Maximierung der Plausibilität bzw. Wahrscheinlichkeit der gesamten empirischen Datenmatrix. Für diese Wahrscheinlichkeit der gesamten empirischen Daten (gegeben die Modelparameter) wird der Begriff *Likelihood* verwendet (vgl. Abschnitt 4.4.1), wodurch sich die Bezeichnung *Maximum Likelihood Methode* ableitet.

Nach Fisher (1922) ist die Likelihood L eines Datensatzes definiert als das Produkt der Wahrscheinlichkeiten der einzelnen Datenpunkte, gegeben ein Modell g mit Parametern Θ (vgl. Gleichung 4.42).

$$L = \prod_{v=1}^{n} \prod_{i=1}^{k} p\left(X_{vi} = x_{vi} | g\left(x_{iv} | \Theta\right)\right) \tag{4.42}$$

Das Prinzip Likelihood-basierter Schätzverfahren zur Bestimmung der Modellparameter besteht nun darin, diejenigen Werten für die Parameter Θ eines Modell g zur Erklärung der empirischen Datenpunkte x_{iv} zu bestimmen, welche deren Likelihood (Wahrscheinlichkeit) maximiert (Fisher, 1922, S. 323). Wie zum Beispiel von M. S. Johnson (2007) zusammengefasst, lassen sich innerhalb der Likelihood-basierten Parameterschätzmethoden drei Verfahren unterscheiden, welche im Rahmen der IRT in den Sozialwissenschaften häufig angewendet werden. Dies sind die verbundene Maximum Likelihood (*Joint Maximum Likelihood* – JML), die bedingte Maximum Likelihood (*Conditional Maximum Likelihood* – CML) sowie die Randwert Maximum Likelihood (*Marginal Maximum Likelihood* – MML) Methode (vgl. auch Molenaar, 1995, mit Schwerpunktsetzung auf die Itemparameterschätzung). Bei diesen Likelihood-basierten Methoden wird zunächst von der jeweiligen Schätzgleichung ausgegangen, welche sich als ersten Ableitung der Modellgleichung, jeweils nach einem der beiden Parameter (Personen- und Itemparameter), darstellen lässt. Daraus resultiert (am Beispiel des Rasch-Modells) ein Gleichungssystem mit mehreren unbekannten Größen (vgl. z. B. Rost, 2004, S.301), welches ohne weitere Annahmen zunächst nicht explizit gelöst werden kann und daher einen iterativen Schätzprozess nach sich zieht – in der Regel einen Algorithmus vom

Typ Newton–Raphson (Linacre, 2004b), sowie Fischer (1974, S. 255-257) für eine genaue Darstellung des Algorithmus. Das Prinzip besteht darin, die durch die Modellformulierung definierten Randsummen (die Modellparameter) der Datenmatrix in einem iterativen Prozess so zu bestimmen, dass die Wahrscheinlichkeit für die vorliegenden Daten unter der Bedingung der Modellformulierung und deren (geschätzten) Parametern maximiert wird (Bortz & Schuster, 2010, S. 92). Dieser Schätzalgorithmus besteht in der Regel pro Iteration aus zwei Schritten. Dem Schritt der *Parameterschätzung* [*Estimation*] und dem Schritt der *Maximierung* [*Maximazation*] und wird als *EM–Algorithmus* bezeichnet (Dempster, Laird & Rubin, 1977). Die in jedem einzelnen Durchgang des iterativen Prozesses geschätzten Parameter werden dazu verwendet, die Antwortkategoriewahrscheinlichkeiten der empirischen Datenmatrix zu bestimmen und in *einer* Größe, der *Likelihood*, bezogen auf die empirischen Daten zusammenzufassen. Die *Likelihood* drückt damit in einer Zahl aus, wie gut sich die einzelnen Antworten von allen Personen in den Daten über die geschätzten Modellparameter (Randsummen) erklären lassen. Die am Ende des iterativen Prozesses erzielte *Likelihood* kann dann die Basis für unterschiedliche Prinzipen für die Testung der Modellgültigkeit darstellen (Hoyt, 1945, sowie Abschnitt 4.4.1).

Bei der *JML*-Methode werden die beiden Parametergruppen (Personen- und Itemparameter) gleichzeitig – daher auch Ausdruck „*joint*“ – geschätzt. Unter den freilich theoretischen, asymptotischen Bedingungen, dass sich die Anzahl der Personen in der Stichprobe dem Umfang der Population annähert und die Anzahl der Items sich ebenfalls der Anzahl aller möglichen Items zur Erfassung des jeweiligen Merkmals annähert, führt diese Methode zu konsistenten Parameterschätzungen (Haberman, 1977). Diese hier formulierten (theoretischen) Bedingungen für die Konsistenz der Parameterschätzungen implizieren gewissermaßen eine Symmetrie der analysierten Datenmatrix im Hinblick auf die Personen und Items. Wie aber aus der Formulierung dieser Konsistenzbedingungen leicht zu erkennen ist, kann diese Symmetrieannahme in der Praxis schwer aufrechterhalten werden. Während nämlich die Anzahl der Personen in einer Stichprobe noch vergleichsweise einfach zu erhöhen ist, kann die Anzahl der Items in der Praxis kaum ohne weiteres erhöht werden. Die Struktur der jeweiligen psychometrischen Skala ist in der Regel über eine vor der Erhebung

festgelegte Anzahl von Items bestimmt. In Bezug auf die Itemparameter wird daher auch von sogenannten *strukturellen Modellparametern* gesprochen. Die Anzahl der Personen dagegen kann auch im Laufe der Erhebung im Prinzip immer wieder erhöht werden, weswegen die Personenparameter auch als *inzidentelle Modellparameter* bezeichnet werden. Aus diesen Randbedingungen in der Praxis der empirischen Datenauswertung ergibt sich ein Widerspruch zu der bei der JML-Methode angenommenen Symmetrie der Daten. Diese Asymmetrie führt dann in der Praxis dazu, dass die Schätzungen der Personenparameter bei der JML-Methode stets ungenauer ausfallen als diejenigen der Itemparameter. Dieses Phänomen wird in der Literatur auch als *Incidental Parameter Problem* bezeichnet (vgl. Gustafsson, 1980a; Neyman & Scott, 1948).

Bei der *CML*-Methode besteht die Lösung dieser prinzipiellen Problematik bei der JML-Methode darin, die Schätzung der beiden Parametergruppen voneinander getrennt durchzuführen. Im RM und auch im PCM wird dabei eine vorteilhafte Eigenschaft der Modell- oder Schätzgleichung ausgenützt, nach der sich die Personenparameter unter der Bedingung des Summenwertes der Antwortscores aus den Gleichungen heraus kürzen lassen. Unter der Bedingung einer konstanten *Scoregruppe*, im dichotomen Fall also der Anzahl gelöster bzw. zugestimmter Items, lassen sich daher in einem ersten Schritt lediglich die Itemparameter schätzen (Andersen, 1972). Aus diesem Grund wird die Methode als *bedingte Maximum Likelihood Methode* (*Conditional Maximum Likelihood* – CML) bezeichnet. Auf der Basis der geschätzten Itemparameter können dann im Anschluss die Personenparameter für die einzelnen Scoregruppen, also diejenigen Personen, welche den gleichen Summenwert auf den vorgegebenen Items erhalten haben, bestimmt werden. Die Möglichkeit des Herauskonditionierens der Personenparameter und damit die Möglichkeit der Schätzung von Itemparametern, welche so gewissermaßen unabhängig von den antwortenden Personen sind, stützen sich letztlich auf die Grundannahme der spezifischen Objektivität des RM (und PCM) (Fischer, 1974, S. 233). Neben der getrennten und damit konsistenten Schätzung der Item- und Personenparameter weist diese Methode noch einen weiteren Vorteil auf. So müssen aufgrund der im ersten Schritt heraus konditionierten Personenparameter auch keinerlei a priori Annahmen bezüglich der Verteilung des zu erfassenden Merkmals innerhalb der Personenstichprobe getroffen werden. Es lässt sich zeigen, dass

dieses Prinzip der Parameterschätzung in der Regel weitgehend unabhängig von der Merkmalsverteilung zu konsistenten Schätzern der Modellparameter führt (Andersen, 1973a; Rost, 2004).

Bei der *MML*-Methode wird, zumindest vom Prinzip her, ähnlich wie bei der CML-Methode vorgegangen. Auch hier werden die Itemparameter zunächst in einem separaten Schritt getrennt geschätzt. Im Gegensatz zur CML-Methode basiert die bedingte Schätzung jedoch nicht auf der Anzahl der korrekt gelösten Items, sondern auf der Annahme einer bestimmten Verteilung der Merkmalsausprägung der Personen in der Stichprobe (Thissen, 1982). Dies birgt den Vorteil, dass nicht für jede Rohwertgruppe (Scoregruppe) ein Parameter geschätzt werden muss, sondern lediglich die Verteilungsparameter der Personenparameter (Rost, 2004) Unter der Annahme einer normalverteilten Merkmalsausprägung der Personen gleichen sich die Parameterschätzungen der CML- und der MML-Methode (Thissen, 1982). Auf eine detaillierte Darstellung der MML-Methode wird hier verzichtet, da dieses Verfahren in der vorliegenden Arbeit nicht eingesetzt wird. Stattdessen sei lediglich auf die entsprechende Literatur verwiesen, wie z. B. Bock und Aitkin (1981); M. S. Johnson (2007); Thissen (1982), sowie Molenaar (1995).

Eine weitere, vielleicht praktischere Gemeinsamkeit dieser ML-basierten Schätzmethoden besteht darin, dass sie alle vergleichsweise große Stichprobengrößen benötigen – oder zumindest ernsthaft nur auf Basis größerer Datensätze angewendet werden sollten. Solche Datensätze mit ausreichenden Stichprobengrößen finden sich beispielsweise im Rahmen internationaler Bildungsvergleichsstudien wie PISA (OECD, 2019) und anderen. Unter solchen ausreichenden Stichprobengrößenbedingungen führen die ML-basierten Verfahren üblicherweise zu konsistenten Parameterschätzungen.

4.5.2 Schätzprobleme

Die beschriebenen Likelihood-basierten Schätzmethoden sind weit verbreitet. Einerseits, weil sie, bei genügend großen Stichproben in der Regel zu konsistenten und erwartungstreuen Parameterschätzungen führen, und andererseits, weil eine ganze Reihe von praktisch nutzbaren Software-Implementierungen existieren – so zum Beispiel *WinMira* (von Davier, 2001), *ConQuest* (M. L. Wu, Adams, Wilson & Haldane, 2007) als prominente Beispiele für eigenständige

Programme, sowie zahlreiche Pakete für die statistische Programmumgebung *R* (R Core Team, 2018) wie zum Beispiel für die CML-Schätzung das Paket `eRm` (Mair, Hatzinger, Maier & Rusch, 2018), für die MML-Schätzung das Paket `TAM` (Robitzsch, Kiefer & Wu, 2018) und für die JML-Schätzung das Paket `mixRasch` (Willse, 2014), um nur einige stellvertretend zu erwähnen.

Allerdings stellen die Likelihood-basierten Schätzmethoden auch spezifische Anforderungen an die zu analysierenden Daten. So weisen Kubinger (2005) sowie Kubinger und Draxler (2007b) im Zusammenhang mit der Frage nach der „Testbarkeit“ des Rasch-Modells darauf hin, dass (paradoxerweise) das komplette Fehlen von Antwortmustern, welche dem kumulativen Modellcharakter widersprechen (in diesem Falle würde ein perfektes Skalogramm vorliegen), die Testbarkeit der Modellgeltung erschweren. Kubinger und Draxler (2007b) führen dazu weiter aus, dass „*eine Verletzung der Eindeutigkeitsbedingung bei (bedingten) Maximum-Likelihood-Schätzungen*“ (Kubinger & Draxler, 2007b, S. 133) dazu führen kann, dass die Parameterschätzungen für das Rasch-Modell nicht möglich sind und damit die Testbarkeit, zumindest mit parametrischen Modelltests, nicht erfolgen kann. Vereinfacht dargestellt resultiert eine solche Situation einfach daraus, dass der probabilistische Modellcharakter des RM oder PCM beim Fehlen von dem Modell widersprechenden Antwortmustern verloren geht. Die zu testende Datenmatrix entspricht dann dem Guttman-Modell und impliziert so schon allein aus logischen Gründen die Ablehnung eines probabilistischen Modells.

Glas (1988) zeigt im Zusammenhang mit der CML-Methode zur Schätzung von Modellparametern des Rasch-Modells, dass diese bei adaptiven Testdesigns (Kubinger, 2017), die bei großen Schulleistungsstudien (z. B. PISA – OECD, 2019) eingesetzt werden, mit Verzerrungen bei den geschätzten Modellparametern zu rechnen ist. In diesem Sinne finden auch (Kubinger, Steinfeld, Reif & Yanagida, 2012) bei Einsatz der CML-Methode bei adaptiven Testszenarien mögliche Verzerrungen und empfehlen deren Ausmaß zuvor für das jeweils vorliegende Testdesign zu überprüfen. Adaptive Testdesigns sind im Hinblick auf die resultierende Datenmatrix dahingehend charakterisiert, dass daraus Datenmatrizen mit einem bezogen auf den gesamten Itempool hohen Anteil an fehlenden Werten resultieren. Die auch von Glas (1988) und Kubinger et al. (2012) gefundenen Verzerrungen bei der Likelihood-basierte Parameterschät-

zung lassen sich insofern auch im Hinblick auf die Unvollständigkeit der analysierten Datenmatrix interpretieren. Andrich und Luo (2003) weisen darauf hin, dass allgemein die Parameterschätzer bei Likelihood basierten Schätzmethoden direkt durch die empirischen Häufigkeiten von benachbarten Itemkategorien nachteilig beeinflusst werden können. In Analysedaten mit mehrstufigen Antwortformaten und einem gleichzeitig hohen Anteil fehlender Werte kann dies zu leicht zu Antwortkategoriehäufigkeiten mit dem Wert null führen. Luo und Andrich (2005) folgern, dass in solchen Situationen beim Vorliegen von Nullkategorien die Parameterschätzung der Kategorieschwellenwerte mit den üblicherweise verwendeten CML, MML und JML Methoden unschätzbar werden. Der Grund für diese Situation ist, dass diese Schätzalgorithmen im Prinzip die erwarteten Häufigkeiten jeder Antwortkategorie, auch der Nullkategorien, umfassen.

Neben solchen mehr oder weniger grundlegenden Problemen bei der Likelihood-basierten Schätzung der Modellparameter steigen Schätzprobleme auch mit der Anzahl der Modellparameter – also der Komplexität der Modelle. So steigt die Schwierigkeit der Parameterschätzung tendenziell mit der Zunahme der Anzahl der Modellparameter (z. B. G. Maris & Bechger, 2009; Puchhammer, 1988). Insbesondere die beiden Parameter der unteren und oberen Asymptote im 3– und 4–PL–Modell erweisen sich dabei als schwierig zu schätzen. So zeigt beispielsweise Puchhammer (1988) in einer Simulationsstudie, dass die Parameterschätzung über ML-Verfahren prinzipiell möglich ist, aber Probleme aufwerfen kann. So werden die Rate-Parameter (für die untere Asymptote) bei kleinen Stichproben (z. B. $n = 500$) nur relativ ungenau geschätzt und resultieren in systematischen Verzerrungen der Schätzung der Schwierigkeitsparameter der Items (Puchhammer, 1988). Wegen derartiger Schätzprobleme kann der Parameter für die untere Asymptote im 3–PL–Modell bei Mehrfachwahlaufgaben im Bereich der Leistungsmessung, z. B. über die Anzahl der vorgegebenen Antwortkategorien, a priori festgelegt werden, um so mögliche Schätzprobleme zu minimieren (Linacre, 2004a). Soll der „Rate-Parameter“ für die untere Asymptote dagegen frei geschätzt werden, so können sich in Abhängigkeit der Parametrisierung Probleme bei der Modell Identifikation ergeben, welche im Anschluss zu inhaltlich unterschiedlichen Interpretationen des „Rate-Parameters“ führen können. So zeigen G. Maris und Bechger (2009), dass unterschied-

liche Implementierungen des 3–PL–Modells auf der Basis desselben Datensatzes zu unterschiedlichen Schlussfolgerungen bezüglich des Rate-Verhaltens (Raten oder Nichtraten) bei der Beantwortung der Items führen. Auch bei der praktischen Anwendung parametrischer Modelle für einen Nähe–Distanz-Antwortprozess auf empirische Daten zeigt sich oft, dass die Bestimmung der Modellparameter im Rahmen einer Likelihood-basierten Modellschätzung schwierig ist (z. B. Carter & Zickar, 2011; de la Torre, 2006; M. S. Johnson & Junker, 2003). Bereits Heiser und Meulman (1983) stellen in Bezug auf die Parameterschätzung für Unfoldingmodelle fest, dass sich diese im Hinblick auf die Schätzbarkeit der Modellparameter als sehr sensitiv gegenüber schlecht passenden Daten [„*ill-conditioned data*“] (Heiser & Meulman, 1983, S. 139) erweisen. In gleicher Weise stellen A. Brown und Maydeu-Olivares (2010) fest, dass sich bei Unfoldingmodellen, die einen Nähe–Distanz-Antwortprozess abbilden, die Parameterschätzung mit Likelihood-basierten Methoden schwierig und die Genauigkeit der erzielten Parameterschätzer ungewiss ist; „*Estimation May Not Be as Accurate for Ideal Point Models*“ (A. Brown & Maydeu-Olivares, 2010, S. 490). Mit Bezug auf die ungenaue Schätzung der Parameter bei Unfoldingmodellen schlussfolgern A. Brown und Maydeu-Olivares (2010) weiter, dass solche parametrischen Modelle nutzlos werden, wenn die Parameterschätzung auch in kleinen Datensätzen nur schwer möglich ist: „*An IRT model is useless unless it can be shown that its item characteristic curves (ICC) can be estimated with enough precision in reasonably small samples*“ (A. Brown & Maydeu-Olivares, 2010, S. 490). Neben der parametrischen oder nichtparametrischen Formulierung von Unfoldingmodellen mit einer unimodalen ICC bestehen weitere Methoden und Verfahren zur Abbildung von Nähe–Distanz-Antwortprozessen. Diese weiteren Methoden und Verfahren, welche zur Identifikation der Nähe–Distanz-Relation in den zu analysierenden Daten angewendet werden können, wurden teilweise zunächst in einem nicht-psychometrischen Kontext entwickelt. Dieser Ansatz kann im Vergleich zu der eher modellorientierten Vorgehensweise wie sie bisher beschrieben wurde, als eher *datenorientiert* bezeichnet werden, und hat zunächst das Ziel, die beobachteten Daten auf geeignete Weise zu reorganisieren. Dieses Prinzip wird am Ende dieses Abschnittes (Abschnitt 4.5) im Rahmen der Beschreibung von Verfahren zur Schätzung von Modellparametern in Abschnitt 4.5.4 *Identifikation von Nähe–Distanz-Antwortprozessen als kombinatorisches Problem* beschreiben.

4.5.3 Itemparameterbestimmung durch Pairwise Limited–Information

Zur Bestimmung der Modellparameter (insbesondere der Itemparameter) existieren, neben den drei bisher angesprochenen und z. B. von Molenaar (1995) detailliert dargestellten Likelihood-basierten Schätzmethoden, noch weitere Verfahren. In einer Übersicht zu unterschiedlichen Methoden der Parameterbestimmung im Rasch-Modell unterschiedet Linacre (1999) zunächst grundsätzlich zwischen *iterativen* (Likelihood-basierten) Methoden und *nichtiterativen* Methoden. Für vollständige, dichotome Antwortdaten entwickelt beispielsweise L. Cohen (1979) ein Verfahren zur expliziten Bestimmung der Modellparameter und löst die von Wright und Panchapakesan (1969) aufgestellten Maximum–Likelihood-Gleichungen zur (eigentlich iterativen) Schätzung, approximativ. Dieses als *PROX* bezeichnete Verfahren stützt sich allerdings stark auf die Annahme der Normalverteilung von Item- und Personenparametern. Es kann aufgrund der rechnerischen Einfachheit sogar von Hand bzw. mit einem Taschenrechner durchgeführt werden (Wright & Masters, 1982). Ein weiteres Verfahren zur Bestimmung der Itemparameter stützt sich auf die (bedingten) paarweisen Vergleiche der Lösungs- bzw. Wahlhäufigkeiten der Itemkategorien. Dieser Ansatz wurde bereits von Georg Rasch in Form von theoretischen Überlegungen erwähnt (vgl. Rasch, 1960, S. 172), welche auf einen Austausch mit Gustav Leunbach zurückgehen (Leunbach, 1961; Rasch, 1966b). Ein Vorteil dieses Verfahrens liegt in seiner rechnerischen Einfachheit (Wright & Masters, 1982). Darüber hinaus ermöglicht es im Rahmen der IRT die stabile Bestimmung von Itemparametern für Datensätze, die einerseits einen geringen Umfang von antwortenden Personen oder andererseits einen vergleichsweise hohen Anteil von fehlenden Werten aufweisen (Heine & Tarnai, 2015), sowie Choppin (1983, S.11) und Wright und Masters (1982, S. 60). Aufgrund des Prinzips des paarweisen Vergleichs der Item-(Kategorie)-Häufigkeiten wird dieses Verfahren der Itemparameterbestimmung als *PAIR*-Algorithmus bezeichnet. Derartige [paarweisen] Vergleiche bilden nach einer Definition von Rasch (1966a) die fundamentale Grundlage jeglicher wissenschaftlichen Betätigung. Im Kern besteht das Prinzip darin, die Items jeweils paarweise hinsichtlich der empirisch beobachteten Häufigkeiten der gewählten Kategorien der Antwortskala zu

vergleichen. Dieses Verfahren zur Itemkalibrierung weist Parallelen zu der in Abschnitt 1.3.3 dargestellten *LCJ-Skalierung* und dem von Thurstone (1927a) formulierten *Law of comparative judgement (case five)* auf (vgl. Tabelle 1.1 in Abschnitt 1.3.3). Ferner besteht eine allgemeine Analogie zu Modellen zur Analyse von Entscheidungsverhalten (Luce, 1977a, 1977b; Rasch, 1966b). Wie von David (1988, S. 9) festgestellt, lässt sich diese Methode darüber hinaus auf das bereits von Fechner (1860a) propagierte Prinzip der Quantifizierung der Schwelle der Unterschiedsempfindlichkeit von zwei Reizen unterschiedlicher Intensität für eine menschliche Wahrnehmungsmodalität zurückverfolgen. Der *PAIR*-Algorithmus weist ferner Parallelen zu Methoden der fairen Bewertung von Schachwettkämpfen auf, wie sie bereits zu Beginn des 20. Jahrhunderts diskutiert wurden (z. B. Ahrens, 1901; Drobny, 1900, 1901; Tietz, 1900a, 1900b). Das Ziel bestand dabei darin, die Spielstärke der teilnehmenden Spieler eines Wettbewerbs über die Auswertung der Häufigkeiten ihrer Siege und Niederlagen (in den paarweisen Partien) zu evaluieren. Für eine detailliertere Darstellung der historischen Wurzeln des Verfahrens sowie der Bezüge zur Eigenschaft der spezifischen Objektivität des Rasch-Modells sei hier auf den Aufsatz von Heine und Tarnai (2015, S. 8–10) verwiesen. In Erweiterung dazu beschreibt Glickman (1995) die Bezüge des Verfahrens des paarweisen Vergleichs zu anderen aktuellen Rangordungsverfahren für paarweise Sportwettbewerbe wie z. B. Fußball, Basketball, Tennis oder Hockey. Glickman (2005) diskutiert darüber hinaus Prinzipen zur unvollständigen Auswahl der einzelnen Paarungen, die zum Vergleich herangezogen werden, vor dem Hintergrund der Problematik, dass mit einer zunehmenden Anzahl von zu vergleichenden Objekten ein vollständiger Paarvergleich schnell sehr umfangreich werden kann. Diese von Glickman (2005) diskutierten Aspekte weisen wiederum Bezüge zu unvollständigen Booklet- bzw. Multimatrixdesigns auf, wie sie im Bereich internationaler Schulleistungsstudien angewendet werden (vgl. z. B. Frey, Hartig & Rupp, 2009; Heine, Sälzer, Borchert, Siberns & Mang, 2013; OECD, 2014).

Ebenso wie bei der *CML*- und *MML*-Methode zur Itemparameterschätzung werden beim *PAIR*-Algorithmus die Itemparameter unabhängig von den Personenparametern bestimmt (vgl. Choppin, 1983). Eine detaillierte Ableitung des Algorithmus befindet sich im Anhang A dieser Arbeit, wobei gezeigt wird, dass sich dabei die Personenparameter aus den Gleichungen des Algorithmus

zur Schätzung der Itemparameter herauskürzen lassen. In diesem Sinne nutzt das Verfahren die vorteilhafte Eigenschaft der Separierbarkeit von Item- und Personenparametern des Rasch-Modells (Wright & Masters, 1982, S. 60). Im Unterschied zu den beiden Methoden *CML* und *MML* werden hier allerdings keine iterativen Schätzungen vorgenommen, sondern die Itemparameter explizit berechnet. In Zusammenhang mit der Skalierung von international vergleichenden Schulleistungsstudien wurde dieses Verfahren zur Itemkalibrierung erstmalig von Choppin (1968) vorgeschlagen (vgl. auch Fischer, 1970; Fischer & Scheiblechner, 1970b). Neben der nichtiterativen Methode entwickelte Choppin (1983) auch eine Likelihood-basierte, iterative Methode zur Itemkalibrierung auf der Basis paarweiser Itemvergleiche. Die nichtiterative Variante wendet demgegenüber nicht das Prinzip der Maximierung der Likelihood der Daten an. Bei den explizit berechneten Itemparametern handelt es sich um Parameterwerte, die auf das Prinzip der Methode der kleinsten-Quadrate-Schätzung [*least-square-estimation*] zurückzuführen sind (Choppin, 1982; Garner & Engelhard, 2000; Mosteller, 1951). Der *least-square-estimation* Ansatz verfolgt dabei das Ziel die Abweichungen (der empirischen Daten) vom Modell zu minimieren und wurde bereits von Rasch (1960) als Prinzip zur Bestimmung der Modellparameter vorgeschlagen (Choppin, 1982). Das Verfahren lässt sich ohne weiteres auch für mehrstufige Antwortformate verallgemeinern. Dem paarweisen Vergleich werden dabei dann nicht die Lösungshäufigkeiten der einzelnen Items, sondern die Häufigkeiten der Antwortkategorien des mehrstufigen Antwortformats der Items unterzogen (vgl. z. B. Garner & Engelhard Jr, 2002). Eine detaillierte formale Herleitung des *PAIR*-Algorithmus am Beispiel für dichotome Antwortformate ist in Anhang A gegeben. In Anhang B wird das rechnerische Prinzip der Itemparameterbestimmung anhand eines einfachen Beispiels für dichotome Antwortdaten dargestellt.

Ein Vorteil des *PAIR*-Algorithmus besteht darin, dass dieser recht elegant mit fehlenden Werten in der zu analysierenden Datenmatrix umgehen kann (Heine & Tarnai, 2015; Wright & Masters, 1982). Zur Bestimmung der Itemparameter wird im *PAIR*-Verfahren der (logarithmierte) Quotient der bedingten Itemkategoriewahrscheinlichkeiten herangezogen. Diese wiederum basieren auf den bedingten relativen Kategoriehäufigkeiten in einer Stichprobe, welche zur Kalibrierung der Items herangezogen wird. In Bezug auf etwaige existierende

fehlenden Werte in der zugrunde liegenden Datenmatrix werden nun nur diejenigen Datenpunkte zur Bestimmung der bedingten Itemkategoriewahrscheinlichkeiten herangezogen, für die eine gültige Antwort vorliegt. Nach diesem Prinzip ist es für die Bestimmung der Itemparameter unerheblich, um welche Personen es sich handelt, welche die relativen Kategoriehäufigkeiten produzieren (Choppin, 1983; Rasch, 1966b, 1977). Im Zusammenhang mit dem allgemeinen Prinzip eines paarweisen Vergleichs von Itemkategorien stellt Gifi (1991, S. 344-345) fest, dass dieses Prinzip eine interessante Verallgemeinerung erlaubt, die aufgrund des doppelten paarweisen Vergleichs[5] Antwortverzerrungen in Bezug auf die resultierende Invarianz der Itemrangreihe zulässt.

Der in der vorliegenden Arbeit eingesetzte *PAIR*-Algorithmus und die daraus resultierenden least-square Itemparameter können auch als so genannte *Limited-Information*-estimates (LI) angesehen werden (Bolt, 2005; Christoffersson, 1975; Forero & Maydeu-Olivares, 2009; Lance, Cornwell & Mulaik, 1988; Maydeu-Olivares, 2001, 2005; Maydeu-Olivares & Joe, 2005; McDonald & Mok, 1995). Im Gegensatz zur LI-Methodik stützen sich ML-basierte Verfahren oft auf das Prinzip der vollständige Informationsmaximierung [*Full Information Maximum Likelihood*] (FIML) im Rahmen deren Implementierung über den den Erwartungswert-Maximierungs-Algorithmus (EM – Bock & Aitkin, 1981; Bock, Gibbons & Muraki, 1988) zur Schätzung der Parameter unterschiedlicher Modelle anhand der Daten (z. B. Bolt, 2005; Cai, 2010; Cai & Hansen, 2013; Cai, Yang & Hansen, 2011; Forero & Maydeu-Olivares, 2009; Gibbons et al., 2007; Maydeu-Olivares & Joe, 2005). Wie Forero und Maydeu-Olivares (2009) feststellen, leitet sich dabei der Begriff der *vollständigen Information* aus dem Prinzip ab, bei der Schätzung der Modellparameter diejenige Information zu verwenden, welche auf der Berücksichtigung der vollständigen – also aller möglichen – Antwortmuster basiert.

Um Probleme bei der Parameterschätzung und Modelltestung vor dem Hintergrund von sparsamen Kontingenztabellen zu überwinden, haben Maydeu-Olivares und Joe (2005) die Verwendung der LI-Methodik zur Schätzung und

[5]Wie in den Anhängen A und B zu dieser Arbeit dargestellt, findet für jede Itemkategorie ein *doppelter* paarweiser Vergleich statt – die jeweilige Itemkategorie steht beim Vergleich einmal an erster und einmal an zweiter Stelle, was in der symmetrisch, reziproken Matrix jeweils den Einträgen oberhalb bzw. der unterhalb der Diagonalen entspricht (z. B. Nishisato, 1978a, 1978b).

Modellprüfung vorgeschlagen, die nur univariate und bivariate Informationen verwenden (vgl. auch Cai, Maydeu-Olivares, Coffman & Thissen, 2006; Joe & Maydeu-Olivares, 2010; Maydeu-Olivares, 2001, 2006; Maydeu-Olivares & Joe, 2006). Maydeu-Olivares und Joe (2005) zeigen, dass das Prinzip der *Limited-Information*-estimates, welche auf den bivariaten Assoziationen der Items basieren, in der Praxis im Vergleich zu den (theoretisch) *asymptotisch optimalen* ML-basierten Prozeduren eine hohe Effizienz aufweist – „*They show that bivariate information methods have high efficiency relative to asymptotically optimal procedures such as maximum likelihood*“ (Maydeu-Olivares, Hernández & McDonald, 2006, S. 452). In einer weiteren vergleichenden Untersuchung zur Schätzgenauigkeit der beiden Ansätze (Limited- vs. Full-Information), konnten Forero und Maydeu-Olivares (2009) zeigen, dass sich vergleichbare IRT-Modellparameter Schätzungen ergeben – insgesamt lieferte die LI-Methode dabei etwas genauere Parameterschätzungen und die FIML-Methode liefert etwas genauere Standardfehler (Forero & Maydeu-Olivares, 2009).

Der *PAIR*-Algorithmus und dessen Implementierung im *R*-Paket `pairwise` (Heine, 2019) ist insofern der LI-Methodik zuzuordnen, da daraus resultierende Itemparameter ebenfalls nur auf der Information der jeweils bivariaten, paarweisen Itemkategoriehäufigkeiten aufbauen (vgl. z. B. Millsap & Maydeu-Olivares, 2009). Der *PAIR*-Algorithmus wird im Anhang A dieser Arbeit aus den Grundgleichungen des Raschmodells abgeleitet und in Anhang B wird anhand eines einfachen Beispiels auf dessen rechnerische Implementierung eingegangen. Eine praktische Anwendung des *PAIR*-Algorithmus bei der Skalierung findet sich in den Arbeiten von Gebhardt, Heine, Zeuch und Förster (2015), Gebhardt, Heine und Sälzer (2015), Heine et al. (2018), Heine und Tarnai (2015) und Sälzer und Heine (2016). Der Beitrag von Heine et al. (2018) diskutiert zusätzlich die allgemeine Verbindung des *PAIR*-Algorithmus mit dem Prinzip LI-Methodik.

4.5.4 Identifikation von Nähe–Distanz-Antwortprozessen als kombinatorisches Problem

Betrachtet man die in Abbildung 4.12 in Abschnitt 4.3.1 als Beispiel gegebene grafische Veranschaulichung der aus dem Nähe–Distanz-Antwortprozess (Unfolding) resultierenden Daten, so lässt sich die Herausforderung bei der Skalierung von Datenmatrizen mit Unfolding- oder Idealpunktmodellen auch als kombinatorische Aufgabe der gleichzeitigen zeilen- und spaltenweisen Umsortierung interpretieren. Das Ziel bei einer solchen Permutation der Ordnung der Zeilen und Spalten in den Daten besteht darin, die Items und auch die Personen mit ihren jeweils unterschiedlichen Merkmalsausprägungen in eine mit dem Prinzip der minimalen Distanz korrespondierende Rangreihe zu bringen. Werden die Antwortdaten (im dichotomen Falle) dahingehend kodiert, dass eine Eins („1“) die Zustimmung zu einer Frage repräsentiert und eine Null („0“) die Ablehnung der betreffenden Frage repräsentiert, sollten sich bei optimaler Reorganisation der Daten jeweils „rechts und links“ von den in der Diagonalen stehenden Einsen nur Nullen stehen. Im polytomen Falle ist analog zu diesem Prinzip die Reorganisation der Daten dahingehend zu optimieren, dass in der Diagonale der Datenmatrix jeweils die hohen Werte der kodierten Antwortkategorien stehen, welche zu den *beiden* Rändern hin abfallen. Dieses Prinzip der Datenreorganisation wurde auch von Guttman (1950) und später anderen Autoren (z. B. Cliff, Collins, Zatkin, Gallipeau & McCormick, 1988; Hoijtink, 1991; Leenen & Van Mechelen, 2004) innerhalb der psychometrischen Methodik, unter dem Begriff *Parallelogramm-Analyse* diskutiert. Das Prinzip Beobachtungsdaten bei der Auswertung und Skalierung durch geeignete Verfahren dahingehend so zu permutieren (sortieren), dass sich in der Diagonalen der sortierten Datenmatrix die hohen Werte (für starke Zustimmung) befinden, welche zu den Rändern hin abfallen, wird auch in anderen wissenschaftlichen Disziplinen verfolgt. Derartige Methoden der Umsortierung werden zum Beispiel im Bereich der Archäologie als *Seriation* oder als *Ordination* bezeichnet. Die beiden Begriffe sind Sammelbegriffe für multivariate Techniken, die einen potentiell mehrdimensionalen Raum von Datenpunkten so anpassen, dass bei einer Projektion auf einen zweidimensionalen Raum ein verborgenes Muster, das die Daten besitzen, bei visueller Betrachtung sichtbar wird (D. G. Kendall, 2004; Liiv, 2010).

Das bei der Ordination oder Seriation angewendete Prinzip hat Parallelen zu parametrischen und nichtparametrischen Modellen mit eingipflig, unimodaler Itemcharakteristik (vgl. Abschnitt 4.3), bei denen oberhalb und unterhalb (bzw. „rechts und links“) von einer bestimmten Merkmalsausprägung die Zustimmungswahrscheinlichkeiten zu einem Item ein Minimum erreichen. So beschreibt zum Beispiel Hubert (1974) die Parallelen zwischen der Datenauswertung nach dem Prinzip der *Seriation* in der Archäologie mit den Prinzipien der Auswertung von Unfoldingdaten in der Psychometrie. Während es in der Archäologie bei der Anwendung der Methode der Seriation darum geht, archäologische Fundstücke kultureller Artefakte mehr oder weniger eindeutig entlang eines zeitlichen Kontinuums anzuordnen, besteht die Aufgabe im Rahmen der psychometrischen Messung darin, Personen anhand ihrer Antworten auf eine Reihe von Items entlang eines Kontinuum einer postulierten, latenten Einstellungsdimension anzuordnen. Der entscheidende gemeinsame Aspekt der unterschiedlichen Anwendungen der Methode der Seriation (bzw. der Unfoldingmodelle) besteht darin, dass beispielsweise im Anwendungsfalle der psychometrischen Messung die Zustimmungswahrscheinlichkeit zu einem (dichotomen) Item und analog dazu die Auftretenshäufigkeit bestimmter Archäologischer Fundstücke (in einer bestimmten Epoche), jeweils nur an einem bestimmten Punkt des Kontinuums (bei der Archäologie der historisch zeitliche Verlauf und bei der Psychometrie die latente Einstellungsdimension) ihr Maximum erreicht. Vor und nach diesem Punkt auf dem Kontinuum strebt nach dieser Modellvorstellung die Zustimmungswahrscheinlichkeit zu einem Item bzw. die Auftretenshäufigkeit des betreffenden Fundstückes jeweils gegen null, sodass ein Funktionsgraph der Zustimmungswahrscheinlichkeit bzw. Auftretenshäufigkeit einen eingipfligen Verlauf entlang der x-Achse aufweist (vgl. auch Abbildung 1.1). D. G. Kendall (2004) beschreibt die Ursprünge der Seriation im Rahmen archäologischer Fragestellungen und Liiv (2010) gibt in einem neueren Aufsatz einen historischen Überblick über die Methode der Seriation und deren Anwendung in unterschiedlichen Wissenschaftsdisziplinen.

Das Prinzip der Reorganisation von Zeilen und Spalten in Datenmatrizen durch kombinatorische Algorithmen wird in der Regel mit einem starken Schwerpunkt auf Aspekte der Datenvisualisierung angewendet. Ein dahingehend klassisches Beispiel von Bertin (1977, S. 33) bezieht sich auf die Analyse

von französischen Siedlungen anhand kultureller Indikatoren (vgl. auch Romer, 1975) im Rahmen einer soziologischen Fragestellung nach den Auswirkungen und Indikatoren für den Wandel der Siedlungscharakteristik auf einem (latenten) angenommenen (Eigenschafts-)Kontinuum mit den Endpolen *ländlich* vs. *urban* (vgl. Abbildung 4.20).

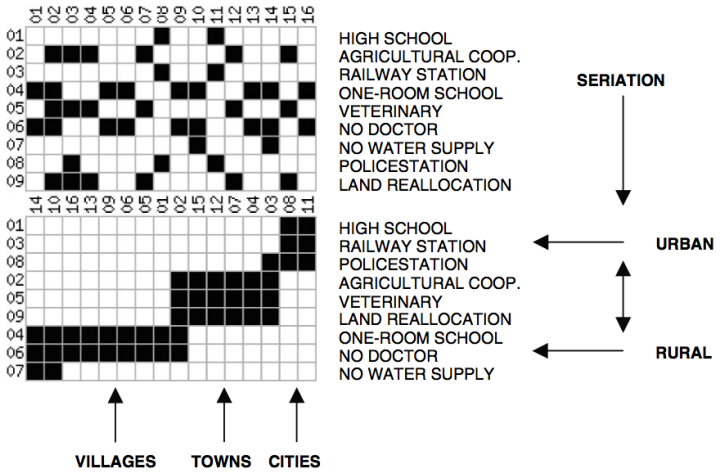


Abbildung 4.20 Darstellung des Prinzips der Seriation; Beispiel entnommen aus Liiv (2010).

Die in Abbildung 4.20 aus Liiv (2010) entnommene, vereinfachte Darstellung des Prinzips der visuellen Datenanalyse von Bertin (1977, S. 33), veranschaulicht hier das Prinzip der Reorganisation der beobachteten Daten. Die einzelnen beobachteten manifesten Indikatoren zu Merkmalen der Siedlungen (Items) werden zunächst in einer zweidimensionalen Datenmatrix gesammelt, welche zu Beginn eine (scheinbar) zufällige Struktur aufweist (vgl. Abbildung 4.20 obere Matrix). Nach einer geeigneten Reorganisation der Zeilen und Spalten der Datenmatrix lässt sich aber eine systematische Struktur erkennen, die so eindeutige Rückschlüsse zur Klassifikation der Siedlungen auf dem latenten Kontinuum *Urbanität* erlaubt.

Für den Bereich der Auswertung von Fragebogendaten im Rahmen der psychologischen Diagnostik konnten bereits Hubert und Arabie (1987) die Nützlichkeit der Methodik der Seriation am Beispiel deren Anwendung bezogen auf das hexagonale Modell der beruflichen Interessenorientierungen von Hol-

land (1997) zeigen. Für den Bereich der Differentiellen Psychologie und Persönlichkeitspsychologie illustrieren Waller et al. (1996) die Nützlichkeit von nichtlinearen Skalierungsmethoden bei der Entwicklung von psychometrischen Skalen zur Erfassung von Persönlichkeitsmerkmalen am Beispiel der Skala „negative Emotionalität“. Zur Bestimmung der reorganisierten Rangreihung der Items und Personen im Rahmen der Seriation bzw. Ordination der Daten können unterschiedliche Verfahren eingesetzt werden. In einer neueren Arbeit beschreiben Brusco und Steinley (2006) verschiedene Methoden und kombinatorische Algorithmen bei der Seriation, die dazu geeignet sind, als Lösung eine Reorganisation der Zeilen und Spalten einer Matrix zu erzeugen, welche eine inhaltliche Interpretation erlauben.

Bei der praktischen Anwendung des Prinzips einer Seriation der Daten (Hubert, 1974, 1976) kann auch auf Methoden der *Multidimensionalen Skalierung* (MDS – Kruskal, 1964a, 1964b; Torgerson, 1967) zur Bestimmung der Objektkoordinaten zurückgegriffen werden. Dieses Prinzip hat Analogien zu einer Idee von Tucker und Messick (1963), wonach individuelle Unterschiede im Antwortverhalten durch die MDS modelliert werden können. Auch Hill (1974) stellt fest, dass die gleichzeitige Reorganisation der analysierten Datenmatrix nach Zeilen und Spalten durch die Anwendung von Methoden der Multidimensionalen Skalierung effektiv erreicht werden kann. Die Multidimensionale Skalierung (MDS) bietet nach Gediga (1998) eine Möglichkeit, das Unfoldingprinzip im Sinne einer Skalierung nach Coombs (1950) näherungsweise zu behandeln, ohne dabei allzu strenge Annahmen bezüglich einer bestimmten parametrischen Modellierung der Item Characteristic Curve (ICC) zu machen – „*MCA optimizes a general-purpose criterion, not a model- specific one.* (Warrens & Heiser, 2006, S. 235) – vgl. auch van Schuur (2006). Nach Warrens und Heiser (2006) lässt sich ein eindimensionales Unfoldingmodell – also der Nähe–Distanz-Antwortprozess – durch die Anwendung der nichtmetrischen Multidimensionalen Skalierung (nMDS) modellieren, wobei dabei nur die Rangreihe der Koordinaten der ersten Dimension interpretiert werden.

Aus einer etwas erweiterten Perspektive umfasst der Begriff *Multidimensionale Skalierung* (MDS) eine Reihe von Methoden, um „verborgene“ Strukturen in (potentiell mehrdimensionalen) Datenstrukturen zu entdecken. Grundlegend basieren alle diese Verfahren auf einer *Nähe–Distanz-Matrix*, welche

sich im Falle von Fragebogendaten aus der Wechselbeziehung bzw. Interaktion der Personen mit den Items ableitet. Im Falle einer Itemanalyse von Fragebogendaten kann eine derartige (rechteckige) Distanzmatrix beispielsweise aus den bedingten Itemkategoriehäufigkeiten abgeleitet werden, wie sie auch im *PAIR*-Algorithmus (vgl. Abschnitt 4.5.3) zur Bestimmung der Itemschwierigkeiten zugrunde gelegt wird (vgl. auch Abbildung B.4 in Anhang B für eine Darstellung dieser aus den (Antwort-)Häufigkeiten abgeleiteten Matrix).

Die Distanzen werden im Rahmen der Skalierung dann auf eine räumliche Darstellung mit niedriger Dimension (typischerweise zwei oder drei Dimensionen) abgebildet (de Leeuw & Mair, 2009). Meulman et al. (1998) beschreiben die historischen Wurzeln von unterschiedlichen Ansätzen zur Skalierung basierend auf Nähe–Distanz-Informationen unter dem Überbegriff der optimalen Skalierung [*optimal scaling*]. Abdi und Valentin (2007) weisen darauf hin, dass die Multiple Korrespondenzanalyse [*Multiple Correpondence Analysis* – MCA] (vgl. Greenacre, 2010; Greenacre & Blasius, 1994) als Erweiterung der Korrespondenzanalyse [Correpondence Analysis] (CA) eine der optimalen Skalierung (oder MDS) äquivalente Methode darstellt, welche auch unterer englischsprachigen Begriffen wie *appropriate scoring*, *dual scaling*, *homogeneity analysis*, *scalogram analysis*, oder *quantification method* bekannt ist. Die Vorteile der optimalen Skalierung oder Multidimensionalen Skalierung bestehen darin, auch nicht normalverteilte oder unvollständig Daten mit Nominalskalenniveau im Rahmen der multivariaten Auswertung zu berücksichtigen, wobei auch nichtlineare Beziehungen zwischen den zu analysierenden Variablen bestehen können (Meulman et al., 1998).

Innerhalb der Vielzahl unterschiedlicher Verfahren lassen sich grob zwei zentrale Aspekte zu deren Kategorisierung identifizieren. Diese beiden Aspekte beziehen sich einerseits auf die Struktur der Daten und andererseits auf deren Skalenniveau. So unterscheidet Young (1984) zunächst zwischen multidimensionaler Skalierung (MDS) und multidimensionaler Entfaltung [*Multi Dimensional Unfolding*] (MDU) und begründet dies mit den strukturellen Unterschieden der analysierten Daten. Während die MDS immer auf quadratischen Datenmatrizen basiert, welche die Distanzen *einer* Gruppe von Objekten untereinander repräsentieren (2-way, 1-mode), werden bei der multidimensionalen Entfaltung [*Unfolding*] rechteckige Datenmatrizen (z. B. Personen $\times$

Items) analysiert (2-way, 2-mode), also Distanzen zwischen *zwei* Gruppen von Objekten (vgl. Carroll & Arabie, 1980; Jacoby, 1991; Young, 1984, zur Klassifikation von Datenstrukturen). Beide Verfahren teilen jedoch das Prinzip eines dahinterliegenden Nähe–Distanz-Modells (Young, 1984). Des Weiteren setzten die in frühen Ansätzen zur MDS verwendeten Algorithmen bezüglich der analysierten Distanzen ein metrisches oder Intervallskalenniveau voraus (L. V. Jones & Thissen, 2006). Allerdings dürfte sich die Annahme, dass beispielsweise Präferenzen für bestimmte Antwortkategorien wie metrische Entfernungen aufzufassen sind, in den meisten Anwendungsfällen als zu restriktiv erweisen. Zur Umgehung dieses Problems entwickelten Shepard (1962a, 1962b) und Kruskal (1964a, 1964b) eine Methode, die als nicht-metrische multidimensionale Skalierung (nMDS) bekannt ist. Bei der nichtmetrischen MDS wird, im Gegensatz zur MDS, nur die ordinale Information (Ranginformation) in den Distanzmatrizen zur Abbildung der räumlichen Konfiguration verwendet.

Das algorithmische Problem der MDS besteht darin, eine Reihe von Objekten so in einem niedrigdimensionalen Raum anzuordnen, dass die Abstände zwischen den Objekten (in diesem niedrigdimensionalen Raum) möglichst den Distanzen in der empirischen *Nähe–Distanz-Matrix* entsprechen. Es geht also darum möglichst eine optimale *Konfiguration* (im niedrigdimensionalen Raum) zu finden, bei der die Summe der quadrierten Differenzen zwischen den optimal skalierten Objektabständen und den Distanzen in der *Nähe–Distanz-Matrix* (mit einer Größe $i \times j$)[6] minimal ist.

$$DIF_{min} = \sqrt{\left(\frac{\sum_{i,j} \left(f(p_{i,j}) - d_{i,j} \right)^2}{\sum {d_{i,j}}^2} \right)} \tag{4.43}$$

Formal ausgedrückt, wenn p der Vektor der (empirischen) Distanzen ist (das obere oder untere Dreieck der Nähe–Distanz-Matrix) und $f(p)$ eine monotone Transformation von p sowie d die Objektabstände im niedrigdimensionalen Raum, dann müssen für diesen Raum Koordinaten so gefunden werden, dass die Summe der quadrierten Differenzen minimal wird (vgl. Gleichung 4.43).

[6] Die beiden Indizes i und j sind hier als *generische* Laufindizes der (zweidimensionalen) Distanzmatrix zu verstehen. Im Falle der MDS entspricht die Größe dieser *quadratischen* Matrix der Anzahl der Objekte k (Items) – mit $i = j = 1, ..., k$. Im Falle des MDU entspricht die Größe dieser *rechteckigen* Matrix der Anzahl der Personen n und Items k – mit $j = 1, ..., k$ und $i = 1, ..., n$.

Die Algorithmen für die MDS minimieren diese quadrierten Differenzen in einem iterativen Verfahren, um eine finale MDS-Lösung zu erhalten. In diesem Sinne eignet sich die MDS dazu, das Prinzip des Nähe–Distanz-Antwortprozesses abzubilden, ohne eine spezifischen ICC anzunehmen. Die Differenzen zwischen den empirischen Distanzen d_{ij} und den geschätzten Distanzen $\hat{d}_{ij}$ (den Objektabständen im niedrigdimensionalen Raum) lassen sich auch zur Beurteilung der Modellpassung für die finale MDS-Lösung heranziehen. Ein direktes, distanzbasiertes Maß für die Passung der MDS-Lösung ist der sogenannte STRESS-Index. Nach Kruskal (1964a) ist der STRESS-Index als globales Maß für die Passung der MDS-Lösung die Wurzel aus der *Summe* der quadrierten Differenzen zwischen den empirischen (d_{ij}) und den für den niedrigdimensionalen Raum geschätzten ($\hat{d}_{ij}$) Distanzen zwischen allen Objekten i und j (vgl. Gleichung 4.44).

$$STRESS_1 = \sqrt{\left(\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2} \right)} \tag{4.44}$$

Kruskal (1964a) gibt für den $STRESS_1$-Index Hinweise zur Interpretation der Anpassungsgüte der MDS-Lösung, die hier in der Tabelle 4.6 wiedergegeben sind.

Tabelle 4.6 Werte des $STRESS_1$-Index zur Beurteilung der MDS-Lösung.

$STRESS_1$	Beurteilung der Passung
>.20	schlecht [*poor*]
.10	ausreichend, mäßig, angemessen [*fair*]
.05	gut [*good*]
.025	exzellent [*excelent*]
.00	perfekt [*perfect*]

Anmerkungen: Hinweise zur Beurteilung der MDS-Lösung nach Kruskal (1964a); englische original Bezeichnungen in eckigen Klammern (Für die deutsche Übersetzung vgl. auch Gigerenzer, 1981, S. 349).

Analog zu den residuumbasierten Fit-Statistiken im Rasch-Modell (vgl. Abschnitt 4.4.2) ist auch der STRESS-Index prinzipiell für jeden einzelnen Datenpunkt (z. B. Personen × Items) definiert. Demnach lassen sich die Differenzen zwischen den empirischen und geschätzten Objektdistanzen, wie sie im Rahmen einer multidimensionalen Entfaltung (MDU) auf der Basis einer rechteckigen Distanzmatrix (2-way, 2-mode) bestimmt wurden, auf die gleiche Weise wie die Antwortmatrix anordnen. Zur Ableitung des STRESS-Index für eine lokale Modellpassung können die Differenzen dann, wie bei den residuumbasierten Fit-Statistiken im Rasch-Modell (vgl. Abschnitt 4.4.2), entweder über die Personen (spaltenweise) oder die Items (zeilenweise) aggregiert werden. Für die Analysen im empirischen Teil der vorliegenden Arbeit wird für alle Analysen nach dem Prinzip der MDS das *R*-Paket `smacof` eingesetzt (vgl. de Leeuw & Mair, 2009).

Die unterschiedlichen Formen dieses Skalierungsprinzips (MDS / MDU) ergeben sich, neben den entweder quadratischen (2-way, 1-mode) oder rechteckigen (2-way, 2-mode) Distanzmatrizen, im Wesentlichen auch durch unterschiedliche Distanzmetriken um die Distanzen zwischen den Objekten in der Konfiguration auszudrücken. Das grundlegende gemeinsame Prinzip der Multidimensionalen Skalierung (MDS) teilt mit den Techniken zur Skalierung von Thurstone (z. B. Thurstone, 1927b, 1927c, 1928, 1929; Thurstone & Chave, 1929) die Idee, dass Ähnlichkeitsdaten (wie sie aus verschiedenen experimentellen Verfahren und menschlichen Beurteilungen resultieren können) räumlich dargestellt werden können. Während die Skalierungstechnik von Thurstone jedoch Objekte (Personen und Items) mit reellen Zahlen auf einer einzelnen Dimension darstellt, repräsentiert die MDS Objekte meistens als Punkte in einem zwei- oder auch höherdimensionalen Raum. Im Zuge dieser Erweiterung der Dimensionalität sind die meisten Verfahren für die MDS eher datenanalytisch und basieren weniger auf einem psychologischen Prozessmodell als die Skalierungsverfahren von Thurstone. Insofern bildet die Multidimensionalen Skalierung (MDS) eine Brücke zwischen den ursprünglichen Techniken und Methoden der psychologischen Skalierung und datenanalytisch explorativ strukturfindenden Verfahren.

4.6 Modelle und Methoden zur Analyse von Datenstrukturen

Eine gemeinsame Charakteristik der beiden in Abschnitt 1.3 *Skalierung von Fragebogendaten* vorgestellten Prinzipien zur Skalierung besteht darin, dass sich der Zusammenhang der einzelnen Items auf eine dimensional aufgefasste, kontinuierliche, latente Merkmalsvariable zurückführen lässt. Diese Annahme impliziert, dass in Bezug auf die Personen eine Merkmalsdimension zur Beschreibung der Personen existiert, für die im Sinne einer quantitativen Skala unterschiedliche Grade der Ausprägung (auf dieser Dimension) messbar sind.

Das Skalenniveau (vgl. Abschnitt 1.2) dieses gemessenen Ausprägungsgrades kann dabei entweder ordinal sein, wie bei der Mokken-Analyse (vgl. Abschnitt 4.2.2) und auch dem nichtparametrischen Unfoldingmodell nach Coombs (1950, vgl. auch Abschnitt 4.3.1), oder ein metrisches, Intervallskalenniveau aufweisen wie beim Rasch-Modell und dem *Partial Credit Model* (vgl. Abschnitt 4.2.3) und auch den parametrischen Modellvarianten zur Modellierung eines Nähe–Distanz-Antwortprozesses (vgl. Abschnitt 4.3.2).

Neben der entweder probabilistischen oder deterministischen Natur der Modelle handelt es sich dabei einerseits um parametrische Modelle, welche den Antwortprozess über bestimmte Parameter (Ausprägung der Personen und Items) sowie die eigentliche Modellformulierung abbilden. Andererseits bestehen auch entsprechende nichtparametrische Ansätze, welche entweder eine unimodale oder monoton ansteigende ICC postulieren, aber dabei keine parametrischen Spezifikationen bezügliche der Funktionskurve dieser ICC vornehmen (Mokken & Lewis, 1982; Post, 1992; Post, van Duijn & van Baarsen, 2001; Sijtsma & Molenaar, 2002). Die bisher dargestellten Modelle sind demnach formal gesehen Modelle, welche die Zustimmung zu einer Antwortkategorie der Items jeweils in Abhängigkeit einer angenommenen latenten Variablen für die Merkmalsausprägung der Personen und der Schwierigkeit der Items nach einem spezifischen, angenommenen Antwortprozess modellieren, welcher sich aus der Beziehung zwischen den Personen und Items ergibt. Dies kann entweder eine *Dominanz-Relation* (vgl. Abschnitt 4.2) oder andererseits eine *Nähe–Distanz-Relation* (vgl. Abschnitt 4.3) sein.

Neben diesen beiden, den Antwortprozess auf der Basis bestimmter Annah-

men erklärenden Modellgruppen, lassen sich auch explorative, Antwortmuster orientierte, (strukturfindende) psychometrische Verfahren oder Modelle unterscheiden. Während sich die erklärenden Modelle letztlich immer zu einer entsprechenden Skalierungstechnik, wie sie in Kapitel 1 dargestellt wurden, in Beziehung setzen lassen, besteht bei Antwortmuster orientierten Verfahren keine entsprechend a priori zugeordnete Skalierungstechnik in Bezug auf die intendierte Verrechnung der Antworten auf die einzelnen Items. Diesen Verfahren liegt letztlich die Analyse von überzufällig häufigen (oder seltenen) Antwortmustern, auf eine Reihe von Items, zugrunde. Dabei wird im Sinne einer Analyse der *Kookkurrenzen* (Brehm & Feger, 2001) von Antwortkategorien, welche durch die antwortenden Personen ausgewählt wurden, nach typischen *Konfigurationen* oder Zusammenhängen gesucht und diese ggf. in übergeordneten Gruppen zusammengefasst. In der entsprechenden Literatur werden diese Konfigurationen oder Antwortmuster oft auch als *pattern* bezeichnet [engl.: *pattern* ≡ deut.: *Muster*]. Die in einer Datenmatrix vorhandenen Information wird dabei genutzt, um die antwortenden Personen entweder probabilistisch (Analyse Latenter Klassen – LCA; z. B. Formann, 1984) oder aber deterministisch (Konfigurationsfrequenzanalyse – KFA; Lienert, 1971) bestimmten, im Hinblick auf Größe und Zusammensetzung, zunächst unbekannten Gruppen zuzuordnen.

Die in den folgenden Abschnitten vorgestellten Modelle setzen bezüglich der latenten Personeneigenschaft insofern lediglich ein nominales Skalenniveau voraus. Bei diesen Modellen geht es also nicht um die Messung gradueller Unterschiede auf einer Merkmalsdimension zwischen einzelnen Personen. Vielmehr soll lediglich eine *Unterschiedsrelation* zwischen einzelnen Personen oder Personengruppen erfasst werden. Die primäre Messcharakteristik derartiger Modelle besteht also nicht in einer *Quantifizierung* der Personen auf einer metrisch definierten (latenten) Merkmalsdimension, sondern in der (nominalen) *Klassifikation* einzelner Personen oder Personengruppen anhand beobachteter manifester Antwortreaktionen.

4.6.1 Die Konfigurationsfrequenzanalyse

Im Gegensatz zu den bisher vorgestellten Methoden zur Skalierung 1.3 *Skalierung von Fragebogendaten* verfolgt die Konfigurationsfrequenzanalyse (KFA)

[*configural frequency analysis* – CFA] ein anderes Ziel. Die bisher vorgestellten Verfahren zur Skalierung einzelner Fragebogen-Items haben als gemeinsames Ziel, eine geeignete Methode zur *Verrechnung* der einzelnen Items zu einem Gesamtmesswert der Merkmalsausprägung zu finden (z. B. Torgerson, 1961). Demgegenüber verfolgt die KFA das Prinzip der Klassifikation einzelner Antwortmuster (*pattern*), welche sich aus den einzelnen Antworten ergeben. Blasius und Lautsch (1990) ordnen demnach die KFA in die Gruppe derjenigen multivariaten Verfahren ein, bei denen es „*keine Voraussetzungen bezüglich des Skalenniveaus*“ (Blasius & Lautsch, 1990, S. 110) der Antwortskalen der einzelnen Items gibt.

Die Konfigurationsfrequenzanalyse (KFA) geht auf G. A. Lienert zurück, der sie ursprünglich zur Identifikation von bedeutsamen klinischen Symptomclustern einführte (Lienert, 1971). Ursprünglich wollte G. A. Lienert die KFA daher nur für exploratorische Zwecke einsetzen. Allerdings lassen sich durch verschiedene Erweiterungen auch gezielt Hypothesen bezüglich der beobachteten Merkmals- oder Antwortmuster, beziehungsweise *Konfigurationen* testen (z. B. Stemmler, 2014; Stemmler & Heine, 2017).

Die KFA analysiert dementsprechend mehrdimensionale Kontingenztabellen, welche sich bei der multivariaten Betrachtung von zwei- oder mehrkategorialen Variablen mit mindestens nominalem Skalenniveau ergeben. Bei der KFA wird im Gegensatz zu den bereits beschriebenen psychometrischen Modellen keine latente Variable dimensional kontinuierlicher Natur postuliert. Stattdessen werden die Ausprägungskombinationen betrachtet, welche sich aus den in die Analyse einbezogenen Variablen sozusagen als *Konfigurationen* verschiedener Merkmale ergeben. Analysiert werden die absoluten Häufigkeiten dieser *Konfigurationen* – also deren *Frequenz*, woraus sich der Name des Verfahrens ableitet. Im Vergleich zu den, unter verschiedenen Modellannahmen (zur Verteilung der Konfigurationen) berechneten, erwarteten Häufigkeiten lassen sich so die beobachteten Häufigkeiten auf signifikant über- und unter frequentierte Konfigurationen (pattern) testen. Lienert bezeichnete Konfigurationen, die überzufällig hohe Häufigkeiten aufweisen als *Typen* und Konfigurationen, die überzufällig niedrige Häufigkeiten aufweisen als *Anti–Typen* (Lienert, 1971; Lienert & Krauth, 1975).

Beim einfachen *Haupteffektmodell* (KFA erster Ordnung) besteht die zu testende Modellannahme (H_0) darin, dass keine signifikanten Zusammenhänge zwischen den Variablen und *eine* gemeinsame Multinomialverteilung vorliegen. Die erwarteten Zellhäufigkeiten (pattern-Häufigkeiten) lassen sich am Beispiel einer mehrdimensionalen Kontingenztabelle mit vier Variablen i, j, k und l (mit gleicher Anzahl an Kategorien m) demnach über deren Randsummen nach der Gleichung 4.45 bestimmen.

$$e_{ijkl} = \frac{o_{i...} o_{.j..} o_{..k.} o_{...l}}{n^3} \tag{4.45}$$

Die χ^2 – verteilte Prüfstatistik des globalen χ^2 – Tests auf Unabhängigkeit bestimmt sich für dieses Beispiel nach Gleichung 4.46; mit Freiheitsgraden nach Gleichung 4.47.

$$\chi^2 = \sum_{ijkl=1}^{m} \frac{(o_{ijkl} - e_{ijkl})}{e_{ijkl}} \tag{4.46}$$

$$df = (m-1)^4 \tag{4.47}$$

Die Berechnung der erwarteten Zellhäufigkeiten in diesem als *Haupteffektmodell* bezeichneten Basismodell unter H_0 lassen sich auch als loglineares Modell darstellen. Aus der Berechnung der erwarteten Zellhäufigkeiten über die Randsummen nach der Gleichung 4.45, lässt sich dasselbe Modell (für vier Variablen A, B, C und D) gemäß Gleichung 4.48 darstellen. Die darin enthaltenen Parameter λ erklären die Effekte der Variablen im Modell auf die erwarteten Häufigkeiten (Stemmler, 2014).

$$\ln e_{ijkl} = \lambda_0 + \lambda_i A_i + \lambda_j B_j + \lambda_k C_k + \lambda_l D_l \tag{4.48}$$

Die Berechnung der erwarteten Zellhäufigkeiten in diesem Basismodell unter H_0 bezieht sich dabei immer auf die gesamte Kontingenztabelle. Dies entspricht der Annahme und zugrunde liegenden Nullhypothese, dass die Häufigkeiten der Typen oder Antitypen derselben Population wie alle anderen (möglichen) Konfigurationen angehören. Diese Annahme einer gemeinsamen Population und damit gemeinsamer (Multinomial-)Verteilung kann allerdings bei der lokalen Signifikanztestung für einzelne *Typen* und *Antitypen* verletzt sein, wenn beispielsweise extreme lokale Zellhäufigkeiten (Ausreißer) vorliegen.

Solche Grenzen der KFA wurden erstmals in den Siebziger Jahren beobachtet und thematisiert (Langeheine, 1980; Wermuth, 1973). Die Problematik der Annahme einer gemeinsamen Population wird in einer interessanten Erweiterung der KFA von Victor und Kieser (1991) adressiert. Um dem Problem *struktureller* extremer Zellfrequenzen Rechnung zu tragen, welche möglicherweise die Ergebnisse der Signifikanztestung der anderen Zellen beeinflussen, schlug Victor (1989) vor, die Existenz bestimmter Konfigurationen als „Typen“ innerhalb der Definition des Basismodells mit einzubeziehen (Kieser & Victor, 1999; Victor, 1983; Victor & Kieser, 1991). Ein entsprechendes Datenbeispiel wurde dazu von Kieser und Victor (1999, S. 969) eingeführt, das hier in Tabelle 4.7 wiedergegeben ist. Die Tabelle 4.7 gibt die Zellhäufigkeiten (Kontigenzen) von zwei Variablen mit jeweils drei Kategorien in einer 3×3 Tabelle wieder. In der Tabelle 4.7 wird deutlich, dass zwei Konfigurationen („11“ und „33“)[7] sehr geringe bzw. sehr hohe Häufigkeiten aufweisen.

Tabelle 4.7 Beispiel mit zwei Ausreißern in den Zellhäufigkeiten bei der KFA für eine 3×3 Kontingenztabelle.

	1	2	3
1	1	10	10
2	10	10	10
3	10	10	370

Anmerkungen: Zwei Indikatoren mit jeweils drei Kategorien; Beispiel nach Kieser und Victor (1999, p.969).

Nach Victor (1989) besteht das Hauptproblem bei der Annahme einer gemeinsamen Population und (Multinomialverteilung) in solch einem Fall in der angenommenen vollständigen Unabhängigkeit in Bezug auf die gesamte Kontingenztabelle innerhalb des Haupteffektmodells. Diese Annahme impliziert, dass im Rahmen des Haupteffektmodells die Berechnung der Häufigkeiten lediglich für die meisten möglichen Konfigurationen nach einer gemeinsamen multinomialen Verteilung korrekt sind (vgl. auch Krauth & Lienert, 1973; Lie-

[7] Die Indizierung der Zellen der Kontingenztabelle hat hier zwei Stellen: „11“ steht für die erste Zeile und die erste Spalte und „33“ steht für die dritte Zeile und die dritte Spalte mit den jeweils beobachteten Häufigkeiten.

nert & Krauth, 1975). Betrachtet man die Daten in Tabelle 4.7, ist es plausibel das Ergebnis einer Signifikanztestung im Rahmen einer KFA erster Ordnung dahingehend zu antizipieren, dass hier wohl ein Typ und ein Antityp vorliegt. Ein Antityp in Zelle „11“ und ein Typ in Zelle „33“.

Die Berechnung einer KFA erster Ordnung ergibt jedoch ein überraschendes Ergebnis: Die Zellen „11“, „13“, „31“ und „33“ sind die einzigen Zellen, welche die Bedingung der Unabhängigkeit erfüllen, demgegenüber werden alle anderen Zellen als Typen oder Antitypen ausgewiesen. Das gewählte Basismodell der Unabhängigkeit für die gesamte Kontingenztabelle ist hier offenbar nicht geeignet, um den offensichtlichen Typ in der Konfiguration „33“ und den Antityp in der Konfiguration „11“ zu erkennen und darüber hinaus die erwarteten Häufigkeiten für die anderen Zellen zu schätzen. Das Problem ist dabei, dass die Annahme der Unabhängigkeit hier auf die gesamte Kontingenztabelle angewendet wird, was der impliziten Annahme einer gemeinsamen Multinomialverteilung entspricht. Um den offensichtlichen Typ statistisch zu erfassen, schlug Victor (1989) vor, die Ausreißerzelle mit dem Muster „33“ als sogenannte *strukturelle Null* zu behandeln. Strukturelle Nullen sind „selbstevidente“ bzw. a priori evidente Merkmalskombinationen, wie z. B. auch Zellen einer Kontingenztabelle, die sich aufgrund logischer Ableitung und theoretischer (Vor-)Überlegungen ergeben. Für derartige Konfigurationen hat sich in der Literatur zur KFA der Begriff *Victor-Typ* etabliert (Stemmler, 2014). Nachdem die Existenz eines Victor-Typs angenommen wurde, muss geprüft werden, ob der verbleibende Rest der Tabelle unabhängig ist. Erweist sich der Rest der Kontingenztabelle als unabhängig, wird dies als *Quasi-Unabhängigkeit* bei Vorhandensein eines Typs bezeichnet. Dieser und vergleichbare Ansätze der KFA, welche sich im Wesentlichen durch die Art der Formulierung des entsprechenden Basismodells unterscheiden, werden allgemein unter dem Begriff funktionale KFA [*functional CFA*] (z. B. von Eye & Mair, 2008) subsumiert. Übergreifend unterscheidet von Eye (2004) vier Arten von Basismodellen der KFA.

Das erste Modell ist das Modell der Unabhängigkeit für die gesamte Kontingenztabelle (Haupteffektmodell), das für eine KFA erster Ordnung verwendet wird. Dieses Modell macht die Annahme, dass jede beobachtete Merkmalskonfiguration aus derselben Population stammt. Der zweite Ansatz, der Victor-Ansatz zur KFA, wurde von Victor (1989) vorgestellt. Bei diesem An-

satz basiert die zugrunde liegende Nullhypothese auf der Annahme, dass die Konfigurationshäufigkeiten der Typen oder Antitypen aus einer anderen Population stammen. Der dritte Ansatz ist der funktionale Ansatz zur KFA (von Eye & Mair, 2008). Dabei wird eine iterative Prozedur angewendet, bei der einzelne Zellen nacheinander ausgeblendet werden, bis das Basismodell (der Unabhängigkeit) passt oder bis keine weiteren Zellen mehr ausgeblendet werden können. Im Vergleich zu „Standard"-KFA zeigt sich, dass die funktionelle KFA sparsamer ist, das heißt, es müssen weniger Typen und Antitypen ausgewählt werden. Die funktionale KFA und der Victor-Ansatz sind ähnlich. Beide Ansätze können mit dem in dieser Arbeit eingesetzten *R*-Paket `confreq` (Heine, Alexandrowicz & Stemmler, 2019) ausgeführt werden. Einen weiteren Ansatz stellt die sogenannte zwei-Stichproben-KFA [*two sample CFA*] dar (z. B. Stemmler & Bingham, 2003). Das hinter diesem Ansatz stehende Prinzip besteht darin, die vorliegende Stichprobe über ein manifestes Teilungskriterium (z. B. Geschlecht, oder andere dichotome Merkmale) in zwei Teilstichproben zu unterteilen. Bei der *two sample CFA* wird nun untersucht, inwieweit bestimmte Konfigurationen bestehen, die hinsichtlich ihrer unterschiedlichen Häufigkeiten Unterschiede zwischen den beiden Teilstichproben konstituieren. Eine praxisorientierte Einführung und ein Überblick zu unterschiedlichen Varianten der KFA findet sich bei Stemmler und Heine (2017). Für die Analysen in der vorliegenden Arbeit wird im Rahmen der Anwendung der KFA das *R*-Paket `confreq` (Heine et al., 2019) eingesetzt, welches die hier beschriebenen Aspekte der KFA rechnerisch implementiert.

4.6.2 Die Latent-Class-Analysis

Die Latent-Class-Analysis (LCA) geht zurück auf die 1950 von Lazarsfeld begründete latente Struktur Analyse [*Latent Structure Analysis*] (Lazarsfeld, 1950, 1959). Die LCA setzt, ebenso wie die KFA, qualitative Variablen mit einem mindestens nominalen Skalenniveau voraus, welche gleichzeitig an mehreren Untersuchungseinheiten beobachtet werden. Die Analysen stützen sich auf die Modellannahme, dass ein zwischen den Variablen bestehender statistischer Zusammenhang, auf die Existenz latenter, d. h. nicht direkt beobachtbarer, Eigenschaften der Untersuchungseinheiten zurückzuführen ist. Im Gegensatz zum bereits beschriebenen Rasch-Modell (vgl. Abschnitt 4.2.3) muss diese latente

Variable dabei nicht unbedingt quantitative Unterschiede auf einem eindimensionalen Merkmalskontinuum der einzelnen Personengruppen ausdrücken. Im Hinblick auf die Analyse von Antwortdaten aus Fragebogenverfahren können diese latenten Eigenschaften, ebenso wie die einzelnen manifesten Indikatorvariablen, qualitative Unterschiede zwischen den einzelnen Personen oder Personengruppen zum Ausdruck bringen. Diese Unterschiede können in unterschiedlichen Antworttendenzen, im Sinne einer konsistenten Präferenz für bestimmte Antwortkategorien, oder aber auch in (unregelmäßigen) typischen, idiosynkratischen Antwortmustern begründet liegen. Mit der LCA werden somit die Verhaltenstendenzen verschiedener Teilgruppen bei der Beantwortung von Fragebogeninventaren systematisiert, ohne dabei die Personen hinsichtlich ihrer Merkmalsausprägung zu quantifizieren. Die LCA klassifiziert die Personen anhand ihrer Antworten in unterschiedliche latente Klassen (Teilgruppen). Die Zuordnung zu den latenten Klassen erfolgt dabei nicht deterministisch, sondern auf Basis der geschätzten Antwortwahrscheinlichkeiten nach einem probabilistischen Prinzip. Die Anzahl der latenten Klassen wird im Rahmen des Verfahrens nicht geschätzt, sondern muss a priori festgesetzt werden. Jede Person wird anhand ihres Antwortmusters mit einer bestimmten Wahrscheinlichkeit jeder der latenten Klasse zugeordnet, wobei in der Regel für eine der latenten Klassen eine maximale Zuordnungswahrscheinlichkeit bestimmbar ist. Insofern besteht die (ideale) Modellvorstellung bei der LCA in der Annahme disjunkter und exhaustiver Personenklassen von zunächst unbekannter Größe. Die zu schätzenden Modellparameter beziehen sich auf diese Klassenzuordnungswahrscheinlichkeit der einzelnen Personen, die relative Größe der einzelnen latenten Klassen und auf die Antwortkategoriewahrscheinlichkeiten innerhalb der jeweiligen latenten Klassen. In diesem Sinne kann die LCA als Methode zur modellbasierten Datenclusterung angesehen werden, da durch die Anwendung der LCA einzelne Untersuchungseinheiten (Personen) innerhalb der Gesamtstichprobe zu Teilgruppen zusammengefasst werden können (Fraley & Raftery, 2002). Die Teilgruppen entsprechen dabei den latenten Klassen und die einzelnen Personen können anhand ihrer maximalen Klassenzuordnungswahrscheinlichkeit, gemäß einem auf die Daten passenden LCA Modell, den Klassen zugeordnet werden.

Ausgehend von den Überlegungen zur latenten Strukturanalyse (Lazarsfeld, 1959), nimmt die LCA ein konstantes Muster von Antwortkategoriewahrscheinlichkeiten für alle Personen innerhalb einer Klasse an (Formann, 1984). Innerhalb einer latenten Klasse besteht daher gemäß der Modellannahme eine stochastische Unabhängigkeit aller beobachteten Variablen der Untersuchungseinheiten bzw. der Items für eine Skala. Die Wahrscheinlichkeiten der einzelnen Antwortpattern $p\left(\underline{x}\right)$, bezogen auf die analysierte Stichprobe, lassen sich gemäß der Modellannahmen formal nach Gleichung (4.49) allgemein darstellen.

$$p\left(\underline{x}\right) = \sum_{g=1}^{G} \pi_g \prod_{i=1}^{k} \pi_{ixg} \tag{4.49}$$

Die unbedingte Wahrscheinlichkeit eines Antwortpatterns $p(\underline{x})$ in der latenten Klasse g von insgesamt G Klassen ergibt sich danach als Summe des Produktes der bedingten Itemkategoriewahrscheinlichkeiten π_{ixg} aller beantworteten Items über alle Klassen g; wobei sich die Summe der Kategoriewahrscheinlichkeiten für jedes Item zu einem Wert von $p_i = 1$ summieren muss (vgl. Gleichung 4.50).

$$\sum_{x_i=0}^{m} \pi_{ixg} = 1 \tag{4.50}$$

Bei der LCA handelt es sich um ein *Mischverteilungsmodell* – die entmischende latente „Personenvariable“ θ ist dabei die (bedingte) Zuordnungswahrscheinlichkeit des Antwortmusters (pattern) der betreffenden Person zur jeweiligen latenten Klasse. Im Rahmen der Modellschätzung erfolgt die Zuordnung der pattern so, dass die bedingten Patternwahrscheinlichkeiten in einer der latenten Klassen maximiert wird. Neben den bedingten und unbedingten Wahrscheinlichkeiten als Modellparameter besteht in dem Modell der LCA auch noch der (Klassen-)Parameter G. Dieser Parameter, also die Anzahl der latenten Klassen, wird nicht geschätzt, sondern muss a priori festgelegt werden. Das hinsichtlich der Anzahl der latenten Klassen am besten passende Modell kann dann im Anschluss an die Schätzung der Klassenzuordnungswahrscheinlichkeiten über den Vergleich Likelihood-basierter, informationstheoretischer Kriterien wie AIC und BIC (vgl. Abschnitt 4.5) vorgenommen werden. Eine vertiefende Einführung in die LCA gibt Formann (1984).

4.7 Zusammenfassung und Ausblick auf die Anwendung psychometrischer Modelle

Neben den in diesem Kapitel besprochenen Modellen wurden im Verlauf der Entwicklung der psychometrischen Forschung noch eine ganze Reihe weitere, sehr spezifische Modelle entwickelt, welche hier nicht alle dargestellt sind. In einer Übersicht diskutiert van der Ark (2001) einige weitere Modelle und stellt Kriterien zur Wahl des geeigneten (polytomen) IRT-Modells im Hinblick auf die jeweilige Forschungsfragen auf. Eine Darstellung der Gemeinsamkeiten und Unterschiede der einzelnen Modelle und eine Einordnung in drei unterscheidbare Modellklassen schlagen Thissen und Steinberg (1986) vor. In Ergänzung dieser bestehenden Taxonomien der unterschiedlichen IRT-Modelle zeigen die Tabellen 4.2 und 4.3 in den Abschnitten 4.2 und 4.3 jeweils für beide Antwortprozesse einen Vorschlag zur Einteilung und Klassifikation der unterschiedlichen IRT-Modelle in *parametrische* v.s. *nichtparametrische*, *deterministische* v.s. *probabilistische*, und *dichotome* vs. *polytome* bezogen auf die *Antwortskala* der einzelnen Items.

Ein wichtiger Aspekt bei der Wahl eines psychometrischen Antwortmodells kann der Umstand sein, dass bei den meisten aktuell eingesetzten psychometrischen Skalen zur Erfassung individueller Unterschiede in der diagnostischen Praxis meist die summative Verrechnung der Itemscores (Anzahl gelöster Aufgaben oder Items denen zugestimmt wurde) als Maß für die erfasste Merkmalsausprägung herangezogen wird. Insofern rechtfertigt sich die fast ausschließliche und routinemäßige Anwendung kumulativer Modelle für einen Dominanz-Antwortprozess bei der Überprüfung zur Skalierbarkeit aus der Absicht, diese summative Verrechnungsvorschrift anhand empirischer Daten für die jeweilige psychometrische Skala zu überprüfen. Berücksichtigt man diesen Umstand als zentrale Grundlage für die Modellwahl, so muss festgestellt werden, dass z. B. für dichotome Items das Rasch-Modell das einzige psychometrische Modell ist, welches diese *ungewichtete*, summative Verrechnung über dessen Modellpassung an empirische Daten in spezifisch objektiver Weise überprüft – „*The Rasch model is the only latent trait model for a dichotomous response that is consistent with 'number right' scoring*“ (Wright, 1977, S. 102). Gleiches gilt unter bestimmten Voraussetzungen, nämlich einer nachgewiesenen aufsteigenden

Rangreihe der Antwortkategorien, für die von Masters (1982) vorgeschlagene Modellerweiterung für polytome Antwortformate. Einschränkend muss allerdings darauf hingewiesen werden, dass diese streng aufsteigende Rangreihe der Antwortkategorien nicht etwa in Form einer Modellrestriktion oder entsprechende Modellformulierung in dem *Partial Credit Model* von Masters (1982) (fest) implementiert ist. Vielmehr können die Schwellwerte der Itemkategoriegrenzen in diesem Modell prinzipiell frei variieren, was bei dessen Anwendung auf empirische Daten nicht selten zu Überschneidungen bei den entsprechenden Schwellenparameterprofilen über eine Reihe von Items einer Skala resultiert. Bei allen anderen Modellen mit zusätzlichen Modellparametern implizieren eben gerade diese zusätzlichen Parameter komplexere Verrechnungsmodelle für die einzelnen Itemantworten (Wright, 1977). So führen beispielsweise bereits unterschiedliche Trennschärfen der Items, wie sie im *Generalized Partial Credit Model* (GPCM – auch 2–PL–Modell) von Muraki (1992) in Form eines zusätzlichen variablen Modellparameters α_i möglich sind, zu sich überschneidenden ICCs. Derartige Überschneidungen implizieren dabei aber, dass möglicherweise eine Person auf einem schweren Item eine höhere Lösungswahrscheinlichkeit haben kann als eine andere Person auf einem eigentlich leichteren Item – was dem Prinzip einer spezifisch objektiven Messung widerspricht (vgl. Abbildung 4.7 in Abschnitt 4.2.4). Für eine gültige Verrechnung der einzelnen Itemantworten müssen, bei Geltung des 2–PL–Modells, die einzelnen Itemscores mit deren Trennschärfe gewichtet werden. Wright (1977) stellt in diesem Zusammenhang fest, dass diejenigen (summativen) Verrechnungsmodelle, welche die Skalierung der Personen und Items auf ungewichtete Summenwerte stützen, im Grunde die Geltung des RM (oder dessen polytome Erweiterung – PCM), zumindest implizit, voraussetzen (vgl. auch Kubinger, 2005).

Andererseits können auch die im Rasch-Modell getroffenen Modellannahmen als zu streng kritisiert werden (z. B. Bryce, 1981; Goldstein & Blinkhorn, 1982). Diese Kritik kann sich zunächst speziell auf die Annahme gleicher Trennschärfen beziehen, welcher mit der Einführung eines zusätzlichen Trennschärfeparameters, wie zum Beispiel im Birnbaum-Modell oder GPCM begegnet werden kann (vgl. Abschnitt 4.2.4). Ausgehend vom Rasch-Modell sind so eine Reihe von Erweiterungen der probabilistischen Modellfamilie zur Abbildung einer *Dominanz-Relation* zwischen Personen und Items entwickelt

worden (vgl. Abschnitt 4.2.4 und Tabelle 4.2 in Abschnitt 4.2.5). Die Unterschiede zwischen diesen Modellerweiterungen begründen sich meist durch deren unterschiedliche Anzahl von Modellparametern zur Beschreibung der Daten. Dabei kann grundsätzlich beobachtet werden, dass sich mit einer zunehmenden Anzahl von Modellparametern die Modellanpassung an die empirischen Daten in der Regel immer verbessern lässt (z. B. Rost, 2004, S. 330) – vgl. dazu auch Molenaar (1997a, S. 40) und Aitkin und Aitkin (2011, S. 42). Allerdings nimmt mit steigender Anzahl von Modellparametern einerseits die Problematik der Modellidentifikation und Parameterschätzung, insbesondere bei kleinen Datensätzen (vgl. auch Abschnitt 4.5), und andererseits auch die Problematik der inhaltlichen Interpretation dieser zusätzlichen Parameter zu (z. B. Glas, 2009; G. Maris & Bechger, 2009). So zeigen G. Maris und Bechger (2009), dass im 3–PL–Modell bei gleichen Diskriminationsparametern die übrigen Modellparameter – der Schwierigkeitsparameter, der „Rate-Parameter“ und der Personenparameter – auf nicht triviale Weise nicht identifiziert sind. Darüber hinaus zeigen G. Maris und Bechger (2009), dass zwei unterschiedliche Parametrisierungen des 3–PL–Modells (bei denselben Daten) zu unterschiedlichen Schlussfolgerungen bezüglich des Antwortverhaltens bei der Lösung von Items eines Fähigkeitstests (einerseits Raten oder Nichtraten) führen können. Glas (2009) stellt dazu in seinem Kommentar zum Artikel von G. Maris und Bechger (2009) einleitend fest, dass das Problem der Abhängigkeit inhaltlicher Interpretationen von der Art der Modellparametrisierung „*omnipresent bei der psychometrischen Modellbildung ist*“; [„ *... the problem raised by Maris and Bechger is omnipresent in psychometric modeling ...* “] (Glas, 2009, S. 91).

Neben solchen Fragen nach der Interpretation und Identifikation der Modellparameter besteht die Problematik einer angemessenen vergleichenden Beurteilung der unterschiedlichen Modelle hinsichtlich ihrer Passung auf empirische Datensätze (C. Brown, Templin & Cohen, 2015). Insofern kann die Wahl eines geeigneten psychometrischen Modells aus dem Bereich der IRT auch eher mit der Frage nach der intendierten Strategie bei dessen Anwendung verknüpft werden. Molenaar (1997a) beschreibt in diesem Zusammenhang zwei grundsätzlich unterschiedliche Strategien bei der Modellauswahl und deren Anwendung. Die nachsichtige Strategie vertritt die Auffassung, dass sich die sparsamen Modelle (mit wenigen Modellparametern) im Hinblick auf ihre wichtigsten Schlussfol-

gerungen (zur Skalierbarkeit) im Allgemeinen als recht robust gegenüber geringen oder mäßigen Verletzungen der Modellannahmen erweisen (vgl. auch Box, 1979, zum Begriff der Robustheit von Modellen). Die strikte Strategie dagegen versucht entweder durch die Hinzunahme von Modellparametern, oder aber durch das Entfernen von Items, die Passung der empirischen Daten an das jeweilige Modell zu optimieren. Die Strategie der Anwendung des sparsamen und restriktiven Rasch-Modells unterscheidet sich gegenüber der Anwendung komplexerer Modelle dahingehend, dass dessen Anwendung nicht auf die Daten abgestimmt ist. Stattdessen begründet sich die Anwendung des Rasch-Modells aus einer Reihe von Anforderungen für eine *spezifisch objektive* Messung (Andrich, 2004; Wright, 1977, 1999). Auch im Rasch-Modell stellt eine adäquate Anpassung des Modells an die Daten zwar eine wichtige, aber gegenüber den primären Anforderungen an eine fundamentale Messung von individuellen Merkmalen, eher sekundäre Anforderung dar. Demgegenüber betont die Strategie der Anwendung von komplexeren Modellen den Vorrang der Anpassung eines Modells an die beobachteten Daten. Ein solches Vorgehen kann daher als eher explorativ bezeichnet werden, wobei versucht wird die beobachteten empirischen Daten möglichst optimal zu modellieren.

Vor dem Hintergrund der unterschiedlichen Perspektiven dieser beiden Strategien – *Messung* vs. *(Daten-)Modellierung* – verläuft in der Psychologie und Psychometrie eine teils vehement geführte Debatte zum angemessenen Auflösungsgrad psychometrischer Modelle zur Abbildung des Antwortverhaltens in Fragebogendaten (z. B. Andrich, 2004; Borsboom & Mellenbergh, 2004; Bryce, 1981; Goldstein, 1980, 2015; Goldstein & Blinkhorn, 1982; Humphry, 2011, 2013; Jerrim, Micklewright, Heine, Sälzer & McKeown, 2018; Linacre & Fisher Jr, 2012; Michell, 2000, 2004, 2008; Panayides, Robinson & Tymms, 2015; Saint-Mont, 2012; Sijtsma, 2012; Sijtsma & Emons, 2013; Vautier, Veldhuis, Lacot & Matton, 2012). Allerdings muss im Zusammenhang mit dieser Debatte auf die in Abschnitt 4.1 dargestellten Hauptmerkmale des Modellbegriffs nach der *Allgemeinen Modelltheorie* von Stachowiak (1973) verwiesen werden. Danach sind Modelle im Wesentlichen durch ihr *Abbildungsmerkmal*, das *Verkürzungsmerkmal* und durch ihr *pragmatisches Merkmal* charakterisiert (Stachowiak, 1973, S. 131-133). So dürfte in den Sozialwissenschaften gemäß dem *Abbildungsmerkmal* und *Verkürzungsmerkmal* ein wahres Modell, unabhängig

von seiner Komplexität, wohl kaum existieren. Vielmehr muss im Hinblick auf das *pragmatische Merkmal* die Frage nach der Nützlichkeit und damit die Frage nach dem (jeweiligen) Zweck der Anwendung eines psychometrischen Modells im Vordergrund stehen (vgl. Abschnitt 4.1). Insofern wird im Vergleich zu komplexeren Modellen ein Rasch-Modell niemals perfekt auf Daten passen, es bietet allerdings eine konzeptuell kohärente Methodik für eine *spezifisch objektive* und vergleichende Messung (vgl. Abschnitte 4.2.3 und 4.2.5). Die Anwendung des Rasch-Modells bildet damit eine objektive Grundlage für die Diagnostik von lokalen Fehlanpassungen des Skalierungsmodells im Sinne einer Identifikation von nicht passenden Antwortmustern in den analysierten Daten.

Allerdings lässt sich die rigorose Anwendung des Rasch–Modells auf jegliche Fragebogenskalen auch ganz grundsätzlich kritisch hinterfragen (z. B. Divgi, 1986; Embretson & Reise, 2000). So stellt sich, insbesondere vor dem Hintergrund der in Kapitel 1 im Abschnitt 1.4 vergleichend dargestellten, unterschiedlichen Antwortprozesse und entsprechender Modelle zur Indexbildung, grundsätzlich die Frage, ob die (ausschließliche) Anwendung von Modellen für Dominanz-Antwortprozesse auf Fragebogendaten, welche über als Likert–Items gedachte Fragen erhoben wurden, immer gerechtfertigt ist. Im Zusammenhang mit Überlegungen zu einem möglichen stochastischen Prozessmodell für kognitives und Einstellungsverhalten, aus dem sich in Folge spezifische IRT–Modell ableiten lassen, merkt Roskam (1985) an: „*In this perspective, it seems rather unlikely that the same process model would be valid for attitudes as well as abilities, and this makes me sceptical about the general validity of any latent trait model.*“ (Roskam, 1985, S. 15). Roskam (1985) formuliert hier also bereits implizit die Hypothese, dass bei der Erfassung von Einstellungen im Vergleich zur Messung von Fähigkeiten (z. B. im Sinne der Intelligenzmessung) durchaus unterschiedliche kognitive Antwortprozesse bestehen, welche dann konsequenterweise durch unterschiedliche psychometrische Antwortmodellen modelliert werden müssten. Dies sind einerseits Modelle für einen *Dominanz*-Antwortprozess und andererseits Modelle für einen *Nähe–Distanz*-Antwortprozess.

Im Kapitel 3.1 *Antwortverhalten, Antwortmuster, Antwortstile, Antwortverzerrung – ein Überblick* werden unterschiedliche empirischen Befunde zu verschiedenen Formen abweichender Antwortmuster bei der Anwendung von

Fragebogenverfahren vorgestellt. Diese Arbeiten basieren in der überwiegenden Mehrzahl auf Modellen für Dominanz-Antwortprozesse und der Annahme deren universeller Gültigkeit bezogen auf alle Personen einer Stichprobe. Allerdings muss eine Analyse von abweichenden Antwortmustern nicht notwendigerweise auf Modelle für Dominanz-Antwortprozesse beschränkt bleiben. So basieren Testmodelle im Allgemeinen auf Annahmen in Bezug auf den Antwortprozess, welche durch ihre Anpassung an die Daten überprüft werden sollen (von Davier, 2009). Überprüft werden kann dabei auch, ob sich die durch das jeweilige Modell formalisierten Annahmen in gleicher Weise für alle Personen (oder Items) in einem Datensatz rechtfertigen lassen. Die Modellierung von Modellabweichungen, welche sich möglicherweise nur für manche Personengruppen innerhalb eines Datensatzes ergeben, können sogenannte *mixture-Modelle* eingesetzt werden (z. B. Rost, 1990, 1991; Rost et al., 1997; von Davier, 2009; von Davier & Yamamoto, 2007; K. Yamamoto, 1989; K. Yamamoto & Everson, 1995). Das grundlegende Prinzip besteht dabei darin, unterschiedliche latente Personenklassen zu postulieren (vgl. auch Abschnitt 4.6) innerhalb derer entweder unterschiedlich restringierte psychometrische Modelle gelten, z. B. in einer allgemeinen Form das HYBRID-Modell von K. Yamamoto (1989); K. Yamamoto und Everson (1995), oder aber die gleichen Modelle – wie zum Beispiel im *mixed-Rasch-Modell*, allerdings dann mit möglicherweise unterschiedlichen Ergebnissen für die Parameterschätzung (Rost, 1990, 1991; Rost et al., 1997).

Manche, im Rahmen der Testung von Dominanz-Antwortmodellen, als abweichend klassifizierte Antwortmuster lassen sich möglicherweise unter Annahme eines anderen Antwort- und Skalierungsmodells, z. B. eines für einen Nähe–Distanz-Antwortprozess, ohne weiteres sehr gut und plausibel erklären (z. B. Formann, 2002; Goodman, 1975). Eine allgemeine Theorie zum unimodalen Nähe–Distanz-Antwortprozess wurde von Coombs und Avrunin (1977a, 1977b) entwickelt. Die durch diesen Antwortprozess in den Daten zu findende Unfoldingstruktur wird dabei als Ergebnis eines nicht linearen Interaktionseffekts zwischen bestimmten Items und Verhaltensmerkmalen der antwortenden Personen definiert (vgl. auch Aschenbrenner, 1981). In diesem Sinne illustrieren beispielsweise Waller et al. (1996) die Nützlichkeit von nichtlinearen Methoden – gegenüber linearen, (faktoranalytischen) Methoden – bei der Entwicklung

von psychometrischen Skalen zur Erfassung von Persönlichkeitsmerkmalen.

Im Zusammenhang mit der faktorenanalytischen Auswertung von Antwortdaten im Rahmen der KTT zeigt sich beispielsweise oft, dass eine solche Auswertung von Daten, welche eigentlich einem Unfoldingmodell folgen, in artifiziellen zusätzlichen Faktoren resultiert (z. B. Maraun & Rossi, 2001; Matschinger & Krebs, 1998; Post et al., 2001; Schönemann, 1970; van Schuur & Kiers, 1994). Schönemann (1970) argumentiert mit Coombs (1967, S. 181-182) im Rahmen der Ableitung einer algebraischen Lösung im metrischen Unfolding zur Bestimmung der Koordinaten von zwei Gruppen von Objekten (also bei einer *2-way, 2-mode* Datenstruktur)[8], dass sich bei der faktorenanalytischen Auswertung der Daten, welche potentiell Präferenzurteile in Form individueller I–Skalen (vgl. Coombs, 1950, und Abschnitt 4.3.1) enthalten, immer ein zweidimensionaler Faktorraum ergibt. Die Argumentation basiert auf der einfachen Überlegung, dass zwar die individuellen I–Skalen von zwei Personen die nahe beieinander auf dem latenten Merkmalskontinuum positioniert sind hohe *positive* Korrelationen aufweisen können, wohingegen allerdings zwischen individuellen I–Skalen von zwei Personen die an den entgegengesetzten Enden auf dem latenten Merkmalskontinuum positioniert sind letztendlich hohe *negative* Korrelationen zu erwarten sind.

> Consider the I scale of an individual A at the extreme left end of the scale and that of another individual very close to him. Clearly, their preference orderings will be almost identical and will correlate close to $+1$. Individual A's I scale will correlate progressively less with the I scales of other individuals as they are farther removed from him on the joint scale. In fact, the correlation will be zero between individual A and the median individual in the distribution, and will ultimately be -1 between him and the individual at the extreme opposite end of the scale. The median individual will have correlations ranging from close to $+1$ with those individuals near him on either side, to zero with the individuals at either end. (Coombs, 1967, S. 181-182)

[8]Diese *2-way, 2-mode* Datenstruktur entsteht typischerweise bei der Anwendung von Fragebogenverfahren. Die beiden Objektgruppen sind die *Personen* und *Items*, deren Distanzen zueinander (aus der Interaktion bei der Beantwortung – in einem gemeinsamen Raum vgl. Abschnitt 4.4.2) in einem gemeinsamen Raum bei der Skalierung dargestellt werden.

In diesem Sinne zeigen Post et al. (2001) an einem realen Datenbeispiel einer elf Items umfassenden, (eigentlich) summativ zu verrechnenden Skala mit positiv und negativ kodierten Items, dass sich die mit einem kumulativen Modell für den Dominanz-Antwortprozess gefundenen zwei Dimensionen (positive vs. negative Items), durch die Skalierung nach einem nichtparametrischen Unfoldingmodell auf einer Dimension darstellen bzw. skalieren lassen. Klinkenberg (2001) merkt im Zusammenhang mit seiner Modellerweiterung des *One Parameter Logistic Model* (OPLM) von Verhelst und Glas (1995) an, dass die eindeutige Trennung von Modellen für einerseits Dominanz- und Nähe–Distanz-Antwortprozesse durch die Polarität und Extremität der Itemformulierungen konfundiert sein kann (vgl. auch Coombs & Coombs, 1976; Hayes & Dunning, 1997; Weinberger, Darkes, Del Boca, Greenbaum & Goldman, 2006). So lässt sich schon rein grafisch zeigen, dass sich der Verlauf der ICC – also die Funktion der Zustimmungswahrscheinlichkeit in Abhängigkeit der Merkmalsausprägung – von Modellen mit eingipfliger ICC derjenigen von Modellen mit monoton steigender ICC für diejenigen Items angleicht, welche eher eine extreme Merkmalsausprägung repräsentieren (vgl. Abbildung 4.21). In dieser aus Stark, Chernyshenko, Drasgow und Williams (2006, S. 27) entnommenen Abbildung sind die hypothetischen ICCs für ein Item, das für eine extreme Merkmalsausprägung steht, jeweils nach den beiden unterschiedlichen Antwortmodellen skizziert. Einerseits nimmt diese hypothetische ICC einen eingipfligen Verlauf für einen angenommenen Nähe–Distanz-Antwortprozess und andererseits einen monoton steigenden Verlauf für einen Dominanz-Antwortprozess. Dabei wird deutlich, dass die beiden hypothetischen ICCs sich über einen weiten Teil des latenten Merkmals (trotz gänzlich unterschiedlicher Antwortprozesse) weitgehend angleichen. Insbesondere im Bereich einer *mittleren* Merkmalsausprägung im Bereich von -2 bis +2 (Logits) überlagern sich die beiden Kurvenverläufe fast vollständig (vgl. Abbildung 4.21). Aus dieser zunächst hypothetischen Betrachtung folgt, dass im Falle von undifferenzierten Urteilen oder *mittleren Antworten* (potentiell induziert durch extreme Aussagen in den Items) die Passung der gegebenen Antworten zu beiden Antwortmodellen möglich wäre, oder aber zumindest die eindeutige Zuordnung zu nur einem der beiden Antwortmodell schwierig oder nicht möglich ist. Denkbar ist auch die Situation oder der mögliche Befund, dass Personen, deren Antwortmuster nach einem

Dominanz-Antwortmodell als *„Mittelkreuzer“* (MRS – vgl. Abschnitt 3.2.4) klassifiziert sind, ebenso gut dem Nähe–Distanz-Antwortprozess (mit eingipfliger Itemcharakteristik) zugeordnet werden könnten.

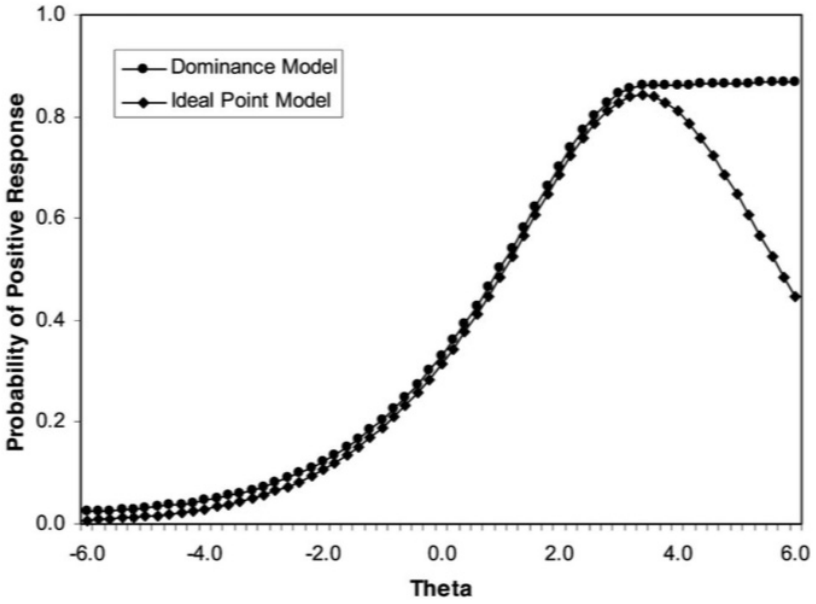


Abbildung 4.21 Darstellung von zwei *Item Characteristic Curves* (ICCs) für zwei Antwortprozesse für Items mit einer hohen Schwierigkeit; Abbildung entnommen aus Stark et al. (2006, S. 27)

Einige Autoren (z. B. Andrich, 1996; Chernyshenko, Stark, Drasgow & Roberts, 2007; Drasgow, Chernyshenko & Stark, 2010b; Roberts, 1995; Stark et al., 2006; van Schuur & Kiers, 1994) argumentieren daher in Anlehnung an Coombs (1952, 1967), dass Modelle mit unimodalen ICCs gegenüber solchen mit monoton steigenden ICCs verwendet werden sollen, wenn die Beantwortung der Items eine Introspektion[9] erfordert (Carter, Lake & Zickar, 2010; Drasgow et al., 2010b) und daher der psychologische Mechanismus der Interaktion zwischen Personen und Items einer *Nähe–Distanz*-Relation folgt – im Gegensatz zu einer *Dominanz*-Relation bei Leistungstests. In diesem Sinne propagieren Stark et al. (2006) die Anwendung von Unfolding- bzw. Idealpunkt-

[9]Wie beispielsweise bei einem Item zur Persönlichkeitsdimension *Extraversion*, die sich wie in Abschnitt 2.1.1 in Kapitel 2 dargestellt, zu einem erheblichen Anteil auf ein Konzept aus dem psychoanalytischen Persönlichkeitsparadigma stützt, welches wiederum geprägt ist von introspektiven Prozessen und Prozessen zur Objektbeziehung.

modellen zur Skalierung von Fragbogendaten zur Erfassung von Einstellungen oder Persönlichkeitseigenschaften (vgl. auch Chernyshenko et al., 2007; Drasgow et al., 2010b). Bereits Coombs und Coombs (1976) weisen drauf hin, dass die Ambiguität der Aussagen in den Items (bei deren introspektiver Evaluation) dazu führen können, dass für manche Personen die Betreffende psychometrische Skala besser nach einem Unfolding- oder Idealpunktmodell zu modellieren ist. In diesem Sinne kritisiert zum Beispiel auch Harvey (2016) – vergleichsweise aktuell – die dogmatische Anwendung von kumulativen Antwortmodellen und schlägt vor, Modelle mit eingipfliger ICC ebenso in Betracht zu ziehen. Eine solche Berücksichtigung von Modellen zum Nähe–Distanz-Antwortprozess bezieht sich – wenn sie überhaupt in Erwägung gezogen wird – dann allerdings meist auf alle Personen eines Datensatzes (z. B. Drasgow, Chernyshenko & Stark, 2010a; Drasgow et al., 2010b; Huang & Mead, 2014; O'Brien & LaHuis, 2011; Spector & Brannick, 2010; Tay, Drasgow, Rounds & Williams, 2009; Weekers & Meijer, 2008) und schließt damit die Möglichkeit aus, dass bei einzelnen Personen(-Gruppen) bezüglich der vorgegebenen Items unterschiedliche implizite Antwortmodelle (Dominanz- vs. Nähe–Distanz-Antwortmodell) bei der Beantwortung *derselben Fragen* vorliegen können. Bei der ausschließlichen Anwendung des einen oder anderen Modells manifestiert sich dies dann in abweichenden Antwortmustern, welche sich aufgrund der jeweils für alle Personen universell getroffenen Modellannahmen nicht erklären lassen.

Die in der vorliegenden Arbeit analysierten Skalen in den drei Konstrukten *Persönlichkeit*, *berufliche Interessenorientierungen* und *musikalischer Präferenzen* orientieren sich, z. B. im Hinblick auf deren Auswertung im Rahmen der psychologischen (Einzelfall-)Diagnostik, auf die kumulative, summative Verrechnung der einzelnen Items. Daher ist die Anwendung von entsprechenden psychometrischen Modellen zur Abbildung eines dabei (implizit) angenommenen *Dominanz*-Antwortprozesses bei der Überprüfung der Skalierbarkeit der einzelnen Dimensionen der Konstrukte, angezeigt. Die entsprechenden Analysen werden in drei Untersuchungen daher im Kapitel 6 *Untersuchungen zum Dominanz-Antwortprozess* durchgeführt.

Wie auch bereits in der Einleitung in Abschnitt 1.4 festgestellt, lassen die in der vorliegenden Arbeit analysierten Skalen aus den Konstrukten *Persönlichkeit*, *berufliche Interessenorientierungen* und *musikalischer Präferenzen* jedoch

im Hinblick auf die beiden unterschiedlichen Antwortmodelle verschiedene Interpretationen zu. Daraus kann die Hypothese abgeleitet werden, dass sich in den Antwortdaten aus der Bearbeitung der hier eingesetzten psychodiagnostischen Fragebogenverfahren unterschiedliche Personengruppen nachweisen lassen, welche dieselben Items bzw. Skalen nach unterschiedlichen impliziten Antwortmodellen beantwortet haben (vgl. dazu auch z. B. Chernyshenko et al., 2007; Drasgow et al., 2010a, 2010b; Hardy & Ford, 2014; Harvey, 2016; Stark et al., 2006; van Schuur, 1995). An diese Hypothese anschließend stellt sich dann die Frage, ob es sich bei einer derartig unterschiedlichen Rezeption der Items durch unterschiedliche Personengruppen um ein über unterschiedliche Konstrukte und Skalen hinweg konsistentes Phänomen handelt. Die entsprechenden Analysen zu differentiellen (impliziten) Antwortmodellen – entweder nach dem *Dominanz*-Antwortprozess oder dem *Nähe–Distanz*-Antwortprozess – werden daher in zwei Untersuchungen im Kapitel 7 *Untersuchungen zum Nähe–Distanz-Antwortprozess* durchgeführt.

Kapitel 5

Stichproben und Instrumente

Dieses Kapitel beschreibt die im empirischen Teil der vorliegenden Arbeit eingesetzte Datengrundlage. Dargestellt werden die den Auswertungen und Analysen in den Kapiteln 6 und 7 zugrunde liegenden Stichproben, sowie die eingesetzten Fragebogeninventare (Instrumente) zu den untersuchten Konstrukten *Persönlichkeit*, *Präferenzen des Musikgeschmacks* und *berufliche Interessenorientierungen*. Die in den beiden Kapiteln 6 und 7 eingesetzten Datenmatrizen beziehen sich auf zwei Personenstichproben, welche jeweils über einen längeren Zeitraum an der Universität der Bundeswehr München erhoben wurden. Bei den Erhebungen wurden, wenn auch teilweise in unterschiedlicher Form, im Wesentlichen immer die gleichen Skalen und Instrumente zur Bearbeitung vorgelegt (vgl. Anhang C). Daher werden die eingesetzten Instrumente und erhobenen Variablen (vgl. Abschnitt 5.1) und die beiden Stichproben (vgl. Abschnitt 5.2) in den folgenden beiden Abschnitten zunächst übergreifend dargestellt.

5.1 Eingesetzte Instrumente und erhobene Variablen

Die Erhebung erfolgte für beide Stichproben in Form eines Online-Fragebogens (vgl. Anhang C). Im ersten Abschnitt dieses Fragebogens wurden, neben demografischen Variablen (*Alter* und *Geschlecht*), Fragen bezüglich der Studiensituation, der militärischen Ausbildung (soweit zutreffend) und Fragen zur Organisation des (studentischen) Alltags erfasst. Diese Variablen wurden in erster

Linie für universitätsinterne Zwecke oder andere laufende Forschungsprojekte erfasst und werden daher in den folgenden Analysen nicht berücksichtigt.

In den darauffolgenden Abschnitten wurden zunächst 25 Fragen zu sechs Persönlichkeitsdimensionen erhoben, von denen sich fünf Dimensionen an das *Big-Five-Inventory* (BFI–K – Rammstedt & John, 2005) anlehnen. Im Anschluss wurden insgesamt vier Dimensionen musikalischer Präferenzorientierungen mit einer in die deutsche Sprache übertragenen Online-Version (Langmeyer et al., 2012) des *Short Test Of Music Preferences* (STOMP– Rentfrow & Gosling, 2003) erfasst. Im letzten Abschnitt folgten die beruflichen Interessenorientierungen, die mit einer online adaptierten Version des *Allgemeinen Interessen-Struktur-Tests* (AIST–R – Bergmann & Eder, 2005) erfasst wurden. Je nach Projektjahrgang wurden zwischen diesen drei Konstrukten unterschiedliche Fragen bzw. Skalen ergänzend eingefügt (vgl. Anhang C), welche nicht in der vorliegenden Arbeit analysiert werden. Sowohl der Umfang der Online-Fragebögen und auch einzelne Items der jeweiligen Skalen der drei Konstrukte wurden von Erhebungsjahr zu Erhebungsjahr im Rahmen wechselnder Forschungsprojekte und im Rahmen der intendierten Skalenoptimierung teilweise leicht variiert. In Anhang C sind die drei für die in dieser Arbeit analysierten Daten relevanten Fragebogen-Versionen enthalten (vgl. Anhang C Abschnitte C.1, C.1 und C.3). Die in den durchgeführten Untersuchungen berücksichtigten Instrumente (BFI–K, STOMP und AIST–R) sollen in den nachfolgenden Abschnitten kurz beschrieben werden.

5.1.1 Die Kurzversion des Big-Five-Inventory BFI–K

Zur Erfassung des Konstrukts *Persönlichkeit* nach dem Big-Five-Modell im Rahmen des Eigenschaftsparadigmas wurde ein an der Universität der Bundeswehr in München entwickelter Onlinefragebogen zur Erfassung von einer plus fünf Persönlichkeitsdimensionen (BFI–K) von Schmolck (2006a) eingesetzt (vgl. auch Schmolck, 2003, 2004, 2005, 2006b). Dieses Instrument ist, bezogen auf die Big-Five-Dimensionen, bis auf einige Itemformulierungen, in weiten Teilen vergleichbar mit dem von Rammstedt und John (2005) publizierten BFI–K, welcher eine aus dem Englischen übersetze Kurzversion (Rammstedt, 1997) des ursprünglich von John et al. (1991) entwickelten, 44 Items umfassenden *Big-Five-Inventory* ist (vgl. auch John & Srivastava, 1999). In Anlehnung an die deutsche Version des BFI–K von Lang, Lüdtke und Asendorpf (2001) wurde die an der Universität der Bundeswehr weiterentwickelte Onlineversion des BFI–K erstmals im Jahre 2004 im Rahmen eines Lehrprojektes an der Universität der Bundeswehr in München eingesetzt (Schmolck, 2003, 2004). Seit 2005 wurden die Items zur Operationalisierung der Persönlichkeit nach dem Big-Five-Modell im Rahmen der laufenden Erhebungen um Items zur Erfassung einer weiteren Persönlichkeitsdimension ergänzt. Diese Dimension *Ehrlichkeit-Bescheidenheit* [Honesty-Humility] – H–Skala ist Bestandteil des HEXACO-Modells von Michael Ashton und Kimbeon Lee (Ashton et al., 2004; K. Lee & Ashton, 2004), dessen andere fünf Faktoren, *Emotionality* (E), *Extraversion* (X), *Agreeableness* (A), *Conscientiousness* (C) und *Openness to Experience* (O), abgesehen von gewissen Bedeutungsverschiebungen den Dimensionen des Big-Five-Modells, entsprechen. Jeweils aufbauend auf die Ergebnisse der Datenauswertung aus den jährlichen Erhebungen an Studenten der Universität der Bundeswehr wurde in den folgenden Jahren eine Version des BFI–K mit jeweils vier Items pro Dimension entwickelt (Schmolck, 2006b). Diese Version wurde für die vorliegenden Untersuchungen zur Erfassung des Konstrukts Persönlichkeit nach Fünf-Faktoren-Modell eingesetzt. Die während der laufenden Erhebungen in den Jahren 2007 bis 2009 und 2010 bis 2011 noch in der Entwicklung befindliche, zusätzliche H–Skala konnte für die hier weiter unten berichteten Analysen nicht eingesetzt werden, da zwischen den einzelnen Erhebungen für diese Dimension noch kein kontinuierlich konsistenter und unveränderter Itempool vorlag.

Die insgesamt 25 Items werden auf einer fünfstufigen Antwortskala beantwortet. Dabei soll eingeschätzt werden, in welchem Ausmaß die jeweilige Aussage für die antwortende Person zutrifft. Den einzelnen Antwortkategorien sind dabei, zur besseren Kennzeichnung von *Ablehnung* und *Zustimmung*, ganzzahlige numerische Werte von -2 bis $+2$ zugeordnet. Zusätzlich sind die Antwortkategorien zu Beginn des Inventars jeweils mit einer qualifizierenden Wortmarke überschrieben. Die Tabelle 5.1 gibt die Darstellungsform der Antwortkategorien in der Onlineversion des modifizierten BFI–K wieder.

Tabelle 5.1 Texte der qualifizierenden Wortmarken der Antwortkategorien des modifizierten BFI–K.

Antwortkategorie	Scoring	Text Wortmarke
-2	0	*„Sehr unzutreffend“*
-1	1	*„Unzutreffend“*
0	2	*„Teils / teils“*
1	3	*„Zutreffend“*
2	4	*„Sehr zutreffend“*

Die originale Reihenfolge bei der Darbietung der 25 Fragen (mit der nicht verwendeten H–Dimension), die theoretische Zuordnung zu den sechs Dimensionen, die Itempolung sowie der Itemwortlaut sind mit den Häufigkeiten der gewählten Antwortkategorien jeweils für beide Stichproben in den Tabellen 5.6 und 5.10 dargestellt.

5.1.2 Der Short Test Of Music Preferences STOMP

Die Originalversion des *Short Test Of Music Preferences* (STOMP) entstand im Rahmen einer umfangreichen Serie von Untersuchen von Sam Gosling und seinem Mitarbeiter Jason Rentfrow (Rentfrow & Gosling, 2003). Über Präferenzeinschätzung von 14 Musikgenres werden in der Version von Rentfrow und Gosling (2003) damit vier Dimensionen des Musikgeschmacks erfasst:

1. Reflective & Complex (RC): Classical, Blues, Folk, Jazz.
2. Intense & Rebellious (IR): Alternative, Rock, Heavy Metal
3. Upbeat & Conventional (UC): Country, Religious, Pop, Soundtracks/Theme Songs
4. Energetic & Rhythmic (ER): Dance/Electronica, Rap/hip-hop, Soul/funk

In der im Rahmen des ESF-Projektes eingesetzten deutschen Adaption des STOMP müssen die jeweiligen Präferenzen für Musikgenres (Items) auf einer siebenstufigen Antwortskala eingeschätzt werden. Wie der Vergleich der in Anhang C dargestellten Varianten des STOMP mit deren jeweiligen Items zeigt, wurde bei der Entwicklung der deutschen Adaption der Itempool hinsichtlich einzelner Genrebezeichnungen und Items teilweise variiert und ergänzt. So können beispielsweise für *Folk*, *Country* und das Item „religiöse Musik“[1] [*Religious*] in Deutschland Bedeutungsunterschiede im Vergleich zum amerikanischen Kontext bestehen. Um solche kulturellen Bedeutungsverschiebungen abzubilden, wurde diese Genrebezeichnungen durch gängige deutschen Genres wie „Populäre Volksmusik“, „Schlager“ und „Neue Deutsche Welle“ ergänzt (vgl. Langmeyer, Guglhör-Rudan & Tarnai, 2012, S. 122). Aufgrund der im Vergleich zum amerikanischen Kontext geringeren Popularität religiöser Musik war auch das in der amerikanischen Version von Rentfrow und Gosling

[1] Insbesondere für die direkte deutsche Übersetzung des Items *Religious* – *religiöse Musik* – dürften hier erhebliche kulturelle Unterschiede in den damit (spontan) verbundenen Assoziationen bestehen. Für den amerikanischen Kontext mag hier z. B. eine naheliegende Verbindung zur vergleichsweise populären und verbreiteten Gospelmusik bestehen, wohingegen für den deutschen Kontext für den Begriff *religiöse Musik* möglicherweise eher eine Verbindung zu traditioneller Kirchenmusik wie z. B. Choralgesang besteht.

(2003) enthaltene Item „religiöse Musik“ [*Religious*] in der deutschen Fragebogenversion für die Erhebungsjahrgängen 2007 und 2008 nicht enthalten.

Um für die Untersuchungen in dieser Arbeit einen möglichst großen Stichprobenumfang zu gewährleisten, können nur diejenigen Items berücksichtig werden, welche im gesamten Erhebungszeitraum zwischen 2007 und 2011 in gleichlautender Weise eingesetzt werden. Bezogen auf die vier Dimensionen des STOMP sind dies die folgenden Items bzw. Bezeichnungen für Musikgenres:

1. Reflective & Complex (RC): Klassik, Blues, Folk, Jazz
2. Intense & Rebellious (IR): Alternative, Rock, Heavy Metal
3. Upbeat & Conventional (UC): Country, Pop, Filmmusik/Titelmelodien
4. Energetic & Rhythmic (ER): Electronica, Rap/hip-hop, Soul/R&B

Die Liste der Items mit den einzelnen musikalischen Genrebezeichnungen ist dabei mit der folgenden einleitenden Erklärung überschrieben (vgl. auch Anhang C Abschnitte C.1, C.1 und C.3):

„Geben Sie für jedes der unten aufgeführten Musikgenre an, wie gerne oder ungern Sie diese Art von Musik hören. Verwenden Sie für Ihre Antworten die folgende 7-stufige Skala:“

Die Antwortkategorien der einzelnen Items sind mit ganzzahlig, numerischen Werten von -3 bis $+3$ überschrieben. Die jeweiligen Endpunkte der Antwortskala sind mit ergänzende Wortmarken, welche von *„Mag ich überhaupt nicht“* über *„Neutral / keine Meinung“* bis *„Mag ich sehr“* reichen, überschrieben (vgl. Tabelle 5.2).

Die Items zu den insgesamt 13 Musikgenres[2] und deren Zuordnung zu den vier Dimensionen sind mit den Häufigkeiten der gewählten Antwortkategorien jeweils für beide Stichproben in den Tabellen 5.7 und 5.11 dargestellt.

[2]Die englische Originalversion von Rentfrow und Gosling (2003) umfasst wie beschrieben 14 Musikgenres bzw. Items.

Tabelle 5.2 Texte der qualifizierenden Wortmarken der Antwortkategorien des modifizierten STOMP.

Antwortkategorie	Scoring	Text Wortmarke
−3	0	*„Mag ich überhaupt nicht“*
−2	1	
−1	2	
0	3	*„Neutral / keine Meinung“*
1	4	
2	5	
3	6	*„Mag ich sehr“*

5.1.3 Der Allgemeine Interessen-Struktur-Test AIST–R

Der *Allgemeine Interessen-Struktur-Test* (AIST) wurde von Bergmann und Eder (1999, 2005) zur Operationalisierung des Modells der beruflichen Interessenorientierungen (Holland, 1997) als Inventar zur Erfassung von sechs Umwelt- und Interessenorientierungen entwickelt. Eine Revision des entwickelten Itempools im Jahr 2005 (Bergmann & Eder, 2005), bezog sich hauptsächlich auf Anpassungen von Formulierungen und Inhalten der Items zu zeitgemäßen Tätigkeiten sowie der Neunormierung des Tests (Muck, 2007). So wurden z. B. für den Bereich der *praktisch technischen Orientierung (Realistic)*, als Anpassung an den technischen Fortschritt, auch Fragen zur Computernutzung ergänzt.

Der Test umfasst insgesamt 60 Fragen, welche jeweils auf einer fünfstufigen Antwortskala beantwortet werden. Die fünf Skalenpunkte der Antwortskala sind dabei, neben einer Nummerierung von 1 bis 5, jeweils mit qualifizierenden Wortmarken überschrieben. Die Tabelle 5.3 gibt die aufsteigende Nummerierung der Antwortkategorien und die jeweilige Wortmarke wieder.

Tabelle 5.3 Texte der qualifizierenden Wortmarken der Antwortkategorien des AIST–R.

Antwortkategorie	Scoring	Text Wortmarke
1	0	*„Das interessiert mich gar nicht; das tue ich nicht gerne“*
2	1	*„Das interessiert mich wenig“*
3	2	*„Das interessiert mich etwas“*
4	3	*„Das interessiert mich ziemlich“*
5	4	*„Das interessiert mich sehr; das tue ich sehr gerne“*

Die insgesamt 60 Fragen des AIST–R besteht aus einer Liste mit verschiedensten Tätigkeitsbeschreibungen. Für jede der beschriebenen Tätigkeiten soll angegeben werden, wie sehr sich die antwortende Person für diese interessiert bzw. interessieren würde. Jeweils zehn der Fragen beziehen sich dabei auf eine der von Holland (1997) angenommenen Dimensionen beruflicher Interessenorientierungen. Somit werden mit den 60 Fragen sechs Dimensionen beruflicher Interessen erfasst (vgl. Abschnitt 2.2 zum theoretischen Hintergrund des Konstruktes): R: *Praktisch-technische Interessen (Realistic)*, I: *Intellektuell-forschende Interessen (Investigative)*, A: *Künstlerisch-sprachliche Interessen*

(Artistic), S: *Soziale Interessen (Social)*, E: *Unternehmerische Interessen (Enterprising)* sowie C: *Konventionelle Interessen (Conventional)*. Die Darbietungsreihenfolge der Fragen erfolgt dabei, bezüglich der sechs Dimensionen, in systematisch durchmischter Form. Die erste Frage bezieht sich auf die Dimension *Realistic*, die zweite auf die Dimension *Investigative* und so weiter. Die Reihenfolge der Fragen, die theoretische Zuordnung zu den sechs Dimensionen, sowie der Itemwortlaut sind mit den Häufigkeiten der gewählten Antwortkategorien jeweils für beide Stichproben in den Tabellen 5.8 und 5.12 dargestellt. Die Formulierungen der 60 Fragen entsprechen denen aus der überarbeiteten Version des AIST, welcher 2005 in einer revidierten Fassung als AIST–R von Bergmann und Eder (2005) herausgegeben wurde. Die Reliabilität der einzelnen Skalen zu den sechs Dimensionen erreicht als interne Konsistenz (Cronbach, 1951; Kuder & Richardson, 1937), bezogen auf die Normstichprobe, Werte zwischen $r_\alpha = .82$ (*Investigative*, *Artistic*) und $r_\alpha = .87$ (*Social*, *Enterprising*).

5.2 Stichproben und Erhebung

Die in den einzelnen Kapiteln des empirischen Teils der vorliegenden Dissertation berichteten Untersuchungen, stützen sich auf zwei überwiegend studentische Stichproben. Diese wurden jeweils im Rahmen der sozialwissenschaftlichen Methodenausbildung an der Universität der Bundeswehr jeweils im zweiten Studien–Trimester erhoben. Die erste Stichprobe umfasst dabei die Jahre 2007, 2008 und 2009 und die zweite Stichprobe die Jahre 2010 und 2011. Die in einigen der im Folgenden berichteten Untersuchungen teils getrennte Behandlung und Analyse dieser beiden Stichproben hat dabei die folgenden Gründe.

Zunächst unterscheiden sich die Onlineplattformen, welche für Erhebung der Daten in den beiden Stichproben eingesetzt wurde. So wurde im ersten Zeitraum eine rein HTML–basierte Webseite eingesetzt, welche auf den internen Servern der Universität der Bundeswehr gehostet wurde. Für den zweiten Erhebungszeitraum wurde dagegen die Plattform *UNIPARK* (Globalpark AG, 2010) eingesetzt.

Zum anderen wurden für den zweiten Erhebungszeitraum die Modalitäten für die Bearbeitung der eingesetzten Inventare verändert. Während im ersten Erhebungszeitraum das Auslassen einzelner Items nicht möglich war, wurde die Beantwortung einzelner Items für den zweiten Erhebungszeitraum freigestellt. Die Teilnahme an der Erhebung war grundsätzlich für alle Universitätsangehörige der Universität der Bundeswehr zugänglich, sodass die beiden Stichproben unterschiedliche Studiengänge und auch Personen welche nicht studieren umfasst (vgl. Tabelle 5.4).

Tabelle 5.4 Häufigkeiten der Studiengänge in den Stichproben I und II.

	Studiengang	*Häufigkeiten*	
		Stichprobe I[b]	Stichprobe II[c]
1	Pädagogik	296	125
2	Sozialwisenschaft	66	78
3	Sportwissenschaft	90	21
4	Luft- und Raumfahrttechnik	29	66
5	Bauingenieurwesen / Geodäsie[a]	14	47
6	Maschienenbau (FH)	20	34
7	Elektortechnik und Technische Informatik (FH)	12	33
8	Elektrotechnik (Univ.)	20	31
9	Informatik / Wirtschaftsinformatik[a]	29	48
10	Betriebswirtschaft (FH)	21	29
11	Wirtschafts- und Organisationswissenschaft	64	75
12	Mathematical Engineering	1	15
13	kein Student	72	7

Anmerkungen: [a] Studiengänge jeweils zusammengefasst; [b] $n = 734$; [c] $n = 609$.

5.2.1 Stichprobe I (2007 – 2009)

Insgesamt umfasste die über drei Studienjahrgänge erhobene Stichprobe $n = 734$ Personen. Dabei waren $n = 662$ Studienteilnehmer Studenten der Universität der Bundeswehr aus den Studienjahrgängen 2007, 2008, 2009. Schwerpunktmäßig sind in den Daten Studienfächer der Sozialwissenschaften (mit Sport) vertreten, welche 61.5 % der Gesamtstichprobe ausmachten. Die Studienteilnahme war für die Studierenden der Universität der Bundeswehr im Rahmen des „Scheinerwerbs“ für die Methodenveranstaltung verpflichtend. Neben den studentischen Studienteilnehmern nahmen auch wissenschaftliche Mitarbeiter, Angestellte der Universität sowie Studierende an anderen Universitäten freiwillig an der Erhebung teil, welche insgesamt 9.8 % der Gesamtstichprobe ausmachten.

Tabelle 5.5 Kategorie Häufigkeiten der Variable *Alter*; Stichprobe I.

Kategorie	Alter	Absolute Häufigkeiten
1	unter 21 Jahre	73
2	21 Jahre	123
3	22 Jahre	121
4	23 Jahre	143
5	24 Jahre	81
6	25 Jahre	51
7	26 Jahre	41
8	über 26 Jahre	101

Anmerkungen: $n = 734$.

Das Alter der Studienteilnehmer wurde als kategoriale Variable mit acht Kategorien erfasst. Tabelle 5.5 gibt eine Übersicht über die Kategorie Häufigkeiten der Variable Alter der Studienteilnehmer mit dem Modus der acht Kategorien bei einem Alter von 23 Jahren. Der Median für die Verteilung der Kategorien $2 \equiv$ *21 Jahre* bis $7 \equiv$ *26 Jahre* liegt ebenfalls bei einem Alter von 23 Jahren.

Aufgrund der hauptsächlich aus Studenten der Universität der Bundeswehr

rekrutierten Stichprobe konnte eine Gleichverteilung der Geschlechter nicht erreicht werden. So nahmen 73 % männliche und 27 % weibliche Teilnehmer teil. Aus diesem Grund wird in den Analysen der vorliegenden Arbeit auf nach Geschlecht vergleichende oder replizierende Analysen verzichtet. Die Erhebung wurde über einen Onlinefragebogen realisiert, der sich in mehrere, bereits beschriebene, Abschnitte gliederte (vgl. Abschnitt 5.1 und Anhang C).

Die Tabelle 5.6 gibt neben den Antwortkategoriehäufigkeiten der noch nicht umgepolten Items die Itemformulierungen für die einzelnen BFI–K-Dimension wieder. Die Tabelle 5.7 gibt die Itemformulierungen und die Antwortkategoriehäufigkeiten für den STOMP wieder und die Tabelle 5.8 die Itemformulierungen und die Antwortkategoriehäufigkeiten für den AIST–R.

Tabelle 5.6 Antwortkategorie Häufigkeiten der Items des BFI–K; Stichprobe I.

	Polung	Dim.	fehlend	-2	-1	0	1	2	Itemformulierung
bfi04	+	N[1]	0	105	282	222	105	20	leicht nervös und unsicher wird
bfi07	+	N[1]	0	72	221	189	184	68	sich viele Sorgen macht
bfi13	-	N[1]	0	16	75	224	299	120	ruhig bleibt, selbst in Stresssituationen
bfi22	-	N[1]	0	10	108	188	306	122	emotional ausgeglichen und nicht leicht aus der Fassung zu bringen ist
bfi03	+	E[2]	0	12	92	197	274	159	aus sich herausgeht, gesellig ist
bfi11	+	E[2]	0	3	37	162	373	159	begeisterungsfähig ist, andere mitreißen kann
bfi15	-	E[2]	0	118	235	190	150	41	eher zurückhaltend und reserviert ist
bfi17	-	E[2]	0	190	269	164	86	25	eher still und wortkarg ist
bfi02	+	O[3]	0	2	27	121	344	240	gerne Überlegungen anstellt, mit Ideen spielt
bfi08	-	O[3]	0	123	224	143	142	102	nur wenig künstlerische Interessen hat
bfi16	+	O[3]	0	62	160	178	235	99	künstlerische und ästhetische Eindrücke schätzt
bfi24	+	O[3]	0	8	74	158	316	178	eine aktive Vorstellungskraft hat, phantasievoll ist
bfi12	-	A[4]	0	263	303	125	38	5	oft Krach mit anderen hat
bfi18	+	A[4]	0	11	58	204	336	125	rücksichtsvoll und einfühlsam zu anderen ist
bfi21	+	A[4]	0	10	65	182	326	151	lieber kooperiert als konkurriert
bfi25	-	A[4]	0	59	195	181	216	83	schroff und abweisend zu anderen sein kann
bfi01	+	C[5]	0	1	16	93	354	270	zuverlässig und gewissenhaft arbeitet
bfi09	+	C[5]	0	0	10	130	369	225	Aufgaben gründlich erledigt
bfi14	-	C[5]	0	83	193	217	175	66	bequem ist und zur Faulheit neigt
bfi19	-	C[5]	0	140	181	172	184	57	dazu neigt, unordentlich zu sein

Anmerkungen: Antwortkategorien von -2 bis 2; **nicht umgepolten** Items des modifizierten BFI–K; [1] Neurotizismus [2] Extraversion [3] Offenheit [4] Verträglichkeit [5] Gewissenhaftigkeit; $n = 734$.

Tabelle 5.7 Antwortkategorie Häufigkeiten der Items des STOMP; Stichprobe I.

	Polung	Dimension	miss	-3	-2	-1	0	1	2	3	Iteminhalt
stomp01	+	RK[1]	2	48	93	80	114	233	117	47	Klassik
stomp02	+	RK[1]	2	89	103	119	142	195	60	24	Blues
stomp05	+	RK[1]	2	154	159	114	164	92	36	13	Folk
stomp11	+	RK[1]	2	101	85	116	134	183	75	38	Jazz
stomp10	+	IR[2]	2	105	101	52	148	123	111	92	Alternative
stomp12	+	IR[2]	2	9	15	25	57	143	227	256	Rock
stomp14	+	IR[2]	2	145	97	82	85	122	100	101	Heavy Metal
stomp03	+	UC[3]	2	145	151	124	130	125	45	12	Country
stomp13	+	UC[3]	2	17	22	40	105	215	230	103	Pop
stomp15	+	UC[3]	2	10	27	34	101	192	210	158	Filmmusik/Titelmelodien
stomp04	+	ER[4]	2	158	102	80	94	94	100	104	Electronica
stomp06	+	ER[4]	2	181	90	67	65	126	118	85	Rap/hip-hop
stomp07	+	ER[4]	2	94	68	87	84	174	119	106	Soul/R&B

Anmerkungen: [1] Reflective & Complex (RK), [2] Intense & Rebellious (IR), [3] Upbeat & Conventional (UC), [4] Energetic & Rhythmic (ER); $n = 734$.

Tabelle 5.8 Antwortkategorie Häufigkeiten der Items des AIST–R; Stichprobe I.

	Polung	Dim.	fehlend	0	1	2	3	4	Iteminhalt
aist01	+	R[1]	0	62	167	189	177	139	[7]
aist07	+	R[1]	0	31	104	201	255	143	[7]
aist13	+	R[1]	0	136	209	149	146	94	[7]
aist19	+	R[1]	0	21	74	178	283	178	[7]
aist25	+	R[1]	0	296	160	90	104	84	[7]
aist31	+	R[1]	0	249	229	142	90	24	[7]
aist37	+	R[1]	0	284	193	132	80	45	[7]
aist43	+	R[1]	0	231	218	156	93	36	[7]
aist49	+	R[1]	0	218	276	133	86	21	[7]
aist55	+	R[1]	0	96	188	231	163	56	[7]
aist02	+	I[2]	0	98	174	219	177	66	[7]
aist08	+	I[2]	0	72	169	234	206	53	[7]
aist14	+	I[2]	0	75	137	207	208	107	[7]
aist20	+	I[2]	0	17	99	259	257	102	[7]
aist26	+	I[2]	0	197	218	127	142	50	[7]
aist32	+	I[2]	0	29	134	272	243	56	[7]
aist38	+	I[2]	0	212	192	163	111	56	[7]
aist44	+	I[2]	0	385	166	83	57	43	[7]
aist50	+	I[2]	0	27	116	247	249	95	[7]
aist56	+	I[2]	0	182	189	164	138	61	[7]
aist03	+	A[3]	0	136	224	151	127	96	[7]
aist09	+	A[3]	0	190	231	160	104	49	[7]
aist15	+	A[3]	0	219	199	160	95	61	[7]
aist21	+	A[3]	0	46	139	179	230	140	[7]
aist27	+	A[3]	0	98	131	225	188	92	[7]
aist33	+	A[3]	0	181	203	146	114	90	[7]
aist39	+	A[3]	0	61	133	223	199	118	[7]
aist45	+	A[3]	0	280	145	118	97	94	[7]
aist51	+	A[3]	0	255	192	118	91	78	[7]
aist57	+	A[3]	0	203	236	125	109	61	[7]
aist04	+	S[4]	0	152	213	187	137	45	[7]
aist10	+	S[4]	0	7	39	111	310	267	[7]
aist16	+	S[4]	0	17	67	172	334	144	[7]
aist22	+	S[4]	0	41	102	192	286	113	[7]
aist28	+	S[4]	0	164	251	203	90	26	[7]
aist34	+	S[4]	0	18	94	215	292	115	[7]
aist40	+	S[4]	0	15	64	193	275	187	[7]
aist46	+	S[4]	0	174	206	178	126	50	[7]
aist52	+	S[4]	0	149	241	201	109	34	[7]
aist58	+	S[4]	0	25	100	193	290	126	[7]
aist05	+	E[5]	0	5	25	112	328	264	[7]
aist11	+	E[5]	0	23	60	177	270	204	[7]
aist17	+	E[5]	0	46	125	228	238	97	[7]
aist23	+	E[5]	0	85	217	217	173	42	[7]
aist29	+	E[5]	0	34	104	179	283	134	[7]
aist35	+	E[5]	0	20	85	209	289	131	[7]
aist41	+	E[5]	0	43	134	240	245	72	[7]
aist47	+	E[5]	0	11	68	229	320	106	[7]
aist53	+	E[5]	0	36	151	243	241	63	[7]
aist59	+	E[5]	0	49	101	231	238	115	[7]
aist06	+	C[6]	0	221	261	160	73	19	[7]
aist12	+	C[6]	0	101	239	254	109	31	[7]
aist18	+	C[6]	0	192	281	182	66	13	[7]
aist24	+	C[6]	0	19	99	264	267	85	[7]
aist30	+	C[6]	0	59	184	224	197	70	[7]
aist36	+	C[6]	0	170	266	181	91	26	[7]
aist42	+	C[6]	0	90	240	225	141	38	[7]
aist48	+	C[6]	0	87	226	234	142	45	[7]
aist54	+	C[6]	0	67	140	225	233	69	[7]
aist60	+	C[6]	0	175	256	189	89	25	[7]

Anmerkungen: [1] Realistic [2] Investigative [3] Artistic [4] Social [5] Enterprising [6] Conventional; $n = 734$; [7] Die Inhalte der Items des AIST–R unterliegen urheberrechtlichen und copyright-bezogenen Bestimmungen und sind daher hier nicht abgedruckt. Der genaue Wortlaut der Items kann bei Bergmann und Eder (2005) eingesehen werden.

5.2.2 Stichprobe II (2010 – 2011)

Die zweite Stichprobe wurde während der Studienjahre 2010 und 2011 erhoben und umfasst insgesamt $n = 609$ Personen. Dabei sind $n = 602$ Studienteilnehmer Studenten der Universität der Bundeswehr. Im Gegensatz zur ersten Stichprobe ist hier der Anteil der Sozialwissenschaften (mit Sport), mit lediglich 36.7 % der Gesamtstichprobe, wesentlich geringer vertreten. Ebenso war im Gegensatz zur ersten Stichprobe die Teilnahme nicht verpflichtend. Allerdings wurde den Studierenden der sozialwissenschaftlichen Fächer der Universität der Bundeswehr die Teilnahme im Rahmen der Methodenveranstaltung nahegelegt. Der Anteil nicht studentischer Teilnehmer an der Erhebung fällt, im Vergleich zur ersten Stichprobe, mit 1.1 % sehr gering aus. Neben der, auch für die studentischen Teilnehmer bestehenden Freiwilligkeit der Teilnahme insgesamt, war es darüber hinaus möglich, bei der Online-Bearbeitung der einzelnen Skalen einzelne Items nicht zu beantworten.

Tabelle 5.9 Kategorie Häufigkeiten der Variable *Alter*; Stichprobe II.

Kategorie	Alter	Absolute Häufigkeiten
1	unter 21 Jahre	66
2	21 Jahre	137
3	22 Jahre	98
4	23 Jahre	102
5	24 Jahre	75
6	25 Jahre	61
7	26 Jahre	27
8	über 26 Jahre	43

Anmerkungen: $n = 609$

Wie in der ersten Stichprobe wurde das Alter der Teilnehmer wieder als kategoriale Variable mit acht Kategorien erfasst. Tabelle 5.9 zeigt die Kategorie Häufigkeiten der Variable Alter. Der Modus der acht Kategorien liegt hier mit 21 Jahren zwei Jahre unter dem Modus der ersten Stichprobe. Der Median für die Verteilung der Kategorien $2 \equiv$ *21 Jahre* bis $7 \equiv$ *26 Jahre* liegt dagegen wie

bei der Stichprobe I ebenfalls bei einem Alter von 23 Jahren.

Vergleichbar mit der ersten Stichprobe fallen auch die Anteile der Geschlechter der Stichprobe, mit 81 % männlichen und 19 % weiblichen Teilnehmern, aus.

Tabelle 5.10 Antwortkategorie Häufigkeiten der Items des BFI–K; Stichprobe II.

	Polung	Dim.	fehlend	-2	-1	0	1	2	Itemformulierung
bfi04	+	N[4]	3	205	131	193	66	11	leicht nervös und unsicher wird
bfi07	+	N[4]	2	158	105	163	127	54	sich viele Sorgen macht
bfi13	-	N[4]	3	41	20	197	248	100	ruhig bleibt, selbst in Stresssituationen
bfi22	-	N[4]	3	58	32	172	255	89	emotional ausgeglichen und nicht leicht aus der Fassung zu bringen ist
bfi03	+	E[3]	2	58	26	217	187	119	aus sich herausgeht, gesellig ist
bfi11	+	E[3]	3	35	16	188	269	98	begeisterungsfähig ist, andere mitreißen kann
bfi15	-	E[3]	3	166	115	180	116	29	eher zurückhaltend und reserviert ist
bfi17	-	E[3]	3	202	170	141	81	12	eher still und wortkarg ist
bfi02	+	O[2]	2	19	8	144	265	171	gerne Überlegungen anstellt, mit Ideen spielt
bfi08	-	O[2]	3	164	111	135	128	68	nur wenig künstlerische Interessen hat
bfi16	+	O[2]	3	128	79	147	187	65	künstlerische und ästhetische Eindrücke schätzt
bfi24	+	O[2]	3	57	23	152	246	128	eine aktive Vorstellungskraft hat, phantasievoll ist
bfi12	-	A[5]	3	276	220	87	20	3	oft Krach mit anderen hat
bfi18	+	A[5]	3	41	22	198	262	83	rücksichtsvoll und einfühlsam zu anderen ist
bfi21	+	A[5]	3	45	29	167	247	118	lieber kooperiert als konkurriert
bfi25	-	A[5]	3	138	82	169	159	58	schroff und abweisend zu anderen sein kann
bfi01	+	C[1]	1	14	8	89	327	170	zuverlässig und gewissenhaft arbeitet
bfi09	+	C[1]	3	17	11	109	302	167	Aufgaben gründlich erledigt
bfi14	-	C[1]	3	101	76	212	156	61	bequem ist und zur Faulheit neigt
bfi19	-	C[1]	3	150	125	153	133	45	dazu neigt, unordentlich zu sein

Anmerkungen: Antwortkategorien von -2 bis 2; **nicht umgepolten** Items des modifizierten BFI–K; [1] Neurotizismus [2] Extraversion [3] Offenheit [4] Verträglichkeit [5] Gewissenhaftigkeit; $n = 609$.

Die Tabelle der Kategorie Häufigkeiten der noch nicht umgepolten Items des BFI–K (Tabelle 5.10) zeigt, im Vergleich zur ersten Stichprobe, dass trotz Freiwilligkeit der Itembearbeitung der Anteil fehlender Werte relativ gering ausfällt. In gleicherweise zeigt sich dies für den STOMP in Tabelle 5.11 und für den AIST–R in Tabelle 5.12.

Tabelle 5.11 Antwortkategorie Häufigkeiten der Items des STOMP; Stichprobe II.

	Polung	Dimension	miss	-3	-2	-1	0	1	2	3	Iteminhalt
stomp01	+	RK[1]	1	46	68	45	92	195	100	62	Klassik
stomp02	+	RK[1]	2	89	84	79	146	148	43	18	Blues
stomp05	+	RK[1]	2	88	124	103	139	104	32	17	Folk
stomp11	+	RK[1]	3	63	84	79	127	144	75	34	Jazz
stomp10	+	IR[2]	3	72	55	77	106	107	94	95	Alternative
stomp12	+	IR[2]	3	18	14	11	30	83	177	273	Rock
stomp14	+	IR[2]	1	100	67	84	57	100	84	116	Heavy Metal
stomp03	+	UC[3]	2	102	122	96	126	114	33	14	Country
stomp13	+	UC[3]	3	40	25	25	85	169	173	89	Pop
stomp15	+	UC[3]	2	22	23	7	88	161	175	131	Filmmusik/Titelmelodien
stomp04	+	ER[4]	1	61	55	93	63	81	124	131	Electronica
stomp06	+	ER[4]	2	96	79	130	67	107	77	51	Rap/hip-hop
stomp07	+	ER[4]	2	82	79	69	92	133	94	58	Soul/R&B

Anmerkungen: [1] Reflective & Complex (RK), [2] Intense & Rebellious (IR), [3] Upbeat & Conventional (UC), [4] Energetic & Rhythmic (ER); $n = 609$.

Tabelle 5.12 Antwortkategorie Häufigkeiten der Items des AIST–R; Stichprobe II.

	Polung	Dim.	fehlend	0	1	2	3	4	Iteminhalt
aist01	+	R[1]	2	46	79	146	173	163	[7]
aist07	+	R[1]	3	22	57	124	233	170	[7]
aist13	+	R[1]	3	91	121	137	159	98	[7]
aist19	+	R[1]	3	13	38	151	248	156	[7]
aist25	+	R[1]	3	179	125	99	90	113	[7]
aist31	+	R[1]	3	160	141	156	111	38	[7]
aist37	+	R[1]	3	183	129	108	113	73	[7]
aist43	+	R[1]	3	179	152	128	116	31	[7]
aist49	+	R[1]	3	181	176	145	85	19	[7]
aist55	+	R[1]	3	64	108	188	194	52	[7]
aist02	+	I[2]	2	69	117	155	180	86	[7]
aist08	+	I[2]	3	52	126	195	179	54	[7]
aist14	+	I[2]	3	42	84	153	224	103	[7]
aist20	+	I[2]	3	13	63	186	247	97	[7]
aist26	+	I[2]	3	168	190	117	92	39	[7]
aist32	+	I[2]	3	23	98	209	211	65	[7]
aist38	+	I[2]	3	131	137	136	132	70	[7]
aist44	+	I[2]	3	260	133	83	73	57	[7]
aist50	+	I[2]	3	23	73	159	255	96	[7]
aist56	+	I[2]	3	117	147	119	134	89	[7]
aist03	+	A[3]	3	118	179	134	127	48	[7]
aist09	+	A[3]	3	169	202	124	83	28	[7]
aist15	+	A[3]	3	236	137	109	68	56	[7]
aist21	+	A[3]	3	56	118	174	176	82	[7]
aist27	+	A[3]	3	108	126	147	156	69	[7]
aist33	+	A[3]	3	179	170	117	84	56	[7]
aist39	+	A[3]	3	77	127	158	163	81	[7]
aist45	+	A[3]	3	223	118	99	95	71	[7]
aist51	+	A[3]	3	221	167	96	73	49	[7]
aist57	+	A[3]	3	165	192	117	91	41	[7]
aist04	+	S[4]	3	165	169	143	100	29	[7]
aist10	+	S[4]	2	14	39	129	267	158	[7]
aist16	+	S[4]	3	26	84	149	245	102	[7]
aist22	+	S[4]	3	42	109	152	212	91	[7]
aist28	+	S[4]	3	164	190	157	79	16	[7]
aist34	+	S[4]	3	29	68	192	227	90	[7]
aist40	+	S[4]	3	16	62	150	236	142	[7]
aist46	+	S[4]	3	174	176	129	90	37	[7]
aist52	+	S[4]	3	160	175	147	98	26	[7]
aist58	+	S[4]	3	33	93	158	220	102	[7]
aist05	+	E[5]	3	8	24	81	302	191	[7]
aist11	+	E[5]	3	16	65	133	245	147	[7]
aist17	+	E[5]	3	52	103	175	196	80	[7]
aist23	+	E[5]	3	96	177	170	125	38	[7]
aist29	+	E[5]	3	36	83	155	226	106	[7]
aist35	+	E[5]	3	17	65	162	268	94	[7]
aist41	+	E[5]	3	56	115	205	160	70	[7]
aist47	+	E[5]	3	26	76	167	257	80	[7]
aist53	+	E[5]	3	46	102	209	199	50	[7]
aist59	+	E[5]	3	46	88	182	203	87	[7]
aist06	+	C[6]	3	195	197	128	65	21	[7]
aist18	+	C[6]	3	155	240	153	49	9	[7]
aist12	+	C[6]	3	83	209	188	102	24	[7]
aist24	+	C[6]	3	9	69	206	231	91	[7]
aist30	+	C[6]	3	62	124	175	177	68	[7]
aist36	+	C[6]	3	132	194	166	81	33	[7]
aist42	+	C[6]	3	91	200	180	107	28	[7]
aist48	+	C[6]	3	99	198	162	112	35	[7]
aist54	+	C[6]	3	44	103	192	209	58	[7]
aist60	+	C[6]	2	145	175	165	96	26	[7]

Anmerkungen: [1] Realistic [2] Investigative [3] Artistic [4] Social [5] Enterprising [6] Conventional; $n = 609$; [7] Die Inhalte der Items des AIST–R unterliegen urheberrechtlichen und copyright-bezogenen Bestimmungen und sind daher hier nicht abgedruckt. Der genaue Wortlaut der Items kann bei Bergmann und Eder (2005) eingesehen werden.

5.3 Deskriptive Befunde zur internen Konsistenz nach KTT

Zur Absicherung der Korrektheit der vorgenommenen Rekodierungen der negativ gepolten Items in den einzelnen BFI–K-Skalen werden für sämtliche Skalen und beide Stichproben Berechnungen zur internen Konsistenz (Cronbach, 1951) vorgenommen. Zur Berechnung der internen Konsistenz werden nur diejenigen Personen berücksichtigt, welche jeweils sämtliche Items einer Skala beantwortet haben. Der Umfang der Datengrundlage variiert daher in einem sehr geringen Maße über Skalen und Konstrukte und ist in den jeweiligen Anmerkungen zu den Tabellen 5.13, 5.14 und 5.15 angegeben.

5.3.1 BFI–K

Für eine erste Analyse des Antwortverhaltens der beiden Stichproben auf den Skalen des BFI–K werden für sämtliche Skalen vergleichende Berechnungen zur internen Konsistenz (Cronbach, 1951), vorgenommen. Die Ergebnisse zur internen Konsistenz der aufbereiteten BFI–K-Daten für beide Stichproben zeigt Tabelle 5.13.

Tabelle 5.13 Interne Konsistenz für fünf BFI–K-Dimensionen und zwei Stichproben.

Dimension	Stichprobe I	Stichprobe II
Neurotizismus	$r_\alpha = 0.69$	$r_\alpha = 0.62$
Extraversion	$r_\alpha = 0.82$	$r_\alpha = 0.78$
Offenheit	$r_\alpha = 0.71$	$r_\alpha = 0.61$
Verträglichkeit	$r_\alpha = 0.57$	$r_\alpha = 0.49$
Gewissenhaftigkeit	$r_\alpha = 0.70$	$r_\alpha = 0.69$

Anmerkungen: r_α = interne Konsistenz nach Cronbach (1951); Stichprobe I: $n = 734$; Stichprobe II: $n = 606$.

Die Skalen weisen für beide Stichproben vergleichbare Werte für den Koeffizienten der internen Konsistenz auf. Allerdings zeigt sich, dass die Koeffizienten für die Stichprobe II tendenziell eher geringfügig niedriger ausfallen

als diejenigen für die Stichprobe I. Zwischen den rekodierten negativ formulierten Items und den jeweils anderen Items der einzelnen Skalen ergaben sich keinerlei negative Korrelationen.

5.3.2 STOMP

Die Analyseergebnisse des Antwortverhaltens der beiden Stichproben auf den Skalen des STOMP mit Berechnungen zur internen Konsistenz (Cronbach, 1951) nach der klassischen Testtheorie für beide Stichproben zeigt Tabelle 5.14.

Tabelle 5.14 Interne Konsistenz für vier STOMP-Dimensionen und zwei Stichproben.

Dimension	Stichprobe I	Stichprobe II
Reflective & Complex	$r_\alpha = 0.68$	$r_\alpha = 0.64$
Intense & Rebellious	$r_\alpha = 0.64$	$r_\alpha = 0.60$
Upbeat & Conventional	$r_\alpha = 0.40$	$r_\alpha = 0.37$
Energetic & Rhythmic	$r_\alpha = 0.56$	$r_\alpha = 0.56$

Anmerkungen: r_α = interne Konsistenz nach Cronbach (1951); Stichprobe I: $n = 732$; Stichprobe II: $n = 606$.

Wie beim BFI–K weisen die Skalen für beide Stichproben vergleichbare Werte für den Koeffizienten der internen Konsistenz auf. Allerdings zeigt sich auch hier, dass die Koeffizienten für die Stichprobe II tendenziell eher geringfügig niedriger ausfallen als diejenigen für die Stichprobe I. Auffallend ist der recht niedrige Wert für die interne Konsistenz für die Skala *Upbeat & Conventional* welcher in gleicher Weise für beide Stichproben einen Wert von $r_\alpha 0 \leq 0.4$ aufweist.

5.3.3 AIST–R

Die Berechnungen zur internen Konsistenz (Cronbach, 1951) nach der klassischen Testtheorie für beide Stichproben zum Antwortverhalten der beiden Stichproben auf den Skalen des AIST–R zeigt Tabelle 5.15.

Tabelle 5.15 Interne Konsistenz für sechs AIST–R-Dimensionen und zwei Stichproben.

Dimension	Stichprobe I	Stichprobe II
Realistic	$r_\alpha = 0.85$	$r_\alpha = 0.85$
Investigative	$r_\alpha = 0.80$	$r_\alpha = 0.81$
Artistic	$r_\alpha = 0.85$	$r_\alpha = 0.85$
Social	$r_\alpha = 0.87$	$r_\alpha = 0.88$
Enterprising	$r_\alpha = 0.88$	$r_\alpha = 0.88$
Conventional	$r_\alpha = 0.84$	$r_\alpha = 0.82$

Anmerkungen: r_α = interne Konsistenz nach Cronbach (1951); Stichprobe I: $n = 734$; Stichprobe II: $n = 606$.

Die Skalen für beide Stichproben weisen vergleichbare Werte für den Koeffizienten der internen Konsistenz auf und erreichen im Vergleich mit den Skalen der anderen Konstrukte die höchsten Werte für die interne Konsistenz. Die hier anhand der beiden Stichproben empirisch bestimmten Werte für die interne Konsistenz der AIST–R Skalen sind vergleichbar mit den Werten aus der Normierung des Tests. So wird für die interne Konsistenz – ermittelt über Kuder-Richardson, Formel 20 (vgl. Kuder & Richardson, 1937) – ein Wertebereich zwischen $r_\alpha = .82$ (Investigative, Artistic) und $r_\alpha = .87$ (Social, Enterprising) angegeben (vgl. Bergmann & Eder, 2005; Muck, 2007).

Kapitel 6

Untersuchungen zum Dominanz-Antwortprozess

In diesem Kapitel der vorliegenden Arbeit werden insgesamt drei Untersuchungen durchgeführt. Die Gemeinsamkeit der drei Untersuchungen liegt dabei in der grundlegenden Annahme einer summativen (kumulativen) Skalierbarkeit der untersuchten psychometrischen Skalen in den drei Konstrukten *Persönlichkeit*, *berufliche Interessenorientierungen* und *musikalische Präferenzen*. Zur Überprüfung dieser Annahme zur Skalierung werden bei den einzelnen Auswertungen entsprechende psychometrische Modelle angewendet. Erweisen sich die Modelle zur Beschreibung der untersuchten Daten als angemessen, so kann die summative Verrechnung der Antworten auf die Items entweder für die gesamte Stichprobe, oder aber mit unterschiedlichen Modellparametern für bestimmte einzelne Teilgruppen der Stichprobe als gültige Auswertungspraxis angesehen werden. Die Untersuchungen gehen darüber hinaus auf drei grundlegende Fragestellungen im Zusammenhang mit individuell unterschiedlich ausgeprägtem Antwortverhalten auf psychodiagnostischen Skalen ein.

Im Abschnitt 6.1. *Extreme und mittlere Antworttendenz im Bereich beruflicher Interessenorientierungen und Musikgeschmack* wird der Frage nachgegangen, inwieweit sich bei der Erfassung beruflicher Interessenorientierungen mit einem Fragebogen zur Selbstbeurteilung verschieden ausgeprägte, typische Antworttendenzen finden lassen, welche auf den unterschiedlichen Gebrauch der mehrstufig abgestuften Antwortskala der einzelnen Items zurückzuführen sind. Ferner wird in der Untersuchung in Abschnitt 6.1 der Frage nachgegan-

gen, ob es sich bei den Antworttendenzen um eine übergeordnete Verhaltensdisposition oder (Meta-)Eigenschaft der jeweiligen Personen handelt, welche sich immer in der gleichen und konsistenten Art und Weise bei der Beantwortung von Fragebogenskalen zeigt – unabhängig von der jeweils zu erfassenden Merkmalsdimension beruflicher Interessenorientierungen des AIST–R (vgl. Bergmann & Eder, 2005) oder musikalischer Präferenzen des STOMP (vgl. Rentfrow & Gosling, 2003).

In Abschnitt 6.2. *Untersuchung zur Skalierbarkeit des BFI–K* wird zunächst versucht frühere Befunde aus der Literatur zu unterschiedlichen Antworttendenzen bei der Beantwortung von Persönlichkeitsinventaren zu replizieren. Auf Basis der sich daraus ergebenden Befunde, sowie der Befunde aus der Untersuchung in Abschnitt 6.1 wird in Abschnitt 6.3 analysiert, inwieweit sich die unterschiedlich ausgeprägten Antworttendenzen auf die empirisch gefundenen Zusammenhänge zwischen den beiden Konstrukten *Persönlichkeit* und *berufliche Interessenorientierungen* auswirken.

Die bei den drei Untersuchungen jeweils zugrunde gelegten, unterschiedlichen Datenmatrizen beziehen sich auf die zwei in Kapitel 5. *Stichproben und Instrumente* beschriebenen Personenstichproben, welche über einen längeren Zeitraum an der Universität der Bundeswehr erhoben wurden. Die Hauptfragestellungen der jeweiligen Untersuchungen werden anhand der zusammengefügten Daten beider Stichproben untersucht, um für die dabei eingesetzten Auswertungsverfahren eine breitere Datengrundlage bereitstellen zu können. In einigen der drei in den Abschnitten 6.1 bis 6.3 durchgeführten Analysen werden diese beiden Stichproben jedoch zusätzlich getrennt und vergleichend analysiert. Dabei sollen die Ergebnisse aus der ersten Stichprobe (vgl. Abschnitt 5.2.1), jeweils mit den Daten aus der zweiten Stichprobe (vgl. Abschnitt 5.2.2) repliziert werden.

6.1 Extreme und mittlere Antworttendenz im Bereich beruflicher Interessenorientierungen und Musikgeschmack

Einleitung

Die hier vorgestellte Untersuchung geht der übergeordneten Frage nach, ob sich die in den Handanweisungen des AIST–R (vgl. Bergmann & Eder, 2005) und des STOMP (vgl. Rentfrow & Gosling, 2003) geforderte, summative Verrechnung der einzelnen *Itemscores* der einzelnen Dimensionen *beruflicher Interessenorientierungen* oder *musikalischer Präferenzen* zu jeweils einem Summenwert empirisch rechtfertigen lässt. Wie bereits im Abschnitt 3.1 ausgeführt, ist die allgemeine Gültigkeit einer solchen Verrechnung der *Itemscores* im Rahmen eines kumulativen Antwortmodells an bestimmte Voraussetzungen geknüpft, die sich einerseits auf die Items und andererseits auf die Personen (welche die Scores produzieren) beziehen (Rost, 2002). In der allgemeinen Einführung zu dieser Arbeit wurde in Kapitel 4 *Psychometrische Modellierung* bereits dargelegt, dass sich diese Voraussetzungen durch die Überprüfung der Passung von Antwortmodellen der *Item-Response-Theory* (IRT) auf die Daten testen lässt. Geht man dabei zunächst von der Annahme einer grundsätzlich summativen Skalierbarkeit der untersuchten psychometrischen Skalen aus (also deren Items), so stellt sich in Folge die Frage, ob sich die damit assoziierten psychometrischen Modelle für alle Personen einer Stichprobe in gleicher Weise mit übergreifend gültigen Modellparametern anpassen lassen.

Wie in Abschnitt 3.2 dargestellt ist diese universelle Gültigkeit von solchen psychometrischen Antwortmodellen mit universell geltenden Modellparametern auf der Basis der empirischen Befunde aus der Literatur nicht immer gegeben. Insbesondere bei mehrstufigen Antwortskalen für die einzelnen Items, wie sie auch beim AIST–R und STOMP verwendeten werden, stützt sich der Befund unterschiedlich ausgeprägter Antworttendenzen, im Sinne einer Tendenz entweder zur Mitte (*middle response style* – MRS) oder zu den Extremen (*extreme response style* – ERS) einer vorgegebenen Antwortskala, auf eine recht breite empirische Basis (vgl. Abschnitt 3.2.4 im Kapitel 3. *Theoretischer Hintergrund zu Antwortmustern*).

Trotz der recht umfänglichen empirischen Befunde zu unterschiedlich ausgeprägten Antworttendenzen bei verschieden Konstrukten, existieren allerdings bisher keine Untersuchungen, welche sich spezifisch mit dem Phänomen MRS und ERS bei der Erfassung der *beruflichen Interessenorientierung* oder *musikalischer Präferenzen* auseinandersetzen. Lediglich ein neuerer Aufsatz von Wetzel und Hell (2013) befasst sich mit der Problematik differentieller Itemschwierigkeiten, so genanntem *differential item functioning* (DIF – z.B. Hambleton & Swaminathan, 1985; Mellenbergh, 1982), für unterschiedliche Personengruppen bei der Erfassung beruflicher Interessenorientierungen. Zur Operationalisierung der beruflichen Interessen wird dabei ebenfalls der AIST–R von Bergmann und Eder (2005) eingesetzt. Die Analysen von Wetzel und Hell (2013) zielen dabei aber auf die Untersuchung differenzieller Effekte bei der Bestimmung der Itemschwierigkeiten bezogen auf die beiden Geschlechtergruppen ab. Im Vergleich dazu, soll bei der hier durchgeführten Untersuchung die Prävalenz von MRS und ERS als differentielle Antworttendenz im Sinne einer (zusätzlichen) Personeneigenschaft untersucht werden. Solche von Baumeister und Tice (1988) sowie Britt (1993) als Metaeigenschaften [*metatraits*] (vgl. Abschnitt 3.3) bezeichneten Personenmerkmale können zunächst unabhängig von manifesten Merkmalen wie Geschlecht oder anderen gruppendefinierenden Merkmalen bestehen.

In der vorliegenden Studie soll daher untersucht werden, ob sich, unabhängig von solchen manifesten Variablen auch bei der Erfassung beruflicher Interessen auf den sechs Skalen des AIST–R und den vier Dimensionen musikalischer Präferenzen des STOMP, Personengruppen mit unterschiedlich ausgeprägter Tendenz zu mittleren oder extremen Antwortkategorien identifizieren lassen. Ferner soll überprüft werden, ob sich diese Tendenz zur Mitte oder zu den extremen Antwortkategorien innerhalb des Konstrukts der beruflichen Interessenorientierungen nach dem Modell von Holland (1997) konsistent und unabhängig von der jeweiligen Interessenausprägung zeigt. Darüber hinaus werden in dieser Untersuchung bei der Skalierung der Daten unter der Annahme eines kumulativen Antwortmodells zwei verschiedene Methoden zur Bestimmung der Itemschwierigkeiten im Rahmen der IRT eingesetzt und hinsichtlich ihrer Übereinstimmung in Bezug auf die dabei erzielten Ergebnisse, überprüft. Zum einen wird hierbei eine iterative Methode eingesetzt, die *Conditional Maxi-*

mum Likelihood-Schätzung (CML – z. B. Baker & Kim, 2004; Fischer, 1974) und eine nichtiterative Methode, die Methode des paarweisen Itemvergleichs (*PAIR* – Heine & Tarnai, 2015).

Daten

Für die hier zu berichtenden Analysen werden die Antwortdaten zu den 60 Items aus dem AIST–R sowie die Antwortdaten zu den 13 Items aus dem STOMP verwendet. Zur Untersuchung der Frage, ob sich auf den einzelnen Skalen des AIST–R und STOMP Antworttendenzen entweder zur Mitte oder zu den Extremen finden lassen, werden die in Abschnitt 5.2 beschriebenen Stichproben aus dem Erhebungszeitraum 2007 – 2009 (Stichprobe I) und die Stichproben aus dem Erhebungszeitraum 2010 – 2011 (Stichprobe II) als Gesamtdatensatz analysiert. Diese Zusammenlegung der beiden Stichproben erfolgt, um bei der Anwendung der CML-Methode möglicherweise auftretenden Schätzproblemen zu begegnen, welche sich aus den teilweise geringen Kategoriehäufigkeiten für manche Antwortkategorien der vorliegenden beiden einzelnen Stichproben, ergeben könnten (vgl. dazu auch Heine & Tarnai, 2015; Luo & Andrich, 2005).

Die gesamte Datengrundlage für die Auswertung umfasst somit $n = 1343$ Personen sowie $k = 60$ Items des AIST–R und $k = 13$ Items des STOMP. Nachdem die Analysen zur Skalierung für jede der sechs bzw. vier Dimensionen des AIST–R und STOMP getrennt erfolgen, teilt sich diese gesamte empirische Datengrundlage für den AIST–R in sechs einzelne Datenmatrizen mit jeweils $k = 10$ Spalten (Items) auf. Für den STOMP teilt sich empirische Datengrundlage in drei einzelne Datenmatrizen mit $k = 3$ Spalten (*Intense & Rebellious*, *Upbeat & Conventional* und *Energetic & Rhythmic*) sowie eine Datenmatrix mit $k = 4$ Spalten (*Reflective & Complex*). Vor den Analysen werden aus diesen Datenmatrizen jeweils diejenigen Fälle herausgenommen, welche sämtliche Items (*unit non responder*) der betreffenden Dimension ausgelassen haben. Für den AIST–R reduziert sich daher die jeweilige Anzahl der Zeilen in den Datenmatrizen von ursprünglich $n = 1343$ Fällen für die Dimensionen *Realistic*, *Investigative Social* und *Conventional* auf $n = 1341$ Fälle; und für die AIST–R-Skalen *Artistic* und *Enterprising* auf $n = 1340$ Fälle. Beim STOMP bestehen drei unit non responder, sodass hier für die vier

Datenmatrizen jeweils $n = 1340$ Personen erhalten bleiben. Der Wertebereich innerhalb der Datenmatrizen umfasst für jede Spalte beim AIST–R den ganzzahligen Bereich von 0 bis 4 und beim STOMP den ganzzahligen Bereich von 0 bis 6, was den fünf bzw. sieben Antwortkategorien der Items entspricht.

Methode

Zur Überprüfung der zur Auswertung des AIST–R und STOMP geforderten kumulativen Verrechnung der einzelnen Itemscores der jeweiligen Dimensionen zu einem Summenwert (vgl. Bergmann & Eder, 2005; Rentfrow & Gosling, 2003) wird jede Dimension getrennt unter Annahme des *Partial Credit* Modells (PCM – Masters, 1982) für polytome Itemantwortskalen skaliert. Zur Identifikation möglicher Antworttendenzen auf den einzelnen Skalen des AIST–R und STOMP wird darüber hinaus im Rahmen der probabilistischen Testtheorie für jede Skala eine Skalierung mit *mixed-Rasch-Modellen* vorgenommen (Rost et al., 1997). Die getesteten Modellannahmen sind jeweils das mixed-Rasch-Modell für polytome Antwortformate (Rost, 1991), wobei jeweils Modelle mit zwei bis vier latenten Klassen berechnet werden. Die Modell- und Parameterschätzung wird mit dem Programm *WinMira* (von Davier, 2001) nach der *Conditional Maximum Likelihood* Methode (CML) vorgenommen. Zur Beurteilung der relativen Modellpassung können die aus der Modellschätzung resultierenden informationstheoretischen Kriterien AIC und BIC (Akaike, 1974; Schwarz, 1978) der unterschiedlichen Modelle (Klassen-Lösungen) herangezogen werden. Nach Rost (2004) kann als Auswahlkriterium dieser beiden Kriterien der Modellpassung gelten, dass der AIC bei einer kleinen Anzahl von Items mit großen Patternhäufigkeiten und der BIC bei einer großen Anzahl von Items und kleinen Patternhäufigkeiten vorzuziehen ist. Aufgrund der Länge der für die einzelnen Dimensionen des AIST–R verwendeten Skalen (zehn Items pro Interessendimension bei fünf Antwortkategorien) und der mit drei bis vier Items vergleichsweise kurzen Skalen mit allerdings *sieben* Antwortkategorien wird hier für beide Konstrukte das *Bayes Information Criterion* (BIC – Schwarz, 1978) verwendet, welches das Einfachheitskriterium bei der Modellselektion bei *kleinen Patternhäufigkeiten*, im Verhältnis zu Anzahl der möglichen pattern, stärker berücksichtigt. Eine anschließende visuelle Inspektion der Schwellenparameterprofile der ausgewählten Modelle (Klassen-Lösungen),

welche sich aus den geschätzten Itemkategorieschwierigkeiten ergeben, dient dann der Qualifizierung der Antworttendenz in der jeweiligen latenten Klasse. Auf Basis der visuellen Inspektion der Schwellenparameterprofile wird dann für die entsprechenden Dimensionen jeweils eine dichotome *Indikatorvariable* gebildet. Als Kriterium zur Vergabe der beiden Werte (1 = mittlere Antworttendenz und 2 = extreme Antworttendenz) auf dieser Indikatorvariablen zu den jeweils gefundenen Personenklassen, werden die Grafiken der Schwellenparameter der jeweiligen Dimension und latenten Klasse herangezogen. Derjenigen Personenklasse, bei der eher große Abstände der Schwellenparameter vorliegen, wird der Wert 1 auf der Indikatorvariablen zugeordnet (Tendenz zur Mitte – MRS). Bei eher eng verlaufenden Abständen der Schwellenparameter wird dieser Personenklasse der Wert 2 zugeordnet (Tendenz zu den extremen Antwortkategorien – ERS). Mit diesen so neu gebildeten Indikatorvariablen wird im Anschluss überprüft, ob sich die Zuordnung der Personen zu den beiden Antworttendenzen in konsistenter Weise über diejenigen Dimensionen des AIST–R oder STOMP nachweisen lässt, für die sich die 2-Klassen-Lösung als das am besten passende Modell zur Beschreibung der Daten erwiesen hat. Die Indikatorvariablen werden dazu einer Latent-Class-Analysis (LCA) zweiter Ordnung unterzogen (vgl. Abschnitt 4.6.2 in Kapitel 4 *Psychometrische Modellierung*). Erwartet wird hierbei, dass sich bei einer konsistenten Antworttendenz (über unterschiedliche Skalen) eine 2-Klassen-Lösung für die LCA zweiter Ordnung ergibt.

Für den AIST–R werden diejenigen Skalen, welche sich auf Basis dieser Analysen als eindimensional (Personenhomogenität) erweisen, zur Itemkalibrierung zwei unterschiedliche Herangehensweisen zur Bestimmung der Item- bzw. Kategorieschwierigkeiten angewendet. Zum einen wird, wie bereits beschrieben, die in *WinMira* (von Davier, 2001) implementierte *Conditional Maximum Likelihood* Methode (CML) angewendet und zum anderen die in dem *R*-Paket `pairwise` (Heine, 2019) implementierte Methode des paarweisen Item- bzw. Itemkategorievergleichs (*PAIR*) angewendet (vgl. Abschnitt 4.5). Die weitere Prüfung der Angemessenheit einer eindimensionalen Skalierung erfolgt durch die Anwendung von Likelihood-basierten globalen Modelltests. Dabei wird die globale Modellpassung der eindimensionalen Skalierung durch den *Conditional Log–Likelihood* Test (Andersen, 1973b) überprüft. Als Teilungskriterium für

die Aufteilung der Analysedaten in zwei Teildatensätze wird hier der Median des *raw–scores* zur Aufteilung der Analysedaten herangezogen. Das Vorhandensein von lokalen Modellverletzungen auf der Ebene der Items wird anhand von *weighted-mean-square* Fit-Statistiken (*INFIT* und *OUTFIT* – T. G. Bond & Fox, 2015; Masters, 1982, vgl. auch Abschnitt 4.4.2) überprüft.

Ergebnisse

Skalierung des AIST–R

Die Überprüfung der Passung einer eindimensionalen Skalierung für alle Personen der vorliegenden Stichprobe (Personenhomogenität) erfolgte für die CML–Methode (Programm *WinMira*) durch die vergleichende Beurteilung der relativen Modellpassung von *mixed-Rasch-Modellen* (1–4 latente Klassen) anhand informationstheoretischer Kriterien. Der hier durchgeführte Modellvergleich anhand des *Bayes Information Criterion* (BIC – Schwarz, 1978) indiziert für vier der sechs Dimensionen des AIST–R eine 2-Klassen-Lösung (*Realistic*, *Artistic*, *Social* und *Conventional*). Für die Dimension *Investigative* ergibt sich die 3-Klassen-Lösung und für die Dimension *Enterprising* lässt sich die 1-Klassen-Lösung als das am besten passende Modell identifizieren (vgl. Tabelle 6.1).

Die jeweiligen relativen Klassengrößen (Klassenwahrscheinlichkeiten, p_g) der nach dem Kriterium BIC jeweils am besten passenden mixed-Rasch Modelle (mit 2 latenten Klassen) betragen für die Dimension *Realistic* $p_{g=1} = 0,496$; $p_{g=2} = 0,504$, die Dimension *Artistic* $p_{g=1} = 0.662$; $p_{g=2} = 0.338$, die Dimension *Social* $p_{g=1} = 0.473$; $p_{g=2} = 0.527$, die Dimension *Conventional* $p_{g=1} = 0.616$; $p_{g=2} = 0.384$ und für die 3-Klassen-Lösung der Dimension *Investigative* $p_{g=1} = 0.274$; $p_{g=2} = 0.410$; $p_{g=3} = 0.316$.

Die Überprüfung der Passung der eindimensionalen Skalierung auf Basis der mit der *PAIR*-Methode geschätzten Modellparameter erfolgt über den *Likelihood-Quotienten-Test* von Andersen (1973b). Für die jeweils am Median der summierten Itemscores aufgeteilte Stichprobe ergibt sich dabei lediglich für die AIST–R-Dimension *Enterprising* ein nicht signifikantes Ergebnis für den globalen Modelltests ($\chi^2 = 87.04; df = 79; p = 0.2510$). Für die fünf anderen Dimensionen ergibt sich jeweils ein signifikantes Ergebnis für den Andersen-Test. Dieser Befund weist im Einklang mit den Ergebnissen des relativen Mo-

Tabelle 6.1 Relativer Modellvergleich – Informationstheoretische Kriterien – für sechs Dimensionen des AIST–R.

Klassen	*Realistic*				*Investigative*			
	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-18049	79	36255	36666	-18428	79	37014	37425
2	-17634	157	35582	**36399**	-18048	157	36411	37227
3	-17550	235	**35571**	36793	-17713	235	**35896**	**37118**
4	-17550	313	35726	37353	-17699	313	36024	37652
	Artistic				*Social*			
	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-18678	79	37514	37924	-16841	79	33841	34252
2	-18142	157	36597	**37414**	-16287	157	32888	**33704**
3	-17867	235	**36204**	37427	-16091	235	32651	33873
4	-18558	313	37742	39370	-15990	313	**32607**	34234
	Enterprising				*Conventional*			
	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-16589	79	**33336**	**33747**	-17687	79	35531	35942
2	-16512	157	33338	34155	-17245	157	34805	**35621**
3	-16440	235	33350	34572	-17085	235	34641	35863
4	-16458	313	33541	35169	-16976	313	**34577**	36205

Anmerkungen: Mixed-Rasch-Modelle (PCM) für 1 – 4 latente Klassen; CML-Schätzung mit *WinMira*; LL = Log-Likelihood; np = Anzahl freie Modellparameter.

dellvergleichs über das Informationstheoretische Kriterium BIC darauf hin, dass hier das eindimensionale Skalierungsmodell nicht angemessen ist.

Für die AIST–R-Dimension *Investigative* ergibt sich die 3-Klassen-Lösung als das am besten passende Modell. Bei der Inspektion der entsprechenden Schwellenparameterprofile lässt sich die Klasse-3 eindeutig als Personengruppe mit einer Tendenz zu extremen Antwortkategorien identifizieren (vgl. Abbildung 6.3 unten). Die Personengruppe mit einer Tendenz zu eher mittleren Antwortkategorien teilt sich in zwei latente Klassen auf (Klasse-1 und Klasse-2), wobei sich diese Aufteilung auf die unterschiedlich ausfallenden, extremen Itemkategorieparameterschätzungen für zwei Items zurückführen lässt (vgl. Abbildung 6.3 Klasse-1 links und Klasse-2 rechts). Dies sind die Items

aist56: *„Herausfinden, was man mit einem Computerprogramm alles tun kann“* (Schwelle zwischen den beiden untersten Ablehnungskategorien) und *aist32*: *„Über längere Zeit an der Lösung eines Problems arbeiten“* (Schwelle zwischen den beiden obersten Zustimmungskategorien).

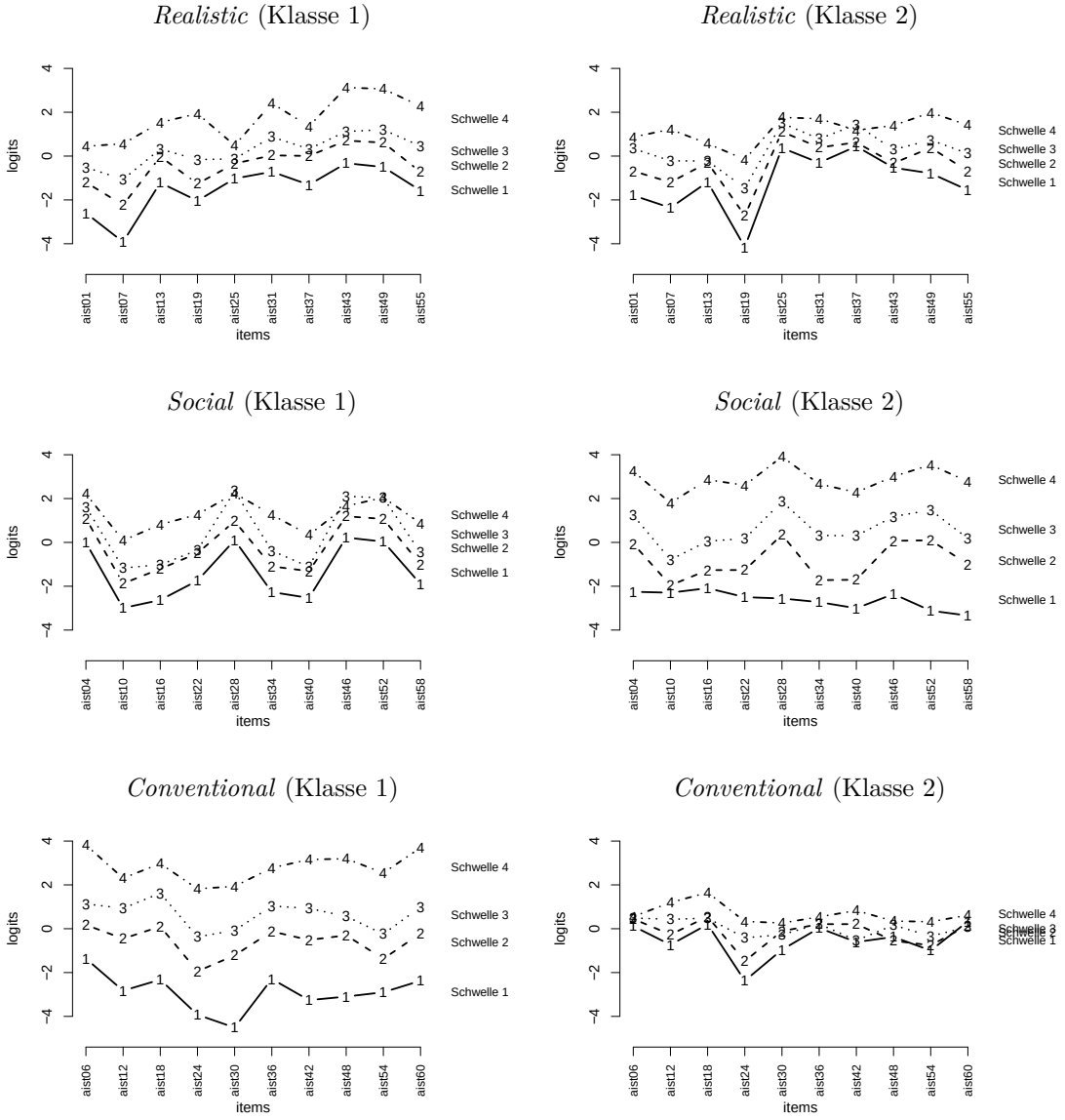


Abbildung 6.1 Darstellung der Schwellenparameterprofile der 2-Klassen-Lösung (*WinMira*) für die AIST–R-Skalen *Realistic* (oben), *Social* (mitte) und *Conventional* (unten); jeweils Klasse-1 links und Klasse-2 rechts.

Für die drei Dimensionen *Realistic*, *Social* und *Conventional* lässt sich je-

weils eine der beiden Klassen der 2-Klassen-Lösung anhand deren Schwellenparameterprofile eindeutig als Personengruppe mit eher extremen Antwortverhalten klassifizieren (vgl. Abbildung 6.1 *Realistic* Klasse-2 , *Social* Klasse-1 und *Conventional* Klasse-1). Die jeweils andere Klasse lässt sich anhand des Profils der Schwellenparameter als Personengruppe mit einer Tendenz zu eher mittleren Antwortkategorien klassifizieren.

Bei der Skalierung der Dimension *Artistic* ergibt sich die 2-Klassen-Lösung als das am besten passende Modell. Die Inspektion der beiden Schwellenparameterprofile (vgl. Abbildung 6.2 oben) zeigt, dass sich diese Aufteilung in zwei latente Klassen aus den unterschiedlichen Itemkategorieschwierigkeiten des Items *aist21*: „*Dinge tun, bei denen es auf Kreativität und Phantasie ankommt*“, resultiert.

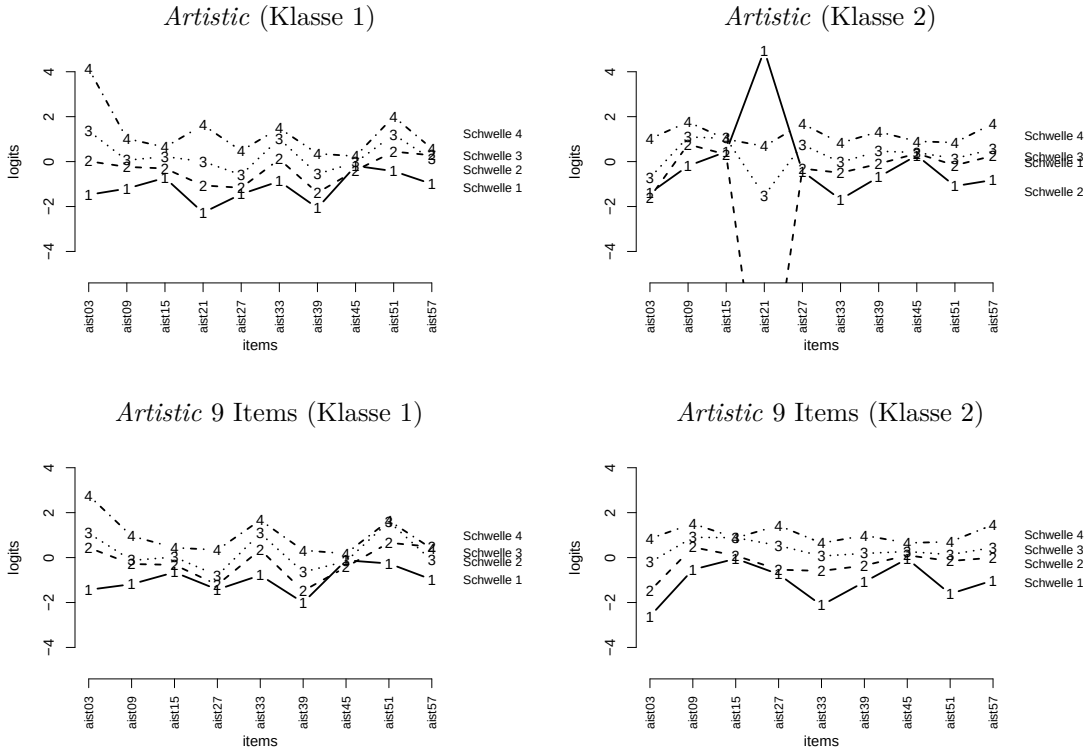


Abbildung 6.2 Darstellung der Schwellenparameterprofile der 2-Klassen-Lösung (*WinMira*) für die AIST–R-Skala *Artistic* (Skala mit 10 Items oben und Skala mit 9 Items unten, nach entfernen des Items *aist21*: *'Dinge tun, bei denen es auf Kreativität und Phantasie ankommt'*).

Diese extremen Werte für die Itemparameter aus der CML-Schätzung in

der Klasse-2 für das Item *aist21* der Dimension *Artistic* deuten auf Schätzprobleme (vgl. Luo & Andrich, 2005) bei der Bestimmung der Itemparameter in dieser latenten Klasse hin. Im Hinblick auf die Zielsetzung der vorliegenden Untersuchung – Identifikation von unterschiedlichen Antworttendenzen der antwortenden Personen – wurde das betreffende Item aus der Skala entfernt und eine erneute Skalierung vorgenommen. Dabei resultiert nach dem Kriterium BIC erneut eine 2-Klassen-Lösung. Die Schwellenparameterprofile der beiden Klassen nach Skalierung mit dem 2 Klassen mixed-Rasch-Modell für die nun neun Items umfassende Skala weisen keine ausgeprägten Unterschiede hinsichtlich der Schwellenabstände auf (vgl. Abbildung 6.2 unten). Aus diesem Grund können hier für die Dimension *Artistic* keine eindeutigen Zuweisungen der latenten Klassen zu den beiden unterschiedlich ausgeprägten Antworttendenzen vorgenommen werden.

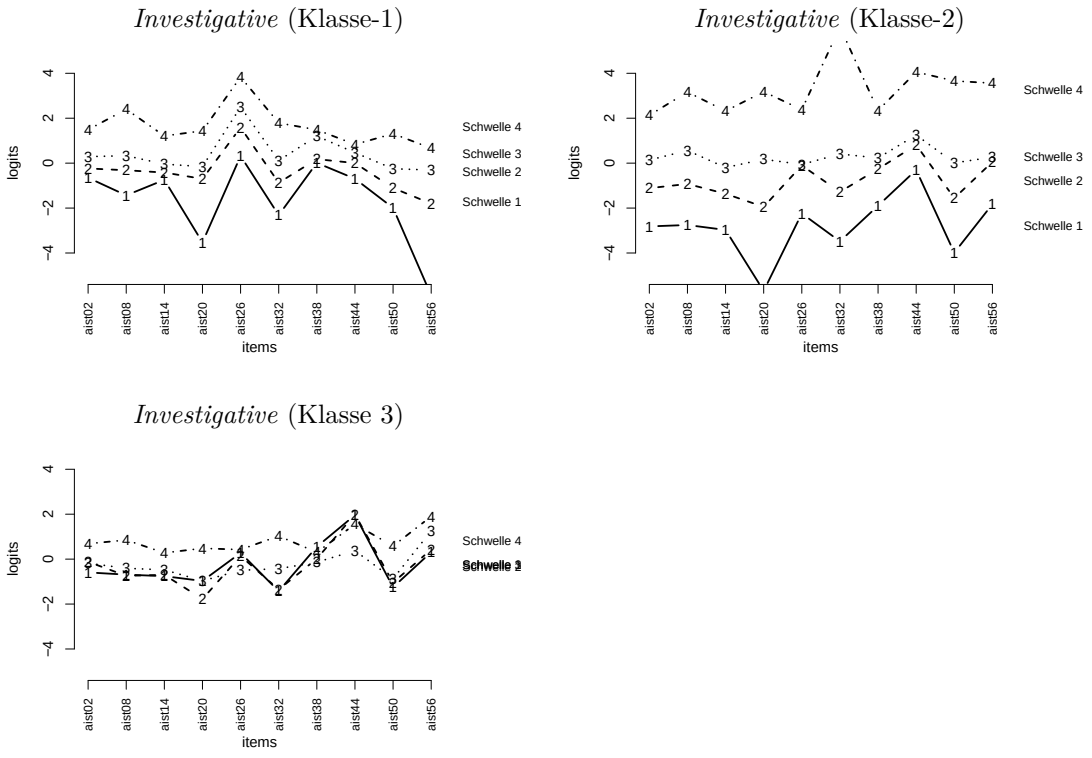


Abbildung 6.3 Darstellung der Schwellenparameterprofile der 3-Klassen-Lösung (*WinMira*) für die AIST–R-Skala *Investigative*; Klasse 1 links, Klasse 2 rechts, Klasse 3 unten.

Eine zusammenfassende Übersicht zur Anzahl latenter Klassen des jeweils

am besten passenden Modells, dessen jeweilige Klassengrößen, sowie deren inhaltliche Interpretation anhand der Inspektion der Schwellenparameterprofile im Hinblick auf die Antworttendenz, findet sich in der Tabelle 6.2.

Tabelle 6.2 Übersicht zur Klassifikation der Antworttendenzen anhand der Schwellenparameterprofile – für sechs Dimensionen des AIST–R.

Realistic			*Investigative*		
Klasse	p_{Klasse}[1]	Antworttendenz[2]	Klasse	p_{Klasse}[1]	Antworttendenz[2]
1.	0.49596	*mittel*	1.	0.27371	*mittel*
2.	0.50404	*extrem*	2.	0.41058	*mittel*
–		–	3.	0.31571	*extrem*

Artistic[3]			*Social*		
Klasse	p_{Klasse}[1]	Antworttendenz[2]	Klasse	p_{Klasse}[1]	Antworttendenz[2]
1.	0.56974[3]	*nicht differenzierbar*	1.	0.47307	*extrem*
2.	0.43026[3]	*nicht differenzierbar*	2.	0.52693	*mittel*

Enterprising[4]			*Conventional*		
Klasse	p_{Klasse}[1]	Antworttendenz[2]	Klasse	p_{Klasse}[1]	Antworttendenz[2]
1.	1	–	1.	0.61629	*mittel*
–	–	–	2.	0.38371	*extrem*

Anmerkungen: [1] Relative Klassengröße; [2] Antworttendenz nach Interpretation der Schwellenparameterprofile; [3] nach erneuter Skalierung nach entfernen des Items *aist21*: *'Dinge tun, bei denen es auf Kreativität und Phantasie ankommt'*; [4] 1-Klassen-Lösung; *Reaistic*, *Investigative Social* und *Conventional*: $n = 1341$; *Artistic* und *Enterprising*: $n = 1340$.

Die Kalibrierung der Items unter der Annahme eines eindimensionalen (kumulativen) Skalierungsmodells (PCM – Masters, 1982) der AIST–R-Skala *Enterprising* resultiert für beide Methoden der Itemparameterbestimmung (*CML* und *PAIR*) in vergleichbaren Ergebnissen (vgl. Abbildung 6.4).

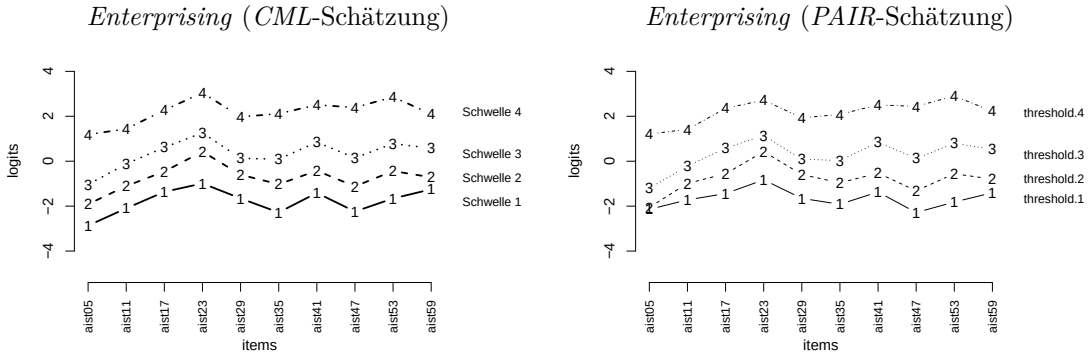


Abbildung 6.4 Darstellung der Schwellenparameterprofile der eindimensionalen Skalierung für die AIST–R-Skala *Enterprising*; *CML*-Schätzung links und *PAIR*-Schätzung rechts.

Die Inspektion des Personen-Itemparameter-Plots (Wright–Map) für die Dimension *Enterprising* des AIST–R zeigt zunächst, bezogen auf die mittlere Itemschwierigkeit (*item location*), ein vergleichsweise ausgewogenes *'Test-Targeting'* an (vgl. Abbildung 6.5). Eine genauere Betrachtung der Itemkategorieschwierigkeiten ergibt, dass lediglich die Schwierigkeiten für einzelne Itemkategorien auf der Logitmetrik (vgl. Abbildung 6.5 rechts) im Vergleich zur Verteilung der Personenparameter (vgl. Abbildung 6.5 links) an den äußeren Rändern dieser Verteilung liegen.

Aufgrund des guten Targeting (Passung) zwischen den Itemschwierigkeiten und der Verteilung der Merkmalsausprägungen der Personen der vorliegenden Stichprobe, stehen die *INFIT* (*inlier-sensitive*) Item-Fit-Statistiken im Vordergrund der Betrachtungen (vgl. Tabelle 6.3). Folgt man den in T. G. Bond und Fox (2015) gegebenen Empfehlungen zur Interpretation der Koeffizienten der (unstandardisierten) *mean square* Fit-Statistiken (*INFIT* und *OUTFIT* – *vgl. Abschnitt 4.4.2*) so zeigt sich, dass lediglich das Item *aist23*: *für eine Sache Werbung betreiben* zu lokalen Modellverletzung des eindimensionalen Skalierungsmodells führt. Demgegenüber indizieren aber die z-standardisierten

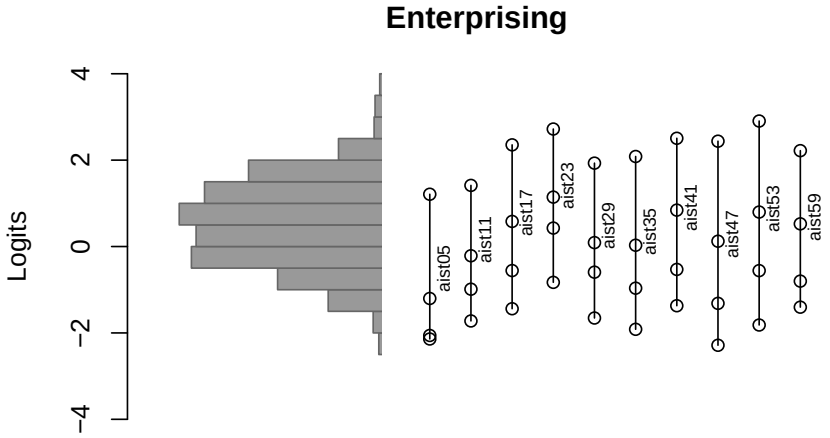


Abbildung 6.5 Darstellung des Personen- Itemparameter Plots (Wright-Map) nach eindimensionaler Skalierung der AIST–R-Skala *Enterprising* (*PAIR*–Algorithmus).

Tabelle 6.3 Item-Fit-Statistiken für die AIST–R-Skala *Enterprising* (*PAIR*-Algorithmus).

	χ^2	*df*	*p*	OUTFIT.MSQ	OUTFIT.ZSTD	INFIT.MSQ	INFIT.ZSTD
aist23	1564.47	1339.00	0.00	1.25	5.95	1.21	5.51
aist59	1084.95	1339.00	1.00	0.89	−2.92	0.90	−2.76
aist53	1090.29	1339.00	1.00	0.89	−2.91	0.90	−2.71
aist17	1079.11	1339.00	1.00	0.89	−3.09	0.90	−2.69
aist11	1411.68	1339.00	0.08	1.13	3.02	1.09	2.32
aist35	1370.62	1339.00	0.27	1.10	2.52	1.09	2.26
aist47	1141.03	1339.00	1.00	0.93	−1.76	0.94	−1.59
aist41	1159.70	1339.00	1.00	0.95	−1.44	0.96	−1.13
aist29	1275.54	1339.00	0.89	1.03	0.80	1.04	1.05
aist05	1151.07	1339.00	1.00	0.94	−1.37	0.96	−0.82

Anmerkungen: Item-Fit-Statistiken nach eindimensionaler Skalierung mit *PAIR*-Algorithmus; MSQ = mean-square; ZSTD = z-standardisiert; $n = 1340$.

Koeffizienten für einige Items signifikanten Modellabweichungen (vgl. Tabelle 6.3).

Die übergreifende Klassifikation der Antworttendenzen über die einzelnen Dimensionen des AIST–R hinweg, beziehen sich auf die Dimensionen *Realistic*, *Investigative*, *Social* und *Conventional*, für die sich in den vorangegangenen Schritten der Auswertung eine klare Zuordnung der latenten Klassen zu den beiden Antworttendenzen anhand der Schwellenparameterprofile ergeben hat. Mit der daraus resultierenden Datenmatrix mit vier dichotomen Indikatorvariablen wird die LCA zweiter Ordnung berechnet, wobei jeweils eine bis fünf latente Klassen als Modellparameter vorgegeben werden. Die Ergebnisse des relativen Modellvergleichs anhand der informationstheoretischen Kriterien indizieren hier (übereinstimmend) für die LCA zweiter Ordnung die 3-Klassen-Lösung als das am besten passende Modell (vgl. Tabelle 6.4)

Tabelle 6.4 Relativer Modellvergleich – Informationstheoretische Kriterien – für LCA zweiter Ordnung über vier Indikatorvariablen.

Klassen	R I S C[a]			
	LL[b]	np[c]	AIC	BIC
1	-3565.50	4	7139.00	7159.80
2	-3455.37	9	6928.73	6975.54
3	-3418.20	14	**6864.40**	**6937.20**
4	-3415.78	19	6869.56	6968.37
5	-3415.78	24	6879.39	7004.20

Anmerkungen: Latente-Klassen Analyse (LCA) für 1 – 4 latente Klassen; CML-Schätzung mit *WinMira*; [a] vier (dichotome) Indikatorvariablen: R–*Realistic*, I–*Investigative*, S–*Social*, C–*Conventional* [b] LogLikelihood;
[c] Anzahl freie Modellparameter; $n = 1340$.

Die Kreuztabellierung des Klassifikationsergebnisses aus der LCA zweiter Ordnung mit den 16 (2^4) möglichen Mustern (*pattern*) der vier Indikatorvariablen ist in Tabelle 6.5 wiedergegeben. Es zeigt sich, dass sich die drei latenten Klassen jeweils durch typische Muster (*pattern*) der Antworttendenz auf den vier untersuchten Dimensionen des AIST–R charakterisieren lassen. So werden Personen mit einer mittleren Antworttendenz auf allen vier Dimensionen (R–*Realistic*, I–*Investigative*, S–*Social*, C–*Conventional*) der Klasse 1 zugeordnet, wohingegen Personen mit einer extremen Antworttendenz auf diesen vier Dimensionen der Klasse 2 zugeordnet werden. Die Klasse 3 wird durch die *pattern* „1121“, „1122“, „1221“ sowie „1222“ charakterisiert.

Tabelle 6.5 Kreuztabellierung der Personenklassifikation durch die LCA zweiter Ordnung mit Mustern der Indikatoren zur Antworttendenz für vier AIST–R-Dimensionen.

p^a	0.93	0.56	0.57	0.90	0.62	0.54	0.66	0.75	0.80	0.68	0.46	0.73	0.87	0.98	0.93	0.97
	Muster der Indikatorvariablen (R I S C)[b]															
Klassen	1111	1112	1121	1122	1211	1212	1221	1222	2111	2112	2121	2122	2211	2212	2221	2222
1	245	48	0	0	25	0	0	0	170	0	77	0	0	0	0	0
2	0	0	0	0	0	15	0	0	0	61	0	55	85	70	66	93
3	0	0	154	112	0	0	28	36	0	0	0	0	0	0	0	0

Anmerkungen: CML-Schätzung mit *WinMira*; [a] mittlere Klassenzuordnungswahrscheinlichkeit [b] vier (dichotome) Indikatorvariablen: R–*Realistic*, I–*Investigative*, S–*Social*, C–*Conventional* – **2 = Extreme Antworttendenz** und **1 = Mittlere Antworttendenz**; $n = 1340$.

Diese *pattern* bzw. Muster der Indikatorvariablen für die Antworttendenz lassen sich, bezogen auf die vier untersuchten AIST–R-Skalen, als eher *inkonsistente* Antworttendenz beschreiben. Dies trifft insbesondere für die Muster „1122“ und „1221“ zu. Die mittlere Zuordnungswahrscheinlichkeit aller Personen mit dem Muster „1122“ zur Klasse 3 beträgt hier $p = .90$. Berücksichtigt man die absoluten Häufigkeiten der in Klasse 3 vertretenen pattern, so lässt sich, trotz eigentlich *inkonsistenter* Antworttendenz eine Tendenz zu eher mittleren Antwortkategorien feststellen – vgl. pattern „1121“ mit der höchsten absoluten Häufigkeit von $n = 154$ Personen mit diesem Muster der Antworttendenz auf den vier untersuchten AIST–R-Dimensionen in dieser latenten Klasse.

Die Personen mit den jeweils häufigsten Mustern in den beiden latenten Klassen 1 und 2 zur Antworttendenz weisen hier auch die höchste mittlere Zuordnungswahrscheinlichkeit zur jeweiligen Klasse auf ($p = .93$ für *pattern* „1111“ in Klasse 1 und $p = .97$ für *pattern* „2222“ in Klasse 2).

Skalierung des STOMP

Die informationstheoretischen Kriterien zur modellvergleichenden Skalierung der vier Dimensionen des STOMP mit dem mixed-Rasch-Modell mit einer unterschiedlichen Anzahl von latenten Klassen ist in Tabelle 6.6 angegeben.

Tabelle 6.6 Relativer Modellvergleich – informationstheoretische Kriterien – für vier Dimensionen des STOMP.

Klassen	*Reflective & Complex*				*Intense & Rebellious*			
	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-9363	47	18820	19064	-6852	35	13774	**13956**
2	-9189	93	18564	**19048**	-6765	69	13668	14027
3	-9060	139	18399	19121	-6724	103	**13655**	14190
4	-8984	185	**18338**	19300	-6706	137	13685	14397
	Upbeat & Conventional				*Energetic & Rhythmic*			
	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-6772	35	**13614**	**13796**	-7545	35	15160	15342
2	-6746	69	13631	13989	-7359	69	14855	**15214**
3	-6710	103	13626	14162	-7300	103	14806	15341
4	-6695	137	13665	14377	-7265	137	**14805**	15517

Anmerkungen: Mixed-Rasch-Modelle (PCM) für 1 – 4 latente Klassen; CML-Schätzung mit *WinMira*; LL = LogLikelihood; np = Anzahl freie Modellparameter.

Anhand des informationstheoretischen Kriteriums BIC ergeben sich demnach für die Dimensionen *Upbeat & Conventional* und *Intense & Rebellious* jeweils die 1-Klassen-Lösung als das am besten passende Modell. Für die beiden Dimensionen *Reflective & Complex* und *Energetic & Rhythmic* muss das eindimensionale Skalierungsmodell für den Dominanz-Antwortprozess zugunsten einer 2-Klassen-Lösung, verworfen werden.

Die Abbildung 6.6 zeigt die jeweiligen Schwellenparameterprofile für die 2-Klassen-Lösung der beiden Dimensionen *Reflective & Complex* und *Energetic & Rhythmic* des STOMP. Lediglich für die Dimension *Energetic & Rhythmic* lassen sich die beiden Klassen anhand der Schwellenparameterprofile eindeutig einer der beiden Antworttendenzen zuordnen – mittlere Antworttendenz (MRS) und extreme Antworttendenz (ERS). So lässt sich für die Dimension *Energetic & Rhythmic* die Klasse-1 mit einer Klassengröße von $p_{g=1} = 0.279$ anhand des Schwellenparameterprofils als *„Mittelkreuzer“* identifizieren. Dem-

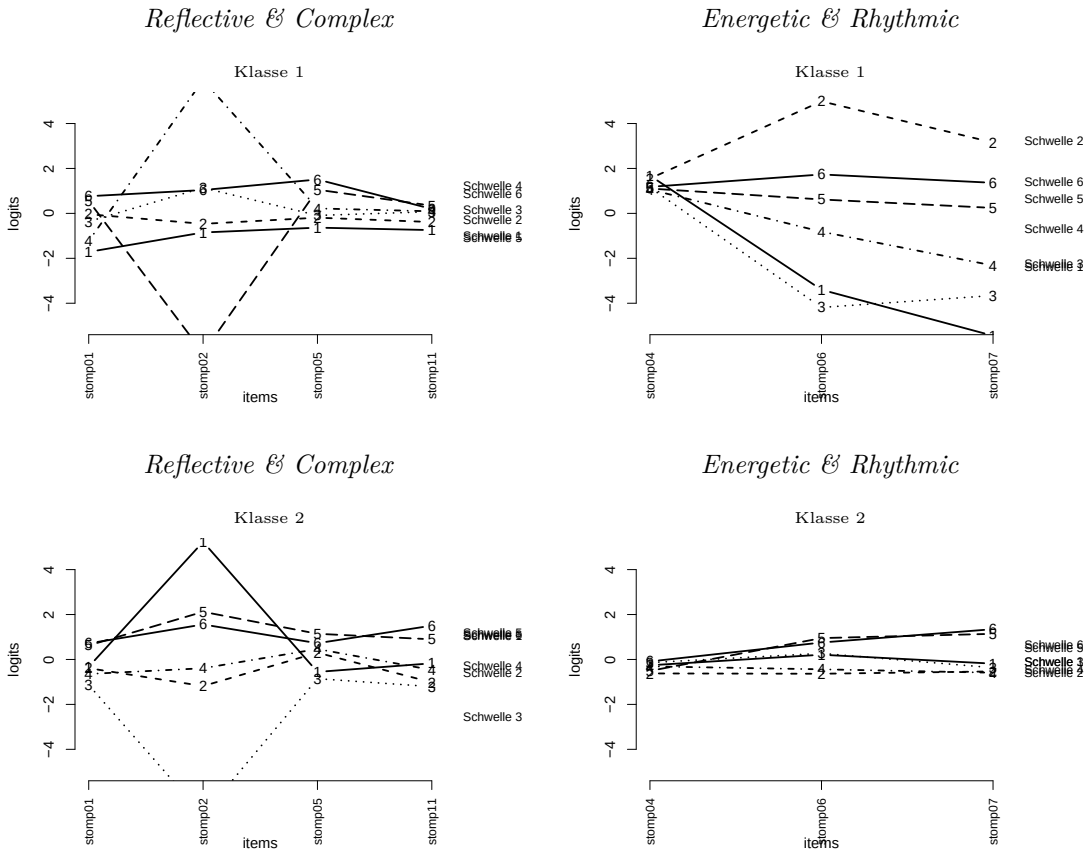


Abbildung 6.6 Darstellung der Schwellenparameter Profile der 2-Klassen-Lösung (*WinMira*) für die STOMP-Skalen *Reflective & Complex* (links) und *Energetic & Rhythmic* (rechts).

gegenüber weist die Klasse-2 mit einer Größe von $p_{g=2} = 0.721$ eher eine Tendenz zu extremen Antwortkategorien auf.

Diskussion

Mit der hier berichteten Untersuchung wird zunächst der Frage nachgegangen, ob sich das Phänomen unterschiedlich ausgeprägter Antworttendenzen zu mittleren oder extremen Antwortkategorien auf mehrstufigen Antwortskalen (ERS vs. MRS) auch für die einzelnen Dimensionen des AIST–R und STOMP nachweisen lässt. Zusätzlich wird die Äquivalenz der beiden Itemparameterschätzmethoden *CML* und *PAIR* überprüft. Durch die Anwendung einer LCA zweiter Ordnung auf die (dichotomen) Indikatorvariablen zur jeweiligen Ant-

worttendenz auf vier der sechs Dimensionen des AIST–R wird die Frage nach der Konsistenz der Antworttendenz innerhalb des Konstrukts *berufliche Interessenorientierungen* untersucht.

Der relative Modellvergleich anhand des informationstheoretischen Kriteriums BIC (Schwarz, 1978) legt nahe, dass sich lediglich für die AIST–R-Dimension für die unternehmerische Orientierung (*Enterprising*) das eindimensionale Skalierungsmodell als passend erweist (vgl. Tabelle 6.1). Für alle anderen Dimensionen des AIST–R muss die Annahme der Personenhomogenität der untersuchten Stichprobe dagegen verworfen werden. Für die praktisch technische Orientierung (*Realistic*), die soziale Orientierung (*Social*) sowie für die konventionelle Orientierung (*Conventional*) erweist sich die 2-Klassen-Lösung als das am besten passende Modell. Die Interpretation der Schwellenparameter-Plots deutet hierbei auf unterschiedlich ausgeprägte Itemkategorieschwierigkeiten in den beiden latenten Personengruppen hin. In Übereinstimmung mit den Befunden anderer Untersuchungen zum unterschiedlichen Gebrauch mehrstufig ordinaler Antwortskalen (vgl. z. B. Arce-Ferrer, 2006; E. Austin et al., 1998; E. J. Austin et al., 2006; Gollwitzer et al., 2005; Meisenberg & Williams, 2008; Naemi et al., 2009; Rost et al., 1999) weisen die Unterschiede in den Itemkategorieschwierigkeiten darauf hin, dass innerhalb der untersuchten Gesamtstichprobe unterschiedliche Antworttendenzen bei der Bearbeitung der einzelnen Fragen vorliegen. Dieser allgemeine, hier auch für den Bereich der *beruflichen Interessenorientierungen* gezeigte Befund stützt die Vermutung, dass sich die unterschiedlich ausgeprägten Antworttendenzen im Sinne einer Präferenz für entweder extreme oder aber mittlere Antwortkategorien weitgehend unabhängig von den Inhalten einzelner Dimensionen bzw. den gesamten Konstrukten zeigen.

Die Ergebnisse zur Untersuchung der Konsistenz der Antworttendenzen auf den vier AIST–R-Dimensionen zeigen zunächst ein nicht ganz eindeutiges Bild. So indizieren die Ergebnisse der LCA zweiter Ordnung hier eine 3-Klassen-Lösung (im Vergleich zu einer 2-Klassen-Lösung) als das am (relativ) besten passende Modell. Allerdings zeigt die ergänzend durchgeführte Kreuztabellierung des Klassifikationsergebnisses aus der LCA mit den 16 (2^4) möglichen *„pattern“* (Mustern) der vier Indikatorvariablen (vgl. Tabelle 6.5), dass die drei Klassen über die jeweils konstituierenden pattern vergleichsweise ein-

deutig entweder einer mittleren oder extremen Antworttendenz auf den vier AIST–R-Skalen zuordnen lassen (vgl. Tabelle 6.5). Insgesamt kann aus diesen Befunden gefolgert werden, dass sich die beiden unterschiedlichen Antworttendenzen (MRS vs. ERS) als weitgehend konsistentes Personenmerkmal über die vier AIST–R-Skalen beschreiben lassen.

Für den STOMP erweist sich für die beiden Dimensionen *Reflective & Complex* und *Energetic & Rhythmic* jeweils die 2-Klassen-Lösung als das am besten passende Modell. Davon lassen sich für die Dimension *Energetic & Rhythmic* die beiden Klassen anhand der Schwellenparameterprofile jeweils einer mittleren Antworttendenz (MRS) oder extremen Antworttendenz (ERS) zuordnen. Die Schwellenparameterprofile der beiden latenten Klassen der Dimension *Reflective & Complex* weisen für das Item *stomp*02 (Blues) jeweils auffällige Schwellenparametervertauschungen auf (vgl. Abbildung 6.6 links). Inhaltlich interpretiert geht die 2-Klassen-Lösung für die Dimension *Reflective & Complex* daher weniger auf die über alle Items konsistent unterschiedliche Interpretation der siebenstufigen Antwortskala zurück. Vielmehr bestehen innerhalb der analysierten Stichprobe offenbar unterschiedliche Auffassungen bezüglich der Präferenz bzw. polaren Zuordnung des Items *Blues* zu der Dimension *Reflective & Complex*, was auf der Ebene der mixed-Rasch Analysen eine 2-Klassen-Lösung bedingt.

6.2 Untersuchung zur Skalierbarkeit des BFI–K

Einleitung

Mit der zweiten Studie soll, in Analogie zu den Analysen in der ersten Studie in Abschnitt 6.1, die Skalierbarkeit der fünf BFI–K-Skalen nach einem kumulativen Antwortmodell, untersucht werden. Das bereits in Abschnitt 6.1 dargestellte Phänomen unterschiedlich ausgeprägter Antworttendenzen bei mehrstufig ordinalen Antwortskalen für das Konstrukt Persönlichkeit nach dem Big-Five-Paradigma (z. B. Piedmont & Aycock, 2007) erweist sich dabei als häufiger Befund empirischer Untersuchungen (z. B. E. J. Austin et al., 2006; Berg & Collier, 1953; Damarin & Messick, 1965; Eid & Zickar, 2007; Heine, 2010; Minkov, 2017; Naemi et al., 2009; Rost, 2002; Rost et al., 1997, 1999; Warr & Coffman, 1970, sowie Abschnitt 3.2.4 in Kapitel 3. *Theoretischer Hintergrund zu Antwortmustern*). Obwohl oft belegt, besteht in der Literatur durchaus Uneinigkeit bezüglich der Bedeutung von derartigen Antworttendenzen (z. B. Damarin & Messick, 1965; Hurley, 1998; Meisenberg & Williams, 2008; Plieninger, 2017; Rorer, 1965; Rundquist, 1966). Darüber hinaus ergeben sich bei empirischen Untersuchungen an unterschiedlichen Stichproben und unterschiedlichen Inventaren immer wieder uneinheitliche Befunde hinsichtlich der Persönlichkeitsdimensionen, für die das Phänomen der unterschiedlich ausgeprägten Antworttendenz gefunden wird (vgl. E. J. Austin et al., 2006; Berg & Collier, 1953; Heine, 2010; Minkov, 2017; Naemi et al., 2009; Rost et al., 1999; Warr & Coffman, 1970). Mit der vorliegenden Studie soll daher für den hier eingesetzten, modifizierten BFI–K untersucht werden, ob sich das Phänomen unterschiedlich ausgeprägte Antworttendenzen (ERS vs. MRS) auch in den hier vorliegenden Stichproben zu beobachten ist.

Daten

Analysiert werden die fünf Skalen des BFI–K zur Erfassung von fünf Dimensionen der Persönlichkeit – *Neurotizismus*, *Extraversion*, *Offenheit*, *Verträglichkeit* und *Gewissenhaftigkeit* (vgl. Tabelle 2.1 in Abschnitt 2.1.3). Zur Überprüfung der Skalierbarkeit werden die Antwortdaten der beiden Stichproben aus den beiden Erhebungszeiträumen (vgl. Abschnitt 5.2) zunächst getrennt analysiert.

Für jede der beiden Stichproben ergeben sich fünf Datenmatrizen zu den fünf Persönlichkeitsdimensionen. Diese umfassen jeweils $k = 4$ Items, wobei in jeder Skala positiv und negativ formulierte Items enthalten sind (vgl. Tabellen 5.6 und 5.10 in Abschnitt 5.2). Wie bei den Analysen in der ersten Studie in Abschnitt 6.1 werden aus den jeweils fünf Datenmatrizen diejenigen Fälle herausgenommen, welche alle Items der betreffenden Dimension nicht beantwortet haben (*unit-non-responder*). Aufgrund des verbindlichen Erhebungsmodus bei der Stichprobe I reduziert sich der Datenumfang daher nicht, sodass hier für alle fünf Dimensionen $n = 734$ Fälle zur Skalierung herangezogen werden können. Für die Stichprobe II reduziert sich der Datenumfang durch unit-non-responder für die Dimensionen *Neurotizismus*, *Extraversion* und *Offenheit* auf $n = 607$, für die Dimension *Verträglichkeit* auf $n = 606$ und für die Dimension *Gewissenhaftigkeit* auf $n = 608$ Fälle.

Methode

Die Analysen zur Skalierbarkeit der einzelnen Skalen des BFIK–K werden sowohl einzeln für die beiden Stichproben getrennt und auch als Gesamtdatensatz (Stichprobe I und II kombiniert) durchgeführt. Die Möglichkeit zur getrennten Analyse der einzelnen Stichproben ergibt sich durch die im Vergleich zum AIST–R in Studie 6.1 kürzeren Skalen des BFI–K. Wie bereits in Abschnitt 5.1 beschrieben, umfasst der BFI–K für jede der fünf Dimensionen lediglich 4 Items – wobei jedes Item fünf Antwortkategorien aufweist. Dadurch ergeben sich hier bei den Analysen für jede BFI–K-Dimension lediglich $m^k = 5^4 = 625$ mögliche Antwortmuster. Das Verhältnis zwischen der Anzahl der möglichen Antwortmuster und dem jeweiligen Stichprobenumfang n (der Stichproben) fällt daher im Hinblick auf die inferenzstatistische Bewertung der Skalierbarkeit der einzelnen Skalen im Vergleich zu den Analysen in der Studie 6.1 zur Skalierung des AIST–R wesentlich günstiger aus. Dieses zunächst getrennte Vorgehen erlaubt so eine Replikation bzw. Kreuzvalidierung der erzielten Ergebnisse zur kumulativen Skalierbarkeit der fünf Skalen des BFI–K.

Analog zu den Analysen für den AIST–R in Abschnitt 6.1 werden die fünf Skalen des BFI–K zunächst hinsichtlich ihrer eindimensionalen Skalierbarkeit untersucht. Wie in der ersten Studie wird dazu die Modellpassung über den Vergleich des eindimensionalen *Partial Credit* Modells (Masters, 1982) mit

mixed-Rasch-Modellen (vgl. Rost, 1990; Rost et al., 1997) mit dem Programm *WinMira* (von Davier, 2001) realisiert. Als Entscheidungsgrundlage zur Wahl des jeweils angemessenen Skalierungsmodells werden, ebenso wie in Abschnitt 6.1, die unterschiedlichen Modelle (Klassen-Lösungen) anhand des informationstheoretischen Kriteriums BIC (Schwarz, 1978) miteinander verglichen.

Zur näheren Qualifizierung von sich dabei unter Umständen ergebenden latenten Personengruppen werden die nach ihrer maximalen Klassenzuordnungswahrscheinlichkeit aufgeteilten Teilgruppen im Hinblick auf ihre jeweils vorherrschenden Antwortmuster untersucht. Dazu wird auf die in Abschnitt 4.6.1 in Kapitel 4 beschriebene *zwei Gruppen KFA* zurückgegriffen.

Vor dem Hintergrund der Befunde der getrennten Analyse der beiden Stichproben, werden die Daten aus beiden Stichproben schließlich als gemeinsamer Datensatz mit dem *R*-Paket `pairwise` (Heine, 2019) skaliert. Durch die Anwendung der dort implementierten, nichtiterativen Itemparameterschätzmethode (vgl. Heine & Tarnai, 2015) für die kombinierte Datenmatrix aus den beiden Stichproben, soll überprüft werden, ob hierbei stabile Itemparameter, auch vor dem Hintergrund unterschiedlicher Antwortmuster in den beiden Stichproben, zu erzielen sind. Auf Grundlage der dadurch bestimmten Itemparameter werden für alle Personen der beiden Stichproben über die gewichtete Likelihood-Schätzung (Warm, 1989) Personenparameter für die fünf Dimensionen des BFI–K geschätzt.

Ergebnisse

Die Beurteilung der eindimensionalen Skalierbarkeit der einzelnen Skalen des BFI–K stützt sich auf die vergleichende Betrachtung der angewendeten Skalierungsmodelle. Der anhand des Bayesschen informationstheoretischen Kriteriums (BIC – Schwarz, 1978) vorgenommene relative Modellvergleich zwischen der eindimensionalen Skalierung und der Modellierung mit *mixed-Rasch-Modellen* mit einer unterschiedlichen Anzahl latenter Personengruppen, indiziert für die Daten der ersten Stichprobe für alle Dimensionen des BFI–K eine 1-Klassen-Lösung. Für die zweite Stichprobe erweist sich lediglich für die Dimension *Extraversion* die 2-Klassen-Lösung anhand des BIC als das besser passende Modell (vgl. Tabelle 6.7). Die relativen Klassengrößen (Klassenwahrscheinlichkeiten, p_g) des nach dem Kriterium BIC am besten passenden

mixed-Rasch Modells (mit 2 latenten Klassen) fallen mit $p_{g=1} = 0.571$ und $p_{g=2} = 0.429$ etwa gleich groß aus.

Die Inspektion der Schwellenparameter-Plots aus den Ergebnissen der Skalierung zeigt für die Dimensionen *Neurotizismus*, *Offenheit*, *Verträglichkeit* und *Gewissenhaftigkeit* auffallende Unterschiede zwischen den beiden Stichproben. So ergeben sich für diese vier Dimensionen bei der Stichprobe I weitgehend parallele Linienzüge ohne Schwellenvertauschungen (vgl. Abbildungen 6.7 und 6.8, jeweils linke Seite). Demgegenüber weisen die Linienzüge der Schwellenparameter-Plots für die Stichprobe II jeweils bei einigen Items Vertauschungen der Itemkategorieschwellen auf (vgl. Abbildungen 6.7 und 6.8, jeweils rechte Seite).

Tabelle 6.7 Relativer Modellvergleich über Informationstheoretische Kriterien für fünf Dimensionen des BFI–K.

	Stichprobe I				Stichprobe II			
Klassen	Neurotizismus ($n = 734$)				Neurotizismus ($n = 607$)			
	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-3862	31	7785	**7928**	-3279	31	6620	**6757**
2	-3819	61	7759	8040	-3193	61	6508	6777
3	-3780	91	7743	8161	-3162	91	6507	6908
4	-3745	121	7732	8289	-3150	121	6541	7074
Klassen	Extraversion ($n = 734$)				Extraversion ($n = 607$)			
	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-3504	31	7071	**7213**	-3080	31	6223	6359
2	-3452	61	7026	7306	-2979	61	6079	**6348**
3	-3428	91	7038	7457	-2942	91	6066	6467
4	-3404	121	7049	7606	-2936	121	6115	6648
Klassen	Offenheit ($n = 734$)				Offenheit ($n = 607$)			
	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-3797	31	7656	**7798**	-3340	31	6742	**6879**
2	-3726	61	7574	7854	-3276	61	6675	6943
3	-3750	91	7682	8100	-3221	91	6624	7025
4	-3680	121	7603	8159	-3184	121	6610	7144
Klassen	Verträglichkeit ($n = 734$)				Verträglichkeit ($n = 606$)			
	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-3779	31	7620	**7762**	-3184	31	6429	**6566**
2	-3747	61	7615	7896	-3143	61	6407	6676
3	-3715	91	7612	8030	-3111	91	6403	6804
4	-3683	121	7608	8165	-3088	121	6417	6950
Klassen	Gewissenhaftigkeit ($n = 734$)				Gewissenhaftigkeit ($n = 608$)			
	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-3555	31	7173	**7315**	-3039	31	6141	**6278**
2	-3569	61	7259	7540	-2953	61	6027	6296
3	-3506	91	7194	7613	-2918	91	6017	6418
4	-3620	121	7483	8039	-2889	121	6020	6553

Anmerkungen: Mixed-Rasch-Modelle (PCM) für 1 – 4 latente Klassen; CML-Schätzung mit WinMira; LL = LogLikelihood; np = Anzahl freie Modellparameter.

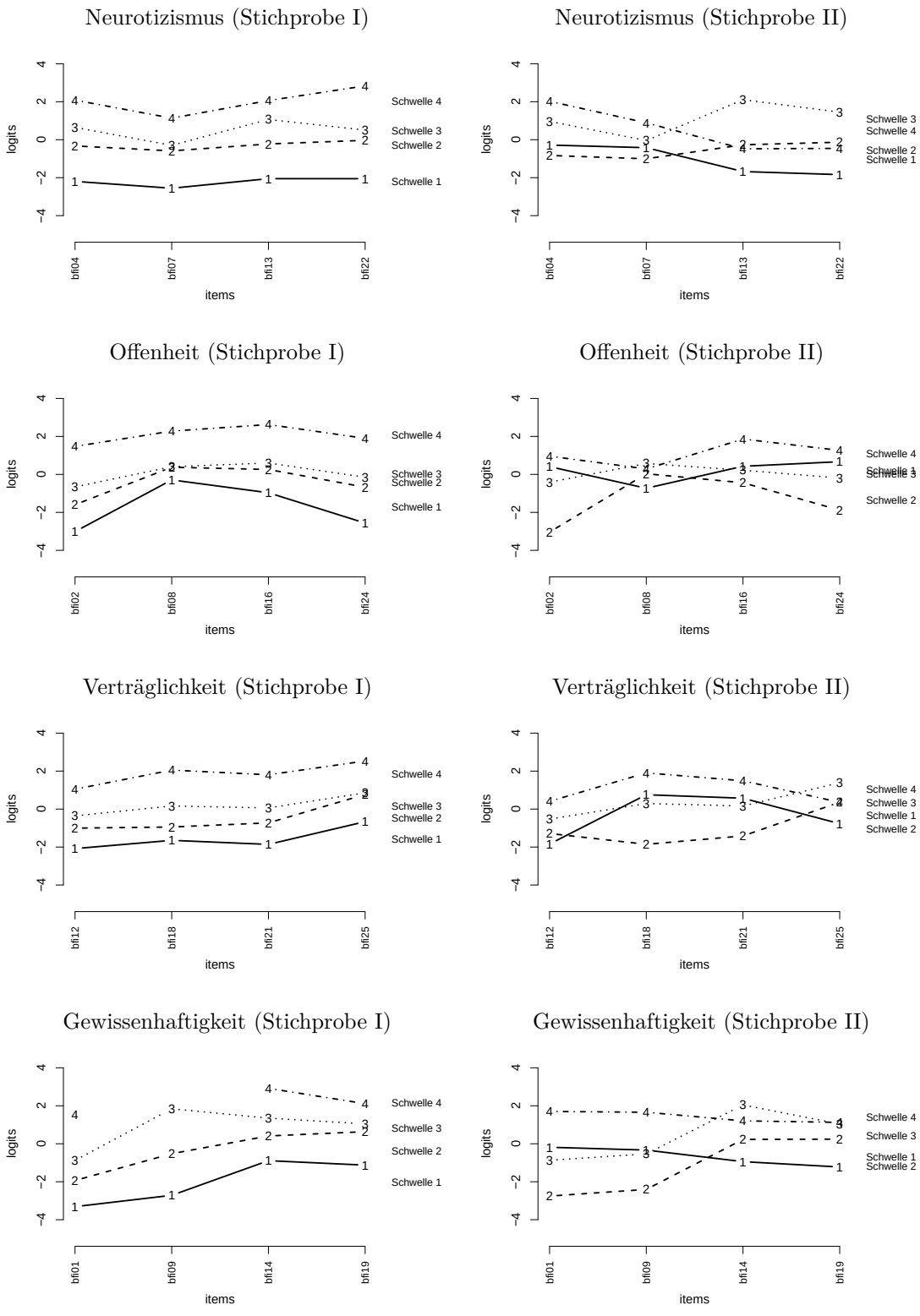


Abbildung 6.7 Darstellung der Schwellenparameterprofile der 1-Klassen-Lösung (*WinMira*) für vier BFI–K-Skalen; jeweils Stichprobe I links und Stichprobe II rechts.

Die visuelle Inspektion der Schwellenparameter-Plots aus der Skalierung der Dimension *Extraversion* für die Stichprobe II (2-Klassen-Lösung) erlaubt zunächst keine eindeutige Klassifikation der beiden latenten Klassen im Hinblick auf eine Qualifizierung der Antworttendenz (vgl. Abbildung 6.8 rechte Seite).

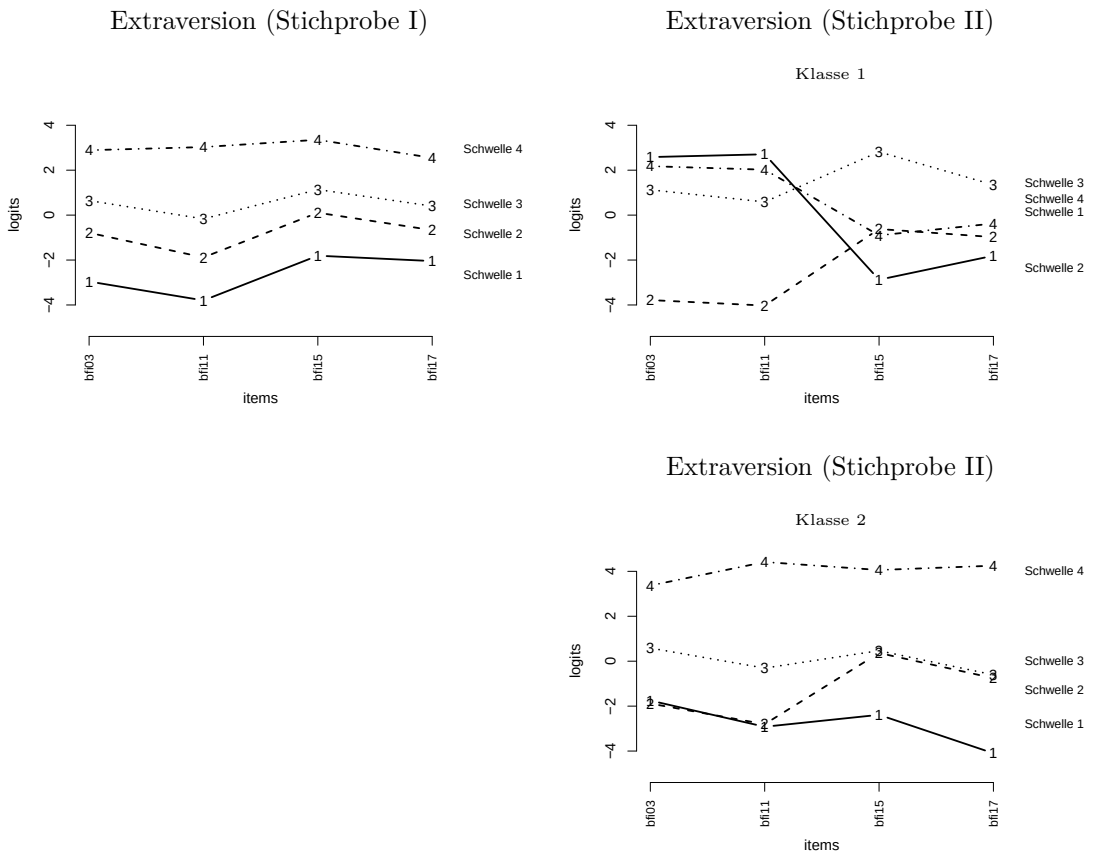


Abbildung 6.8 Darstellung der Schwellenparameterprofile der 1- und 2-Klassen-Lösung (*WinMira*) für die BFI–K-Skala *Extraversion*; jeweils Stichprobe I links und Stichprobe II rechts.

Zur weiteren Untersuchung der Qualität des Antwortverhaltens dieser beiden, anhand ihres Antwortverhaltens in der Dimension *Extraversion* in latenten Klassen klassifizierten Personen-Teilgruppen, wird daher eine *zwei Gruppen KFA* (z. B. Stemmler, 2014; Stemmler & Heine, 2017) mit dem *R*-Paket `confreq` (Heine et al., 2019) durchgeführt. Als gruppierender Faktor wird dabei das Klassifikationsergebnis aus der mixed-Rasch Modellierung zugrunde gelegt. Die Basis für eine explizite Zuordnung in eine der beiden Klassen unterschiedlichen Antwortverhaltens ist die maximale Zuordnungswahrscheinlichkeit. Die

auf dieser Basis neu erstellte dichotome Indikatorvariable bildet die Grundlage für die zwei Gruppen KFA. Die Ergebnisse der *zwei Gruppen KFA* indizieren insgesamt zwölf signifikant diskriminierende Antwortmuster (*pattern*) von den insgesamt $5^4 = 625$ möglichen pattern (vgl. Tabelle 6.8).

Tabelle 6.8 Signifikant diskriminierende Antwortmuster aus *zwei Gruppen KFA* für die BFI–K-Dimension *Extraversion* auf Basis der Personen Klassifikation nach mixed-Rasch Modellierung.

	pattern[a]	Typ[b]	$f_{eKlasse1}$	$f_{bKlasse1}$	$f_{eKlasse2}$	$f_{bKlasse2}$	p[c]	χ^2	df	p_χ[d]
Polung:[e]	+ + − −									
(Klasse 1)	4 4 5 5	+	11.99	21.00	9.01	0.00	0.0000059	14.58	1	0.0001341
	5 4 5 5	+	10.28	18.00	7.72	0.00	0.0000342	12.19	1	0.0004796
	3 4 3 3	+	9.71	17.00	7.29	0.00	0.0000613	11.40	1	0.0007332
	3 3 3 3	+	9.71	17.00	7.29	0.00	0.0000613	11.40	1	0.0007332
(Klasse 2)	5 5 5 5	+	8.56	0.00	6.44	15.00	0.0000024	18.15	1	0.0000204
	5 5 4 4	+	10.28	0.00	7.72	18.00	0.0000002	22.34	1	0.0000023
	5 4 5 4	+	7.42	0.00	5.58	13.00	0.0000140	15.38	1	0.0000880
	5 4 4 4	+	7.42	0.00	5.58	13.00	0.0000140	15.38	1	0.0000880
	4 4 5 4	+	7.42	0.00	5.58	13.00	0.0000140	15.38	1	0.0000880
	4 4 4 5	+	6.85	0.00	5.15	12.00	0.0000336	14.00	1	0.0001827
	4 4 4 4	+	12.56	0.00	9.44	22.00	0.0000000	28.01	1	0.0000001
	3 3 2 3	+	6.85	0.00	5.15	12.00	0.0000336	14.00	1	0.0001827

Anmerkungen: Beobachtete (f_b) und erwartete (f_e) Häufigkeiten für signifikant diskriminierende Antwortmuster für zwei latente Klassen (Stichprobe II); Klasse 1: $n = 346$, Klasse 2: $n = 260$; [a] Antwortmuster auf den umgepoolten Items für *Extraversion* (vlnr.: *bfi03*, *bfi11*, *bfi15*, *bfi17*); [b] Bonferoni adjustiertes Alpha: $\alpha_{bonferoni} = .00008$; [c] p Koeffizient des exakten Tests nach Fischer; [d] p Koeffizient des χ^2-Tests; [e] Formulierung der Items (Polung), die Analysen werden mit dem umgepolten Items durchgeführt.

Bei der genaueren Betrachtung dieser signifikant zwischen den beiden latenten Klassen aus der mixed-Rasch Modellierung diskriminierenden Antwortmustern fällt auf, dass jedes Antwortmuster ausschließlich in einer der beiden Klassen von null verschiedene Häufigkeiten aufweist (vgl. Spalten „$f_{bKlasse1}$“ und „$f_{bKlasse2}$“ in Tabelle 6.8).

Zur Absicherung des Befundes einer für fast alle Skalen des BFI–K eindimensionalen Skalierbarkeit werden die Daten der beiden Stichproben (I und II) zusammen einer erneuten Skalierung unterzogen. Dabei wird, wie für die AIST–R Dimension *Enterprising* in Abschnitt 6.1, auf die nichtiterative Methode zur Itemparameterbestimmung nach der *PAIR*-Methode zurückgegriffen, wie sie in dem *R*-Paket `pairwise` (Heine, 2019) implementiert ist. Die Ergebnisse dieser Parameterbestimmung für die zusammengelegte Gesamtstichprobe und die daran anschließende Testung der Modellgeltung rechtfertigen die

Gültigkeit des eindimensionalen Skalierungsmodells. So ergeben sich für alle Dimensionen des BFI–K zufriedenstellende Reliabilitäten auf Basis der Weighted Liklihood Schäzung der Personenparameter im Bereich von $r_{WLE} = .66$ (*Verträglichkeit*) bis $r_{WLE} = .79$ (*Extraversion*). Neben der Reliabilität wird zur Beurteilung der Gültigkeit des eindimensionalen Skalierungsmodells für die fünf BFI–K Dimensionen ein grafischer Modelltest und der Andersen Modelltest (Andersen, 1973b) durchgeführt.

Im Hinblick auf die in der folgenden Untersuchung (vgl. Studie 6.3) durchzuführenden Analysen zum Zusammenhang der beiden Konstrukte (*Persönlichkeit* und *berufliche Interessenorientierungen*) in Abhängigkeit von Antworttendenzen (Auf dem AIST–R), wird zur Testung der Teilgruppen invarianten Modellgeltung auf die Klassifikationsergebnisse aus der LCA zweiter Ordnung aus der ersten Studie zurückgegriffen (vgl. Tabelle 6.5). Um hierbei trotz der dort gefundenen drei latenten Klassen auf ein dichotomes Teilungskriterium für die Gesamtstichprobe zurückgreifen zu können, werden die Teilgruppen aus der latenten Klasse 3 und der latenten Klasse 1 zusammengelegt. Diese Zusammenlegung ist durch das in der latenten Klasse 3 am häufigsten auftretende Muster der Antworttendenz auf den vier dort analysierten AIST–R Dimensionen ($f_{KL3:'1121'} = 154$ – vgl. Tabelle 6.5 in Studie 6.1) begründet. Die so zur Modelltestung des eindimensionalen Skalierungsmodells für die BFI–K-Dimensionen aufgeteilte Gesamtstichprobe umfasst $n = 895$ und $n = 445$ Personen. Der grafische Modelltest für die fünf BFI–K-Dimensionen nach Kalibrierung der Items in diesen beiden Teilgruppen ist in Abbildung 6.9 dargestellt.

Der Andersen Test (Andersen, 1973b) wird für alle fünf Dimensionen des BFI–K nicht signifikant (vgl. Tabelle 6.9). Dies steht im Einklang mit den Befunden aus dem grafischen Modelltest und stützt die Geltung des eindimensionalen Skalierungsmodells für die fünf BFI–K-Dimensionen.

Die Schwellenparameterplots aus der eindimensionalen Skalierung mit dem *PAIR*-Algorithmus sind in Tabelle 6.10 dargestellt. Die Linienzüge für die einzelnen Dimensionen weisen, wie auch bei den Ergebnissen der Skalierung mit dem Programm *WinMira* (vgl. Tabelle 6.7), teilweise Vertauschungen der einzelnen Itemkategorieschwellen auf.

Tabelle 6.9 Andersen Test für fünf Dimensionen des BFI–K für die Gesamtstichprobe.

	$\chi^2_{And.}$ [a]	*df*	*p*
Neurotizismus	29.70	31	.53
Extraversion	22.60	31	.86
Offenheit	41.92	31	.09
Verträglichkeit	34.82	31	.29
Gewissenhaftigkeit	37.87	31	.18

Anmerkungen: Teilungskriterium: Antworttendenz auf vier Dimensionen des AIST–R (mittlere Antworttendenz: $n = 895$, extreme Antworttendenz: $n = 445$); [a] Modelltest nach Andersen (1973b).

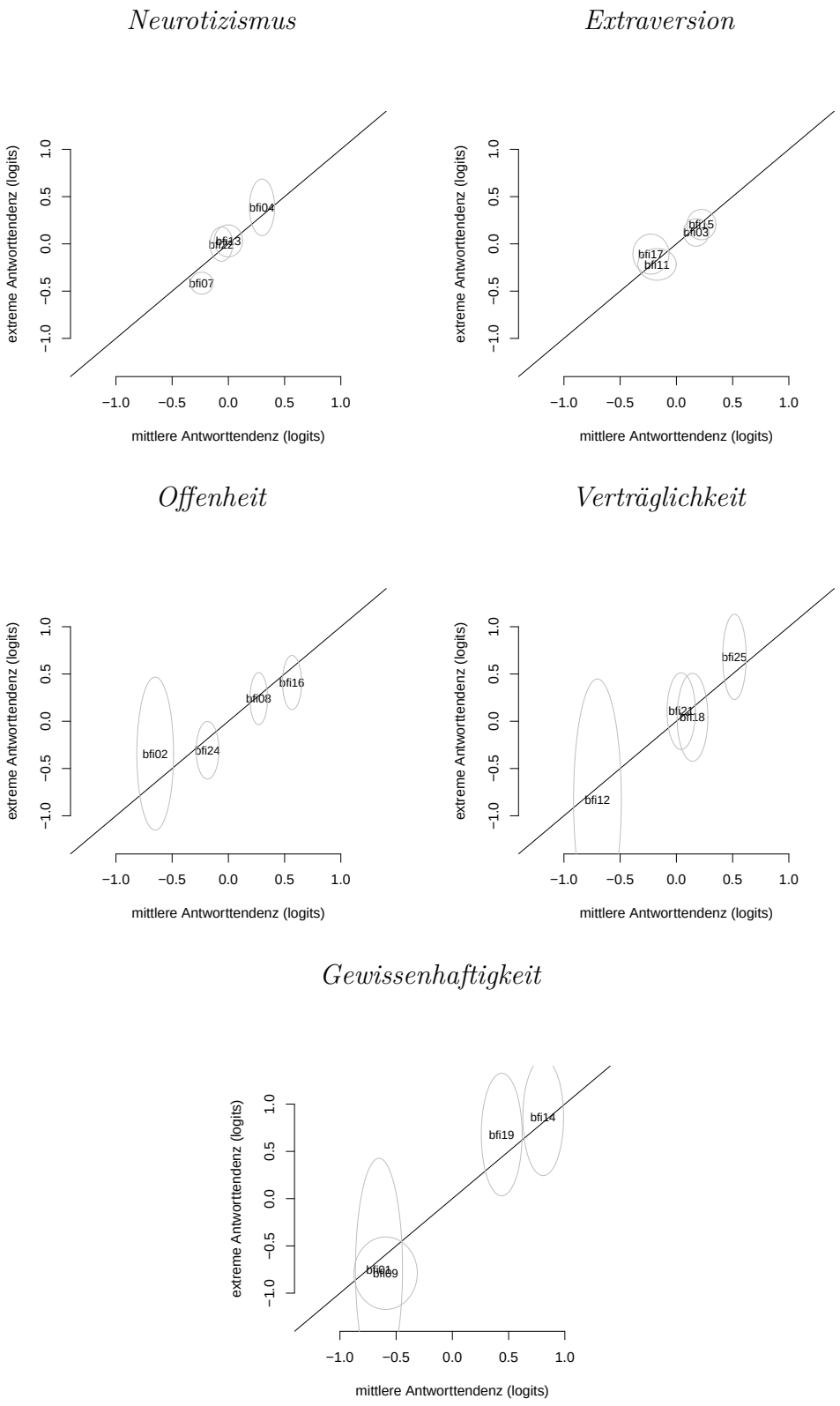


Abbildung 6.9 Grafischer Modelltest für die eindimensionale Skalierung des BFI–K mit fünf Dimensionen der Persönlichkeit (*PAIR*-Algorithmus); Teilungskriterium: Antworttendenz auf vier Dimensionen des AIST–R (mittlere Antworttendenz: $n = 895$, extreme Antworttendenz: $n = 445$).

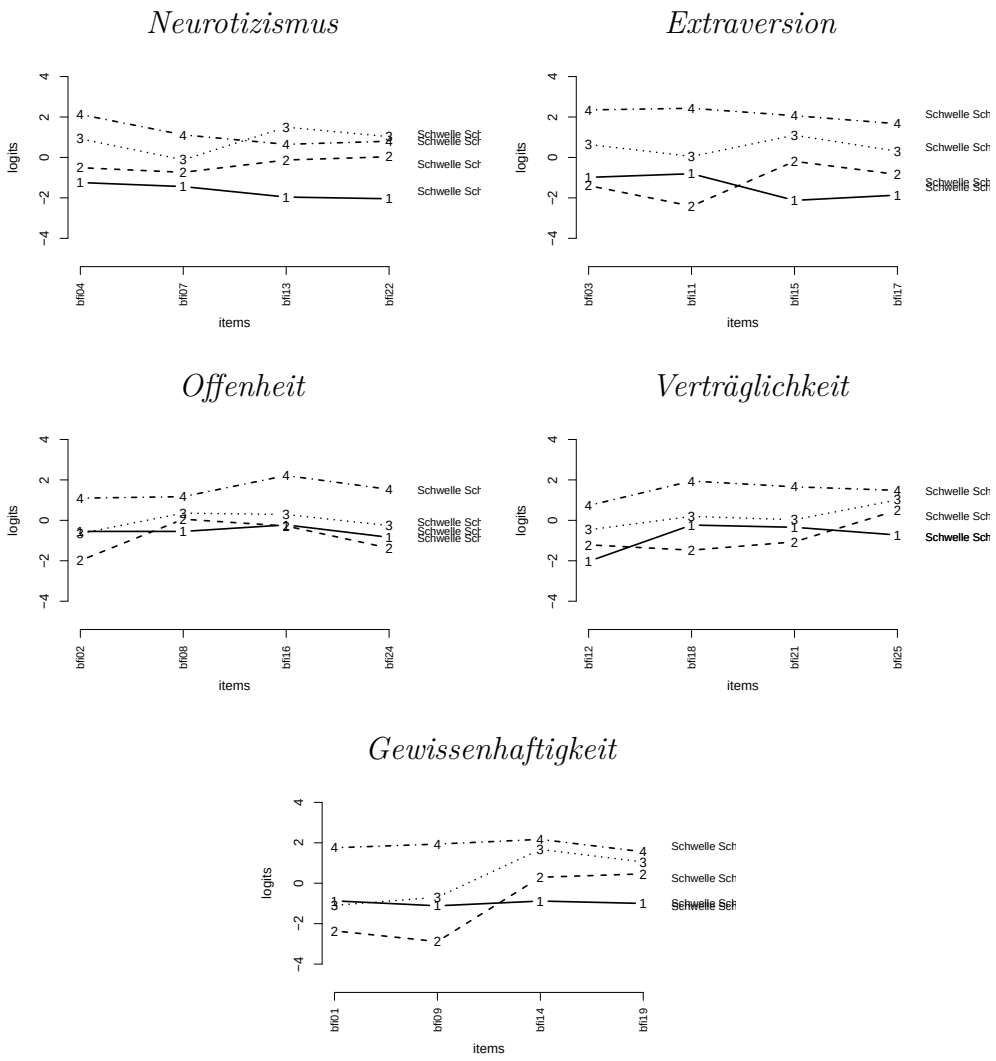


Abbildung 6.10 Darstellung der Schwellenparameterprofile der eindimensionalen Skalierung mit dem *PAIR*-Algorithmus für die BFI–K-Skalen.

Diskussion

In der hier berichteten Untersuchung wird die Skalierbarkeit der fünf BFI–K-Skalen nach einem (summativ) kumulativen Antwortmodell analysiert. Es wird untersucht, inwiefern sich das immer wieder berichtete Phänomen der unterschiedlich ausgeprägten Antworttendenzen (MRS vs. ERS) auf mehrstufigen Antwortskalen zur Erfassung von Persönlichkeitsmerkmalen (z. B. E. J. Austin et al., 2006; Berg & Collier, 1953; Damarin & Messick, 1965; Eid & Zickar, 2007; Heine, 2010; Minkov, 2017; Naemi et al., 2009; Rost et al., 1999; Warr & Coffman, 1970, sowie Abschnitt 3.2.4 in Kapitel 3. *Theoretischer Hintergrund zu Antwortmustern*) auch in der hier eingesetzten, modifizierten Version des BFI–K gefunden werden kann.

Zu der Frage nach der eindimensionalen Skalierbarkeit der untersuchten Dimensionen des Konstrukts Persönlichkeit mit einem Modell für den Dominanz-Antwortprozess ergibt sich zunächst für vier der fünf Dimensionen der Befund, dass nach den informationstheoretischen Kriterium BIC im Rahmen des relativen Modellvergleichs das eindimensionale Modell (PCM – Masters, 1982)) gegenüber dem mixed-Rasch-Modell (Rost, 1990, 1991) zu bevorzugen ist.

Der Befund sich überschneidender Schwellenparameterprofile in der Stichprobe II, sowie auch bei der gemeinsamen Skalierung beider Stichproben mit dem *PAIR*-Algorithmus, kann zunächst nicht direkt als Verletzung der Modellannahmen des *Partial Credit* Modells (Masters, 1982) interpretiert werden, da die Schwellenparameter nach diesem Modell prinzipiell frei variieren können. Allerdings widersprechen derartige Überschneidungen der grundlegenden Annahme einer kontinuierlich ansteigenden, ordinalen Charakteristik der mehrstufigen Antwortskala der Items.

Bei Betrachtung der Skalierungsergebnisse mit der Methode der CML-Schätzung (*WinMira*) für die einzelnen Stichproben, muss lediglich in der Stichprobe II für die Persönlichkeitsdimension *Extraversion* die Annahme der Personenhomogenität verworfen werden (2–Klassen–Lösung). Allerdings lassen die jeweiligen Schwellenparameterprofile der beiden latenten Personenklassen keine eindeutige Interpretation im Hinblick auf unterschiedlich ausgeprägten Antworttendenzen (MRS vs. ERS) zu.

Die vergleichende Analyse der Personenantworten der beiden Stichproben für die Dimension *Extraversion* mit der *zwei Gruppen KFA* weisen allerdings

darauf hin, dass die hier gefundene Personenheterogenität aus der Existenz bestimmter, spezifischer Antwortmuster resultiert, welche hier eine 2–Klassen-Lösung determinieren (vgl. Tabelle 6.8). Die hier für die Stichprobe II bei der mixed-Rasch Modellierung gefundene 2–Klassen Lösung scheint hier insofern eher auf ein methodisches Problem der Parameterschätzung (mit der CML-Methode) vor dem Hintergrund geringer Antwortkategoriehäufigkeiten hinzuweisen. Der Befund, dass in jeweils einer der beiden latenten Klassen die in der 2-Gruppen-KFA identifizierten, diskriminierenden Antwortmuster Häufigkeiten von null aufweisen, legt nahe, dass das an sich probabilistische Modell der Analyse latenter Klassen (LCA) (bzw. des mixed-Rasch-Modells) hier deterministisch zwei Klassen produziert.

Demgegenüber weisen die Ergebnisse der Skalierung der Gesamtstichprobe für alle Dimensionen der BFI–K mit der *PAIR*-Methode eher auf die Gültigkeit des eindimensionalen Skalierungsmodells hin (vgl. Abbildung 6.9 sowie Tabelle 6.9). Die Ergebnisse des grafischen Modelltests sowie des Andersen Tests (Andersen, 1973b), belegen hier die Skalierbarkeit der fünf Dimensionen des BFI–K mit dem eindimensionalen Rasch-Modell für mehrstufige Antwortskalen (Masters, 1982). Insgesamt weisen die Ergebnisse dieser zweiten Studie daher darauf hin, dass sich alle fünf Skalen der modifizierten Form des BFI–K als mit dem Rasch-Modell für mehrstufige Antwortformate (PCM) skalierbar erweisen (vgl. Abbildung 6.9 und Tabelle 6.9).

Diese aus diagnostischer Perspektive an und für sich positiven Ergebnisse stehen somit zunächst im Widerspruch zu anderen Untersuchungen zur Skalierbarkeit von Big-Five-Persönlichkeitsinventaren nach dem Rasch und mixed-Rasch-Modell (z. B. E. J. Austin et al., 2006; Rost et al., 1999).

Aufgrund dieser widersprüchlichen Befundlage ergibt sich die (noch weiter zu überprüfende) Frage, ob es sich vor dem Hintergrund idiosynkratischer Antwortmuster bei der hier zunächst bestätigten, eindimensionalen und personenhomogenen Skalierbarkeit des modifizierten BFI–K um einen belastbaren Befund handelt.

Der Befund, dass sich bei der Itemkalibrierung der BFI–K-Dimensionen anhand der zweiten Stichprobe deutliche Schwellenvertauschungen ergeben (vgl. Abbildungen 6.7 und 6.8), widerspricht der Annahme einer kontinuierlich ansteigenden, ordinalen Charakteristik der mehrstufigen Antwortskala der Items.

Dieser auch mit dem mit dem *PAIR*-Algorithmus replizierte Befund stellt die postulierte summative Verrechnung der einzelnen Itemscores als Maß der Merkmalsausprägung für diese Stichprobe in Frage. Diese Schwellenvertauschungen, auch bei den mit Stichprobe I und II gemeinsam durchgeführten Analysen zur Skalierbarkeit der BFI–K-Dimensionen, kann hier Anlass für weitere Analysen begründen. So kann angenommen werden, dass trotz der durch die relativen Modellvergleiche und den Andersen Test erfolgten Bestätigung einer eindimensionalen Skalierbarkeit der fünf Persönlichkeitsdimensionen, dennoch Personengruppen mit abweichendem Antwortverhalten in den beiden Stichproben bestehen. Die auch bei der Skalierung mit dem *PAIR-Algorithmus* identifizierten Schwellenvertauschungen können auch als Indiz dafür interpretiert werden, dass die in Abhängigkeit der Merkmalsausprägung monoton steigende Itemcharakteristik des *Dominanz*-Antwortprozesses durch idiosynkratisches Antwortverhalten lokal nicht erfüllt ist. Vor diesem Hintergrund erscheint es lohnenswert das Antwortverhalten der vorliegenden Stichproben für die einzelnen Dimensionen des BFI–K und die anderen eingesetzten Konstrukte und Dimensionen genauer zu betrachten. Entsprechende Untersuchungen zu differentiellen Antwortprozessen werden in Studie 7.1 in Kapitel 7 durchgeführt.

6.3 Auswirkungen von Antworttendenzen auf empirische Befunde zum Zusammenhang zwischen Dimensionen der Persönlichkeit und beruflichen Interessenorientierungen

Einleitung

Aufbauend auf den Ergebnissen der beiden Studien in den Abschnitten 6.1 und 6.2 zur summativen Skalierbarkeit der Skalen der beiden Konstrukte *Persönlichkeit* und *berufliche Interessenorientierungen* sollen in der vorliegenden Studie die empirischen Zusammenhänge zwischen den einzelnen Skalen der beiden Konstrukte untersucht werden. Einerseits soll dabei versucht werden die Befunde aus bestehenden Arbeiten zum Zusammenhang der Konstrukte zu replizieren. Andererseits soll der Einfluss unterschiedlicher Antworttendenzen, wie sie sich beim AIST–R in der ersten Studie in Abschnitt 6.1 zeigen, untersucht werden. Es soll mit den hier berichteten Analysen der Frage nachgegangen werden, ob die in der ersten Studie in Abschnitt 6.1 gefundenen, unterschiedlich ausgeprägten Antworttendenzen Einfluss auf die empirischen Zusammenhänge der beiden Konstrukte auf der Ebene der einzelnen Dimensionen haben. Zunächst sollen daher bisherige Ergebnisse und Befunde aus der Literatur zum Zusammenhang der beiden Konstrukte für den englisch- und deutschsprachigen Raum im Überblick beschrieben werden.

Beziehung von Persönlichkeit und beruflichen Interessen

Bereits Holland (1959, 1997) betont in seiner Theorie der beruflichen Interessenorientierungen, dass die Wahl eines Berufs ein Ausdruck der Persönlichkeit einer Person ist und berufliche Interessen somit einen wesentlichen Aspekt der Persönlichkeit einer Person darstellen. In diesem Sinne erklären (berufliche) Interesseninventare gleichzeitig auch einen wichtigen Bereich der Persönlichkeit (vgl. z. B. Krapp & Lewalter, 2001; Todt, 1978; Todt & Schreiber, 1998). Die Beziehungen zwischen den in Abschnitt 2.1 beschriebenen Dimensionen der Persönlichkeit auf der Grundlage des Fünf-Faktor Modells und den beruflichen Interessen nach Holland (vgl. Abschnitt 2.2), sollen hier näher dar-

gestellt werden. Eine der ersten systematischen Untersuchung zu den Zusammenhängen zwischen *Dimensionen der Persönlichkeit* und *beruflichen Interessen* berichten Costa, McCrae und Holland (1984). Die Autoren untersuchten die Beziehung des Neo-Inventars (NEO – McCrae & Costa, 1983a) und des Self-Directed Search (SDS) von Holland (1979) anhand einer Stichprobe von Erwachsenen, welche sich freiwillig als Teilnehmer für eine medizinische Studie gemeldet hatten. Die Korrelationen zwischen den eingesetzten SDS– (berufliche Interessen) und NEO– (Persönlichkeit) Skalen zeigten dabei bedeutsame Assoziationen von intellektuell-forschenden (*Investigative*) und künstlerischen (*Artistic*) Interessen mit der NEO–Dimension *Offenheit für Erfahrung*. Ebenfalls konnten Zusammenhänge zwischen sozialen (*Social*) und unternehmerischen (*Enterprising*) Interessen mit *Extraversion* gefunden werden. Personen mit eher konventionellen (*Conventional*) Interessen zeigten eher geringe Werte in der NEO–Dimension *Offenheit*. Auf Basis dieser Befunde folgern McCrae und Costa (1983a), dass die eingesetzten NEO–Skalen die Interessenskalen vor allem um Maße für die Persönlichkeitsdimension *Neurotizismus* ergänzen. Vergleichbare Ergebnisse konnten auch in einer Reihe weiterer Studien aus dem englischsprachigen Raum festgestellt werden, wobei verschiedenen Operationalisierungen (Inventare) zu den beiden Konstrukten und unterschiedlich zusammengesetzte Stichproben die empirische Basis bildeten. (z. B. de Fruyt & Mervielde, 1997; Gottfredson, Jones & Holland, 1993; Holland, Johnston & Asama, 1994; Larson & Borgen, 2002). Diese positiven Zusammenhänge von *Investigative* und *Artistic* mit *Offenheit für Erfahrung*, *Social* und *Enterprising* mit *Extraversion* sowie der negative Zusammenhang von *Conventional* mit *Offenheit*, konnten auch in Metaanalysen bestätigt werden (Barrick, Mount & Gupta, 2003; Larson, Rottinghaus & Borgen, 2002; Mount, Barrick, Scullen, Rounds & Sackett, 2005). Übereinstimmend finden alle drei Metaanalysen geringe positive Zusammenhänge ($r = .05$ bis $r = .08$) zwischen *Realistic* und *Offenheit für Erfahrungen*. Demgegenüber berichten Tokar und Swanson (1995) in einer Stichprobe von $N = 359$ Erwachsenen einen negativen Zusammenhang ($r = -.24$) zwischen diesen beiden Dimensionen *Realistic* und *Offenheit für Erfahrungen*. Die Ergebnisse zum Zusammenhang zwischen *Neurotizismus* und *Realistic* fallen bei der Betrachtung der drei Metaanalysen uneinheitlich aus. Während Mount et al. (2005) hier geringe positive Zusammenhänge ($r = .07$

und $r = .06$) berichten, finden Larson et al. (2002) dagegen einen geringen negativen ($r = -.09$), aber dennoch signifikanten Zusammenhang. Zumindest hinsichtlich der Richtung des Zusammenhanges, finden Schinka, Dye und Curtiss (1997) übereinstimmend mit den Befunden aus der Metaanalyse von Larson et al. (2002) für die Dimensionen *Realistic* und *Neurotizismus* eine Korrelation von $r = -.18$. de Fruyt und Mervielde (1997) berichten für eine Stichprobe von Studenten der Universität Gent für die Dimensionen *Realistic* und *Neurotizismus* ebenfalls einen negativen Zusammenhang ($r = -.19$). Übereinstimmend mit Larson et al. (2002) werden von de Fruyt und Mervielde (1997) darüber hinaus negative Zusammenhänge zwischen *Neurotizismus* und *Enterprising* ($r = -.33$) und *Conventional* ($r = -.24$) berichtet. Diese negativen Zusammenhänge zwischen *Neurotizismus* und *Realistic*,*Enterprising* sowie *Conventional*, werden dagegen in den beiden Metaanalysen von Barrick et al. (2003); Mount et al. (2005) nicht gefunden. Für den deutschsprachigen Raum untersuchte Bergmann (2001) erstmals die Beziehung zwischen den Persönlichkeitsdimensionen nach dem Fünf-Faktor-Modell und den beruflichen Interessenorientierungen von Holland (1997). Dazu wurden die Antworten von insgesamt $n = 300$ Personen (davon $n = 186$ Frauen und $n = 114$ Männer) ausgewertet. Die Stichprobe setzte sich mit $n = 226$ Personen aus Studierenden der Wirtschaftspädagogik der Universität Linz und aus $n = 74$ Angestellten der österreichischen Finanzbehörde zusammen. Befragt wurden die Probanden anhand der deutschen Übersetzung des *NEO-Five-Factor Inventory* (NEO-FFI – Costa & McCrae, 1985) von Borkenau und Ostendorf (1993) und des *Allgemeinen Interessen-Struktur-Tests* (AIST) von Bergmann und Eder (1999). Auch hier fanden sich für beide Geschlechter vergleichbare Beziehungen zwischen der Dimension *Offenheit für Erfahrungen* und der intellektuell-forschenden (*Investigative*) und der künstlerisch-sprachlichen Orientierung (*Artistic*) sowie auch positive Zusammenhänge zwischen der Persönlichkeitsdimension *Extraversion* und der sozialen (*Social*) und unternehmerischen Orientierung (*Enterprising*). Zusätzlich ergaben sich aber auch positive Zusammenhänge zwischen der Dimension *Verträglichkeit* und der sozialen Orientierung (*Social*) sowie auch zwischen der Persönlichkeitsdimension *Gewissenhaftigkeit* und der konventionellen Orientierung (*Conventional*). Negative Beziehungen fanden sich dagegen zwischen der Dimension *Neurotizismus* und der unternehmerischen Orientierung

(*Enterprising*), der Dimension *Offenheit für Erfahrungen* und der konventionellen Orientierung (*Conventional*) und der Dimension *Gewissenhaftigkeit* und der künstlerisch-sprachlichen Orientierung (*Artistic*).

Tabelle 6.10 Signifikante Korrelationen der Persönlichkeitsdimensionen und der beruflichen Interessenorientierungen nach Geschlecht (W / M) getrennt.

		R	I	A	S	E	C
Neurotizismus (n)	W	.18				−.17	
	M					−.30	
Extraversion (e)	W				.15	.40	
	M				.31	.43	
Offenheit (o)	W	.20	.37	.63			−.22
	M		.38	.52	.24		−.22
Verträglichkeit (a)	W				.25		
	M				.34		
Gewissenhaftigkeit (c)	W	−.20		−.19			.43
	M			−.24			.35

Anmerkungen. Entnommen aus Bergmann (2001, S. 191), $n = 186$ (weiblich), $n = 114$ (männlich); nicht signifikante Korrelationen ($p < 0.05$) sind nicht dargestellt.

Tabelle 6.10 gibt die empirischen Befunde von Bergmann (2001, S. 191) zum Zusammenhang der beiden Konstrukte getrennt nach Geschlecht zusammenfassend wieder.

Trotz der, hinsichtlich des Zusammenhanges bestimmter Dimensionen, relativ einheitlichen Befundlage, ergeben sich bei der Durchsicht der bestehenden empirischen Literatur immer wieder auch widersprüchliche Befunde. In der vorliegenden Untersuchung soll daher der Frage nachgegangen werden, ob sich diese unterschiedlichen Befunde zum Zusammenhang mancher Dimensionen der beiden Konstrukte auf unterschiedliche Antworttendenzen der befragten Personen zurückführen lassen. Auf Basis der Ergebnisse zur Skalierbarkeit des BFI–K (vgl. Abschnitt 6.2), sowie der Befunde aus der Studie in Abschnitt 6.1 zur Skalierbarkeit des AIST–R, soll daher hier untersucht werden, ob der individuell unterschiedlich ausgeprägte Gebrauch der vorgegebenen Antworts-

kalen (MRS vs. ERS) Einfluss auf die empirischen Zusammenhänge zwischen den einzelnen Skalen der Konstrukte *Berufliche Interessen* und *Persönlichkeit* hat.

Darüber hinaus soll der Frage nachgegangen werden, in wie weit sich die Befunde zu den unterschiedlich ausgeprägten Antworttendenzen auf den AIST–R Skalen aus der ersten Studie in Abschnitt 6.1 auf die von Holland (1997) aufgestellten Hypothese der strukturellen Anordnung (Calculus-Hypothese) der sechs Dimensionen der Beruflichen Interessen auswirken. Es gilt hier der Frage nachzugehen, ob die durch ihre verschiedenen Antworttendenzen definierten Teilstichproben unterschiedlich gut an die hexagonale Modellstruktur anzupassen sind.

Im Einzelnen sollen auf Basis der Befunde aus den beiden Untersuchungen in den Abschnitten 6.1 und 6.2 mit der vorliegenden Studie die folgenden beiden Fragen untersucht werden. Erstens soll die Frage beantwortet werden, ob Antworttendenzen den Zusammenhang zwischen den Konstrukten Persönlichkeit und Interessen beeinflussen. Zweitens soll untersucht werden, ob die Passung der empirischen Daten auf die theoretisch angenommene, ideale hexagonale Struktur durch den unterschiedlichen Gebrauch der fünfstufigen Antwortskala (ERS vs. MRS) beeinflusst wird.

Daten

Die Datengrundlage für die vorliegenden Analysen besteht aus den kombinierten Stichproben I und II (vgl. Abschnitt 5.2 in Kapitel 5 *Stichproben und Instrumente*), wie sie auch in den beiden Studien in Abschnitt 6.1 und 6.2 untersucht wurden. Über die beiden Konstrukte (berufliche Interessen und Persönlichkeit) hinweg können hier $n = 1340$ Fälle für die Analyse berücksichtigt werden. Für die Analysen werden die abgeleiteten Variablen berücksichtigt, die sich aus den Ergebnissen in den beiden Studien in Abschnitt 6.1 und 6.2 ergeben haben. Eingesetzt werden einerseits die Schätzer der Merkmausausprägungen für die Personen aus der IRT–Skalierung (Personenparameter) für die erfassten Dimensionen der beiden Konstrukte. Andererseits werden als kategorial, klassifizierende Variablen die (latente) Klassenzuordnung der Personen anhand ihres Antwortverhaltens auf den (vier) AIST–R-Skalen eingesetzt (vgl. Ergebnisse aus den beiden Studien in Abschnitt 6.1 und 6.2).

Methode

Zur Untersuchung des Zusammenhanges zwischen den einzelnen Dimensionen der beiden Konstrukte werden Korrelationskoeffizienten nach Pearson anhand der Personenparameter aus den Ergebnissen der Analysen zur Skalierbarkeit in Abschnitt 6.1 als Indizes der individuellen Merkmalsausprägung berichtet. Für die AIST–R-Dimensionen *Realistic*, *Investigative*, *Social*, *Artistic* und *Conventional* werden hier die Personenparameterschätzer aus der mixed-Rasch-Skalierung eingesetzt, wodurch die Personen trotz unterschiedlichen Antwortverhaltens auf einer gemeinsamen latenten Dimension abgebildet werden [1]. Für die AIST–R-Dimension *Enterprising* und die fünf Dimensionen der Persönlichkeit wird auf die Personenparameterschätzer aus der „einfachen" Rasch-Skalierung (1-Klassenlösung) zurückgegriffen.

Zur Untersuchung der Frage ob ein individuell unterschiedlicher Antwortstil (auf einzelnen Dimensionen des AIST–R) einen Einfluss auf die Skalen Interkorrelationen der beiden Konstrukte untereinander hat, werden die Personen der beiden Stichproben nach ihrem Antwortstil auf den AIST–R-Dimensionen in Teilgruppen aufgeteilt. Hierzu wird auf die Ergebnisse der Latenten–Klassen–Analyse zweiter Ordnung aus der ersten Studie in Abschnitt 6.1, zurückgegriffen (vgl. Tabellen 6.4 und 6.5). Die drei dort identifizierten übergeordneten latenten Klassen über die Indikatorvariablen der Antworttendenz für vier der sechs AIST–R-Skalen, bilden die Grundlage für die Aufteilung der Gesamtstichprobe in drei Teilgruppen. Für diese, nach ihrer Antworttendenz aufgeteilten Teilstichproben, werden dann jeweils die Konstrukt-Interkorrelationen auf der Ebene einzelner Dimensionen getrennt berechnet.

Ferner wird für die Gesamtstichprobe, sowie für die nach ihrer Antworttendenz aufgeteilten Teilgruppen untersucht, inwieweit sich die mit dem AIST–R erhobenen beruflichen Interessen der Stichprobe der von Holland (1997) postulierten circumplexen Modellstruktur anpassen lassen. Aufgrund der im

[1] für die AIST–R-Dimensionen *Realistic*, *Investigative*, *Social* und *Conventional* lässt sich dieses unterschiedliche Antwortverhalten als unterschiedliche Tendenz zu entweder mittleren oder extremen Antwortkategorien qualifizieren (vgl. Tabelle 6.2 in Abschnitt 6.1). Für die AIST–R-Dimension ergibt sich zwar auch eine 2-Klassenlösung als das am besten passende Modell, dessen zwei Klassen allerdings nicht als Personen mit entweder mittlerer oder extremer Antworttendenz qualifiziert werden können.

Hinblick auf die Interessenstruktur selektiven Stichprobe von Studierenden der Universität der Bundeswehr, welche sich vornehmlich aus sozialwissenschaftlichen Studiengängen zusammensetzt, kann erwartet werden, dass die Anpassung an die ideale circumplexe Struktur des hexagonalen Modells von Holland (1997) insgesamt eher nicht optimal ausfallen dürfte. In der vorliegenden Untersuchung steht allerdings die differentielle Modellanpassung in Abhängigkeit der unterschiedlichen Antworttendenzen (MRS vs. ERS) im Vordergrund. Zur Berechnung dieser (differentiellen) Passung wird auf die von Nagy, Marsh, Lüdtke und Trautwein (2009) vorgeschlagene Methode der Strukturgleichungsmodellierung zurückgegriffen. Dabei werden auf Basis der Interkorrelationen der einzelnen Dimensionen beruflicher Interessenorientierungen des AIST–R, zunächst die empirischen Winkel innerhalb des hexagonalen Modells als zu schätzende Parameter eines (nichtlinearen) Strukturgleichungsmodells bestimmt (Nagy et al., 2009) – (vgl. auch Nagy, 2007, S. 94). Durch Restriktion der Modellparameter auf die ideale circumplexe Struktur lässt sich dann in einem weiteren Schritt der Modell-Fit, also die Passung der empirischen Daten auf das (ideale) hexagonale Model der Berufsinteressen von Holland, bestimmen. Zur Einschätzung der Modellpassung werden einerseits Koeffizienten für die absolute und die relative Modellpassung herangezogen (vgl. Hooper, Coughlan & Mullen, 2008). Zur Bewertung der relativen Modellpassung wird der von Bentler (1990) vorgeschlagene *comparative fit index* (CFI) herangezogen. Die Bewertung der absoluten Modellpassung erfolgt über die Betrachtung des *standardized root mean square residual* ($SRMR$). Der $SRMR$ ist definiert als die Quadratwurzel aus der Differenz zwischen den Residuen der Stichproben Kovarianzmatrix und dem angenommenen Kovarianzmodell (Hooper et al., 2008). Ergänzend wird hier der *root mean square error of approximation* ($RMSEA$ – Steiger, 1990; Steiger & Lind, 1980) berichtet. Zur Beurteilung des Ausmaßes der Modellpassung anhand von *cut-off* Kriterien für die oben genannten Koeffizienten wird auf die Befunde von Cheung und Rensvold (2002); Hu und Bentler (1999) und Chen (2007) Bezug genommen (vgl. auch Bühner, 2011). So konnte Cheung und Rensvold (2002) zeigen, dass bei einer Differenz der Modellpassung von $\delta_{CFI} > .01$ die Messinvarianz zwischen unterschiedlichen Gruppen nicht gegeben ist. Ferner schlägt Chen (2007) auf Basis von Ergebnissen aus Simulationsstudien vor, dass eine Differenz des RMSEA

$\delta_{RMSEA} > .03$ und für den CFI eine Differenz $\delta_{CFI} > .01$, darauf schließen lassen, dass ein angemessenes Maß an Messinvarianz gegeben ist.

Hinsichtlich der Güte der Anpassung der empirischen Daten an die Hexagonale Modellstruktur werden die Daten der vorliegenden Untersuchung auch mit denen aus der Normstichprobe (vgl. Bergmann & Eder, 2005) für den AIST–R verglichen. Zur grafischen Veranschaulichung der gefundenen Zusammenhänge zwischen den einzelnen Interessenorientierungen und den fünf Dimensionen der Persönlichkeit werden diese in die hexagonale Darstellung der Berufsinteressen projiziert. Ermöglicht wird diese Projektion der Nähe-Relation der Big-Five-Dimensionen als – innerhalb der circumplex Struktur – *passive Variablen* zu den sechs Dimensionen beruflicher Interessen (*aktive Variablen*) in der grafischen Darstellung. Die in der grafischen Darstellung gezeigten Winkel der Vektoren der Big-Five- und Interessen-Dimensionen werden durch die Anpassung der zu analysierenden Daten an die circumplexe Struktur über die oben bereits erwähnte Methode der (nichtlinearen) Strukturgleichungsmodellierung ermittelt (vgl. Nagy, 2007, S. 94).

Ergebnisse

Anpassung an die Circumplexe Struktur der AIST–R-Daten

Das Ausmaß der Anpassung, einerseits an die theoretisch postulierte, ideale circumplexe Struktur, sowie andererseits an die durch die Normstichprobe des AIST–R vorgegebene Struktur, ist in Tabelle 6.11 wiedergegeben. Die beste Anpassung der vorliegenden empirischen Daten an die Circumplexstruktur wie sie durch die Normstichprobe vorgegeben wird, gelingt für die Teilgruppe der Personen mit einer mittleren Antworttendenz ($\chi^2 = 152.330, df = 13, p = .000$; $CFI = 0.912, RMSEA = 0.089, SRMR = 0.071$) – vgl. Tabelle 6.11). Demgegenüber steht eine vergleichsweise schlechtere Anpassung der Daten dieser Teilgruppe an die theoretisch postulierte, ideale circumplexe Struktur ($CFI = 0.755, RMSEA = 0.149, SRMR = 0.131$ – vgl. Tabelle 6.11 oben).

Bei Betrachtung der absoluten Modellpassung über den Koeffizienten der Wurzel aus der Summe der Residuen (*standardized root mean square residual – SRMR*) ergeben sich für die Anpassung der Daten an die durch die Normstichprobe vorgegebene Struktur für alle Teilgruppen und auch die Ge-

samtstichprobe akzeptable Werte von $SRMR = .070$ bis $SRMR = .092$ (vgl. Tabelle 6.11 unten).

Tabelle 6.11 Anpassung der empirischen Struktur an ideale und Normstichproben Circumplexstruktur nach Antworttendenz.

Anpassung		χ^2	*df*	*p*	*CFI*	*RMSEA*	*SRMR*
ideale circumplexe Struktur	Gesamt[a]	408.507	13	0.000	0.751	0.151	0.139
	'mittel'[b]	400.706	13	0.000	0.755	0.149	0.131
	'extrem'[c]	279.530	13	0.000	0.814	0.124	0.114
	'inkonsistent'[d]	469.595	13	0.000	0.771	0.162	0.162
		χ^2	*df*	*p*	*CFI*	*RMSEA*	*SRMR*
circumplexe Struktur der Normstichprobe	Gesamt[a]	160.669	13	0.000	0.907	0.092	0.070
	'mittel'[b]	152.330	13	0.000	0.912	0.089	0.071
	'extrem'[c]	183.243	13	0.000	0.881	0.099	0.074
	'inkonsistent'[d]	226.577	13	0.000	0.893	0.111	0.092

Anmerkungen: [a] Gesamtstichprobe $n = 1340$

[b] Teilgruppe mit mittlerer Antworttendenz; $n = 565$;

[c] Teilgruppe mit extremer Antworttendenz; $n = 445$;

[d] Teilgruppe mit inkonsistenter Antworttendenz; $n = 330$.

Zusammenhänge zwischen den beiden Konstrukten

Die Analyse der Zusammenhänge der einzelnen Skalen der beiden untersuchten Konstrukte bezogen auf die gesamte Stichprobe zeigt Tabelle 6.12. Die entsprechende Darstellung der Circumplexen Anordnung der beiden Konstrukte zeigt Abbildung 6.11.

Tabelle 6.12 Ergebnisse zu den Interkorrelationen der Skalen des AIST–R und des BFI–K für die Gesamtstichprobe nach mixed-Rasch Skalierung.

		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1.	Realistic	1.00	0.55	-0.11	-0.16	-0.09	0.15	-0.08	**-0.10**	0.02	-0.03	0.01
2.	Investigative	0.55	1.00	0.11	-0.08	-0.02	0.15	-0.01	-0.08	**0.21**	0.04	-0.00
3.	Artistic	-0.11	0.11	1.00	0.46	0.23	0.14	**0.15**	**0.12**	**0.61**	**0.12**	0.07
4.	Social	-0.16	-0.08	0.46	1.00	0.50	0.21	0.06	**0.31**	**0.22**	**0.23**	**0.13**
5.	Enterprising	-0.09	-0.02	0.23	0.50	1.00	0.40	**-0.16**	**0.45**	0.10	0.00	0.09
6.	Conventional	0.15	0.15	0.14	0.21	0.40	1.00	0.03	0.04	-0.02	0.01	**0.24**
7.	Neurotizismus	-0.08	-0.01	0.15	0.06	-0.16	0.03	1.00	-0.21	0.07	-0.05	-0.06
8.	Extraversion	-0.10	-0.08	0.12	0.31	0.45	0.04	-0.21	1.00	0.13	0.06	0.04
9.	Offenheit	0.02	0.21	0.61	0.22	0.10	-0.02	0.07	0.13	1.00	0.11	0.06
10.	Verträglichkeit	-0.03	0.04	0.12	0.23	0.00	0.01	-0.05	0.06	0.11	1.00	0.12
11.	Gewissenhaftigkeit	0.01	-0.00	0.07	0.13	0.09	0.24	-0.06	0.04	0.06	0.12	1.00

Anmerkungen: Gesamtstichprobe (Ausschluss von 3 Fällen); $n = 1340$; Korrelationen $r \geq |.08|$ fallen signifikant aus, $p \leq .01$, (zweiseitig).

Die Zusammenhänge der beiden Konstrukte auf der Ebene der einzelnen Skalen ist für die Personengruppe mit mittlerer Antworttendenz auf den AIST–R-Skalen in Tabelle 6.13, für die Personen mit eher extremen Antworttendenzen in Tabelle 6.14 und für die Personengruppe mit inkonsistenten Antworttendenzen in Tabelle 6.15 wiedergegeben.

Tabelle 6.13 Ergebnisse zu den Interkorrelationen der Skalen des AIST–R und des BFI–K für mittlere Antworttendenzen nach mixed-Rasch Skalierung.

		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1.	Realistic	1.00	0.59	-0.11	-0.15	-0.06	0.17	-0.04	-0.10	0.09	-0.06	-0.04
2.	Investigative	0.59	1.00	0.13	-0.06	-0.02	0.14	-0.01	-0.10	**0.26**	0.08	0.01
3.	Artistic	-0.11	0.13	1.00	0.45	0.22	0.13	**0.12**	**0.17**	**0.59**	**0.13**	0.10
4.	Social	-0.15	-0.06	0.45	1.00	0.45	0.20	0.08	**0.31**	**0.28**	**0.24**	0.10
5.	Enterprising	-0.06	-0.02	0.22	0.45	1.00	0.39	**-0.18**	**0.45**	**0.12**	-0.04	0.07
6.	Conventional	0.17	0.14	0.13	0.20	0.39	1.00	0.06	0.08	-0.01	-0.06	**0.22**
7.	Neurotizismus	-0.04	-0.01	0.12	0.08	-0.18	0.06	1.00	-0.19	0.02	-0.06	-0.02
8.	Extraversion	-0.10	-0.10	0.17	0.31	0.45	0.08	-0.19	1.00	0.16	0.07	0.08
9.	Offenheit	0.09	0.26	0.59	0.28	0.12	-0.01	0.02	0.16	1.00	0.19	0.10
10.	Verträglichkeit	-0.06	0.08	0.13	0.24	-0.04	-0.06	-0.06	0.07	0.19	1.00	0.11
11.	Gewissenhaftigkeit	-0.04	0.01	0.10	0.10	0.07	0.22	-0.02	0.08	0.10	0.11	1.00

Anmerkungen: Teilgruppe mit mittlerer Antworttendenz; $n = 565$; Korrelationen $r \geq |.11|$ signifikant, $p \leq .01$, (zweiseitig).

Tabelle 6.14 Ergebnisse zu den Interkorrelationen der Skalen des AIST–R und des BFI–K für extreme Antworttendenzen nach mixed-Rasch Skalierung.

		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1.	Realistic	1.00	0.46	-0.03	-0.08	-0.08	0.15	**-0.17**	-0.06	0.05	-0.00	0.12
2.	Investigative	0.46	1.00	0.23	0.10	0.04	0.20	0.02	-0.03	**0.21**	0.03	0.06
3.	Artistic	-0.03	0.23	1.00	0.40	0.20	0.17	**0.17**	0.07	**0.60**	0.08	-0.02
4.	Social	-0.08	0.10	0.40	1.00	0.52	0.21	0.06	**0.28**	**0.17**	**0.25**	0.10
5.	Enterprising	-0.08	0.04	0.20	0.52	1.00	0.43	**-0.14**	**0.40**	0.03	0.03	0.11
6.	Conventional	0.15	0.20	0.17	0.21	0.43	1.00	0.07	-0.06	-0.04	0.06	**0.32**
7.	Neurotizismus	-0.17	0.02	0.17	0.06	-0.14	0.07	1.00	-0.27	0.11	-0.02	-0.11
8.	Extraversion	-0.06	-0.03	0.07	0.28	0.40	-0.06	-0.27	1.00	0.11	0.04	0.03
9.	Offenheit	0.05	0.21	0.60	0.17	0.03	-0.04	0.11	0.11	1.00	0.07	0.01
10.	Verträglichkeit	-0.00	0.03	0.08	0.25	0.03	0.06	-0.02	0.04	0.07	1.00	0.12
11.	Gewissenhaftigkeit	0.12	0.06	-0.02	0.10	0.11	0.32	-0.11	0.03	0.01	0.12	1.00

Anmerkungen: Teilgruppe mit extremer Antworttendenz; $n = 445$; Korrelationen $r \geq |.13|$ signifikant, $p \leq .01$, (zweiseitig).

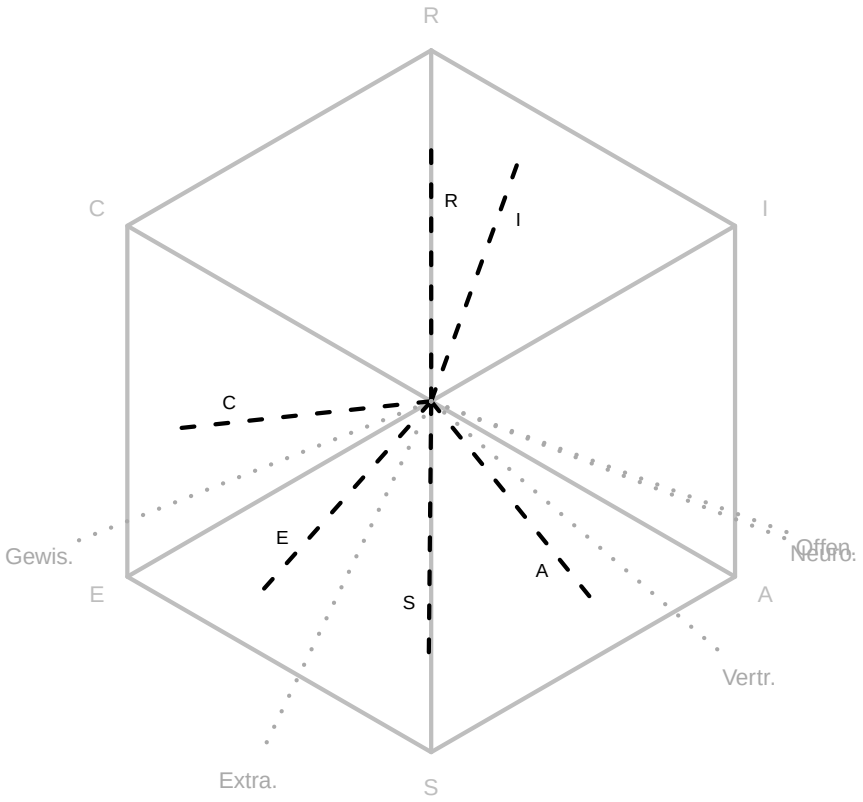


Abbildung 6.11 Darstellung der circumplexen Anordnung für die Gesamtstichprobe; $n = 1340$;
gestrichelte Linien: Sechs Dimensionen beruflicher Interessen (R: *Realistic*, I: *Investigative*, A: *Artistic*, S: *Social*, E: *Enterprising*, C: *Conventional*);
gepunktete Linien: Fünf Dimensionen der Persönlichkeit.

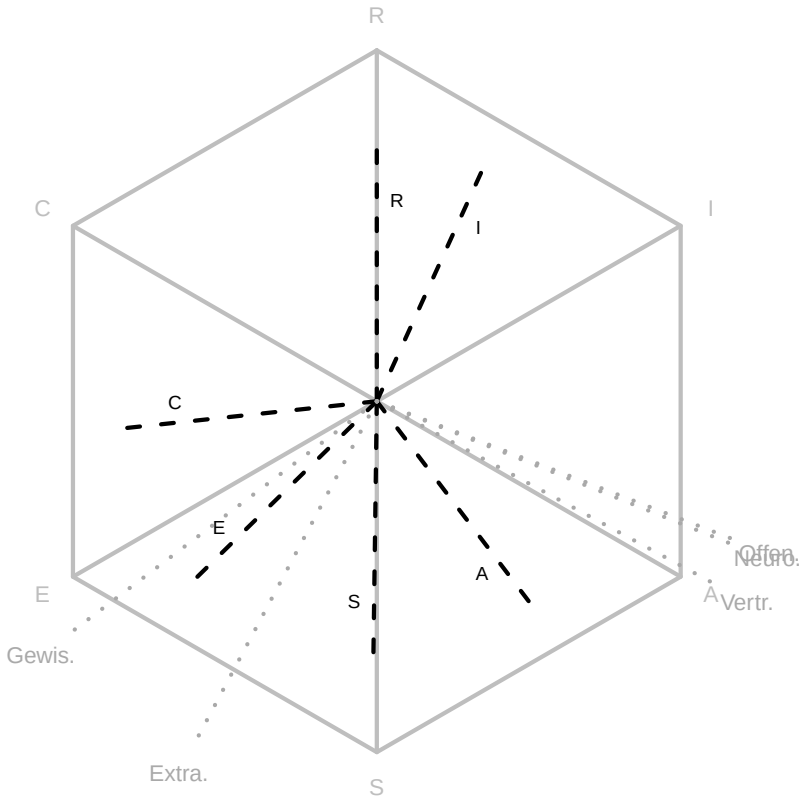


Abbildung 6.12 Darstellung der circumplexen Anordnung für die Teilstichprobe mit mittlerer Antworttendenz; $n = 565$;
gestrichelte Linien: Sechs Dimensionen beruflicher Interessen (R: *Realistic*, I: *Investigative*, A: *Artistic*, S: *Social*, E: *Enterprising*, C: *Conventional*);
gepunktete Linien: Fünf Dimensionen der Persönlichkeit.

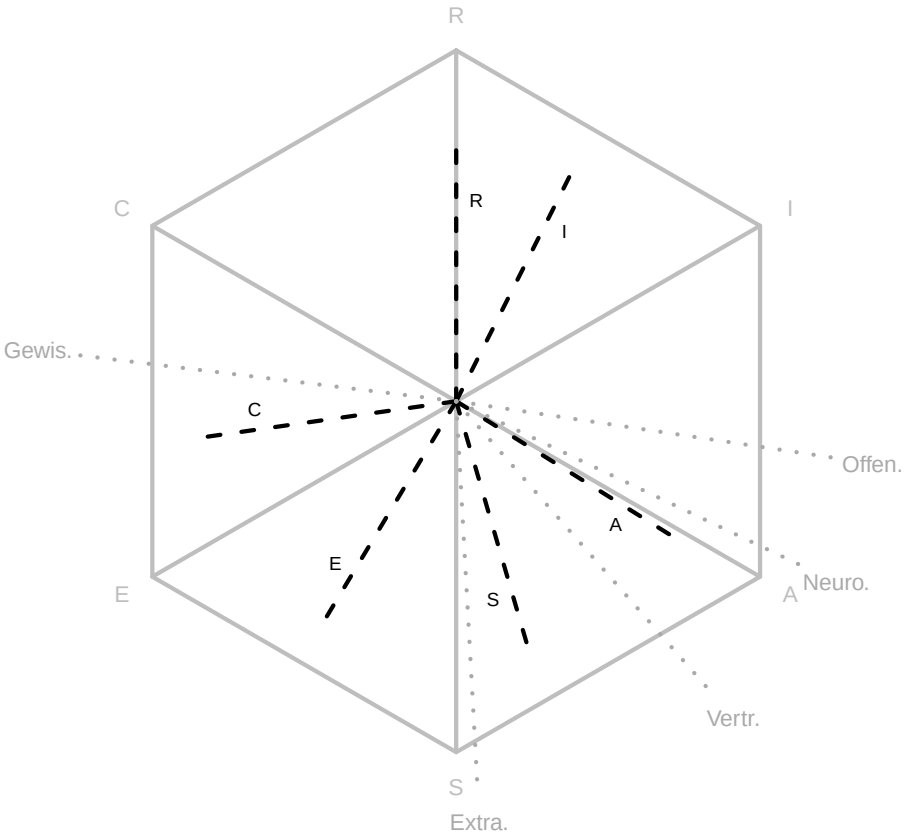


Abbildung 6.13 Darstellung der circumplexen Anordnung für die Teilstichprobe mit extremer Antworttendenz; $n = 445$;
gestrichelte Linien: Sechs Dimensionen beruflicher Interessen (R: *Realistic*, I: *Investigative*, A: *Artistic*, S: *Social*, E: *Enterprising*, C: *Conventional*);
gepunktete Linien: Fünf Dimensionen der Persönlichkeit.

Tabelle 6.15 Ergebnisse zu den Interkorrelationen der Skalen des AIST–R und des BFI–K für inkonsistente Antworttendenzen nach mixed-Rasch Skalierung.

		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1.	Realistic	1.00	0.53	-0.12	-0.08	-0.11	0.13	0.05	-0.06	-0.06	-0.06	0.02
2.	Investigative	0.53	1.00	0.20	0.04	-0.07	0.09	0.04	-0.01	**0.23**	0.01	0.02
3.	Artistic	-0.12	0.20	1.00	0.44	0.31	0.15	0.10	0.12	**0.61**	0.12	0.09
4.	Social	-0.08	0.04	0.44	1.00	0.67	0.33	-0.04	**0.40**	**0.25**	**0.20**	0.10
5.	Enterprising	-0.11	-0.07	0.31	0.67	1.00	0.41	**-0.19**	**0.51**	**0.15**	0.01	0.05
6.	Conventional	0.13	0.09	0.15	0.33	0.41	1.00	-0.06	0.11	-0.00	0.04	**0.20**
7.	Neurotizismus	0.05	0.04	0.10	-0.04	-0.19	-0.06	1.00	-0.19	0.07	-0.09	-0.08
8.	Extraversion	-0.06	-0.01	0.12	0.40	0.51	0.11	-0.19	1.00	0.11	0.09	-0.04
9.	Offenheit	-0.06	0.23	0.61	0.25	0.15	-0.00	0.07	0.11	1.00	0.09	0.06
10.	Verträglichkeit	-0.06	0.01	0.12	0.20	0.01	0.04	-0.09	0.09	0.09	1.00	0.12
11.	Gewissenhaftigkeit	0.02	0.02	0.09	0.10	0.05	0.20	-0.08	-0.04	0.06	0.12	1.00

Anmerkungen: Teilgruppe mit inkonsistenter Antworttendenz; $n = 330$; Korrelationen $r \geq |.15|$ signifikant, $p \leq .01$, (zweiseitig).

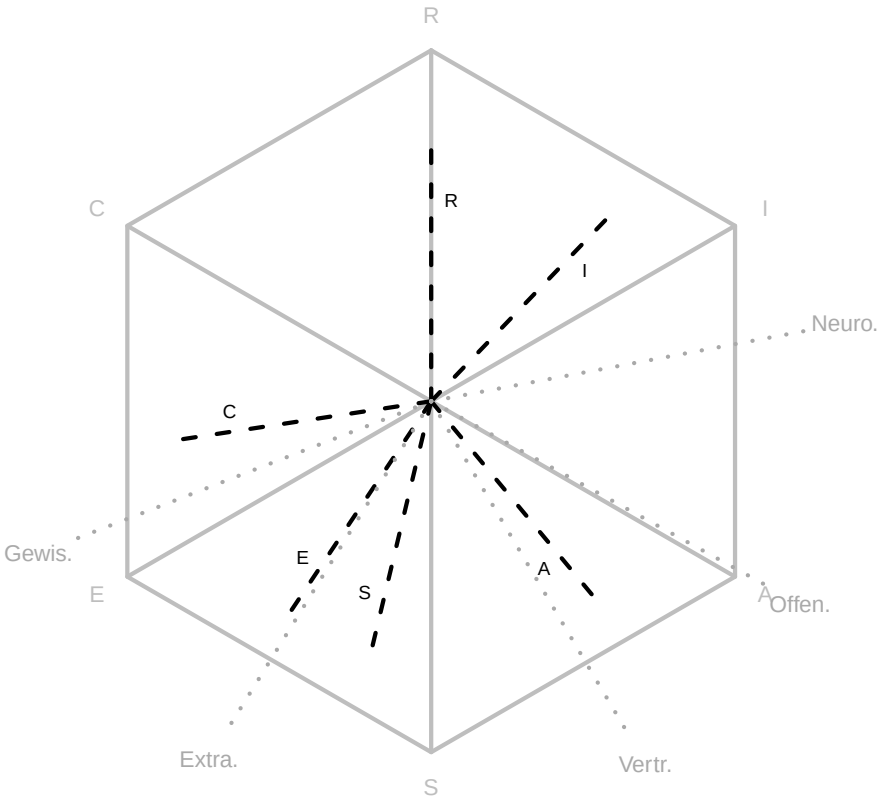


Abbildung 6.14 Darstellung der circumplexen Anordnung für die Teilstichprobe mit inkonsistenter Antworttendenz; $n = 330$;
gestrichelte Linien: Sechs Dimensionen beruflicher Interessen (R: *Realistic*, I: *Investigative*, A: *Artistic*, S: *Social*, E: *Enterprising*, C: *Conventional*);
gepunktete Linien: Fünf Dimensionen der Persönlichkeit.

Diskussion

Ein Ziel der vorliegenden Analysen besteht darin zu überprüfen, ob sich die in früheren Untersuchungen (z. B. de Fruyt & Mervielde, 1997; Gottfredson et al., 1993; Holland et al., 1994; Larson & Borgen, 2002; McCrae & Costa, 1983a) und Metaanalysen (Barrick et al., 2003; Larson et al., 2002; Mount et al., 2005) gefundenen Zusammenhänge zwischen den Dimensionen der Persönlichkeit und beruflicher Interessen replizieren lassen. Daneben soll der potentielle Einfluss des unterschiedlichen Gebrauchs der fünfstufigen Antwortskala (MRS vs. ERS) auf die Passung der Daten auf die theoretisch angenommene hexagonale Struktur der beruflichen Interessenorientierungen nach Holland (1959, 1997) untersucht werden.

Die Ergebnisse der durchgeführten Analysen bestätigen zunächst die in früheren Untersuchungen gefundenen Zusammenhänge. So zeigen sich bezogen auf die Gesamtstichprobe, sowie für die nach ihrer Antworttendenz auf dem AIST–R aufgeteilten Teilgruppen, in Übereinstimmung mit anderen Untersuchungen, positive Zusammenhänge zwischen den Dimensionen *Offenheit* und *Artistic*. Auch die nicht bei allen, so doch aber bei der Mehrzahl der früheren Untersuchungen gefundenen positiven Zusammenhänge zwischen den Dimensionen *Offenheit* und *Social* sowie *Offenheit* und *Investigative* können mit der vorliegenden Studie bei Betrachtung der Gesamtstichprobe sowie für die Teilgruppen repliziert werden. Der positive Zusammenhang zwischen *Extraversion* und den beiden Interessendimensionen *Social* und *Enterprising* lässt sich mit den vorliegenden Daten ebenso replizieren.

Die Analysen der vorliegenden Studie zeigen allerdings auch, dass der unterschiedliche Gebrauch der fünfstufigen Antwortskala des AIST–R in nicht unerheblichem Ausmaße die Interkorrelationen der Skalen der beiden Konstrukte beeinflusst. So ergeben sich für die Teilgruppe mit extremer Antworttendenz (*Extremkreuzer*) substanziell negative Zusammenhänge zwischen den Dimensionen *Neurotizismus* und *Realistic* (vgl. Tabelle 6.14), wohingegen dieser Zusammenhang in den anderen Teilgruppen (nahezu) nicht gefunden werden kann, bzw. für diese Teilgruppen nicht signifikant ist.

Für die gesamte Stichprobe ergibt sich für die Persönlichkeitsdimension *Extraversion*, neben den oben diskutierten Zusammenhängen, noch ein – zwar geringer – aber dennoch signifikanter positiver Zusammenhang zu der AIST–R-

Dimension *Artistc*; und ein geringer negativer Zusammenhang zur Dimension *Realistic*. In der Teilgruppe der Personen mit einer eher konsistenten Präferenz für mittlere Antwortkategorien (*Mittelkreuzer*), wird der negative Zusammenhang zwischen *Extraversion* und *Realistic* nicht signifikant. Für die Teilgruppe mit eher extremer Antworttendenz (*Extremkreuzer*) lässt sich dagegen der für alle anderen Teilgruppen sowie für die Gesamtstichprobe gefundene positive Zusammenhang zwischen *Extraversion* und *Artistc* nicht finden. Ebenso zeigt sich in dieser Teilgruppe kein positiver Zusammenhang zwischen der Persönlichkeitsdimension *Verträglichkeit* und *Artistc*, wie er der Teilgruppe der *Mittelkreuzer* sowie in der Gesamtstichprobe gefunden werden konnte. Vergleichsweise einheitlich sind die Befunde zu den Zusammenhängen zwischen der Persönlichkeitsdimension *Gewissenhaftigkeit* und den sechs AIST–R-Dimensionen. In allen Teilgruppen zeigt sich hier ein positiver Zusammenhang zur Dimension *Conventional*. In der Gesamtstichprobe erreicht der geringe positive Zusammenhang zwischen *Gewissenhaftigkeit* und *Social* ($r = .13$) das Signifikanzniveau. Dieser Zusammenhang erreicht zwar in den anderen Teilgruppen eine mit $r = .10$ vergleichbare Stärke, wobei dieser dort, auch wegen des jeweils geringeren Umfangs der Teilgruppen, nicht signifikant wird.

Beachtenswert sind ebenfalls die differentiellen Befunde zur Anpassung der vorliegenden Daten, einerseits an die theoretisch postulierte, ideale Circumplexstruktur und andererseits an die durch die Normstichprobe des AIST–R vorgegebene Struktur (vgl. Tabelle 6.11). Insgesamt gelingt die Anpassung der Daten sowohl für die Gesamtstichprobe als auch alle Teilgruppen hier am besten an die durch die Normstichprobe des AIST–R vorgegebene Struktur. Auch wenn der Vergleich der einzelnen Modelle mit dem jeweiligen restriktiveren Nullmodell über den komparativen Fit–Index (CFI – Bentler, 1990), für alle getesteten Modelle nicht die von Bühner (2011, S. 427) empfohlene Grenze von $CFI \geq .95$ erreicht (vgl. auch Hu & Bentler, 1999), so ergeben sich bei der Anpassung der Daten an die durch die Normstichprobe vorgegebene Struktur hier deutlich bessere Werte der Anpassung für den CFI. So erreicht der CFI bei der Anpassung der Daten an die durch die Normstichprobe vorgegebene Struktur für die Teilgruppe der Personen mit mittlerer Antworttendenz sowie für die Gesamtstichprobe noch einen nahe an dem akzeptablen Wert von $CFI \geq .95$ liegenden Wert von $CFI = .91$. Demgegenüber liegt das

Ausmaß der Modellpassung, bezogen auf den CFI, an die ideale circumplexe Struktur, für alle Teilgruppen und die Gesamtstichprobe deutlich darunter ($CFI = .75$ bis $CFI = .81$). Auch bei Betrachtung der absoluten Modellpassung ($SRMR$) ergibt sich für die Anpassung der Daten für alle Teilgruppen und die Gesamtstichprobe eine bessere Passung. Der $SRMR$ unterschreitet für alle Teilgruppen und die Gesamtstichprobe hier den von Bühner (2011, S. 427) empfohlenen Wert von $SRMR \leq .11$. Nach einer Simulationsstudie von Cheung und Rensvold (2002), können Differenzen von $\delta_{CFi} > .01$ dahingehend interpretiert werden, dass eine Messinvarianz zwischen unterschiedlichen Gruppen nicht gegeben ist. Wendet man diesen Bewertungsmaststab auf die hier vorliegenden Ergebnisse (vgl. Tabelle 6.11) an, so ergibt sich folgende Interpretation: Sowohl bei der Anpassung an die ideale circumplexe Struktur, als auch bei der Anpassung an die durch die Normstichprobe vorgegebene circumplexe Struktur bestehen zwischen den jeweiligen Teilgruppen der mittel, extrem und inkonsistent antwortenden Personen mit $\delta_{CFi} > .01$ bedeutsame Unterschiede (vgl. Tabelle 6.11). Dies ist, ebenso wie die Befunde zu den differentiell ausfallenden korrelativen Zusammenhängen zwischen den einzelnen Dimensionen der Konstrukte vor dem Hintergrund der Personenparameterschätzung mit dem mixed-Rasch-Modell, ein bedeutender Befund. Die durch die mixed-Rasch-Skalierung vorgenommene Abbildung der Personenklassen mit unterschiedlichem Antwortverhalten auf einer gemeinsamen latenten Dimension ist offenbar nicht ausreichend, um die verzerrenden „Fehler“ aus den Antworttendenzen vollständig zu eliminieren.

Zusammenfassend lässt sich auf Basis der vorgestellten Befunde folgern, dass ein individuell unterschiedlich ausgeprägter Antwortstil Auswirkungen auf die untersuchten Zusammenhänge einzelner Dimensionen der beiden Konstrukte Persönlichkeit und berufliche Interessenorientierungen hat. Insbesondere bei der, in der Gesamtstichprobe größten, Teilgruppe von Personen mit einer Tendenz zu den mittleren Antwortkategorien auf den vier AIST–R-Dimensionen, liegen in der grafischen Darstellung der Zusammenhänge der einzelnen Dimensionen, die (empirischen) Winkel der BFI–K-Dimensionen *Verträglichkeit*, *Offenheit* und *Neurotizismus* eher nahe zusammen – zwischen den Dimensionen *Artistic* und *Investigative* (vgl. Abbildung 6.12). Diese (mittlere) Antworttendenz scheint auf die Zusammenhänge der Dimensionen beider Kon-

strukte für die Gesamtstichprobe einen prägenden Einfluss zu haben, da sich hier ebenfalls ein ähnliches Bild für die grafischen Darstellung der Konstruktzusammenhänge ergibt (vgl. Abbildung 6.11). Interessanterweise scheint darüber hinaus die unterschiedliche Lage der Dimensionen in der circumplexen Anordnung für den Bereich Persönlichkeit stärker von den unterschiedlichen Antworttendenzen – auf den vier AIST–R-Dimensionen – beeinflusst zu sein.

Kapitel 7

Untersuchungen zum Nähe–Distanz-Antwortprozess

Im zweiten Abschnitt des empirischen Teils der vorliegenden Arbeit werden die Ergebnisse der vorangegangenen Untersuchungen einerseits aufgegriffen und andererseits die Prämisse einer universellen Geltung eines Dominanz-Antwortmodells für alle Personen der vorliegenden Stichproben hinterfragt. Das Ziel der Analysen und Untersuchungen besteht darin in den vorliegenden Daten Personen oder Personengruppen zu identifizieren, welche möglicherweise bei der Beantwortung der Fragen in den einzelnen psychometrischen Skalen (implizit) einem unterschiedlichen Antwortmodell folgen. Dieses Ziel stützt sich neben theoretischen Überlegungen zum Antwortprozess (vgl. Abschnitt 4.7 in Kapitel 4, sowie Kapitel 3) bei der Erfassung unterschiedlicher Konstrukte (vgl. Kapitel 2) auch auf den widersprüchlichen Befund aus den Analysen der Untersuchungen in Studie 6.2 zur (kumulativen) Skalierbarkeit des BFI–K. Dabei gelingt einerseits die Anpassung eines eindimensionalen Modells zur Abbildung einer Dominanz-Relation zwischen Items und Personen, wobei dies aber andererseits mit erheblichen Schwellenvertauschungen der eigentlich ordinal aufsteigend geforderten Schwellenparameter des summativen Skalierungsmodells einhergeht. Insgesamt indiziert dies eine Analyse der Daten nach einem Skalierungsmodell, welches auch einen *Nähe–Distanz*-Antwortprozess berücksichtigt.

7.1 Seriation und Multidimensionale Skalierung zur Klassifikation der Personenstichprobe nach impliziten Antwortmodellen

Einleitung

In den theoretischen Kapiteln ist bereits dargestellt, dass zur Indexbildung aus mehreren Items einer psychometrischen Skala unterschiedliche Methoden der Skalierung herangezogen werden können (vgl. Kapitel 1). Diese beiden in Abschnitt 1.4 dargestellten Skalierungsmethoden für *Dominanz*-Antwortprozesse einerseits und *Nähe–Distanz*-Antwortprozesse andererseits lassen sich jeweils mit entsprechenden psychometrischen Antwortmodellen zur Modellierung der Daten beschreiben (vgl. Kapitel 4). Die beiden unterschiedlichen Skalierungsmodelle werden in den entsprechenden bislang vorliegenden empirischen Untersuchungen zur Skalierbarkeit, jeweils in Abhängigkeit der untersuchten Konstrukte, für alle Personen einer entsprechenden Stichprobe zur Modellierung der Daten angewendet (z. B. Jansen, 1981, 1983; Post et al., 2001; van Schuur, 2006, sowie Abschnitt 4.7 in Kapitel 4). Dieses Vorgehen verfolgt das Prinzip, den universellen Geltungsbereich (über alle Personen einer Stichprobe bzw. Population) des jeweils angewendeten Skalierungsmodells nachzuweisen (vgl. Abschnitt 3.3 und Abschnitt 4.4). Demgegenüber stehen – resultierend aus der universellen Anwendung jeweils eines der beiden Skalierungsmodelle – zahlreiche empirische Befunde, bei denen immer wieder einzelne Personen oder Personengruppen gefunden werden, deren Antwortverhalten nicht mit dem jeweils angewendeten Skalierungsmodell zu erklären sind (z. B. Ferrando, 2012; Reise & Waller, 1993; Tellegen, 1988, sowie allgemein die in Kapitel 3 insbesondere in den Abschnitten 3.1 und 3.2 dargestellten Phänomene und empirischen Befunde zu abweichendem Antwortverhalten). Die daraus resultierende fehlende Passung der einzelnen Personen oder Personengruppen (vgl. Abschnitt 4.4.2) stellt eine *lokale Modellverletzung* dar und kann mit entsprechenden Fit-Statistiken identifiziert werden (vgl. Abschnitt 4.4.2). In der bislang vorliegenden empirischen Literatur werden Personen mit fehlender Passung zu den meist zugrunde gelegten summativen Modellen für den *Dominanz*-Antwortprozess in der Regel als *unskalierbare* Personengruppe zusammengefasst (z. B. Bem, 1977; Bem &

Allen, 1974; Bem & Funder, 1978; Dayton & Macready, 1980; Keller, 2012) und mit entsprechenden Mischverteilungsmodellen einer gesonderten latenten Klasse zugeordnet (Formann, 2002; Ponocny & Klauer, 2002; Rost et al., 1997, 1999; Rost & Georg, 1991). Als alternatives Vorgehen schlagen verschieden Autoren auch die Skalierung der Daten nach einem Unfoldingmodell vor, welches geeignet ist, einen *Nähe–Distanz*-Antwortprozess abzubilden (Andrich, 1988; Drasgow et al., 2010a, 2010b; Kyngdon, 2006; Rost & Luo, 1997; van Schuur, 1995; van Schuur & Kruijtbosch, 1995). Wie beispielsweise Post et al. (2001) darlegen, ergibt sich die Wahl des entsprechenden Skalierungsmodells – jeweils für eine gesamte Stichprobe – maßgeblich durch die Inhalte der Items und die theoretische Fundierung der zu erfassenden Eigenschaft oder Einstellung (vgl. auch Mellenbergh, 2001).

Als Erweiterung dieser alternativen Empfehlungen und theoretischen Überlegungen zur Wahl eines Skalierungsmodells wird in der vorliegenden Analyse die Hypothese aufgestellt, dass sich diese beiden unterschiedlichen Antwortprozesse auch als Spezifika einzelner Personen oder Personengruppen darstellen lassen, welche möglicherweise aus deren individuell unterschiedlicher Rezeption der vorgegebenen Skalen (bzw. deren Items) resultiert. Diese Hypothese stützt sich auf die in Abschnitt 3.2.2 dargestellten Befunde zur Interaktion zwischen der Polarität von Items, Merkmalen und Antwortverzerrungen mit den daraus resultierenden methodischen Problemen (vgl. auch Matschinger & Krebs, 1998). So argumentieren beispielsweise Matschinger und Krebs (1998), dass sich der oft gefundene zusätzliche (Methoden-)Faktor, beim Versuch der eindimensionalen Abbildung bipolare gedachter Konstruktdimensionen, auf einen Nähe–Distanz-Antwortprozess bei der Beantwortung der Items zurückzuführen ist (vgl. auch Maraun & Rossi, 2001; Matschinger & Krebs, 1998; Post et al., 2001; Schönemann, 1970; van Schuur & Kiers, 1994, sowie ausführlicher Abschnitt 4.7). Ferner stützt sich die Hypothese auf die beispielsweise von Drasgow et al. (2010b), Carter et al. (2010) und auch Coombs und Coombs (1976) formulierten theoretischen Überlegungen zu interindividuellen Unterschieden bei der Antizipation einer der beiden unterschiedlichen Antwortprozesse. Danach kann einerseits die wahrgenommene Ambiguität der Items (z. B. Coombs & Coombs, 1976) und andererseits das Ausmaß der Introspektion bei der Beantwortung der Items (Carter et al., 2010) individuell unterschiedlich ausfallen,

was insgesamt dazu führen kann, dass für manche Personen die betreffende Skala eher nach einem Unfoldingmodell (im Vergleich zu einem Dominanz-Antwortmodell) zu modellieren ist (z. B. Drasgow et al., 2010b, sowie Abschnitt 4.7).

Es soll daher hier untersucht werden, ob sich in den erhobenen Fragebogendaten hinsichtlich ihrer impliziten Antwortmodelle unterscheidbare Personengruppen finden lassen. Personen oder Personengruppen, welche auf *dieselben* Items der jeweiligen psychometrischen Skala implizit, entweder nach einem *Dominanz*-Antwortprozess oder nach einem *Nähe–Distanz*-Antwortprozess, antworten.

Daten

Die Datenbasis für die Analysen bilden die in Abschnitt 5.2 beschriebenen beiden Stichprobe aus dem Erhebungszeitraum 2007 – 2009 (Stichprobe I) und 2010 –2011 (Stichprobe II), welche $n = 734$ (Stichprobe I) und $n = 609$ (Stichprobe II) Studierende der Universität der Bundeswehr umfasst. Die Personen beantworteten einerseits 60 Items des AIST–R (vgl. Bergmann & Eder, 2005), welcher mit jeweils zehn Items sechs Dimension beruflicher Interessenorientierungen erfasst. Daneben beantworteten die Personen 20 Items des BFI–K (vgl. Schmolck, 2003, 2004, 2005, 2006a, 2006b), welcher fünf Dimensionen der Persönlichkeit erfasst. Die Items beider Inventare werden über eine fünfstufige Antwortskala beantwortet (vgl. Abschnitte 5.1.1 und 5.1.3 in Kapitel 5 *Stichproben und Instrumente*). Zusätzlich werden in diese Analyse auch die Daten aus der Bearbeitung der Fragen zur Musikpräferenz aus dem STOMP (vgl. Abschnitt 5.1.2) mit ausgewertet.

Methode

Die Antwortdaten werden in der vorliegenden Analyse mit zwei unterschiedlichen Verfahren skaliert. Die beiden Verfahren korrespondieren dabei mit den beiden unterschiedlichen Antwortprozessen, wie sie im Kapitel 1.3 *Skalierung von Fragebogendaten* dargestellt wurden. Zur Skalierung der Daten werden dazu unterschiedliche psychometrische Antwortmodelle eingesetzt. Einerseits werden die Daten nach dem *Partial Credit Model* (PCM – Masters, 1982)

skaliert, welches als (summativ) kumulatives Modell zur Modellierung eines Dominanz-Antwortprozesses geeignet ist. Andererseits wird auf die Anwendung des Prinzips der Seriation der Daten (Hubert, 1974, 1976) mit Methoden der *Multidimensionalen Skalierung* (Kruskal, 1964b) zur Bestimmung der Objektkoordinaten zurückgegriffen. Die Methode der *Multidimensionalen Skalierung* (MDS) wird zur Skalierung angewendet, um einen Nähe–Distanz-Antwortprozesses zu modellieren ohne restriktive Annahmen zu einem spezifischen Antwort- und Skalierungsmodell mit unimodaler Wahrscheinlichkeitsfunktion zu machen. Dieses Vorgehen stützt sich auf einen Vorschlag von Warrens und Heiser (2006), nach dem sich ein eindimensionales Unfoldingmodell – also der Nähe–Distanz-Antwortprozess – durch die Anwendung der MDS modellieren lässt (vgl. Abschnitt 4.5.4); wobei dabei nur die Koordinaten der ersten Dimension interpretiert werden. Der Vorteil dieses Vorgehens gegenüber der Anwendung expliziter parametrischer oder auch nichtparametrischer Unfoldingmodelle (vgl. Kapitel 4.3) liegt darin, dass die MDS das grundlegende Prinzip des Nähe–Distanz-Antwortprozesses modelliert, ohne dabei allzu strenge Annahmen bezüglich einer bestimmten parametrischen Modellierung der Item-Response-Funktion (ICC) zu machen (van Schuur, 2006; Warrens & Heiser, 2006). Das hierbei angewendete Prinzip weist Analogien zu dem Prinzip der Seriation zur Reorganisation von Daten auf (D. G. Kendall, 2004, vgl. auch Abschnitt 4.5.4).

Das Vorgehen zur Identifikation der beiden unterschiedlichen Personengruppen besteht in der Anwendung lokaler Personen-Fit-Indizes (vgl. Abschnitt 4.4.2) nach erfolgter Bestimmung der jeweiligen Modellparameter. Je nach angewendetem Skalierungsmodell werden hier die entsprechenden Fit-Indizes zur Beurteilung der Personenpassung herangezogen. Für die Skalierung der Daten mit der MDS wird der zeilenweise (Personen) über alle Spalten (Items) aggregierte Wert des STRESS Koeffizienten (vgl. de Leeuw & Bettonvil, 1986; Kruskal, 1964a, 1964b) verwendet (vgl. auch Abschnitt 4.5.4 zur Bedeutung und Interpretation des STRESS Koeffizienten). Bei der Skalierung der Daten nach dem kumulativen Antwortmodell wird die polytome Verallgemeinerung des Q-Index (Tarnai & Rost, 1990) eingesetzt (vgl. Abschnitt 4.4.3). Zur Beurteilung der Personenpassung zum jeweiligen Antwortprozess, werden die empirischen Dichtfunktionen (Verteilungen) der jeweiligen Indizes bestimmt und

der Median der Verteilung des jeweiligen Passungsindexes als Maß der Zentralen Tendenz berichtet (vergleiche z. B. Abbildung 7.5 und 7.6). Als zum jeweiligen Skalierungsmodell passend werden diejenigen Personen klassifiziert, deren jeweiliger Personen-Fit-Index unter einem spezifischen *cut-off* Kriterium liegt.

Zur Bestimmung eines solchen *cut-off* Kriteriums können prinzipiell zwei unterschiedliche Strategien verfolgt werden. Einerseits können für beide Indizes feste cut-off Grenzen gewählt werden, die sich an Befunden aus der Literatur zu den Verteilungseigenschaften des jeweiligen Index orientieren. Andererseits könne Quartilgrenzen aus der empirischen Verteilung der Personen-Fit-Indizes, basierend auf der analysierten Stichprobe, gewählt werden. In der vorliegenden Untersuchung werden für die Analysen beide Ansätze verfolgt. im Rahmen des Ansatzes mit festen cut-off Grenzen werden für den STRESS-Index dabei die Empfehlung von Kruskal (1964a) zugrunde gelegt, wonach ein Wert $STRESS < .05$ als *gut* einzustufen ist (vgl. Tabelle 4.6). Für den Q-Index werden dabei die Ergebnisse der Untersuchungen zur Verteilung von Q von Tarnai und Rost (1990) zugrunde gelegt. Danach ist ein Wert von $Q < .2$ als Indikator für eine (lokale) Passung zum kumulativen Skalierungsmodell einzustufen (Tarnai & Rost, 1990). im Rahmen des verteilungsbasierten Ansatzes wird als cut-off Grenze das untere Quartil der jeweiligen Verteilung des Personen-Fit-Indexes gewählt.

Durchgeführte werden die Analysen mit den *R*-Paketen `pairwise` (Heine, 2019; Heine & Tarnai, 2015) und `smacof` (de Leeuw & Mair, 2009) für die freie Statistik Umgebung *R* (R Core Team, 2018). Die Bewertung der Klassifikation erfolgt über die visuelle Inspektion der grafischen Darstellung der anhand der Modellparameter reorganisierten Datenmatrizen durch so genannte *Bertin-Plots* (vgl. z. B. Bertin, 1977; de Falguerolles, Friedrich & Sawitzki, 1997, – sowie Abschnitt 4.5.4 für eine Darstellung des Prinzips solch einer grafischen Darstellung).

Anhand der cut-off Grenzen der beiden Indizes für die lokale Modellpassung (STRESS und Q-Index) werden dichotome Indikatorvariablen gebildet. Zur Untersuchung der Eindeutigkeit des Klassifikationsergebnisses der Personen zu den jeweiligen Antwortprozessen werden diese Indikatorvariablen für jede Dimension der drei Konstrukte getrennt kreuztabelliert. Dabei soll überprüft

werden ob die beiden Indikatoren für die Personenpassung zu konvergenten Befunden bezüglich des (impliziten) Antwortmodells der Personen führen.

Ergebnisse

Die empirischen Verteilungen der Dichtefunktion für den Q-Index über alle Skalen der drei untersuchten Konstrukte weisen zunächst darauf hin, dass bezogen auf die festen absoluten cut-off Werte für den Q-Index (cut-off$_Q$ = .2) eher geringe lokale Abweichungen vom damit überprüften Dominanz-Antwortprozess in den Daten vorliegen. So liegt für den BFI–K der Wert für den *Median* der empirischen Verteilung von Q über alle Skalen und beide Stichproben bei sehr niedrigen Werten (vgl. Abbildung 7.1 und 7.2) zwischen $Q_{\text{Median}} = 0.021$ (*Gewissenhaftigkeit*, Stichprobe I) und $Q_{\text{Median}} = 0.162$ (*Offenheit*, Stichprobe II). Für den AIST–R (vgl. Abbildung 7.5 und 7.6) liegen die Werte des Medians in einem vergleichbaren Bereich zwischen $Q_{\text{Median}} = 0.061$ (*Social*, Stichprobe I und Stichprobe II) und $Q_{\text{Median}} = 0.124$ (*Artistic*, Stichprobe II). Schließlich liegen die Werte des Medians für den STOMP (vgl. Abbildung 7.9 und 7.10) über die vier Skalen und beide Stichproben am höchsten in einem Bereich von $Q_{\text{Median}} = 0.055$ (*Upbeat & Conventional*, Stichprobe I) und $Q_{\text{Median}} = 0.279$ (*Energetic & Rhythmic*, Stichprobe I). Ein vergleichbares Bild ergibt sich auch für die empirische Dichtefunktion des Zeilen-*STRESS* als Index für die Personenpassung zu einem Nähe–Distanz-Antwortprozess. So liegt der Wert des Medians des STRESS über beide Stichproben, alle drei Konstrukte und deren Skalen um den von Kruskal (1964a) als *ausreichend, mäßig, angemessen* [fair] (vgl. Tabelle 4.6) bezeichneten Wert von $STRESS = 0.10$; vgl. Abbildungen 7.3 und 7.4, 7.7 und 7.8 sowie 7.11 und 7.12).

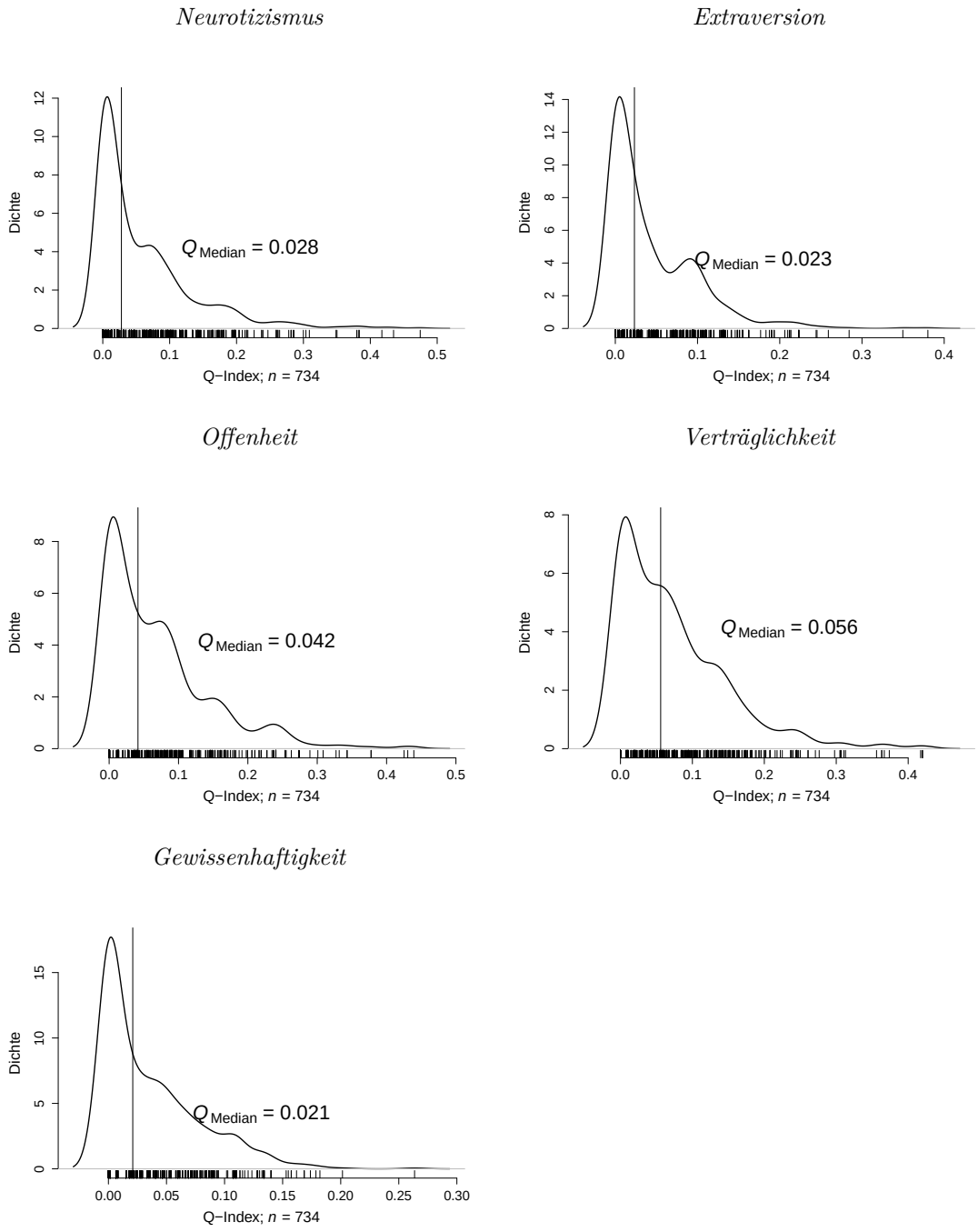


Abbildung 7.1 Stichprobe I: Dichtefunktionen des Q-Index zur Klassifikation der Personenpassung bei kumulativem Antwortmodell (PCM) für fünf Skalen des BFI–K.

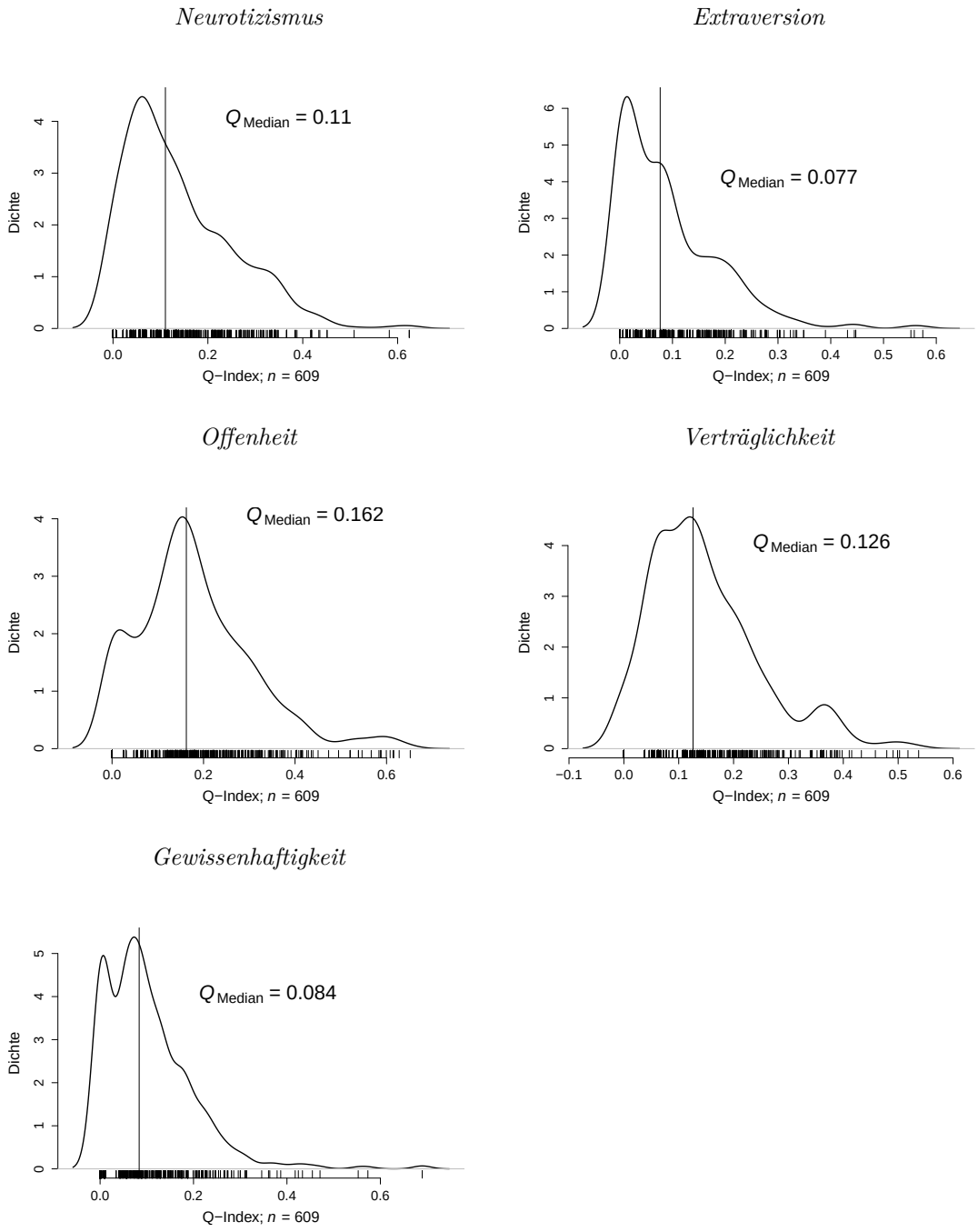


Abbildung 7.2 Stichprobe II: Dichtefunktionen des Q-Index zur Klassifikation der Personenpassung bei kumulativem Antwortmodell (PCM) für fünf Skalen des BFI–K.

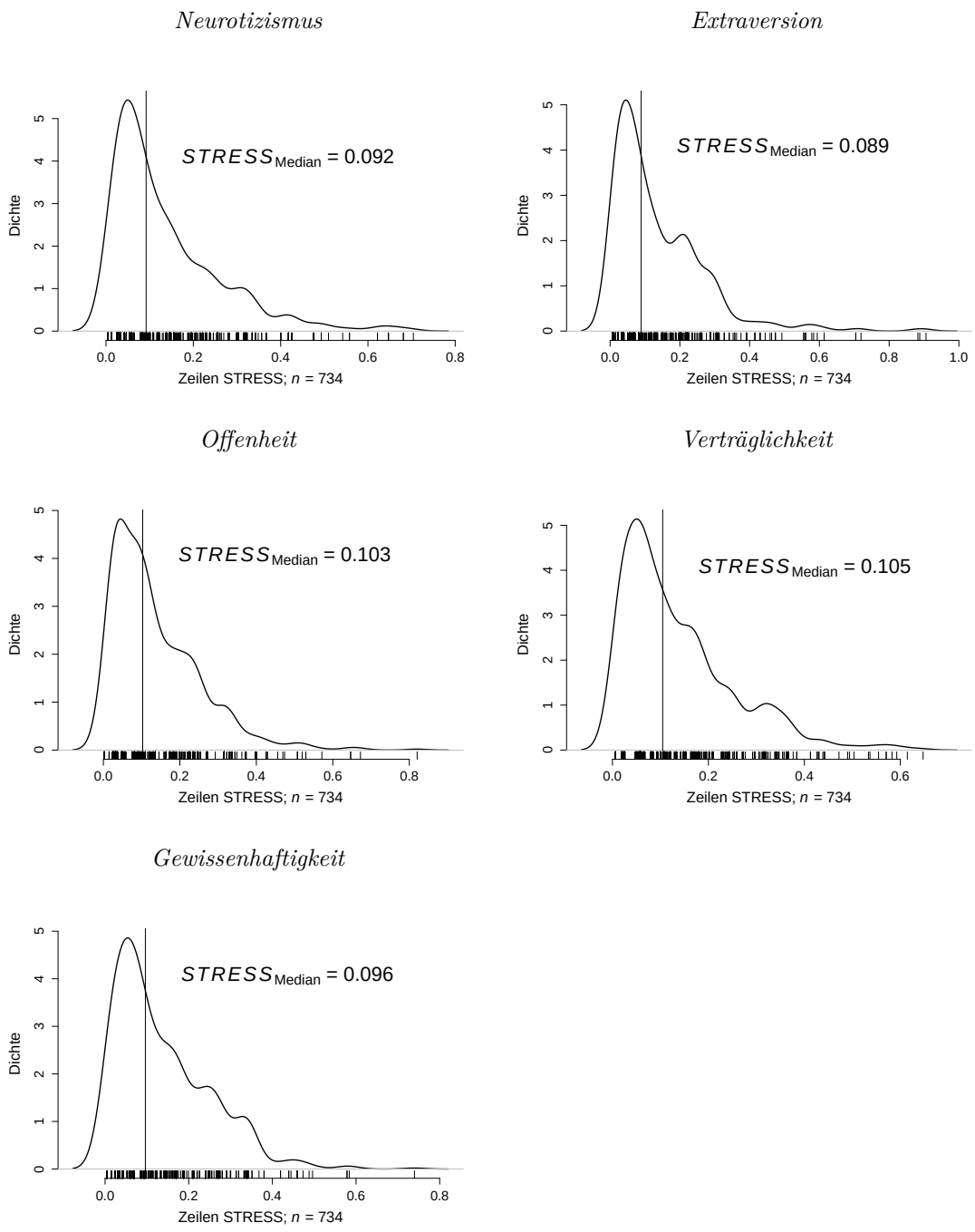


Abbildung 7.3 Stichprobe I: Dichtefunktionen des STRESS zur Klassifikation der Personenpassung gemäß eines *Nähe–Distanz*-Antwortprozesses für fünf Skalen des BFI–K.

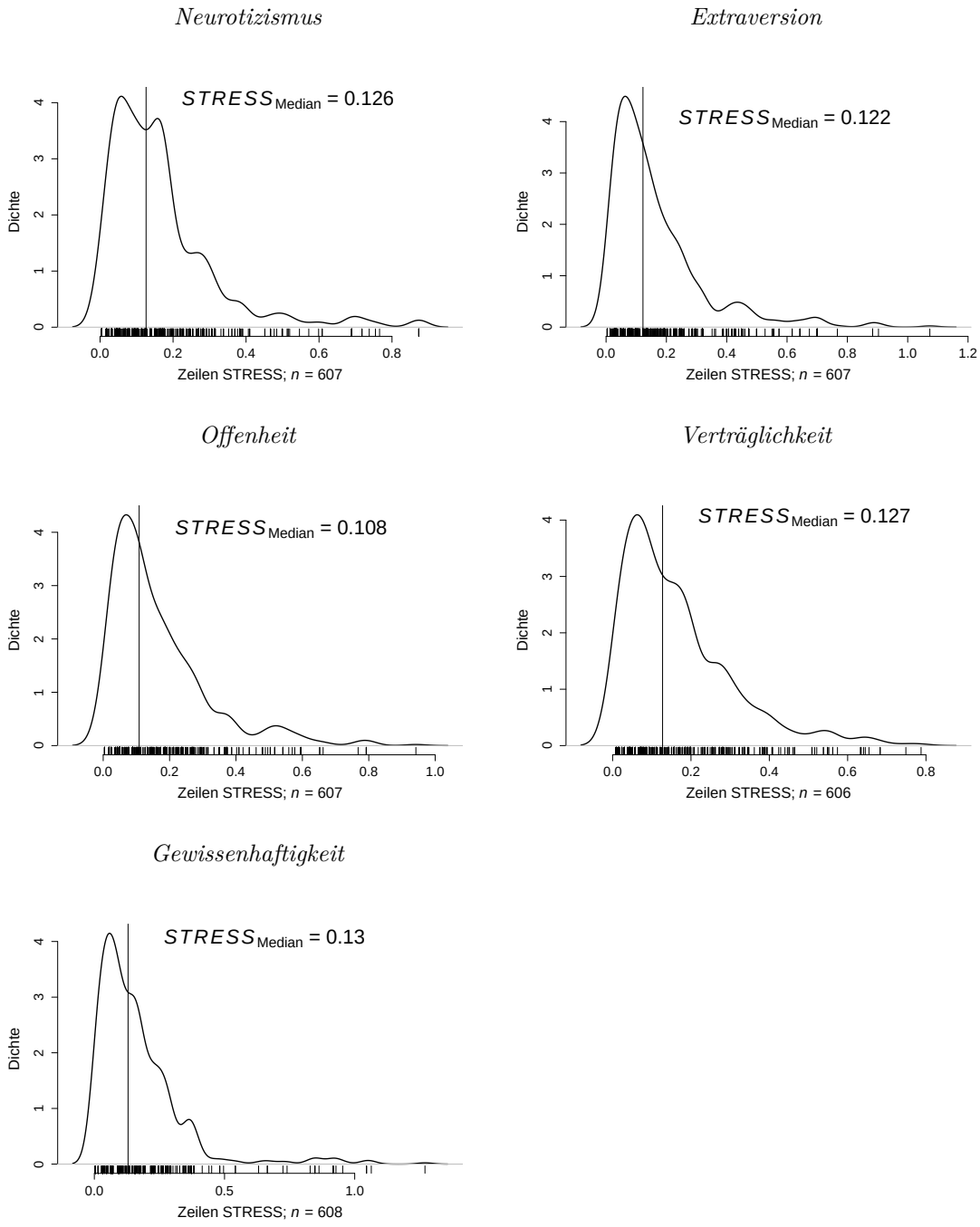


Abbildung 7.4 Stichprobe II: Dichtefunktionen des STRESS zur Klassifikation der Personenpassung gemäß eines *Nähe–Distanz*-Antwortprozesses für fünf Skalen des BFI–K.

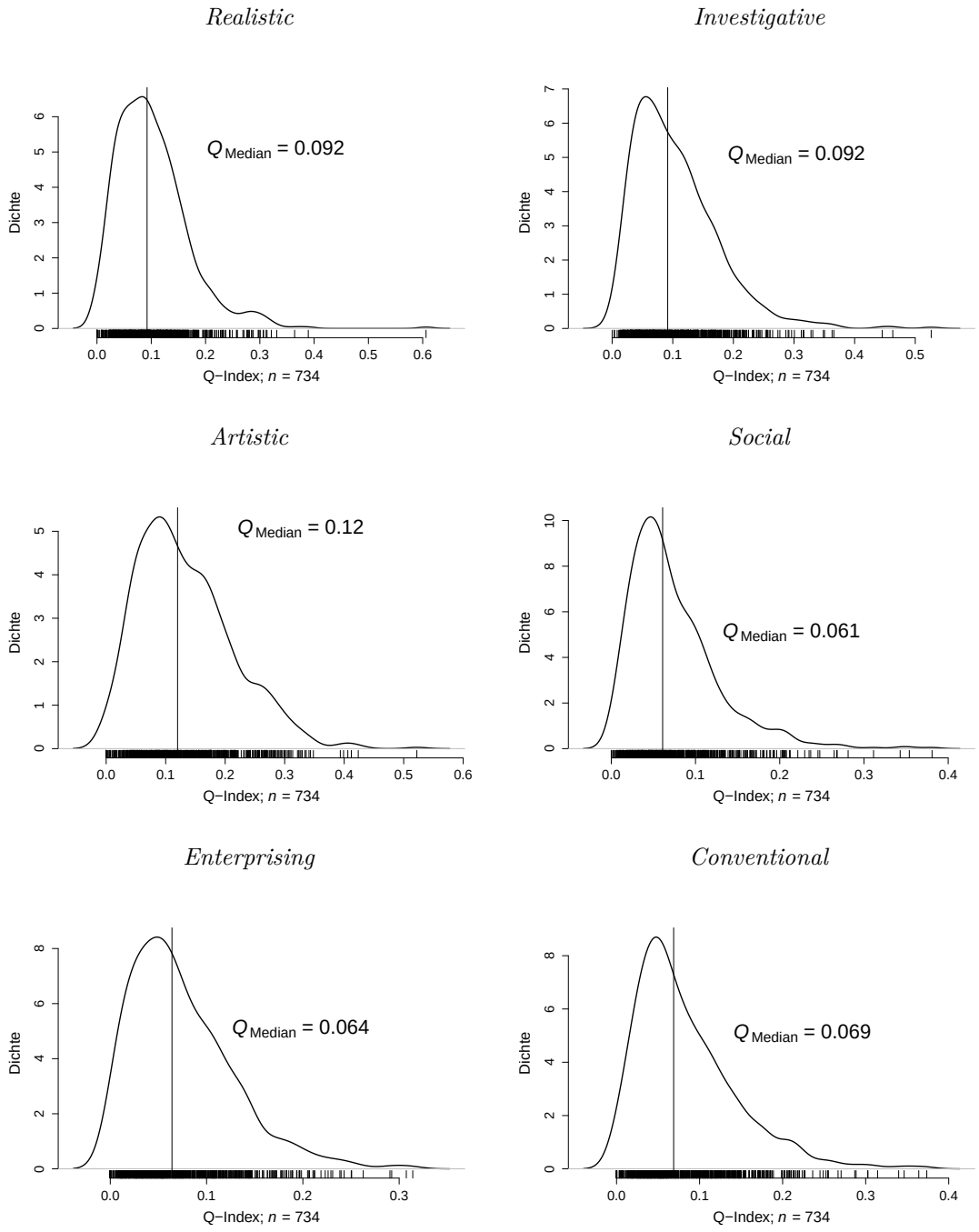


Abbildung 7.5 Stichprobe I: Dichtefunktionen des Q-Index zur Klassifikation der Personenpassung bei *Dominanz*-Antwortprozess (PCM) für sechs Skalen des AIST-R.

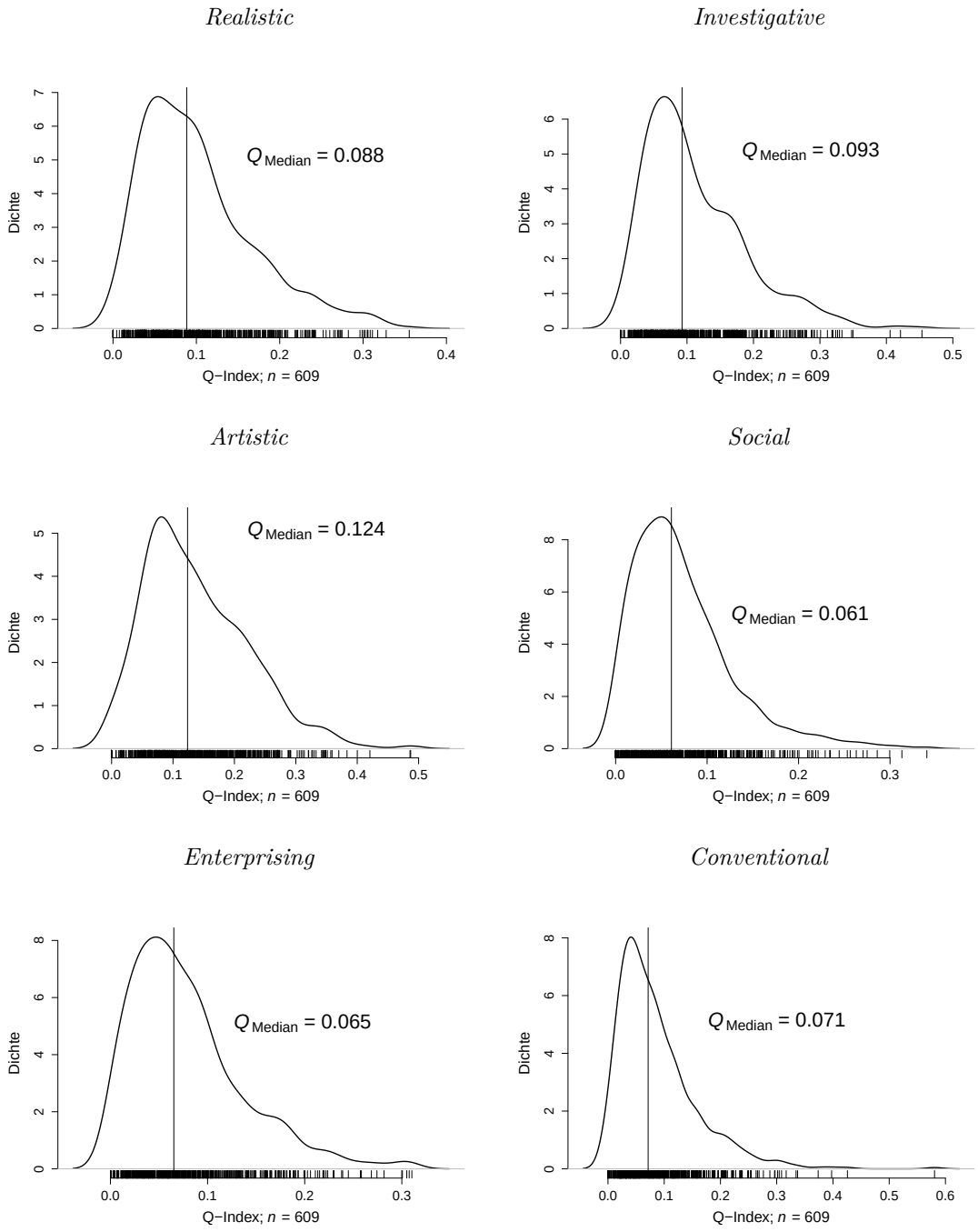


Abbildung 7.6 Stichprobe II: Dichtefunktionen des Q-Index zur Klassifikation der Personenpassung bei *Dominanz*-Antwortprozess (PCM) für sechs Skalen des AIST-R.

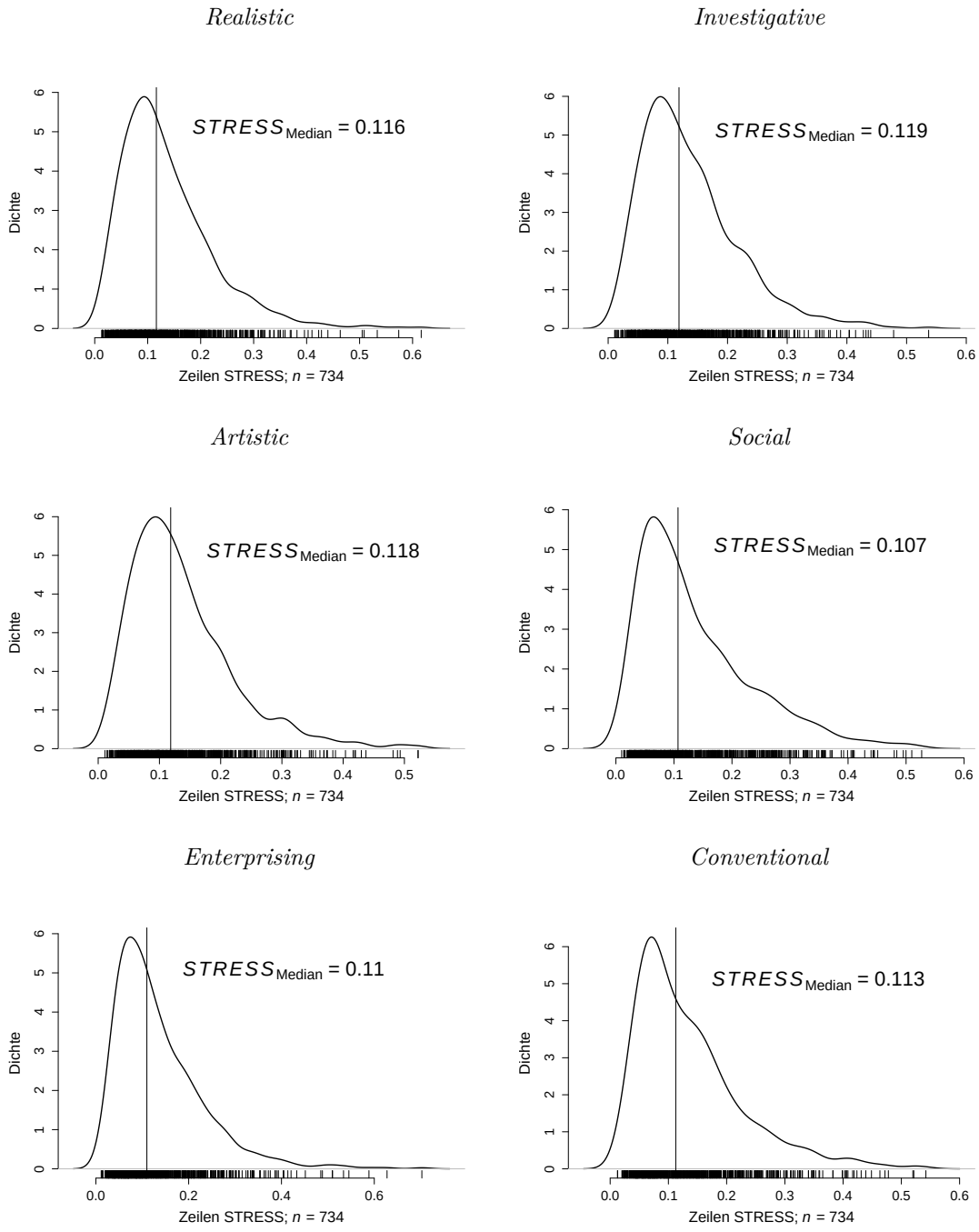


Abbildung 7.7 Stichprobe I: Dichtefunktionen des STRESS zur Klassifikation der Personenpassung gemäß eines *Nähe–Distanz*-Antwortprozesses für sechs Skalen des AIST-R.

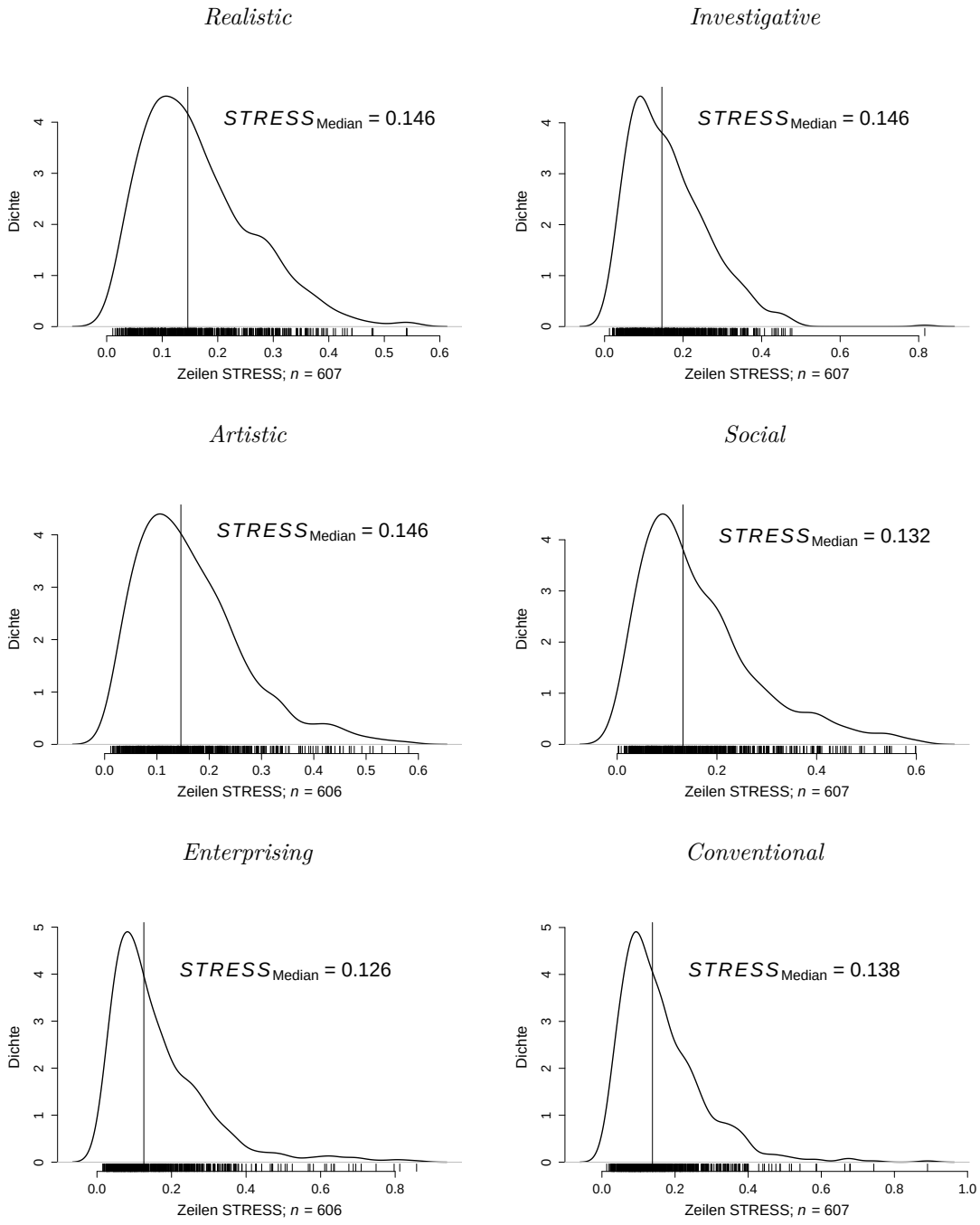


Abbildung 7.8 Stichprobe II: Dichtefunktionen des STRESS zur Klassifikation der Personenpassung gemäß eines *Nähe–Distanz*-Antwortprozesses für sechs Skalen des AIST-R.

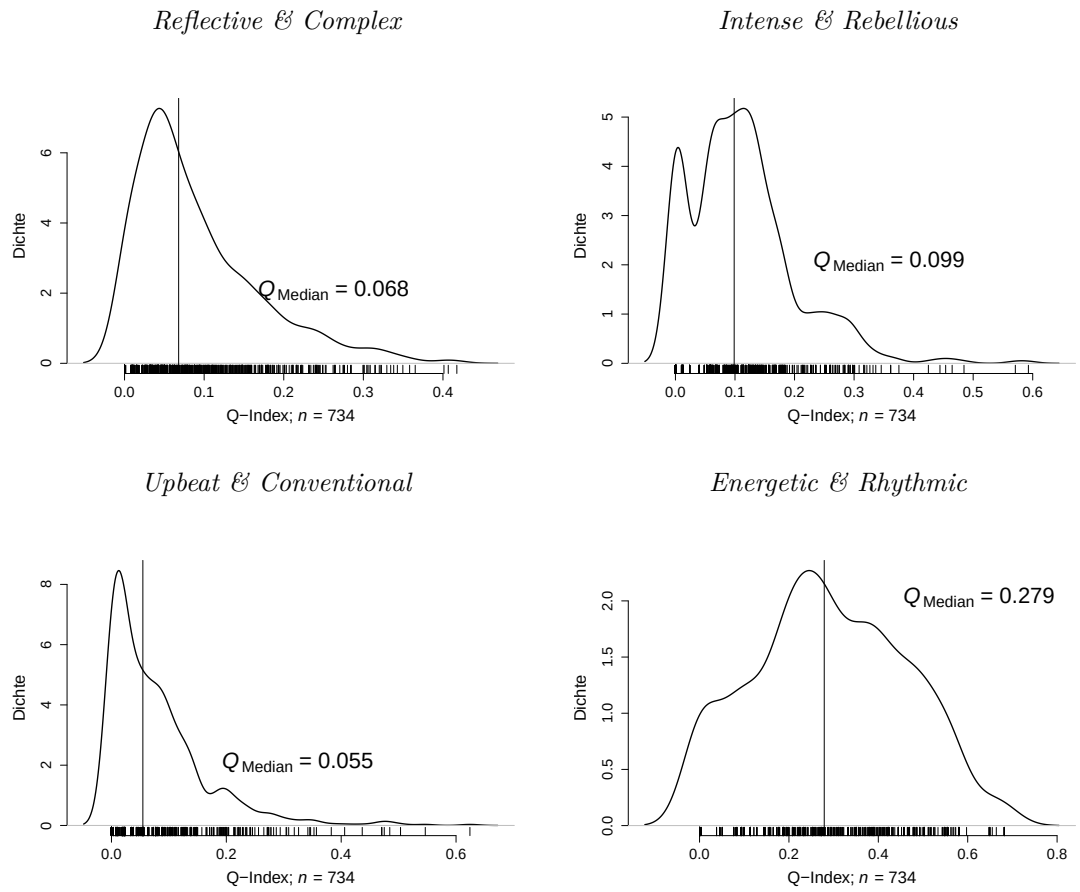


Abbildung 7.9 Stichprobe I: Dichtefunktionen des Q-Index zur Klassifikation der Personenpassung bei *Dominanz*-Antwortprozess (PCM) für vier Skalen des STOMP.

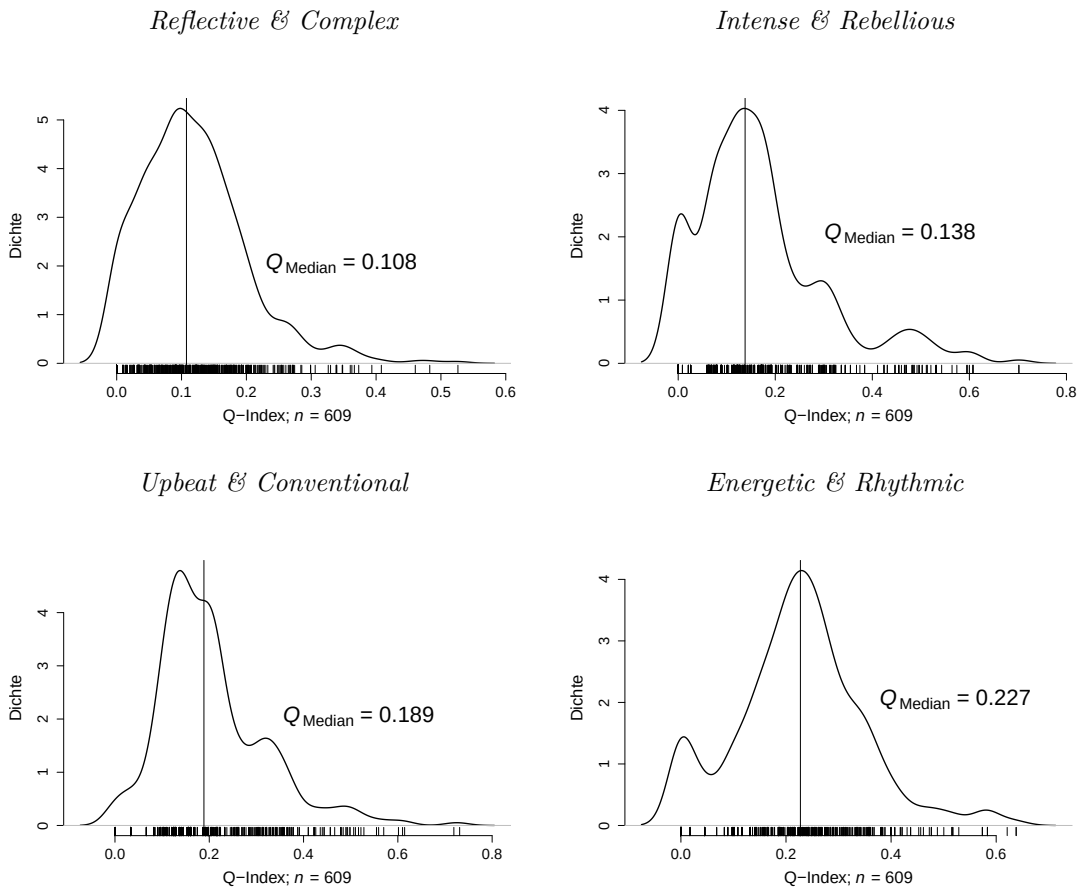


Abbildung 7.10 Stichprobe II: Dichtefunktionen des Q-Index zur Klassifikation der Personenpassung bei *Dominanz*-Antwortprozess (PCM) für vier Skalen des STOMP.

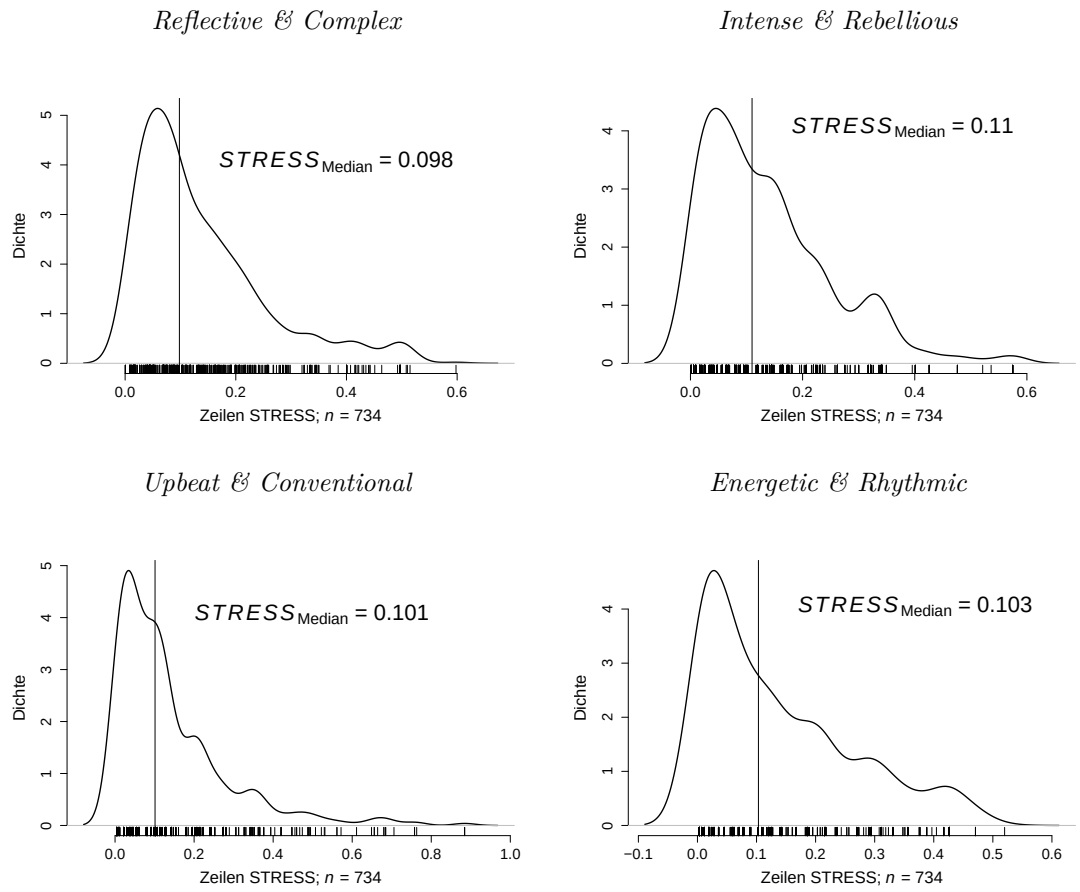


Abbildung 7.11 Stichprobe I: Dichtefunktionen des STRESS zur Klassifikation der Personenpassung gemäß eines *Nähe–Distanz*-Antwortprozesses für vier Skalen des STOMP.

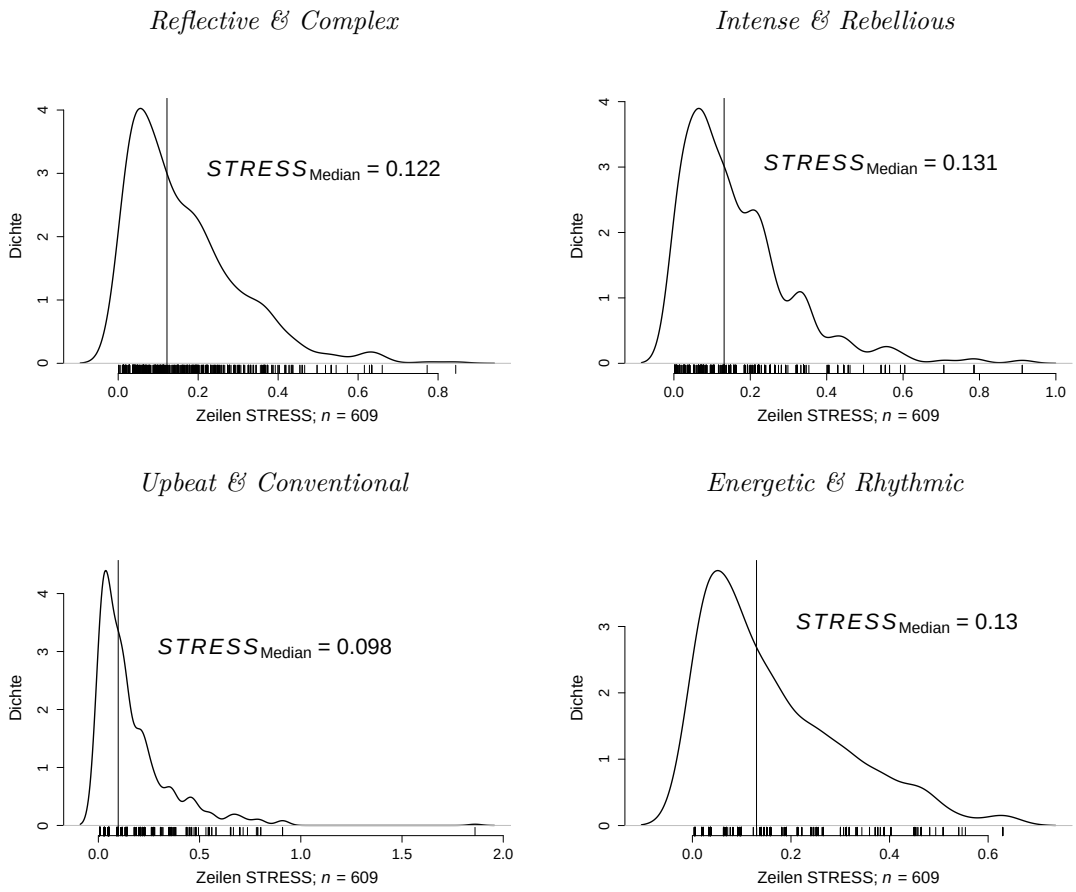


Abbildung 7.12 Stichprobe II: Dichtefunktionen des STRESS zur Klassifikation der Personenpassung gemäß eines *Nähe–Distanz*-Antwortprozesses für vier Skalen des STOMP.

Werden die einzelnen Verteilungen des Q-Index und STRESS-Index für ein Konstrukt jeweils über alle Skalen und beide Stichproben zu einer Gesamtverteilung zusammengefasst, so lässt sich über deren Median die „mittlere“ (lokale) Anpassung an ein Dominanz- oder Nähe–Distanz-Antwortprozess in dem jeweiligen Konstrukt ausdrücken. Dabei fällt der Median der Gesamtverteilung (über die jeweiligen Skalen und beide Stichproben) des Q-Index für den Bereich *Persönlichkeit* am geringsten (BFI–K: $\text{Median}_{Q,\text{BFI–K}} = .063$) und derjenige für die *Musikpräferenzen* (STOMP: $\text{Median}_{Q,\text{STOMP}} = .129$) am höchsten aus. Der Median der Gesamtverteilungen des Q-Index für die *beruflichen Interessenorientierungen* (AIST–R: $\text{Median}_{Q,\text{AIST–R}} = .081$) liegt zwischen diesem Bereich in der Nähe des Medians für Q für den BFI–K. Demgegenüber fällt der Median der Gesamtverteilung des STRESS-Index für den Bereich *Musikpräferenzen* am geringsten aus (STOMP: $\text{Median}_{STRESS,\text{STOMP}} = .106$). Der Median der Gesamtverteilung des STRESS-Index für den Bereich *Persönlichkeit* fällt fast identisch aus (BFI–K: $\text{Median}_{STRESS,\text{BFI–K}} = .109$) und der Median für den STRESS-Index im Bereich *beruflicher Interessenorientierungen* fällt unwesentlich höher aus (AIST–R: $\text{Median}_{STRESS,\text{AIST–R}} = .124$).

Die Klassifikation der Personen nach den beiden impliziten Antwortprozessen kann auf dieser Basis zunächst anhand der festen cut-off Grenzen für beide Indizes ($STRESS = .05$ und $Q = .2$) als Maß für die lokale Modellpassung realisiert werden. Damit kann die Klassifikation zu den beiden Antwortprozessen entweder über den Q-Index für den *Dominanz*-Antwortprozess erfolgen oder aber über den Zeilen STRESS-Index für den *Nähe–Distanz*-Antwortprozess. Die Tabellen 7.1, 7.2 und 7.3 zeigen die entsprechende Kreuztabellierung der Personenklassifikation nach implizitem Antwortprozess für alle Skalen der drei Konstrukte zusammengefasst für beide Stichproben (Stichprobe I und Stichprobe II) nach deren jeweils getrennt erfolgter Skalierung. An dieser Darstellung kann die Eindeutigkeit der Klassifikation – anhand der beiden Indizes Q und $STRESS$ – der Personen nach ihren impliziten Antwortmodellen beurteilt werden.

Bei der Betrachtung der Kreuztabellierung fällt zunächst auf, dass ein erheblicher Anteil der Personen anhand der Personen-Fit-Indizes eindeutig dem Dominanz-Antwortprozess zugeordnet wird. Für den BFI–K variiert der Anteil eindeutig (durch beide Indizes) dem Dominanz-Antwortprozess zugeordneter

Personen über alle fünf Skalen in einem Bereich von $n = 947$ bis $n = 735$, im Gegensatz zu einem Bereich von $n = 3$ bis $n = 26$ für die eindeutige (durch beide Indizes übereinstimmende) Zuordnung zu dem Nähe–Distanz-Antwortprozess (vgl. Tabelle 7.1). Eine noch „einseitigere“ übereinstimmende Zuordnung nach beiden Indizes (Q und $STRESS$) ergibt sich für die Skalen des AIST–R (vgl. Tabelle 7.2). Beim STOMP ergibt sich dagegen ein (je nach Dimension) vergleichsweise größerer Bereich ($n = 18$: *Reflective & Complex* bis $n = 323$: *Energetic & Rhythmic*) von Personen mit eindeutiger Zuordnung zu dem Nähe–Distanz-Antwortprozess (vgl. Tabelle 7.3).

Ferner zeigt die Kreuztabellierung der Personenklassifikation nach beiden Indizes, dass ein nicht unerheblicher Anteil der Personen anhand der beiden Personen-Fit-Indizes (übereinstimmend) *keinem* der beiden Antwortprozesse zugeordnet werden kann. So liegt die Anzahl der Personen in der linken oberen Zelle der jeweiligen 2×2-Tabellen über die beiden Konstrukte Persönlichkeit (BFI–K) und berufliche Interessen (AIST–R) und deren Skalen hinweg, in einem Bereich von $n = 48$ bis $n = 271$ Personen (vgl. Tabellen 7.1 und 7.2). Etwas ungünstiger, im Hinblick auf die Eindeutigkeit der Klassifikation, fällt

Tabelle 7.1 Kreuztabellierung der Personenklassifikation nach implizitem Antwortprozess für den BFI–K.

Neurotizismus				**Extraversion**				**Offenheit**			
	Nähe–Distanz				*Nähe–Distanz*				*Nähe–Distanz*		
	nein	ja	fehlend		nein	ja	fehlend		nein	ja	fehlend
Dominanz	[N]	[U]		*Dominanz*	[N]	[U]		*Dominanz*	[N]	[U]	
nein [N]	179	19	0	nein [N]	94	3	0	nein [N]	242	26	0
ja [K]	847	296	2	ja [K]	924	320	2	ja [K]	735	338	2
fehlend	0	0	0	fehlend	0	0	0	fehlend	0	0	0
Verträglichkeit				**Gewissenhaftigkeit**							
	Nähe–Distanz				*Nähe–Distanz*						
	nein	ja	fehlend		nein	ja	fehlend				
Dominanz	[N]	[U]		*Dominanz*	[N]	[U]					
nein [N]	192	7	0	nein [N]	76	3	0				
ja [K]	870	271	3	ja [K]	947	316	1				
fehlend	0	0	0	fehlend	0	0	0				

Anmerkungen: Getrennte Skalierung jeweils Stichprobe I und Stichprobe II; Absolute Häufigkeiten der klassifizierten Personen; ja = Personen-Fit-Index < cut-off Kriterium ($Q < 0.2$; $STRESS < 0.05$) - [K]: (kumulativ), *Dominanz*-Antwortprozess, [U]: (Unfolding) *Nähe–Distanz*-Antwortprozess, [N]: keine Zuordnung zu einem Antwortprozess; fehlend = keine Klassifikation der Person aufgrund fehlender Werte auf den entsprechenden Items; $n = 1343$.

Tabelle 7.2 Kreuztabellierung der Personenklassifikation nach implizitem Antwortprozess für den AIST–R.

	Realistic				**Investigative**				**Artistic**		
	Nähe–Distanz				*Nähe–Distanz*				*Nähe–Distanz*		
	nein	ja	fehlend		nein	ja	fehlend		nein	ja	fehlend
Dominanz	[N]	[U]		*Dominanz*	[N]	[U]		*Dominanz*	[N]	[U]	
nein [N]	111	3	0	nein [N]	140	3	0	nein [N]	271	9	0
ja [K]	1088	139	2	ja [K]	1074	124	2	ja [K]	938	122	3
fehlend	0	0	0	fehlend	0	0	0	fehlend	0	0	0
	Social				**Enterprising**				**Conventional**		
	Nähe–Distanz				*Nähe–Distanz*				*Nähe–Distanz*		
	nein	ja	fehlend		nein	ja	fehlend		nein	ja	fehlend
Dominanz	[N]	[U]		*Dominanz*	[N]	[U]		*Dominanz*	[N]	[U]	
nein [N]	48	0	0	nein [N]	48	0	0	nein [N]	93	0	0
ja [K]	1101	192	2	ja [K]	1122	170	3	ja [K]	1100	148	2
fehlend	0	0	0	fehlend	0	0	0	fehlend	0	0	0

Anmerkungen: Getrennte Skalierung jeweils Stichprobe I und Stichprobe II; Absolute Häufigkeiten der klassifizierten Personen; ja = Personen-Fit-Index < cut-off Kriterium ($Q < 0.2$; $STRESS < 0.05$) - [K]: (kumulativ), *Dominanz*-Antwortprozess, [U]: (Unfolding) *Nähe–Distanz*-Antwortprozess, [N]: keine Zuordnung zu einem Antwortprozess; fehlend = keine Klassifikation der Person aufgrund fehlender Werte auf den entsprechenden Items; $n = 1343$.

das Ergebnis für den STOMP aus. Die Anzahl der nicht eindeutig zugeordneten Personen liegt hier über alle vier Skalen in einem Bereich von $n = 138$ bis $n = 558$ Personen (vgl. Tabelle 7.3).

Darüber hinaus überrascht auch das Ergebnis, dass ein ebenfalls nicht unerheblicher Teil der Personen nach den beiden Personen-Fit-Indizes und den gewählten cut-off Kriterien *beiden* impliziten Antwortprozessen *gleichzeitig* zugeordnet werden kann. Wie auch bei den nicht eindeutig klassifizierten Personen fällt hier der Anteil der zu beiden Antwortprozessen passenden Personen für die Dimensionen des BFI–K, AIST–R und STOMP unterschiedlich aus. So liegt die Anzahl der in diesem Sinne „zweifach klassifizierten“ Personen für den BFI–K in einem Bereich von $n = 271$ (*Verträglichkeit*) bis $n = 338$ (*Offenheit*), für den AIST–R in einem Bereich von $n = 122$ (*Artistic*) bis $n = 192$ (*Social*) und schließlich für den STOMP in einem Bereich von $n = 85$ (*Energetic & Rhythmic*) bis $n = 377$ (*Upbeat & Conventional*) (vgl. Tabellen 7.1, 7.2 und 7.3).

Tabelle 7.3 Kreuztabellierung der Personenklassifikation nach implizitem Antwortprozess für den STOMP.

Reflective & Complex				**Intense & Rebellious**			
	Nähe–Distanz				*Nähe–Distanz*		
	nein	ja	fehlend		nein	ja	fehlend
Dominanz	[N]	[U]		*Dominanz*	[N]	[U]	
nein [N]	138	18	0	nein [N]	228	37	0
ja [K]	879	305	3	ja [K]	794	281	3
fehlend	0	0	0	fehlend	0	0	0
Upbeat & Conventional				**Energetic & Rhythmic**			
	Nähe–Distanz				*Nähe–Distanz*		
	nein	ja	fehlend		nein	ja	fehlend
Dominanz	[N]	[U]		*Dominanz*	[N]	[U]	
nein [N]	268	45	0	nein [N]	558	323	0
ja [K]	650	377	3	ja [K]	374	85	0
fehlend	0	0	0	fehlend	54	0	3

Anmerkungen: Getrennte Skalierung jeweils Stichprobe I und Stichprobe II; Absolute Häufigkeiten der klassifizierten Personen; ja = Personen-Fit-Index < cut-off Kriterium ($Q < 0.2$; $STRESS < 0.05$) - [K]: (kumulativ), *Dominanz*-Antwortprozess, [U]: (Unfolding) *Nähe–Distanz*-Antwortprozess, [N]: keine Zuordnung zu einem Antwortprozess; fehlend = keine Klassifikation der Person aufgrund fehlender Werte auf den entsprechenden Items; $n = 1343$.

Zur Erstellung einer grafischen Darstellung der nach ihren Randsummen[1] reorganisierten Datenmatrizen wird auf den jeweiligen Personen-Fit-Index des entsprechenden Antwortmodells zur Klassifikation der Personen zurückgegriffen – also auf den *STRESS*-Index beim *Nähe–Distanz*-Antwortprozess und auf den *Q*-Index beim *Dominanz*-Antwortprozess.

Die visuelle Inspektion der resultierenden *Bertin-Plots* (vgl. Abbildungen 7.13, 7.14, 7.17, 7.18, 7.21, 7.22, 7.15, 7.16, 7.19, 7.20, 7.23 und 7.24) für beide Stichproben, über alle Skalen der drei Konstrukte hinweg, deutet insgesamt darauf hin, dass in den hier analysierten Daten die zwei postulierten Antwortprozesse bei den Personen bestehen. Am deutlichsten zeigt sich dies in den grafischen Darstellungen der jeweils reorganisierten Teildatenmatrizen für die AIST–R-Dimension *Realistic*. So ergibt sich für die nach der Merkmalsausprägung der Personen und gleichzeitig nach der Itemschwierigkeit (den Randsummen) aufsteigend sortierten Daten der jeweils identifizierten Teilstichproben für einen *Dominanz*-Antwortprozess (bei Stichprobe I: $n = 677$; bei Stichprobe II: $n = 552$) in den Abbildungen 7.17 und 7.18 (jeweils links oben) ein typisches dreieckiges Muster aus Graustufen. Die *dunklen* Grautöne repräsentieren in diesen grafischen Darstellungen der reorganisierten Datenmatrizen dabei die eher hohen Antwortkategorien (z. B. 4 ≡ *„Das interessiert mich sehr; das tue ich sehr gerne“* und 3 ≡ *„Das interessiert mich ziemlich“*), wohingegen die *hellen* Grautöne eher die niedrigen Antwortkategorien (z. B. 0 ≡ *„Das interessiert mich gar nicht; das tue ich nicht gerne“* und 1 ≡ *„Das interessiert mich wenig“*) repräsentieren.

Für dieselbe AIST–R-Dimension ergibt sich bei der jeweiligen Teilstichprobe für einen *Nähe–Distanz*-Antwortprozess (bei Stichprobe I: $n = 88$; bei Stichprobe II: $n = 54$) ein Muster mit einer (dunkel eingefärbten) Diagonalen, welche die eher höheren Kategorien der Antwortskala der Items repräsentiert (vgl. Abbildung 7.19 und 7.20). Diese Diagonale ist Ausdruck der in dieser Teilstichprobe vorherrschenden *Nähe–Distanz*-Relation zwischen den Personen und Items. Je nach Merkmalsausprägung wählen die Personen die eher hohen Antwortkategorien (dunkle Graustufen) nur derjenigen Items denen sie

[1] Als *Randsummen* werden hier übergreifend die ermittelten *Parameter* oder *Koordinaten* (für Personen und Items) eines Antwortmodells (für den Dominanz oder Nähe–Distanz-Antwortprozess) für einen empirischen (Teil-)Datensatz bezeichnet.

psychisch nahe sind. Eine größere Distanz, sowohl oberhalb als auch unterhalb der individuellen Merkmalsausprägung der Person, führt zu der Wahl eher niedriger Antwortkategorien (helle Grautöne). Auch für die Dimensionen des STOMP (z. B. *Reflective & Complex*) ergibt sich eine ähnlich eindeutige Zuordnung der einzelnen Personengruppen anhand der grafischen Darstellung für die gemäß der Randsummen reorganisierten Datenmatrizen (vgl. Abbildungen 7.21 und 7.22 für den *Dominanz* Antwortprozess sowie 7.23 und 7.24 für den *Nähe–Distanz*-Antwortprozess).

Allgemein lassen sich in den graphischen Darstellungen der nach ihren Randsummen reorganisiert Datenmatrizen in den Abbildungen 7.13, 7.14, 7.17, 7.18, 7.21, 7.22, 7.15, 7.16, 7.19, 7.20, 7.23 und 7.24) für beide Stichproben, über alle Skalen der drei Konstrukte hinweg, die zwei postulierten Antwortprozesse mit unterschiedlicher Deutlichkeit identifizieren.

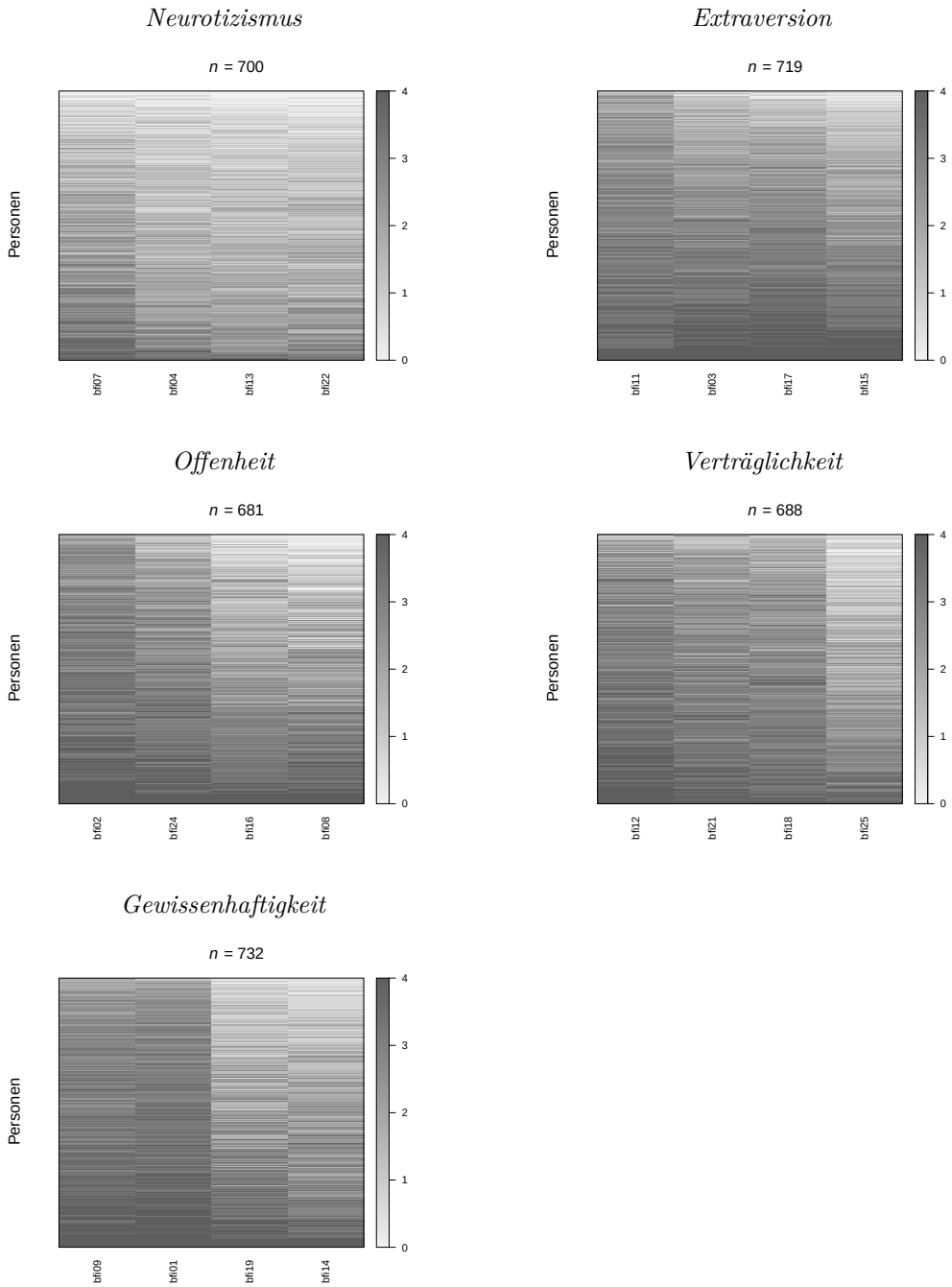


Abbildung 7.13 Stichprobe I: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Dominanz*-Antwortprozess (PCM) für fünf Skalen des BFI–K; Graustufen entsprechen den Antwortkategorien: Dunkel ≡ 4 ≡ *„Sehr zutreffend“* – hell ≡ 0 ≡ *„Sehr unzutreffend“*.

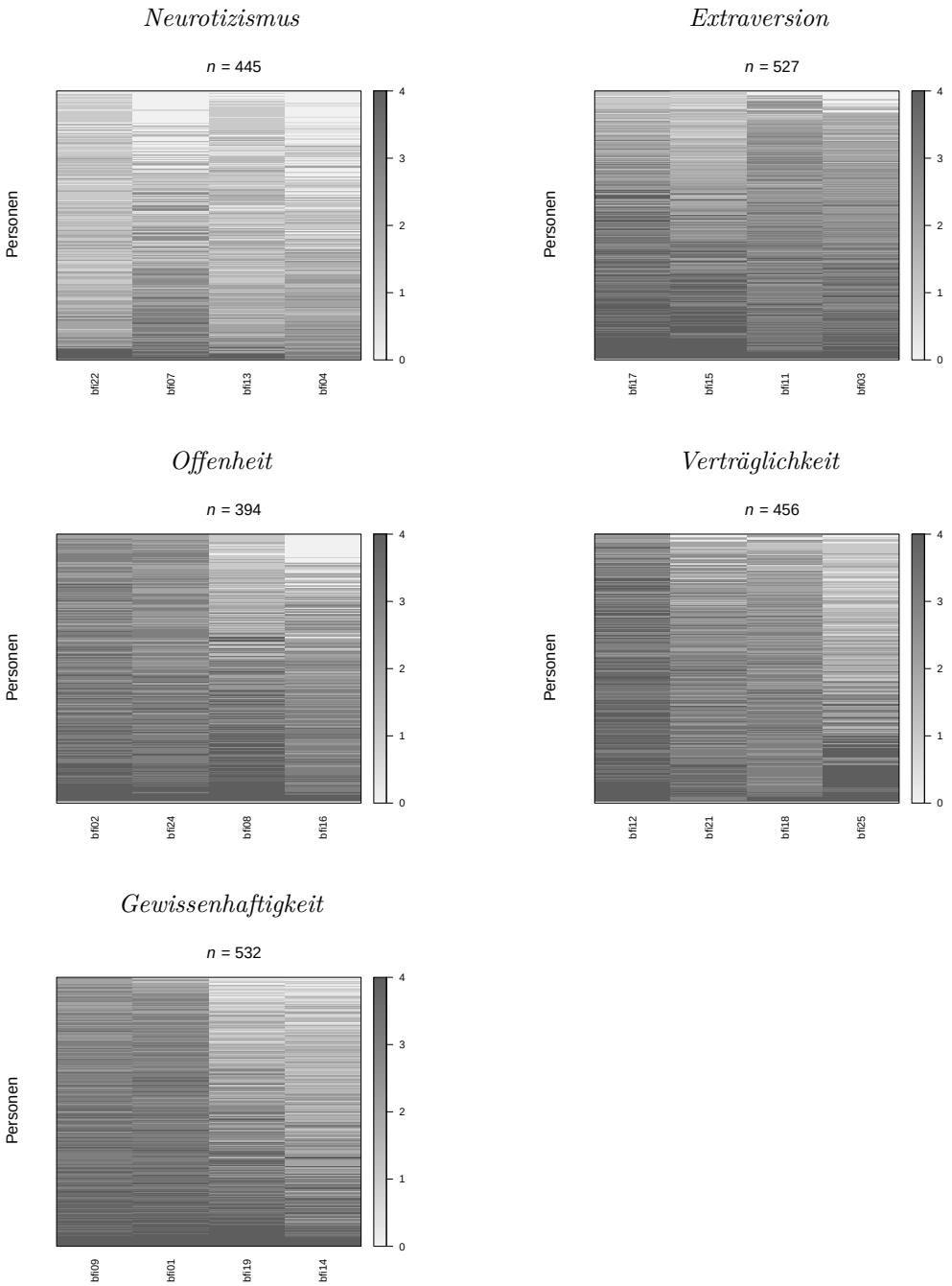


Abbildung 7.14 Stichprobe II: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Dominanz*-Antwortprozess (PCM) für fünf Skalen des BFI–K; Graustufen entsprechen den Antwortkategorien: Dunkel $\equiv 4 \equiv$ *„Sehr zutreffend“* – hell $\equiv 0 \equiv$ *„Sehr unzutreffend“*.

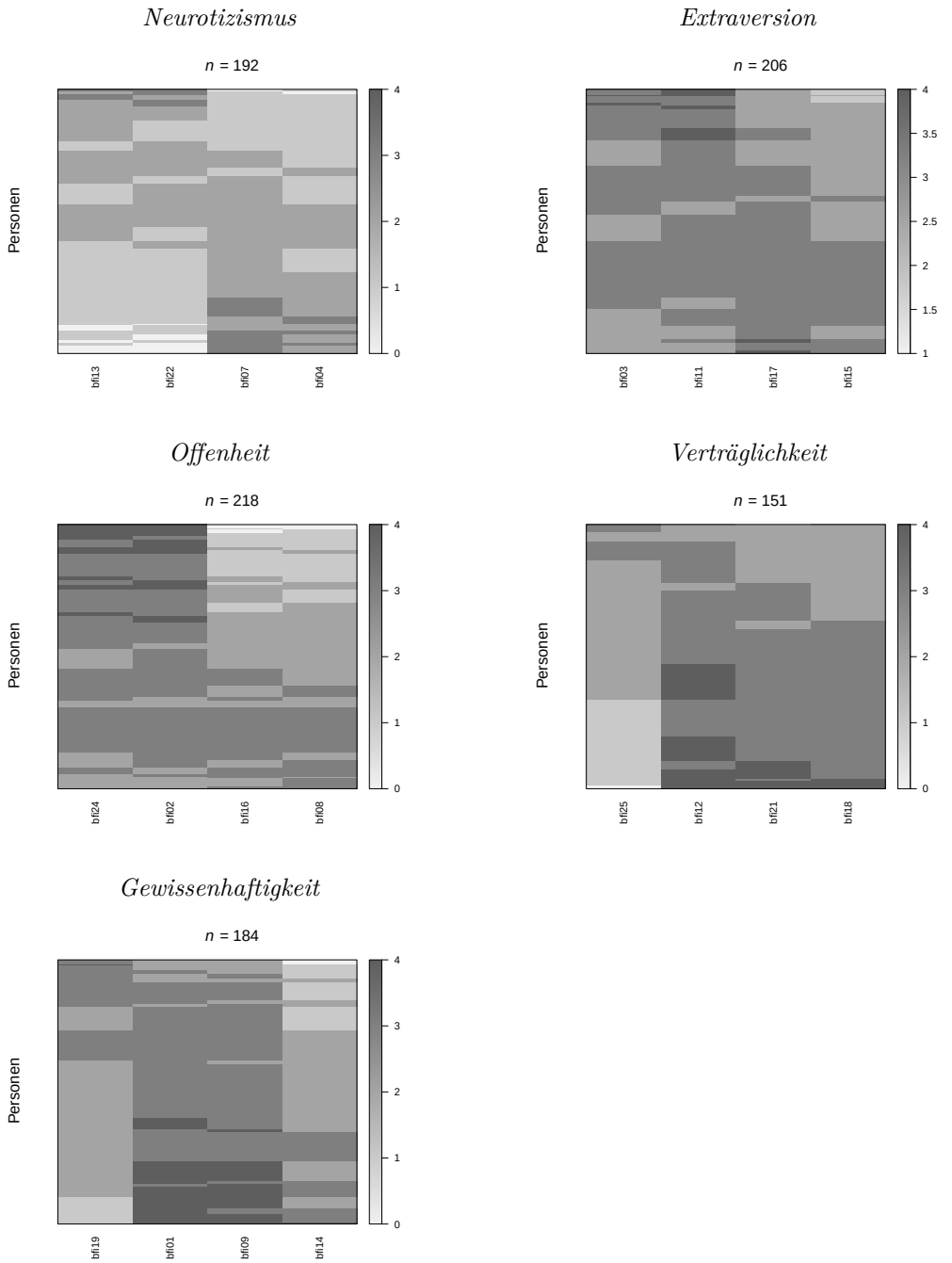


Abbildung 7.15 Stichprobe I: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Nähe–Distanz*-Antwortprozess (MDS) für fünf Skalen des BFI–K; Graustufen entsprechen den Antwortkategorien: Dunkel ≡ 4 ≡ *„Sehr zutreffend“* – hell ≡ 0 ≡ *„Sehr unzutreffend“*.

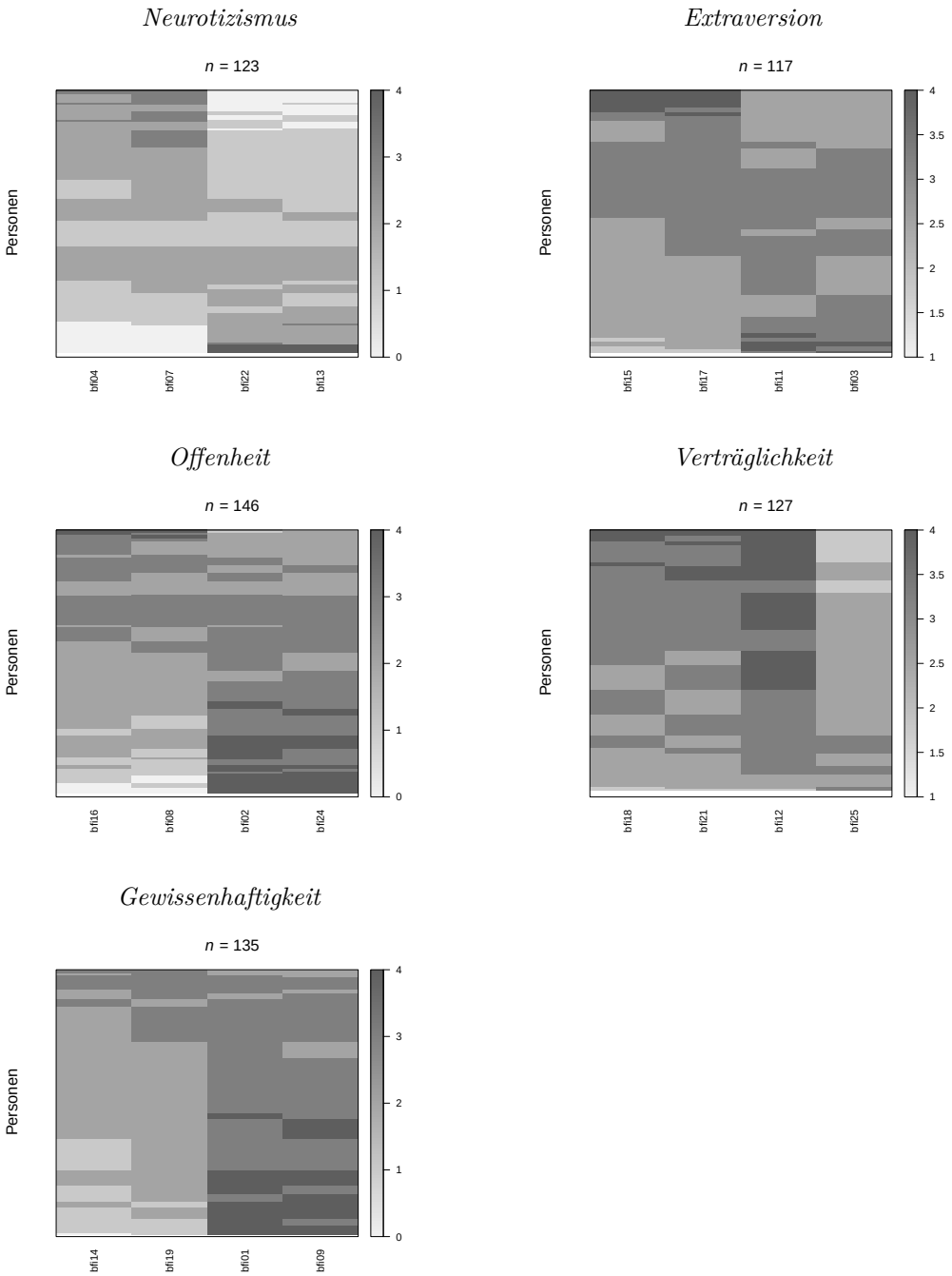


Abbildung 7.16 Stichprobe II: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Nähe–Distanz*-Antwortprozess (MDS) für fünf Skalen des BFI–K; Graustufen entsprechen den Antwortkategorien: Dunkel ≡ 4 ≡ *„Sehr zutreffend“* – hell ≡ 0 ≡ *„Sehr unzutreffend“*.

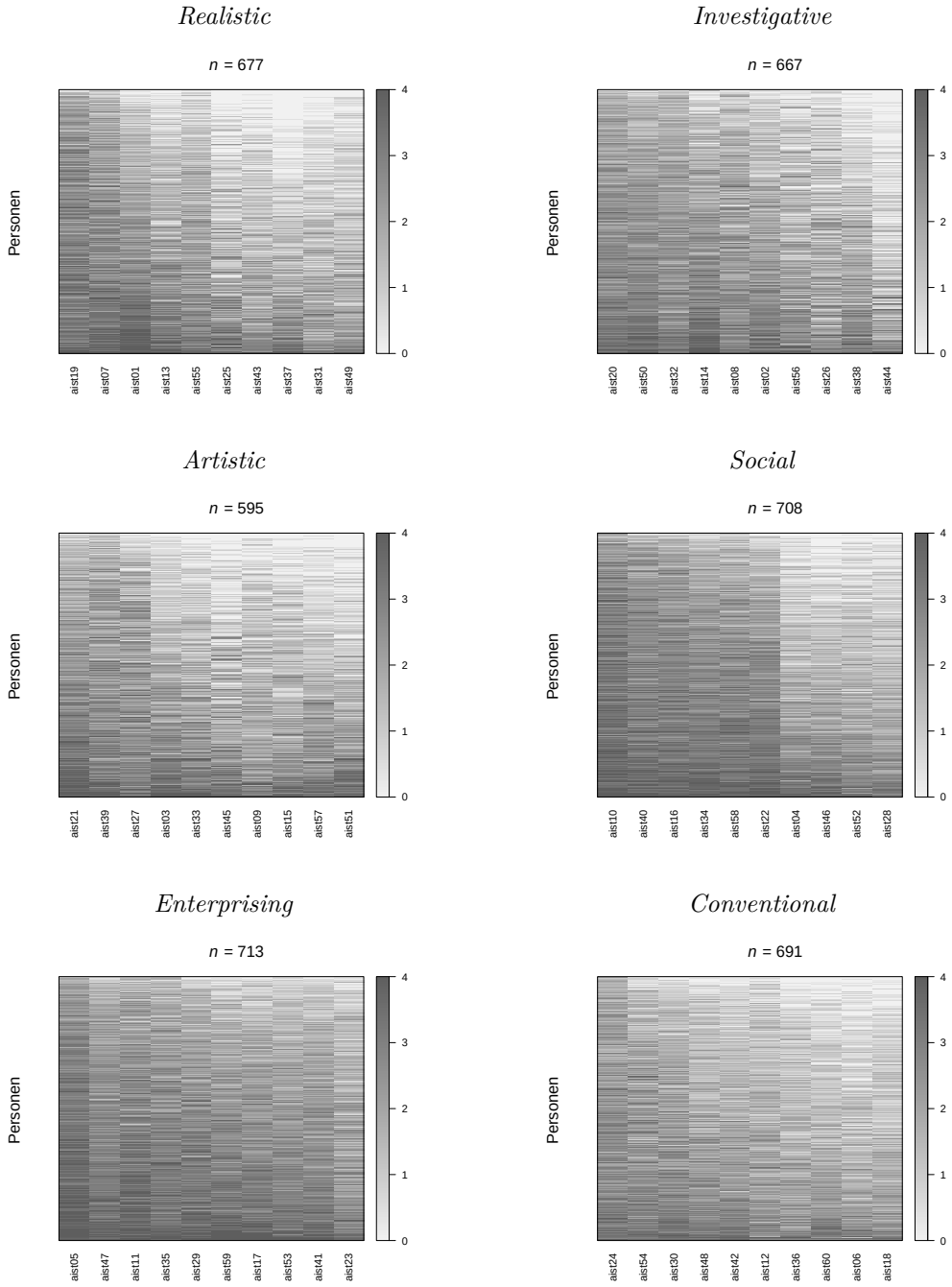


Abbildung 7.17 Stichprobe I: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Dominanz*-Antwortprozess (PCM) für sechs Skalen des AIST-R; Graustufen entsprechen den Antwortkategorien: Dunkel $\equiv 4 \equiv$ „*Das interessiert mich sehr; das tue ich sehr gerne*“ – hell $\equiv 0 \equiv$ „*Das interessiert mich gar nicht; das tue ich nicht gerne*“.

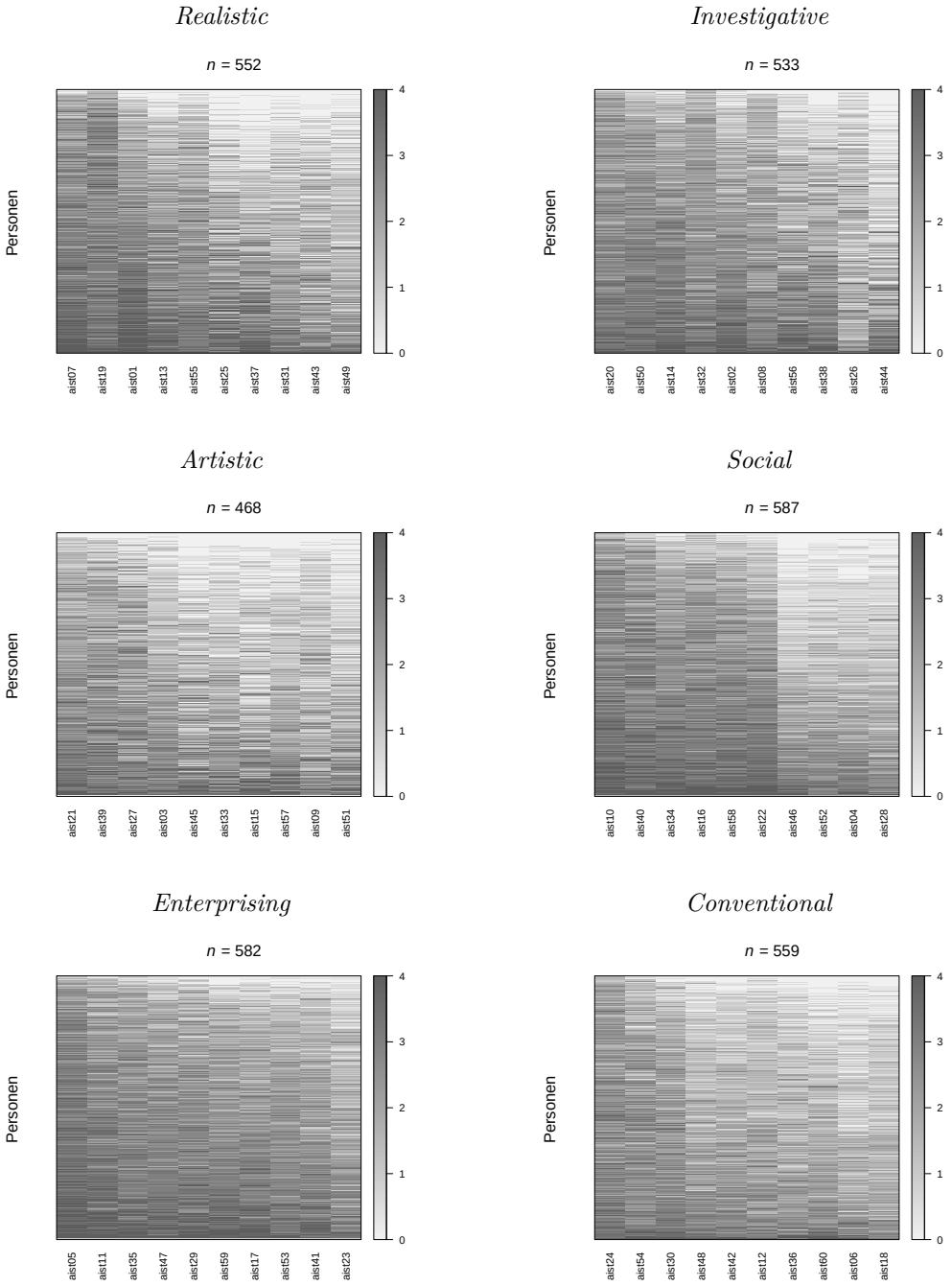


Abbildung 7.18 Stichprobe II: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Dominanz*-Antwortprozess (PCM) für sechs Skalen des AIST-R; Graustufen entsprechen den Antwortkategorien: Dunkel $\equiv 4 \equiv$ *„Das interessiert mich sehr; das tue ich sehr gerne“* – hell $\equiv 0 \equiv$ *„Das interessiert mich gar nicht; das tue ich nicht gerne“*.

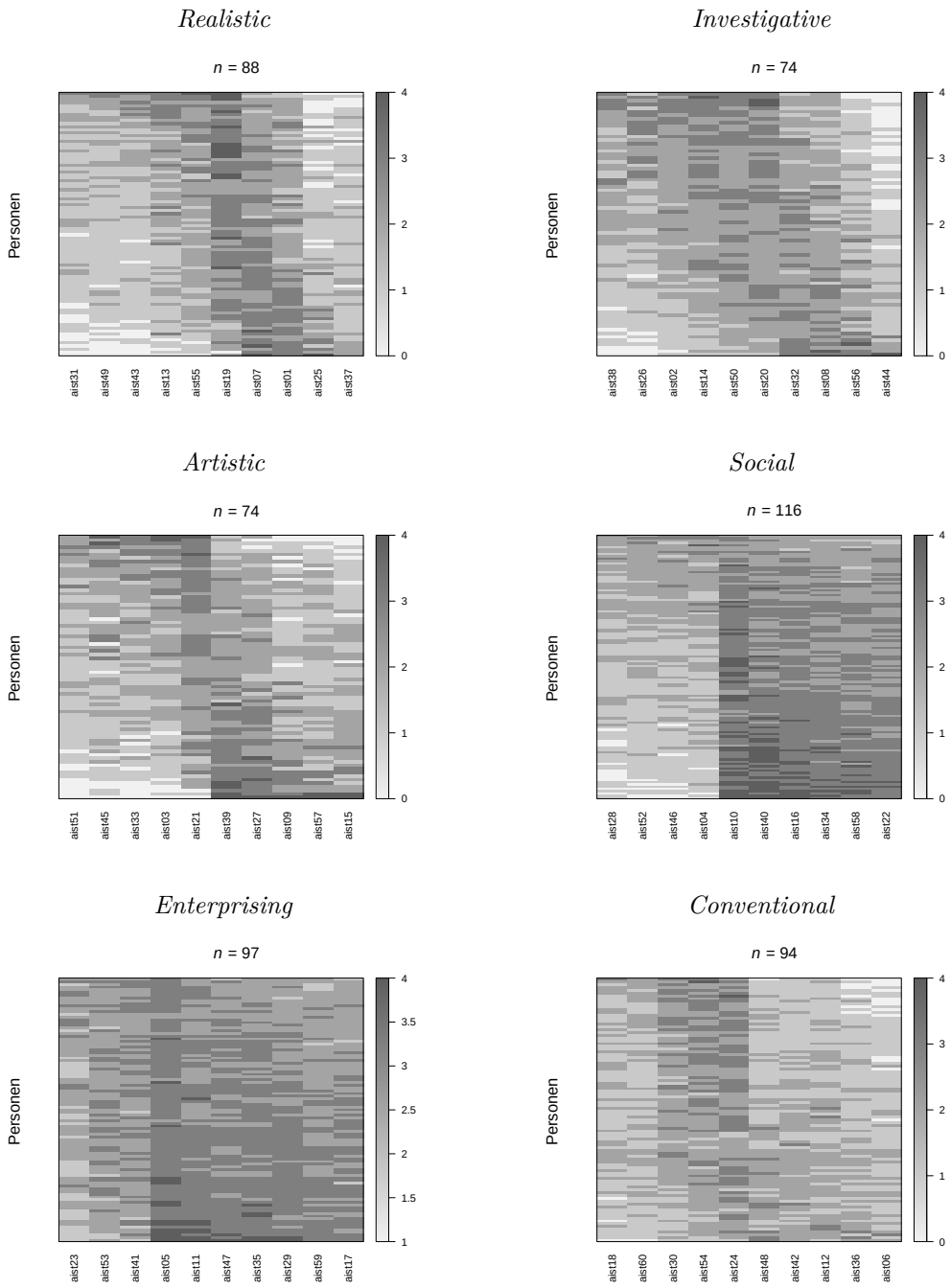


Abbildung 7.19 Stichprobe I: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Nähe–Distanz*-Antwortprozess (MDS) für sechs Skalen des AIST-R; Graustufen entsprechen den Antwortkategorien: Dunkel ≡ 4 ≡ „*Das interessiert mich sehr; das tue ich sehr gerne*“ – hell ≡ 0 ≡ „*Das interessiert mich gar nicht; das tue ich nicht gerne*“.

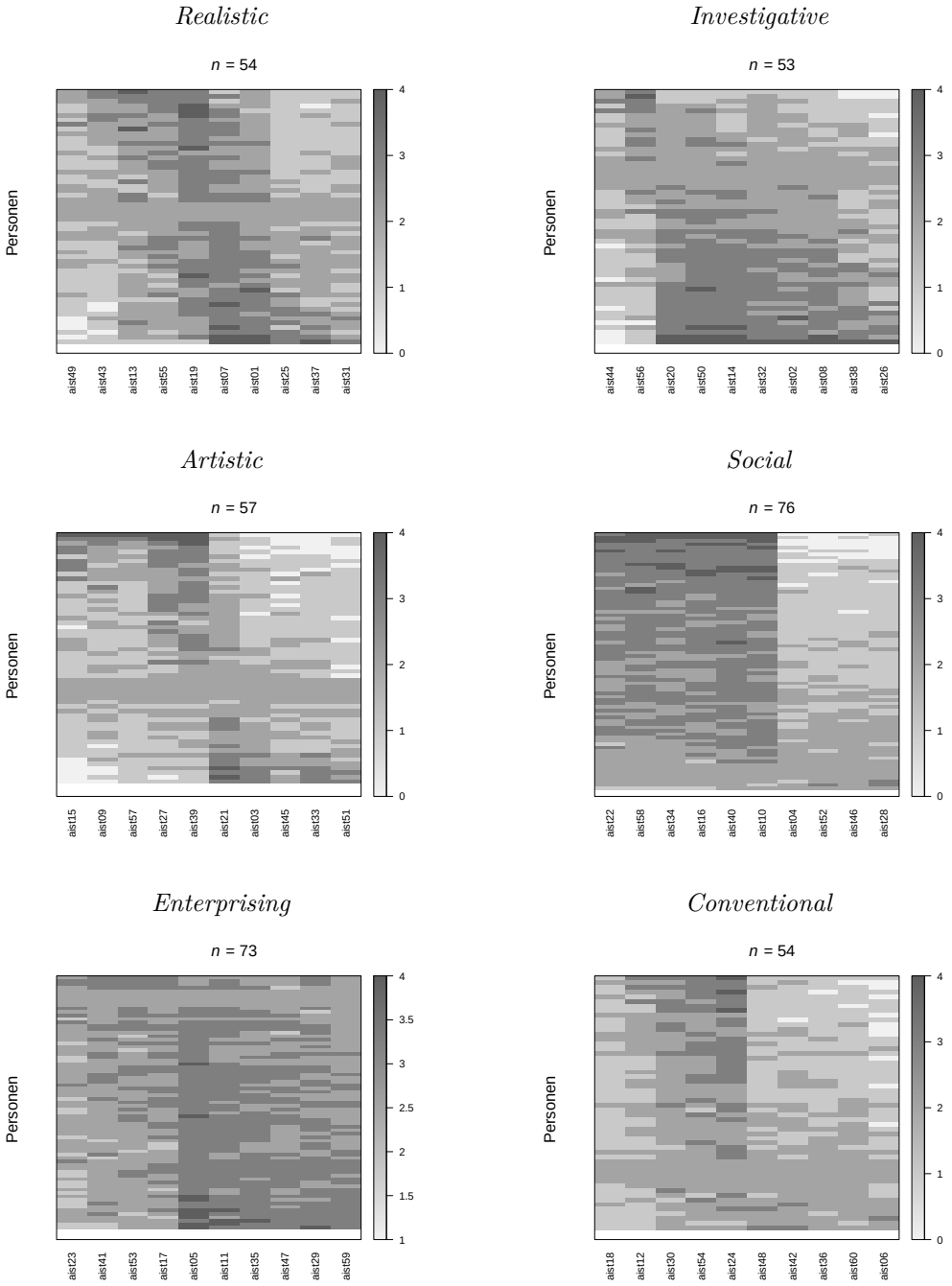


Abbildung 7.20 Stichprobe II: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Nähe–Distanz*-Antwortprozess (MDS) für sechs Skalen des AIST-R; Graustufen entsprechen den Antwortkategorien: Dunkel ≡ 4 ≡ *„Das interessiert mich sehr; das tue ich sehr gerne“* – hell ≡ 0 ≡ *„Das interessiert mich gar nicht; das tue ich nicht gerne“*.

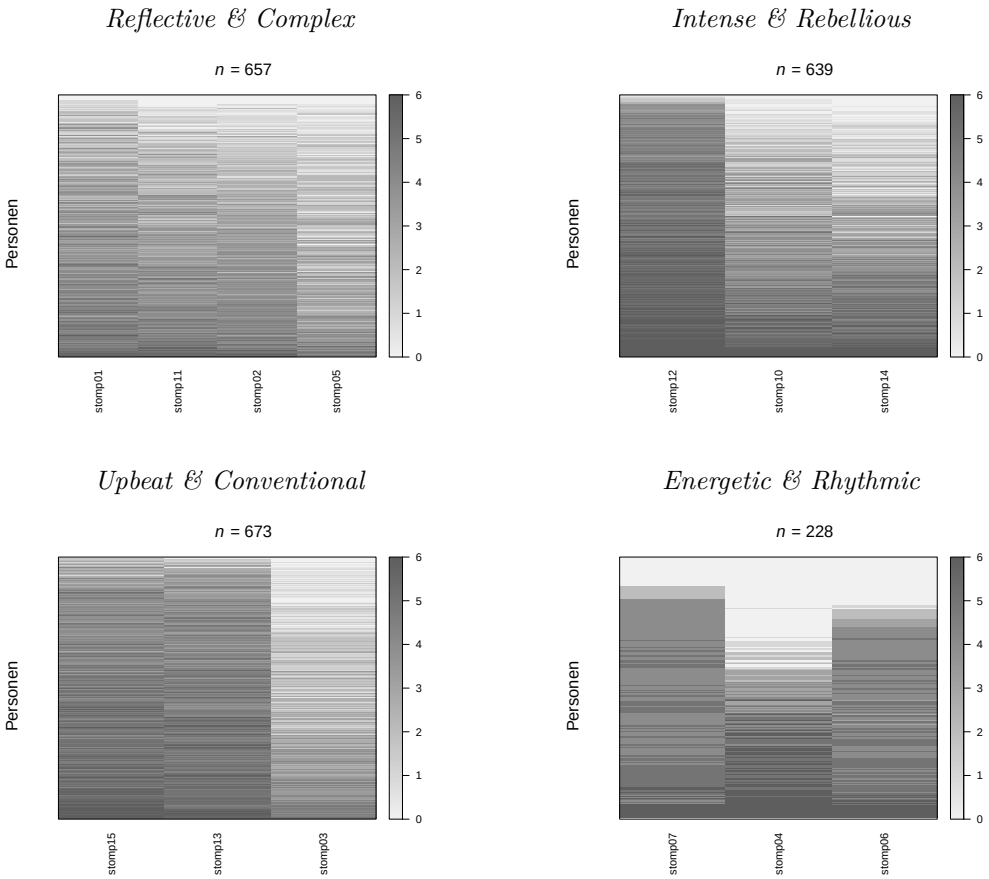


Abbildung 7.21 Stichprobe I: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Dominanz*-Antwortprozess (PCM) für vier Skalen des STOMP; Graustufen entsprechen den Antwortkategorien: Dunkel $\equiv 6 \equiv$ *„Mag ich sehr“* – hell $\equiv 0 \equiv$ *„Mag ich überhaupt nicht“*.

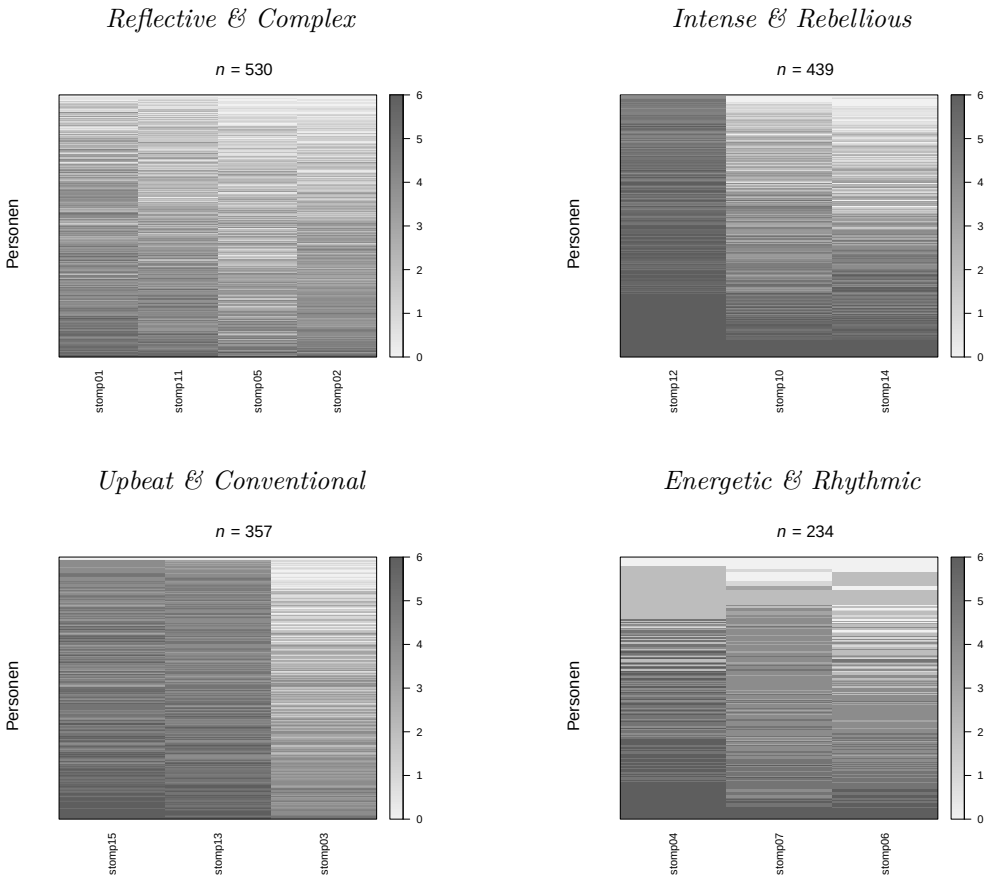


Abbildung 7.22 Stichprobe II: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Dominanz*-Antwortprozess (PCM) für vier Skalen des STOMP; Graustufen entsprechen den Antwortkategorien: Dunkel $\equiv 6 \equiv$ *„Mag ich sehr“* – hell $\equiv 0 \equiv$ *„Mag ich überhaupt nicht“*.

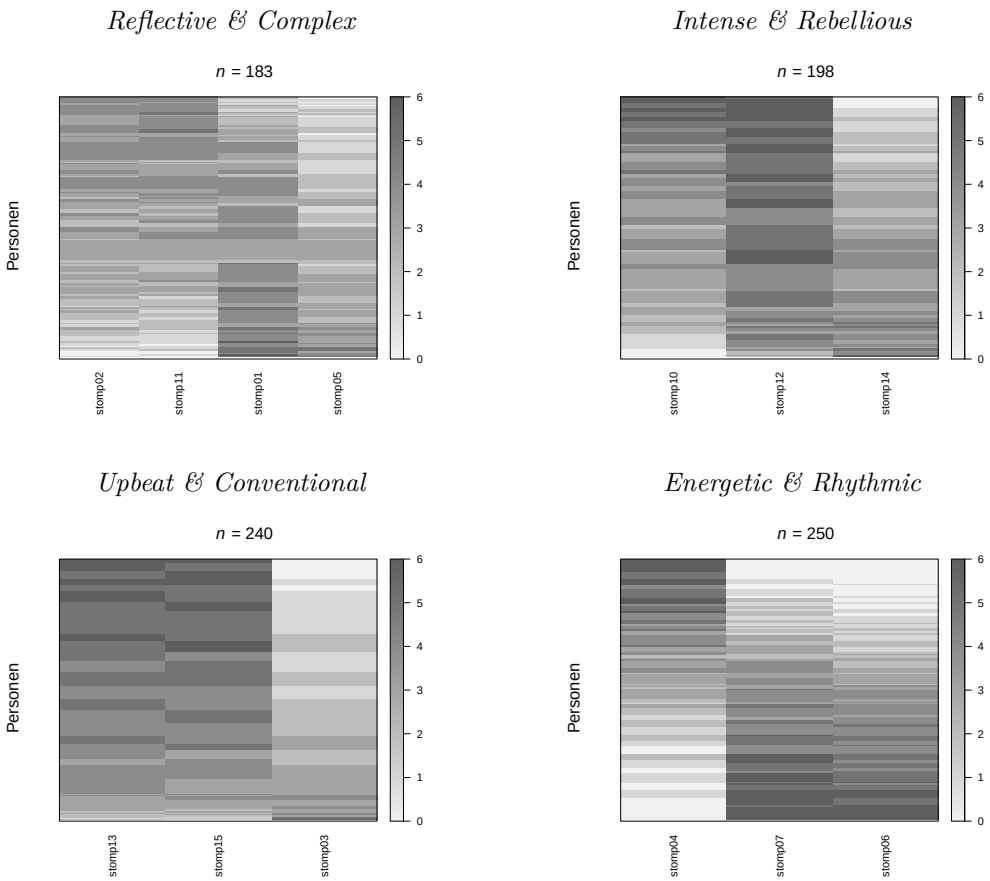


Abbildung 7.23 Stichprobe I: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Nähe–Distanz*-Antwortprozess (MDS) für vier Skalen des STOMP; Graustufen entsprechen den Antwortkategorien: Dunkel ≡ 6 ≡ *„Mag ich sehr“* – hell ≡ 0 ≡ *„Mag ich überhaupt nicht“*.

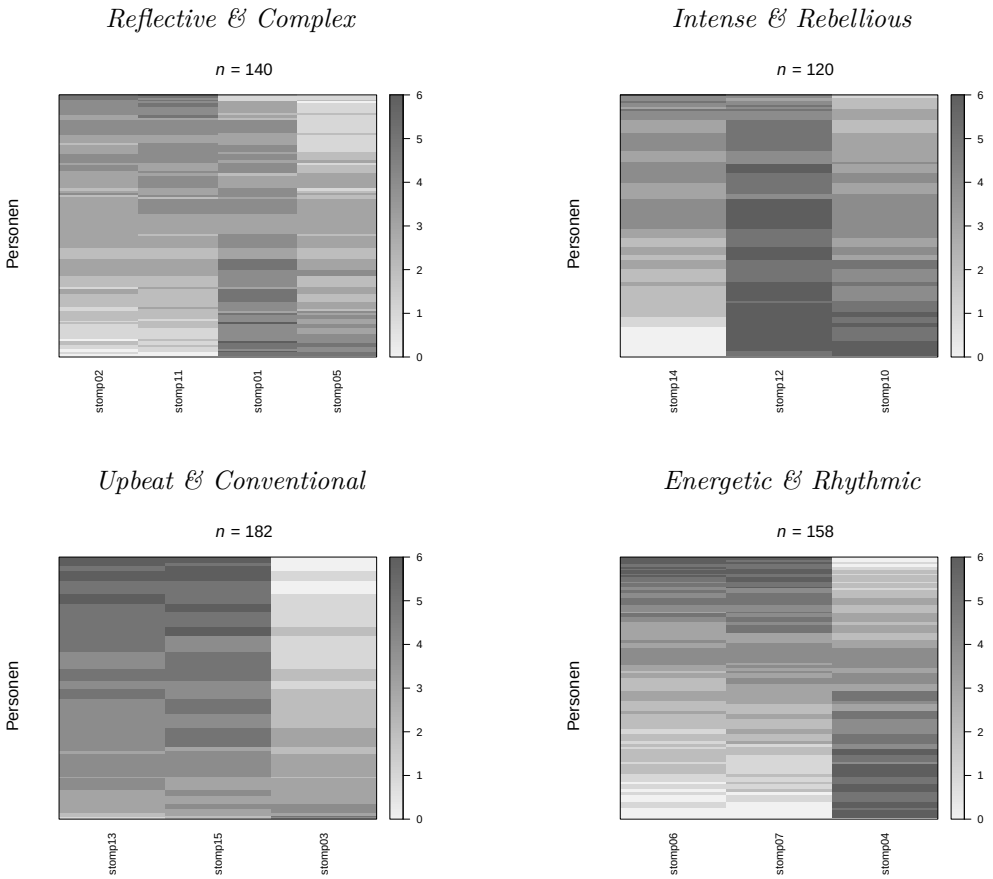


Abbildung 7.24 Stichprobe II: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Nähe–Distanz*-Antwortprozess (MDS) für vier Skalen des STOMP; Graustufen entsprechen den Antwortkategorien: Dunkel ≡ 6 ≡ *„Mag ich sehr“* – hell ≡ 0 ≡ *„Mag ich überhaupt nicht“*.

Die Anwendung der festen cut-off Grenzen ($STRESS = .05$ und $Q = .2$) zur Klassifikation der Personen führt, insbesondere beim Q-Index, zu einer „einseitigen“ Zuordnung der Personen zu einem der beiden Antwortprozesse. So werden beim AIST–R in der Dimension Realistic 1227 Personen (1088+139) nach dem Q-Index dem *Dominanz*-Antwortprozess zugeordnet und 114 Personen ($111 + 3$) als davon abweichend klassifiziert (vgl. Tabelle 7.2). Von diesen 114 nicht dem *Dominanz*-Antwortprozess zugeordneten Personen, lassen sich aber gleichzeitig nach dem $STRESS$-Kriterium lediglich 3 Personen dem *Nähe–Distanz*-Antwortprozess zuordnen (vgl. Tabelle 7.2). Von den 142 Personen ($139 + 3$), welche beim AIST–R in der Dimension Realistic nach dem $STRESS$-Kriterium dem *Nähe–Distanz*-Antwortprozess zugeordnet werden lassen sich aber 139 Personen nach dem Q-Index ebenso gut dem *Dominanz*-Antwortprozess zuordnen (vgl. Tabelle 7.2).

Zur Kreuzvalidierung der Klassifikation werden daher zusätzlich die Daten aus beiden Stichproben (Stichprobe I und II) zunächst nochmals gemeinsam (nach beiden Modellen) Skaliert und zweitens ein verteilungsbasiertes cut-off Kriterium (unteres Quartil der Verteilung) zur Klassifikation herangezogen. Das Ergebnis dieses Vorgehens unterscheidet sich in Bezug auf die resultierende grafische Darstellung der auf dieser Basis reorganisierten Datenmatrizen, zumindest strukturell, kaum. Im Anhang D sind die entsprechenden grafischen Darstellungen der reorganisierten Teildatenmatrizen aus der Gesamtstichprobe jeweils für die Skalen der drei Konstrukte zu beiden Antwortprozessen dargestellt (vgl. Abbildungen D.1, D.2,D.3, D.4,D.5 und D.6).

Die durch die Wahl eins verteilungsbasierten cut-off Kriteriums entstehenden Veränderungen in der Kreuzklassifikation der beiden Indizes ($STRESS$ und Q) werden auf den folgenden Seiten erläutert und sind in den Tabellen 7.4, 7.5 und 7.6 dargestellt.

Die Anwendung der jeweils verteilungsbasierten cut-off Grenzen nach gemeinsamer Skalierung der beiden Stichproben I und II führt erwartungsgemäß zu einer weniger „einseitigen“ Zuordnung der Personen zu einem der beiden Antwortprozesse. Die veränderte Kreuzklassifikation der beiden Indizes ($STRESS$ und Q) ist in den Tabellen 7.4, 7.5 und 7.6 gezeigt.

Tabelle 7.4 Kreuztabellierung der Personenklassifikation nach implizitem Antwortprozess für den BFI–K.

Neurotizismus				**Extraversion**				**Offenheit**			
	Nähe–Distanz				*Nähe–Distanz*				*Nähe–Distanz*		
	nein	ja	fehlend		nein	ja	fehlend		nein	ja	fehlend
Dominanz	[N]	[U]		*Dominanz*	[N]	[U]		*Dominanz*	[N]	[U]	
nein [N]	828	173	0	nein [N]	775	184	0	nein [N]	798	185	0
ja [K]	149	191	2	ja [K]	222	160	2	ja [K]	201	157	2
fehlend	0	0	0	fehlend	0	0	0	fehlend	0	0	0

Verträglichkeit				**Gewissenhaftigkeit**			
	Nähe–Distanz				*Nähe–Distanz*		
	nein	ja	fehlend		nein	ja	fehlend
Dominanz	[N]	[U]		*Dominanz*	[N]	[U]	
nein [N]	839	155	0	nein [N]	754	240	0
ja [K]	161	185	3	ja [K]	244	104	1
fehlend	0	0	0	fehlend	0	0	0

Anmerkungen: Gemeinsame Skalierung Stichprobe I und II; Absolute Häufigkeiten der klassifizierten Personen; ja = Personen-Fit-Index < cut-off Kriterium (unteres 25% Quartil) - [K]: (kumulativ), *Dominanz*-Antwortprozess, [U]: (Unfolding) *Nähe–Distanz*-Antwortprozess, [N]: keine Zuordnung zu einem Antwortprozess; fehlend = keine Klassifikation der Person aufgrund fehlender Werte auf den entsprechenden Items; $n = 1343$.

Bei der Betrachtung der Kreuztabellierung fällt nun allerdings auf, dass ein nicht unerheblicher Anteil der Personen anhand der jeweiligen Personen-Fit-Indizes keinem der beiden Antwortprozesse zugeordnet werden kann. So liegt die Anzahl der Personen in der linken oberen Zelle der jeweiligen 2×2 Tabelle über die beiden Konstrukte Persönlichkeit (BFI–K) und berufliche Interessen (AIST–R) und deren Skalen hinweg, in einem Bereich von $n = 754$ bis $n = 895$ Personen (vgl. Tabellen 7.4 und 7.5).

Etwas günstiger fällt, im Hinblick auf die Eindeutigkeit der Klassifikation, das Ergebnis für den STOMP aus. Die Anzahl der nicht eindeutig zugeordneten Personen liegt hier über alle vier Skalen in einem Bereich von $n = 679$ bis $n = 771$ Personen (vgl. Tabelle 7.6).

Darüber hinaus überrascht zunächst das Ergebnis, dass ein ebenfalls nicht

Tabelle 7.5 Kreuztabellierung der Personenklassifikation nach implizitem Antwortprozess für den AIST–R.

Realistic				**Investigative**				**Artistic**			
	Nähe–Distanz				*Nähe–Distanz*				*Nähe–Distanz*		
	nein	ja	fehlend		nein	ja	fehlend		nein	ja	fehlend
Dominanz	[N]	[U]		*Dominanz*	[N]	[U]		*Dominanz*	[N]	[U]	
nein [N]	829	178	0	nein [N]	846	161	0	nein [N]	836	171	0
ja [K]	177	157	2	ja [K]	160	174	2	ja [K]	169	164	3
fehlend	0	0	0	fehlend	0	0	0	fehlend	0	0	0
Social				**Enterprising**				**Conventional**			
	Nähe–Distanz				*Nähe–Distanz*				*Nähe–Distanz*		
	nein	ja	fehlend		nein	ja	fehlend		nein	ja	fehlend
Dominanz	[N]	[U]		*Dominanz*	[N]	[U]		*Dominanz*	[N]	[U]	
nein [N]	832	175	0	nein [N]	869	138	0	nein [N]	895	112	0
ja [K]	174	160	2	ja [K]	136	197	3	ja [K]	111	223	2
fehlend	0	0	0	fehlend	0	0	0	fehlend	0	0	0

Anmerkungen: Gemeinsame Skalierung Stichprobe I und II; Absolute Häufigkeiten der klassifizierten Personen; ja = Personen-Fit-Index < cut-off Kriterium (unteres 25% Quartil) - [K]: (kumulativ), *Dominanz*-Antwortprozess, [U]: (Unfolding) *Nähe–Distanz*-Antwortprozess, [N]: keine Zuordnung zu einem Antwortprozess; fehlend = keine Klassifikation der Person aufgrund fehlender Werte auf den entsprechenden Items; $n = 1343$.

unerheblicher Teil der Personen beiden impliziten Antwortprozessen *gleichzeitig* zugeordnet werden kann. Wie auch bei den nicht eindeutig klassifizierten Personen fällt hier der Anteil der zu beiden Antwortprozessen passenden Personen für den BFI–K und den AIST–R im Vergleich zu dem STOMP höher aus. Während die Anzahl der in diesem Sinne „zweifach klassifizierten“ Personen für den STOMP in einem Bereich von $n = 57$ bis $n = 121$ liegt (vgl. Tabelle 7.6), liegt diese für die beiden anderen Konstrukte über alle Skalen in einem Bereich von $n = 104$ bis $n = 223$ (vgl. Tabellen 7.4 und 7.5).

Tabelle 7.6 Kreuztabellierung der Personenklassifikation nach implizitem Antwortprozess für den STOMP.

Reflective & Complex				**Intense & Rebellious**			
	Nähe–Distanz				*Nähe–Distanz*		
	nein	ja	fehlend		nein	ja	fehlend
Dominanz	[N]	[U]		*Dominanz*	[N]	[U]	
nein [N]	771	234	0	nein [N]	771	230	0
ja [K]	234	101	3	ja [K]	218	121	3
fehlend	0	0	0	fehlend	0	0	0
Upbeat & Conventional				**Energetic & Rhythmic**			
	Nähe–Distanz				*Nähe–Distanz*		
	nein	ja	fehlend		nein	ja	fehlend
Dominanz	[N]	[U]		*Dominanz*	[N]	[U]	
nein [N]	748	240	0	nein [N]	679	280	0
ja [K]	251	101	3	ja [K]	270	57	0
fehlend	0	0	0	fehlend	54	0	3

Anmerkungen: Gemeinsame Skalierung Stichprobe I und II; Absolute Häufigkeiten der klassifizierten Personen; ja = Personen-Fit-Index < cut-off Kriterium (unteres 25% Quartil) - [K]: (kumulativ), *Dominanz*-Antwortprozess, [U]: (Unfolding) *Nähe–Distanz*-Antwortprozess, [N]: keine Zuordnung zu einem Antwortprozess; fehlend = keine Klassifikation der Person aufgrund fehlender Werte auf den entsprechenden Items; $n = 1343$.

Diskussion

In der vorliegenden Untersuchung wird der Frage nachgegangen, ob sich innerhalb der analysierten Antwortdaten bei unterschiedlichen Konstrukten und deren Dimensionen jeweils zwei, nach ihrem impliziten Antwortmodell unterscheidbare, Personengruppen identifizieren lassen. Zur Anwendung kommen dazu zwei Skalierungsmodelle, welche geeignet sind den jeweiligen Antwortprozess abzubilden. Ferner wird die Eindeutigkeit der erzielten Personenklassifikation untersucht, welche durch die beiden Indizes für die lokale Modellpassung erreicht wird.

Die Verteilungen der jeweiligen Indizes für die lokale Modellpassung weisen sowohl für den STRESS-Index als auch für den Q-Index insgesamt darauf hin, dass eher geringe lokale Abweichungen vom damit jeweils überprüften Antwortmodell in den Daten vorliegen. In der vergleichenden Gesamtschau der empirischen Werte für den Median der Gesamtverteilungen – jeweils beider Indizes – zeigt sich die „mittlere“ (lokale) Anpassung an das jeweilige Antwortmodell. Dabei, lässt sich im Vergleich der drei untersuchten Konstrukte eine gewisse Tendenz erkennen. So steigt bezogen auf den Q-Index zur Identifikation nicht passender Antwortmuster nach dem *Dominanz*-Antwortprozess das Ausmaß lokaler Modellverletzungen über die drei Konstrukte *Persönlichkeit*, *berufliche Interessenorientierungen* und *Musikpräferenzen* zuzunehmend an. Daraus kann gefolgert werden, dass beispielsweise für den STOMP zur Erfassung von Musikpräferenzen im Vergleich zum BFI–K zur Erfassung der Persönlichkeit eine vergleichsweise größere Inkompatibilität zur summativen Skalierung der Items des Fragebogenverfahrens besteht. Diese Schlussfolgerung stützt auch der Befund, dass über die vier Skalen des STOMP diejenigen Personengruppen, welche (eindeutig) einem *Nähe–Distanz*-Antwortprozess zugeordnet werden, im Vergleich zu den beiden anderen Konstrukten (AIST–R und BFI–K) am größten ausfallen (*Reflective & Complex*: $n = 234$; *Intense & Rebellious*: $n = 230$; *Upbeat & Conventional*: $n = 240$; *Energetic & Rhythmic*: $n = 280$ im Vergleich zu *Neurotizismus*: $n = 173$; *Extraversion*: $n = 184$; *Offenheit*: $n = 185$; *Verträglichkeit*: $n = 155$; *Gewissenhaftigkeit*: $n = 240$). In diesem Sinne fällt auch der Median für die Gesamtverteilung des STRESS-Index im Bereich *Musikpräferenzen* am geringsten aus (STOMP: $\text{Median}_{STRESS,\text{STOMP}} = .106$), was dahingehend interpretiert werden kann,

dass über die vier Skalen des STOMP für diesen Bereich eine vergleichsweise größere Kompatibilität zu einem *Nähe–Distanz*-Antwortprozess besteht.

Die in der vorliegenden Untersuchung aufgestellte Hypothese, dass sich diese beiden unterschiedlichen Antwortprozesse nicht nur als Spezifika der Skalen und deren Items darstellen lassen, sondern auch als *Metaeigenschaft* einzelner Personengruppen, wird durch die Befunde aus der visuellen Inspektion der Bertin-Plots, der reorganisierten Datenmatrizen für alle Konstrukte und deren Dimensionen, für beide Stichproben gestützt. So ergeben sich für die grafischen Darstellungen (Bertin-Plots) der nach den jeweiligen Modellparametern reorganisierten Datenmatrizen vergleichsweise eindeutige Muster. Diese beiden Muster visualisieren den jeweiligen Antwortprozess, der über die Indizes zur Personenpassung klassifizierten Personengruppen (vgl. Abbildungen 7.13, 7.14, 7.17, 7.18, 7.21, 7.22, 7.15, 7.16, 7.19, 7.20, 7.23 und 7.24). Dabei zeigen die Bertin-Plots für den Dominanz-Antwortprozess jeweils typische *dreieckige* Muster aus unterschiedlichen Graustufen, welche die gewählten Antwortkategorien repräsentieren (vgl. Abbildungen 7.13, 7.14, 7.17, 7.18, 7.21 und 7.22). Für den Nähe–Distanz-Antwortprozess ergeben sich entsprechend dagegen Muster mit mehr oder weniger deutlich ausgeprägten *Diagonalen* (vgl. Abbildungen 7.15, 7.16, 7.19, 7.20, 7.23 und 7.24).

Zusammenfassend können die Befunde aus der vorliegende Untersuchung dahingehend interpretiert werden, dass insgesamt die Methoden der Multidimensionalen Skalierung mit dem STRESS-Index und die Rasch-Skalierung über den *PAIR*-Algorithmus mit dem Q-Index dazu geeignet sind in den analysierten Daten die beiden unterschiedlichen Antwortprozesse zu identifizieren.

Unbefriedigend oder überraschend bleibt in der vorliegenden Untersuchung der Befund, dass sich, bezogen auf die gemeinsame Skalierung beider Stichproben, ein gewisser Anteil von Personen in den Daten keinem, und ein etwas größerer Anteil von Personen beiden, Antwortprozessen (vgl. Tabellen 7.4, 7.5 und 7.6), zuordnen lässt. Eine mögliche Erklärung für diesen Befund, der in seinem Ausmaß auch in Abhängigkeit der gewählten cut-off Grenzen variiert, könnte darin begründet liegen, dass sich, wie in Abschnitt 4.7 gezeigt, die ICCs für den Dominanz- und Nähe–Distanz-Antwortprozess in bestimmten Bereichen der Verteilung des latenten Merkmals angleichen.

7.2 Konsistenz impliziter Antwortmodelle und Zusammenhänge mit Antworttendenzen

Einleitung

Der zunächst „überraschende“ Befund einer doppelten Klassifikation der Personen zu den *beiden* unterschiedlichen Antwortmodellen aus Studie 7.1 indiziert eine genauere Betrachtung dieser jeweiligen Personengruppen im Hinblick auf die Qualität des Antwortverhaltens. Das dichotome „Personenmerkmal“ der Klassifikation zu einem der beiden (impliziten) Antwortmodelle soll daher in der vorliegenden Analyse weiter untersucht werden. Die hier durchgeführten Analysen verfolgen dabei zwei Hauptfragestellungen – (1) die Frage nach der Konsistenz des impliziten Antwortmodells der Personen und (2) die Frage nach der Verbindung dieses impliziten Antwortverhaltens mit anderen (klassifikatorischen) Merkmalen der Antwortmuster der Personen.

Es stellt sich die Frage, ob das Antwortverhalten nach einem der beiden Antwortprozesse als eine konsistente Reaktionstendenz der Personen beschrieben werden kann. Dementsprechend wird analysiert, ob das Antwortverhalten der Personen nach einem der beiden (impliziten) Antwortmodelle über alle Skalen der drei Konstrukte in konsistenter Weise besteht. Zu klären ist dabei, ob es sich bei der Beantwortung der vorgelegten Fragebogenskalen nach einem der beiden Antwortmodelle um einen Antwortstil [*response style*] handelt (vgl. Kapitel 3), welcher als übergeordnet klassifizierendes Personenmerkmal aufgefasst werden kann. Auf dieser übergeordneten Ebene könnte eine solche Personeneigenschaft als *Metaeigenschaft* [*metatrait*] interpretiert werden (vgl. Abschnitte 3.2.3, 3.3 und 4.4.2).

Ferner stellt sich die Frage, ob Verbindungen zwischen der Zuordnung zu einem der beiden Antwortprozesse und anderen klassifikatorischen Merkmalen der Antwortmuster der Personen bestehen. Die am Ende von Abschnitt 4.7 dargestellte theoretische Betrachtung einer Konvergenz der ICCs für Items, welche eine extreme Merkmalsausprägung repräsentieren, soll hier zum Anlass für eine weitergehende Untersuchung der empirischen Daten genommen werden. Anhand der in Abschnitt 4.7 gezeigten Abbildung aus Stark et al. (2006, S. 27) wird dargestellt, dass sich bei einem Item, welches eine extre-

me Merkmalsausprägung repräsentiert die Kurvenverläufe der ICCs für den Dominanz- und den Nähe–Distanz-Antwortprozess gerade im Bereich einer *mittleren* Merkmalsausprägung der Personen auf dem latenten Kontinuum weitgehend angleichen. Daraus kann gefolgert werden, dass die dichotome Klassifikation der Antworten – Tendenz zur Mitte vs. Tendenz zu Extremen (MRS vs. ERS; vgl. Abschnitt 3.2.4) – nach deren Skalierung mit dem *mixed-Rasch-Modell* (vgl. Abschnitt 4.7) zur Abbildung eines *Dominanz*-Antwortprozesses in zwei latenten Klassen – also sozusagen eines *Dominanz-mixture-Modells* – einen Zusammenhang mit der *eindeutigen* Klassifikation zu einem der beiden Antwortprozesse aufweisen kann. Spezifisch kann auf Basis dieser Überlegung die Hypothese aufgestellt werden, dass Personen deren Antwortmuster nach einem Dominanz-Antwortmodell als *„Mittelkreuzer“* (MRS – vgl. Abschnitt 3.2.4) klassifiziert sind, gleichzeitig sowohl dem *Dominanz*- als auch dem *Nähe–Distanz*-Antwortprozess (mit eingipfliger Itemcharakteristik) zugeordnet werden können. Lässt sich diese Hypothese anhand der empirischen Daten stützen, kann daraus eine Erklärung für den in Studie 7.1 erzielten Befund einer doppelten Klassifikation der Personen zu den beiden unterschiedlichen Antwortprozessen abgeleitet werden. Daher werden hier die Zusammenhänge zwischen den Klassifikationsergebnissen zum impliziten Antwortmodell mit dem Klassifikationsergebnissen zur Antworttendenz nach einem *Dominanz*-Antwortmodell analysiert. Dazu werden die Befunde zur Klassifikation der unterschiedlichen Antworttendenzen aus Studie 6.1 für die vier Skalen des AIST–R (*Realistic*, *Investigative*, *Social* und *Conventional*) sowie die eine Skala des STOMP (*Energetic & Rhythmic*) herangezogen.

Daten

Ebenso wie in der vorangegangenen Analyse in Abschnitt 7.1 beziehen sich die Analysen in der vorliegenden Auswertung auf die in Abschnitt 5.2 beschriebenen beiden Stichprobe aus dem Erhebungszeitraum 2007 – 2009 (Stichprobe I) und 2010 – 2011 (Stichprobe II). Insgesamt liegen also Daten von $n = 1343$ Personen vor. Bei den zur Analyse der Konsistenz impliziter Antwortmodelle verwendeten Daten handelt es sich um insgesamt 30 abgeleitete, dichotome Variablen aus der vorangegangen Analyse in Studie 7.1. Dies sind pro Konstrukt und Skala jeweils zwei dichotome Indikatorvariablen (basierend auf $STRESS$

und Q mit verteilungsbasierten cut-off Grenzen) zur Passung zu einem der beiden impliziten Antwortmodelle wie sie in Studie 7.1 identifiziert wurden (jeweils insgesamt 2×15 dichotome Variablen zur Personenpassung zu den beiden impliziten Antwortmodellen: AIST–R, 2×6 Indikatorvariablen; BFI–K, 2×5 Indikatorvariablen; STOMP, 2×4 Indikatorvariablen). Hinzu genommen werden die jeweiligen Indikatorvariablen zur Klassifikation der Personen im Hinblick auf ihre Antworttendenz auf vier der sechs Skalen des AIST–R, sowie einer der vier Skalen des STOMP aus Studie 6.1.

Methode

Zur Untersuchung der Konsistenz des impliziten Antwortmodells wird das klassifizierte Reaktionsmuster auf den Indikatorvariablen aus Studie 7.1 zur Zuordnung der Personen zum Nähe–Distanz- und Dominanz-Antwortprozess analysiert. Als Analysestrategie werden dazu zwei Methoden eingesetzt. Einerseits werden die Indikatorvariablen für jedes der drei Konstrukte einer Analyse latenter Klassen (LCA - vgl. Abschnitt 4.6.2) unterzogen. Die Berechnung der LCA-Modelle erfolgt dabei jeweils für alle Dimensionen des jeweiligen Konstruktes (*berufliche Interessenorientierungen, Persönlichkeit* und *Präferenzen des Musikgeschmacks*) getrennt und andererseits konstruktübergreifend für alle Dimensionen der drei Konstrukte gemeinsam. Ergänzend werden mit den Indikatorvariablen für jedes der drei Konstrukte (einzeln) Konfigurationsfrequenzanalysen (KFA) gerechnet (vgl. Abschnitt 4.6.1). Die Ergebnisse aus den beiden Methoden werden kreuztabelliert, sodass die bei den konstruktspezifisch berechneten LCA-Modellen identifizierten latenten Klassen über die in der KFA gefundenen signifikanten Reaktionsmuster qualifiziert werden können. Das Ziel besteht darin, zusätzlich zu den Befunden aus der LCA zweiter Ordnung über die signifikanten Muster der KFA die skalenübergreifende Konsistenz des impliziten Antwortmodells zu beschreiben.

Die Befunde aus der Klassifikation der Antworttendenzen auf der Skala *Energetic & Rhythmic* des STOMP und die Befunde zu den unterschiedlichen Antworttendenzen auf vier der sechs Skalen des AIST–R aus Studie 6.1 (vgl. dort Tabelle 6.2) werden mit dem Klassifikationsergebnis aus den Personen-Fit-Indizes zu den beiden impliziten Antwortmodellen aus Studie 7.1 in Bezug gesetzt. Methodisch wird hier auf die Konfigurationsfrequenzanalyse (vgl. Ab-

schnitt 4.6.1) zurückgegriffen. Analysiert werden dabei jeweils (in einzelnen Analysen) die Indikatorvariablen für die vier Skalen des AIST–R und die Indikatorvariable aus dem Ergebnis der Skalierung des STOMP zur Klassifikation der Personen im Hinblick auf ihre Antworttendenz (vgl. Studie 6.1) mit jeweils zwei dichotomen Indikatorvariablen zur Passung zum jeweiligen Antwortmodell aus Studie 7.1.

Für die KFA-Analysen wird das *R*-Paket `confreq` (Heine et al., 2019) für die freie Statistikumgebung *R* (R Core Team, 2018) eingesetzt. Für die Analyse latenter Klassen auf den Indikatorvariablen wird das *R*-Paket `poLCA` (Linzer & Lewis, 2011) für die freie Statistikumgebung *R* (R Core Team, 2018) eingesetzt.

Ergebnisse

Konsistenz impliziter Antwortmodelle oder Antwortprozesse

Die Ergebnisse aus der *Latent-Class-Analysis* (LCA) über die jeweils 15 Indikatorvariablen zur skalenweisen Klassifikation der Personen zum Dominanz- oder Nähe–Distanz-Antwortprozess über alle drei Konstrukte zeigt Tabelle 7.7. Die resultierenden informationstheoretischen Kriterien AIC und BIC weisen dabei für beide Gruppen der Indikatorvariablen (Dominanz- oder Nähe–Distanz-Antwortprozess) übereinstimmend auf die (relativ) beste Passung der 2-Klassen-Lösung hin.

Für die jeweils nach Konstrukt getrennt gerechneten LCA Modelle über die sechs (AIST–R), fünf (BFI–K) und vier (STOMP) Indikatorvariablen zur Personenpassung zum Dominanz- und Nähe–Distanz-Antwortprozess, erweist sich nach dem informationstheoretischen Kriterium BIC für den AIST–R und den BFI–K jeweils die 2-Klassen-Lösung als das am besten passende Modell. Demgegenüber indiziert der BIC für die Analysen der gesamten Indikatorvariablen zu den vier Skalen des STOMP jeweils die 1-Klassen-Lösung als das am besten passende Modell (vgl. Tabelle 7.8).

Tabelle 7.7 Relativer Modellvergleich – informationstheoretische Kriterien – LCA über Indikatorvariablen für die Personenpassung zum Dominanz- und Nähe–Distanz-Antwortprozess über drei Konstrukte.

	Jeweils 15 Indikatorvariablen (AIST–R, BFI–K, STOMP)							
	Dominanz [a], $r = 0.87$ [*]				*Nähe–Distanz* [b], $r = 0.90$[*]			
Klassen	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-10987	15	22004	22081	-11379	15	22789	22867
2	-10803	31	21668	**21828**	-11078	31	22218	**22379**
3	-10789	47	21672	21914	-11052	47	22197	22441
4	-10773	63	21672	21997	-11027	63	22179	22507

Anmerkungen: Latent-Class-Analysis (LCA) für 1 – 4 latente Klassen; ML–Schätzung mit R-Paket `poLCA`; LL = LogLikelihood; np = Anzahl freie Modellparameter; [a] Klasse 2: ($p = 0.26$): konsistent *Dominanz*, Klasse 1: ($p = 0.74$): konsistent nicht *Dominanz*; [b] Klasse 1 ($p = 0.28$): konsistent *Nähe–Distanz*, Klasse 2 ($p = 0.72$): konsistent nicht *Nähe–Distanz*); [*] Reliabilität der Klassifikation (gemittelte modale Klassenzuordnungswahrscheinlichkeit).

Tabelle 7.8 Relativer Modellvergleich – informationstheoretische Kriterien – LCA über Indikatorvariablen für die Personenpassung zum Dominanz- und Nähe–Distanz-Antwortprozess für drei Konstrukte.

	Jeweils 6 Indikatorvariablen AIST–R							
	Dominanz				*Nähe–Distanz*			
Klassen	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-4533	6	9078	9109	-4521	6	9054	9086
2	-4371	13	8768	**8835**	-4286	13	8598	**8666**
3	-4363	20	**8766**	8870	-4271	20	**8582**	8686
4	-4359	27	8773	8913	-4259	27	8572	8713
	Jeweils 5 Indikatorvariablen BFI–K							
	Dominanz				*Nähe–Distanz*			
Klassen	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-3885	5	7781	7807	-3830	5	7671	7697
2	-3873	11	**7769**	**7826**	-3808	11	**7637**	**7694**
3	-3872	17	7777	7865	-3805	17	7644	7732
4	-3868	23	7782	7901	-3802	23	7651	7770
	Jeweils 4 Indikatorvariablen STOMP							
	Dominanz				*Nähe–Distanz*			
Klassen	LL	np	AIC	BIC	LL	np	AIC	BIC
1	-2915	4	5837	**5858**	-3039	4	6085	**6106**
2	-2904	9	**5825**	5872	-3026	9	**6070**	6117
3	-2903	14	5834	5906	-3023	14	6074	6147
4	-2902	19	5841	5939	-3022	19	6082	6181

Anmerkungen: Latent-Class-Modelle (LCA) für 1 – 4 latente Klassen; ML–Schätzung mit R-Paket `poLCA`; LL = LogLikelihood; np = Anzahl freie Modellparameter.

Zur Qualifizierung der beiden in den Konstrukten *berufliche Interessenorientierungen* (AIST–R) und *Persönlichkeit* (BFI–K) gefundenen latenten Klassen, welche sich aus der Analyse der Indikatorvariablen für den Antwortprozess ergeben, wird das Klassifikationsergebnis mit den signifikanten Typen aus der KFA der jeweiligen Indikatorvariablen kreuztabelliert. Die Ergebnisse dieser Analysen sind für die sechs Dimensionen des AIST–R in Tabelle 7.9 und für die fünf Dimensionen des BFI–K in Tabelle 7.10 dargestellt. Die KFA für die sechs Indikatorvariablen des AIST–R ergeben insgesamt zehn signifikante Muster (Typen – vgl. Abschnitt 4.6.1) und die entsprechenden Analysen für den BFI–K ergeben zwei signifikante Muster. Die jeweiligen mittleren Klassenzuordnungswahrscheinlichkeiten für beide Konstrukte und alle signifikanten Muster weisen durchweg hohe Werte zwischen $p = .999$ und $p = .701$ auf (vgl. Tabellen 7.9 und 7.10).

Tabelle 7.9 Kreuztabellierung LCA und KFA mit Indikatoren für die Personenpassung zum Nähe–Distanz-Antwortprozess für sechs Dimensionen des AIST–R.

p[a]	0.993	0.927	0.701	0.995	0.996	0.995	0.996	0.995	0.993	0.999
	Muster der Indikatorvariablen (R I A S E C)[b]									
Klassen	NNNNNN	NNNNNU	UNNNUN	NUUUUU	UNUUUU	UUNUUU	UUUNUU	UUUUNU	UUUUUN	UUUUUU
1[c]	425	51	8	0	0	0	0	0	0	0
2[c]	0	0	0	9	8	7	6	8	8	21

Anmerkungen: [a] mittlere Klassenzuordnungswahrscheinlichkeit – LCA [b] sechs (dichotome) Indikatorvariablen: R–*Realistic*, I–*Investigative*, A–*Artistic*, S–*Social*, E–*Enterprising*, C–*Conventional*, [c] Relative Klassengrößen: $p_{\text{Klasse 1}} = 0.73$, $p_{\text{Klasse 2}} = 0.27$, $n = 1340$; U = Passung zum *Nähe–Distanz*-Antwortprozess und N = keine Passung zum *Nähe–Distanz*-Antwortprozess, nach Dimensionen RIASEC (von unten nach oben); nur signifikante Muster nach KFA; $n = 507$.

Aus der KFA für die Indikatorvariablen (basierend auf dem *STRESS*) zu den Skalen des AIST–R ergeben sich für die latente Klasse 1 drei und für die latente Klasse 2 sieben signifikante Muster. Demnach lässt sich die latente Klasse 2 mit einer relativen Klassengröße von $p = 0.27$ durch das überwiegende Vorherrschen eines impliziten Antwortmodells nach dem Nähe–Distanz-

Antwortprozess charakterisieren. Die latente Klasse 1 mit einer relativen Klassengröße von $p = 0.73$ lässt sich anhand der signifikanten Muster aus der KFA dagegen als Klasse derjenigen Personen beschreiben, deren implizites Antwortmodell eher nicht zum Nähe–Distanz-Antwortprozess passt (vgl. Tabelle 7.9).

Tabelle 7.10 Kreuztabellierung LCA und KFA mit Indikatoren für die Personenpassung zum Nähe–Distanz-Antwortprozess für fünf Dimensionen des BFI–K.

p[a]	0.987	0.918
	Muster der Indikatorvariablen (N E O A C)[b]	
	NNNNN	UUUUU
Klassen		
1[c]	349	0
2[c]	0	9

Anmerkungen: [a] mittlere Klassenzuordnungswahrscheinlichkeit – LCA
[b] fünf (dichotome) Indikatorvariablen: N–*Neurotizismus*, E–*Extraversion*, O–*Offenheit*, A–*Verträglichkeit*, C–*Gewissenhaftigkeit*, C–*Conventional*,
[c] Relative Klassengrößen: $p_{\text{Klasse 1}} = 0.86$, $p_{\text{Klasse 2}} = 0.14$, $n = 1340$;
U = Passung zum *Nähe–Distanz*-Antwortprozess und N = keine Passung zum *Nähe–Distanz*-Antwortprozess, nach Dimensionen NEOAC (von links nach rechts); nur signifikante Muster nach KFA, $n = 358$.

Die KFA der Indikatorvariablen zu den fünf Skalen des BFI–K ergeben zwei signifikante Muster. Diese erweisen sich im Hinblick auf die Konsistenz ihres impliziten Antwortmodells skalenübergreifend als entweder eindeutig *passend* (Muster: „*UUUUU*“; $n = 9$), oder *nicht passend* (Muster: „*NNNNN*“; $n = 349$) zum Nähe–Distanz-Antwortprozess. Durch diese beiden Muster lässt sich die latente Klasse 1 ($p = 0.86$) als Klasse derjenigen Personen beschreiben, deren implizites Antwortmodell *keine Passung* zum Nähe–Distanz-Antwortprozess aufweisen. Die Klasse 2 ($p = 0.14$) umfasst diejenigen Personen, deren implizite Antwortmodelle über alle fünf Skalen eine *Passung* zum Nähe–Distanz-Antwortprozess aufweisen (vgl. Tabelle 7.10).

Assoziationen zwischen implizitem Antwortmodell und Antworttendenz

Die Tabelle 7.11 gibt die Ergebnisse für vier getrennte KFA-Analysen (je Skala) für die Skalen des AIST–R wieder. Dargestellt sind die nach dem exakten Binomialtest signifikanten Muster (Typen) auf den dichotomen Indikatorvariablen zur jeweiligen Passung zum impliziten Antwortmodell (K=Dominanz-Antwortprozess; U=Nähe–Distanz-Antwortprozess) und der dichotomen Indikatorvariablen zur Antworttendenz.

Bei der Betrachtung der Tabelle 7.11 fällt auf, dass sich für alle vier AIST–R-Dimensionen mindesten zwei signifikante Muster für die Indikatorvariablen ergeben bei denen die erwarteten Häufigkeiten von den beobachteten übertroffen werden – also „Typen". Lediglich für die AIST–R-Dimension *Realistic* ergibt sich noch ein zusätzlicher Typ; wobei hier die erwarteten und beobachteten Häufigkeiten nicht sehr stark voneinander abweichen. Die beiden signifikanten Typen sind über alle vier AIST–R-Dimensionen jeweils durch das gleiche, typische Merkmalsmuster gekennzeichnet: „**mittel K U**" und „**extrem N N**". Nach der „Übersetzung" der Kodierung der Indikatorvariablen ergibt sich dabei folgendes inhaltlich zu interpretierendes Muster: Personen die in Studie 6.1 in Bezug auf ihre Antworttendenz als „*Mittelkreuzer*" klassifiziert wurden, passen sowohl nach dem Kriterium *Q-Index < unteres Quartil* zum *Dominanz*-Antwortprozess (**K**), als auch nach dem Kriterium *STRESS-Index < unteres Quartil* zum *Nähe–Distanz*-Antwortprozess (**U**) – insgesamt also das Muster: **mittel K U** (vgl. Tabelle 7.11). Gleichzeitig sind Personen, die in Studie 6.1 in Bezug auf ihre Antworttendenz als „*Extremkreuzer*" klassifiziert werden zu keinem der beiden Antwortprozesse (**N N**) zugeordnet – insgesamt also das Muster: **extrem N N** (vgl. Tabelle 7.11).

Nicht ganz so deutlich fällt dieses Muster für die Analysen der Dimension *Energetic & Rhythmic* des STOMP aus. Die Tabelle 7.12 gibt die Ergebnisse der KFA, über die zwei dichotomen Indikatorvariablen zur Personenklassifikation zu einem der beiden Antwortprozesse und der Indikatorvariable zur Antworttendenz (ERS vs. MRS) nach dem *Dominanz*-Antwortprozess, wieder.

Bei Betrachtung der Tabelle 7.12 zeigt sich, dass, entgegen den entsprechenden Ergebnissen für die Skalen des AIST–R für die Dimension *Energetic & Rhythmic* des STOMP nur das Muster „**mittel K N**" einen signifikanten „Typ" darstellt. Demgegenüber erreicht beim STOMP das Muster „**extrem**

Tabelle 7.11 Signifikante Muster aus der KFA für drei Indikatorvariablen zum impliziten Antwortmodell und zur Antworttendenz für jeweils vier Skalen des AIST–R.

			Realistic				
Muster	*beobachtet*	*erwartet*	*Typ*	χ^2	*df*	p_{χ^2}	$p_{ext.bin.}$
mittel K N	87	123.941	-	11.011	1	0.001	0.000
mittel K U	96	41.314	+	72.387	1	0.000	0.000
extrem N N	449	381.191	+	12.062	1	0.001	0.000
extrem N U	77	127.064	-	19.725	1	0.000	0.000
extrem K N	90	126.559	-	10.561	1	0.001	0.000
extrem K U	61	42.186	+	8.390	1	0.004	0.003
			Investigative				
Muster	*beobachtet*	*erwartet*	*Typ*	χ^2	*df*	p_{χ^2}	$p_{ext.bin.}$
mittel N U	138	173.047	-	7.098	1	0.008	0.002
mittel K N	107	172.359	-	24.784	1	0.000	0.000
mittel K U	166	57.453	+	205.080	1	0.000	0.000
extrem N N	334	235.359	+	41.341	1	0.000	0.000
extrem N U	23	78.453	-	39.196	1	0.000	0.000
extrem K N	53	78.141	-	8.089	1	0.004	0.001
extrem K U	8	26.047	-	12.504	1	0.000	0.000
			Social				
Muster	*beobachtet*	*erwartet*	*Typ*	χ^2	*df*	p_{χ^2}	$p_{ext.bin.}$
mittel N U	105	134.947	-	6.646	1	0.010	0.003
mittel K N	95	134.410	-	11.555	1	0.001	0.000
mittel K U	131	44.803	+	165.833	1	0.000	0.000
extrem N N	443	349.660	+	24.917	1	0.000	0.000
extrem N U	70	116.553	-	18.594	1	0.000	0.000
extrem K N	79	116.090	-	11.850	1	0.001	0.000
			Conventional				
Muster	*beobachtet*	*erwartet*	*Typ*	χ^2	*df*	p_{χ^2}	$p_{ext.bin.}$
mittel N U	104	159.534	-	19.331	1	0.000	0.000
mittel K N	87	158.899	-	32.533	1	0.000	0.000
mittel K U	223	52.966	+	545.844	1	0.000	0.000
extrem N N	458	275.899	+	120.191	1	0.000	0.000
extrem N U	8	91.966	-	76.662	1	0.000	0.000
extrem K N	24	91.601	-	49.889	1	0.000	0.000
extrem K U	0	30.534	-	30.534	1	0.000	0.000

Anmerkungen: Muster (Indikatorvariablen): *mittel – extrem*= Antworttendenz (Studie 6.1), *K*=passend nach kumulativem, *Dominanz*-Antwortprozess, *U*=passend nach Unfolding, *Nähe–Distanz*-Antwortprozess, *N*=nicht passend nach einem der beiden Antwortprozesse; $\alpha_{\text{Bonferroni adj.}} = 0.00625$.

Tabelle 7.12 Muster aus der KFA für drei Indikatorvariablen zum impliziten Antwortmodell und zur Antworttendenz für die Skala *Energetic & Rhythmic* des STOMP.

	Energetic & Rhythmic						
Muster	*beobachtet*	*erwartet*	*Typ*	χ^2	*df*	p_{χ^2}	$p_{ext.bin.}$
mittel N N	143	191.952	-	12.484	1	0.000	0.000
mittel N U	88	68.236	.	5.724	1	0.017	0.010
mittel K N	104	65.520	+	22.599	1	0.000	0.000
mittel K U	14	23.291	.	3.707	1	0.054	0.026
extrem N N	535	514.806	.	0.792	1	0.373	0.131
extrem N U	192	183.006	.	0.442	1	0.506	0.247
extrem K N	166	175.722	.	0.538	1	0.463	0.228
extrem K U	43	62.467	-	6.066	1	0.014	0.005

Anmerkungen: Muster (Indikatorvariablen): *mittel – extrem*= Antworttendenz (nach Skalierung des STOMP), *K*=passend nach kumulativem, *Dominanz*-Antwortprozess, *U*=passend nach Unfolding, *Nähe–Distanz*-Antwortprozess, *N*= nicht passend nach einem der beiden Antwortmodelle; $\alpha_{\text{Bonferroni adj.}} = 0.00625$.

N N“ nicht die Signifikanzgrenze. Allerdings weist auch dieses Muster rein deskriptiv mit 535 beobachteten Fällen eine große absolute Häufigkeit auf. Inhaltlich sind diese beiden Muster nach der „Übersetzung“ der Kodierung der Indikatorvariablen in ähnlicher Weise wie bei den entsprechenden Ergebnissen für die vier untersuchten Skalen des AIST–R zu interpretieren. Demnach passen Personen, die bei der Skalierung der Dimension *Energetic & Rhythmic* nach dem summativen Modell für den *Dominanz*-Antwortprozess als „*Mittelkreuzer*“ klassifiziert wurden nach dem Kriterium *Q-Index* < *unteres Quartil* zum *Dominanz*-Antwortprozess (**K**), können aber nach dem Kriterium *STRESS-Index* < *unteres Quartil* nicht zum *Nähe–Distanz*-Antwortprozess (**N**) zugeordnet werden.

Diskussion

Die vorliegende Untersuchung beantwortet zwei Hauptfragestellungen. Erstens die Frage nach der Konsistenz des impliziten Antwortmodells der Personen bei der Bearbeitung der Skalen der drei Fragebögen AIST–R, BFI–K und STOMP. Zweitens wird die Frage nach einer Verbindung des Klassifikationsergebnisses zu einem der beiden impliziten Antwortmodelle mit dem klassifikatorischen Merkmal der Antworttendenz (MRS vs. ERS) untersucht.

Die Befunde aus der LCA zur konstruktübergreifenden Konsistenz des impliziten Antwortmodells (vgl. Tabelle 7.7) weisen darauf hin, dass es sich bei der Rezeption und Bearbeitung der Skalen nach einem der beiden Antwortmodelle um ein konstrukt- und skalenübergreifendes Merkmal der Personen handelt. So weisen die Ergebnisse aus der LCA der Indikatorvariablen zum impliziten Antwortmodell durchweg auf eine 2-Klassen-Lösung hin (vgl. Tabelle 7.7). Einschränkend muss festgestellt werden, dass die Ergebnisse der nach den drei Konstrukten getrennt durchgeführten Analysen (vgl. Tabelle 7.8) für die Skalen des STOMP – zumindest nach dem Kriterium BIC – auf eine 1-Klassen-Lösung hinweisen, was im Sinne der beiden impliziten Antwortmodelle die Konsistenzhypothese für das Konstrukt *Präferenzen des Musikgeschmacks* nicht stützt. Allerdings weist das Kriterium AIC für die Analysen der Indikatorvariablen zu den STOMP-Skalen konsistent auf eine 2-Klassen-Lösung hin. Insofern mag die hier nach dem Kriterium BIC gefundene 1-Klassen-Lösung auch darauf zurückzuführen sein, dass hier (im Vergleich zu den anderen beiden Konstrukten) lediglich vier Indikatorvariablen in die Analysen mit einfließen und daher eher das Kriterium AIC der angemessenere Koeffizient zur Wahl des am besten passenden LCA-Modells darstellen kann. In diesem Sinne weist auch die konstrukt- und skalenübergreifende LCA (vgl. Tabelle 7.7) in die jeweils 15 Indikatorvariablen mit einfließen auf die 2-Klassen-Lösung hin. Auf der inhaltlichen Ebene kann das skalen- und konstruktübergreifende Merkmal des individuellen impliziten Antwortmodells daher als übergeordnete Personeneigenschaft im Sinne einer *Metaeigenschaft* [*metatrait*] interpretiert werden (vgl. auch Abschnitte 3.2.3 und 4.4.2).

Die Kreuztabellierung der Ergebnisse der LCA mit der ergänzend durchgeführten KFA erweisen sich als geeignet, die beiden identifizierten latenten Klassen in Bezug auf die Konsistenz des impliziten Antwortmodells inhaltlich

zu beschreiben. Die beiden in den Konstrukten (*berufliche Interessenorientierung* – AIST–R und *Persönlichkeit* – BFI–K) identifizierten latenten Klassen (zweiter Ordnung) können durch die signifikanten Muster aus der KFA weitgehend eindeutig jeweils einem der beiden impliziten Antwortmodelle inhaltlich zugeordnet werden. Dieser Befund stützt damit zusätzlich die Hypothese der Konsistenz des impliziten Antwortmodells im Sinne einer übergeordneten Personeneigenschaft bei der Bearbeitung der psychometrischen Skalen in den untersuchten Konstrukten.

Die Ergebnisse aus der KFA über die Indikatorvariablen zur Antworttendenz im STOMP und AIST-R (vgl. Befunde aus Studie 6.1) gemeinsam mit den Indikatorvariablen zum impliziten Antwortmodell (vgl. Befunde aus Studie 7.1) stützen die Hypothese einer systematischen Verbindung zwischen den beiden Phänomenen *Antworttendenz* (ERS vs. MRS) und *implizites Antwortmodell* (mit einem *Dominanz-* vs. *Nähe–Distanz*-Antwortprozess). Für die insgesamt fünf Skalen mit unterschiedlichen Antworttendenzen, nach deren Skalierung mit einem „Dominanz-mixture-Modell“ aus den beiden Konstrukten (*berufliche Interessenorientierung* – AIST–R und *Präferenzen des Musikgeschmacks* – STOMP), ergeben sich signifikante Verbindungen (signifikante Typen) mit der Indikatorvariablen zum impliziten Antwortmodell (vgl. Tabellen 7.11 und 7.12).

Dieser Befund steht im Einklang mit der in Abschnitt 4.7 skizzierten theoretischen Betrachtung einer Konvergenz der ICCs unterschiedlicher (impliziter) Antwortmodelle in einem Bereich eher mittlerer Merkmalsausprägung (vgl. dazu die Abbildung aus Stark et al. (2006, S. 27) in Abschnitt 4.7). Die daraus abgeleitete Folgerung, dass die dichotome Klassifikation der Antworten – Tendenz zur Mitte vs. Tendenz zu Extremen (MRS vs. ERS) – nach deren Skalierung mit einem *Dominanz-mixture-Modell* einen Zusammenhang mit der eindeutigen Zuordnung zu einem der beiden Antwortprozesse aufweisen kann, wird durch die vorliegenden Befunde gestützt. Die spezifisch formulierte Hypothese, dass Personen, deren Antwortmuster nach einem Antwortmodell für den *Dominanz*-Antwortprozess als *„Mittelkreuzer“* (MRS) klassifiziert sind, gleichzeitig sowohl dem *Dominanz*-Antwortprozess als auch dem *Nähe–Distanz*-Antwortprozess (mit eingipfliger Itemcharakteristik) zugeordnet werden können, wird durch die vorliegenden Befunde gestützt. In den Ergebnissen

aus den entsprechenden Konfigurationsfrequenzanalysen (KFA) indizieren die signifikanten Muster jeweils eine Kontingenz zwischen der doppelten Passung der Personen zu beiden impliziten Antwortmodellen und deren Klassifikation als *„Mittelkreuzer“* (MRS). Insofern liefern die Befunde aus dieser Untersuchung auch eine Erklärung für den in Studie 7.1 erzielten Befund der doppelten Zuordnung der Personen zu beiden impliziten Antwortmodellen.

Kapitel 8

Zusammenfassung, Diskussion und Ausblick

8.1 Zusammenfassung der empirischen Untersuchungen

Die vorliegende Dissertation befasst sich mit dem Antwortverhalten von Personen auf Fragebogenverfahren, wie sie in der psychologischen Diagnostik mit unterschiedlichen Zielsetzungen typischerweise eingesetzt werden. Die im empirischen Teil durchgeführten Analysen untersuchen das Antwortverhalten von Personen auf psychometrische Fragebogenskalen in unterschiedlichen Konstrukten, wie sie insbesondere in der Differentiellen Psychologie und Persönlichkeitspsychologie zur Erfassung individueller Merkmale und Einstellungen eingesetzt werden. Stellvertretend für die unterschiedlichsten Konstrukte und Fragebogenskalen, werden hier drei unterschiedliche Typen von Konstrukten mit ihren jeweils verbreiteten Formen der entsprechenden Operationalisierung untersucht.

Das Konstrukt *Persönlichkeit* wird im Rahmen des in der aktuellen Persönlichkeitsforschung vorherrschenden *Eigenschaftsparadigmas* (vgl. Abschnitt 2.1) mit seinen fünf zentralen Dimensionen (Big-Five-Modell) mit dem an der Universität der Bundeswehr (weiter) entwickelten BFI–K erfasst (vgl. Schmolck, 2003, 2004, 2005, 2006a, 2006b). Dieses Instrument ist in weiten Teilen vergleichbar mit dem von Rammstedt und John (2005) publizierten BFI–K, wel-

cher wiederum eine aus dem Englischen übersetze Kurzversion (Rammstedt, 1997) des ursprünglich von John et al. (1991) entwickelten, 44 Items umfassenden *Big-Five-Inventory* (vgl. auch John & Srivastava, 1999) ist. Die in der vorliegenden Arbeit eingesetzte Version des BFI–K wird, wie die meisten anderen Inventare zur Operationalisierung des Big-Five-Modells, mittels faktorenanalytischer Methoden entwickelt, welche ein (kumulatives) Modell für den *Dominanz*-Antwortprozess zur Skalierung implizieren. Wie im Abschnitt 2.1 zum theoretischen Hintergrund des Konstrukts Persönlichkeit dargelegt, stellt das dem Big-Five-Modell zugrunde liegende Eigenschaftsparadigma „nur“ einen möglichen wissenschaftlichen Zugang zum Konstrukt dar, welcher aber im der aktuellen Forschung in der Differentiellen Psychologie und Persönlichkeitspsychologie vorherrschend ist. Einzelne Aspekte des Konstrukts Persönlichkeit, wie beispielsweise das Konzept bzw. die Big-Five-Dimension *Extraversion* mit den beiden antagonistischen Polen *Introversion* und *Extraversion*, haben eine historisch und psychologisch breite konzeptionelle und theoretische Basis, die mit vergleichbaren Konzepten aus anderen Paradigmen der Persönlichkeitsforschung Überschneidungen aufweist – wie zum Beispiel mit der von Jung (1921) formulierten psychoanalytischen Typenlehre (vgl. Abschnitt 2.1.1). Die psycholexikalisch, empirische Fundierung des Eigenschaftsparadigmas (vgl. z. B. Allport & Odbert, 1936), dessen Universalität (vgl. z. B. John et al., 1988; Piedmont & Aycock, 2007; Saucier & Goldberg, 2001), sowie dessen empirische Nützlichkeit im Sinne substanzieller Zusammenhänge mit anderen Merkmalen aus dem Bereich der Differentiellen Psychologie und Korrelaten mit aktuellen biologisch, evolutionären Zugängen zu interindividuellen Unterschieden (vgl. Abschnitt 2.1.2) begründen dessen Bedeutung für den Bereich der aktuellen Persönlichkeitsforschung in der Psychologie.

Das Konstrukt der *beruflichen Interessenorientierungen* nach dem Modell von Holland (1959, 1997) begründet sich demgegenüber eher auf pragmatisch praktische Zielsetzungen aus dem Bereich der Berufsberatung zur Entwicklung diagnostischer Verfahren zur Messung von Interessenprofilen, im Sinne einer Erfassung individueller Neigungen und Präferenzen. So entwickelte Holland (1959, 1963) und Holland et al. (1953) seine Theorie zur Berufswahl zunächst auf der Basis seiner persönlichen Erfahrungen als Berufsberater (vgl. Abschnitt 2.2). Auch wenn beispielsweise Holland (1999) diese beruflichen Interessen-

orientierungen auch als individuelle *Persönlichkeitsorientierungen* angesehen hat, handelt es sich, im direkten Vergleich zum Konstrukt *Persönlichkeit*, dabei eher um bereichsspezifische (beruflicher Kontext) individuelle *Präferenzen*. Zur Erfassung dieser (beruflichen) *Präferenzorientierungen* wird in dieser Arbeit, sowie im deutschsprachigen Raum allgemein, insbesondere im Forschungskontext, der AIST–R von Bergmann und Eder (1999, 2005) eingesetzt. Im Vergleich zu den im BFI–K etablierten Skalen für die fünf Dimensionen des Big-Five-Modells lassen sich die sechs Dimensionen des Holland-Modells im AIST–R dabei eher als *monopolare* Dimensionen individueller Präferenzen interpretieren. So sind auf der inhaltlichen Ebene der Items dementsprechend auch keine negativ formulierten Items (vgl. Abschnitt 3.2.2) in den einzelnen Skalen vorhanden. Konzeptionell kann daher das untere Ende der Merkmalsausprägung der jeweils zehn Items umfassenden Skalen des AIST–R als („natürlicher“) Nullpunkt aufgefasst werden, welcher auf der inhaltlichen Ebene als Fehlen einer Präferenz für die entsprechende Merkmalsdimension interpretiert werden kann – und nicht etwa inhaltlich als „Gegenteil“ einer entsprechend hohen positiven Merkmalsausprägung auf einem bipolaren Merkmalskontinuum. Demgegenüber lassen sich für die einzelnen Skalen des BFI–K (z. B. die Skalen *Neurotizismus* und insbesondere *Extraversion*) im Sinne bipolarer Merkmalsdimensionen auch inhaltlich als antagonistisch geprägte Endpunkte formulieren – z. B. für die Skala *Neurotizismus*: Ängstlichkeit vs. emotionale Stabilität (vgl. Abschnitte 3.2.2 und 2.1.3).

Das dritte in der vorliegenden Arbeit untersuchte Konstrukt – *Präferenzen des Musikgeschmacks* – umfasst in der hier zugrunde gelegten Operationalisierung über die für den deutschen Sparachraum (vgl. Langmeyer, Guglhör-Rudan & Tarnai, 2012) adaptierte Version des STOMP (Rentfrow & Gosling, 2003), analog zur Messintention des AIST–R, ebenso individuelle *Präferenzorientierungen*. Wie beim AIST–R sind hier auf der Ebene der Items ebenfalls keine negativ formulierten Items in den einzelnen Skalen enthalten. Die *Antwortskalen* der Items vom AIST–R und STOMP weisen aber eine um einen neutralen Mittelpunkt symmetrische, Anordnung der Antwortskalen auf, welche von Ablehnung (maximale Distanz) bis Zustimmung (maximale Nähe) reichen. Beim AIST–R ist diese Antwortskala (wie beim BFI–K) fünfstufig und beim STOMP ist sie siebenstufig (vgl. Abschnitt 5.1).

Die Modellierung des Antwortverhaltens im Rahmen der Skalierung derartiger psychometrischer Skalen baut auf zwei unterschiedlichen (impliziten) Antwortmodellen der Personen mit unterschiedlichen *Antwortprozessen* auf. Diese beiden Antwortprozesse korrespondieren mit zwei unterschiedlichen Prinzipien zur Skalierung von Fragebogendaten, wie sie in Abschnitt 1.3 *Skalierung von Fragebogendaten* einleitend dargestellt sind. Aus diesen beiden Skalierungsprinzipien lassen sich jeweils *psychometrische Antwortmodelle* ableiten, welche eine Grundlage zur Beschreibung, Erklärung und Vorhersage des Antwortverhaltens bilden (vgl. Abschnitt 4.1 *Modellbildung zum Antwortverhalten*). Die beiden Antwortprozesse unterscheiden sich dabei vor allem durch die unterschiedliche Beziehung oder *Relation*, die zwischen den Personen (mit einer bestimmten Merkmalsausprägung) und den Items (die eine bestimmte Merkmalsausprägung repräsentieren) besteht. Diese Beziehung kann entweder durch eine *Nähe–Distanz*-Relation oder durch eine *Dominanz*-Relation charakterisiert werden.

Insbesondere die theoretische Fundierung der beiden mit dem AIST–R und STOMP erfassten Konstrukte als auch deren konkrete Operationalisierung können dahingehend interpretiert werden, dass zur Beschreibung des Antwortverhaltens ein psychometrisches Modell zur Modellierung einer *Nähe–Distanz*-Relation (vgl. Abschnitt 4.3) zwischen Personen und Items angemessen sein kann (vgl. z. B. Hubert & Arabie, 1987, für eine Anwendung beim Holland Modell der beruflichen Interessen). Allerdings sind beide Inventare, ebenso wie der BFI–K, mit faktorenanalytischen Modellen entwickelt worden (vgl. Bergmann & Eder, 2005; Rentfrow & Gosling, 2003), welche ein psychometrisches Modell für den *Dominanz*-Antwortprozess (vgl. Abschnitt 4.2) implizieren. Folglich stützt sich auch die empfohlene Auswertung der einzelnen Skalen der drei Inventare, z. B. im Rahmen individualdiagnostischer Fragestellung, auf die Summierung der einzelnen *Itemscores*.

Neben diesen beiden unterschiedlichen Antwortprozessen können individuell unterschiedliche Antworttendenzen oder idiosynkratische Antwortmuster in Fragebogendaten bestehen. Die systematische Durchsicht der Literatur (vgl. Kapitel 3 *Theoretischer Hintergrund zu Antwortmustern*) zeigt, dass sich der überwiegende Anteil an theoretischen Modellen und empirischen Arbeiten bei der Untersuchung von idiosynkratischen Antwortmustern oder Antwort-

tendenzen auf Skalierungsmodelle zu einem *Dominanz*-Antwortprozess stützen (vgl. zusammenfassend Abschnitt 3.3 in dieser Arbeit; sowie Dalal und Carter (2015a) für eine zusammenfassende Übersicht). Lediglich einige theoretische und empirische Arbeiten eher neueren Datums (z. B. Drasgow et al., 2010a, 2010b; Huang & Mead, 2014; O'Brien & LaHuis, 2011; Spector & Brannick, 2010; Tay et al., 2009) beziehen bei der methodischen Diskussion und empirischen Betrachtung von Modellabweichungen auch (unterschiedliche) Modelle zum *Nähe–Distanz*-Antwortprozess mit ein – (vgl. aber auch van Schuur, 2006; van Schuur & Kiers, 1994, sowie Matschinger & Krebs, 1998). Allerdings wird bei entsprechenden Arbeiten dann meist die *gesamte psychometrische Skala* über die *gesamte Stichprobe* (vergleichend) mit einem Unfoldingmodell zur Abbildung des Nähe–Distanz-Antwortprozesses skaliert. Oder aber die Analysen werden dahingehend ausgeweitet, dass einzelne Items betrachtet werden, welche eher dem Nähe–Distanz-Antwortprozess (anstatt einem Dominanz-Antwortprozess) entsprechen.

In der vorliegenden Arbeit werden die einzelnen Skalen der drei Inventare (AIST–R, BFI–K und STOMP) zu den drei Konstrukten (*Persönlichkeit, berufliche Interessenorientierungen und Präferenzen des Musikgeschmacks*) in den entsprechenden Analysen zum Antwortverhalten (vgl. Abschnitte 6.1, 6.2 und 6.3 in Kapitel 6) daher zunächst mit entsprechenden Modellen für den *Dominanz*-Antwortprozess aus der Item-Response-Theory (IRT – vgl. Kapitel 4 *Psychometrische Modellierung*, Abschnitt 4.2) skaliert und untersucht. Angewendet wird zunächst das *Partial Credit Model* (Masters, 1982), um die eindimensionale Skalierbarkeit der Skalen zu überprüfen. Dieses Vorgehen begründet sich einerseits aus der *praktischen Perspektive* auf die in den Testanweisungen für die Auswertung des jeweiligen Inventars gegebenen Hinweise zur summativen Verrechnung der einzelnen Itemscores. Andererseits begründet sich dieses Vorgehen aus einer *theoretisch, methodischen Perspektive* aus dem faktorenanalytisch fundierten Konstruktionsprinzip, dass allen drei Inventaren zugrunde liegt. Vergleichend und ergänzend werden darüber hinaus Analysen mit dem mixed-Rasch-Modell (vgl. Rost, 1990, 1991) vorgenommen, um zu überprüfen, ob hinsichtlich der Antwortmuster und Tendenzen unterschiedliche Personenklassen innerhalb der potentiell heterogenen Stichprobe identifizierbar sind. Diese (insofern auch personenzentrierte) Form der Daten-

analyse stützt sich, wie die eindimensionale Skalierung, auf ein kumulatives Skalierungsmodell für die einzelnen Dimensionen der Konstrukte, welches einen *Dominanz*-Antwortprozess abbildet.

Die Anwendung von Mischverteilungsmodellen wie das mixed-Rasch-Modell oder auch die Latent-Class-Analysis (LCA) zur Skalierung von Fragebogendaten begründet sich durch Befunde aus zahlreichen empirischen Untersuchungen, wonach immer wieder Personen(-Gruppen) mit abweichenden Antwortmustern bei der Beantwortung von Fragebogen zur psychologischen Diagnostik gefunden werden (vgl. Kapitel 3). So zeigen sich bei der Skalierung von Fragebogenskalen mit Modellen für einen Dominanz-Antwortprozess und deren anschließender Modelltestung immer wieder unterschiedliche Formen lokaler Modellabweichungen. Das Forschungsfeld zu solchen Modellabweichungen hat, wie in den einzelnen Abschnitten in Kapitel 3 gezeigt, eine recht lange Historie und entsprechend verläuft die Debatte, zur Systematisierung der Abweichung, deren Ursachen und den dahinterliegenden Modellen zum Antwortprozess sowie den Auswirkungen auf die psychometrische Messung, eher kontrovers. Für einige solcher Antwortverzerrungen (z. B. *Akquieszenz* und *Impression Management*, vgl. Abschnitt 3.2.1), welche eher mit den Inhalten der einzelnen Items einer psychometrischen Skala assoziiert werden, lassen sich Bezüge zu den in der Sozialpsychologie angesiedelten Theorien zum interpersonalen Selbstkonzept, der sozialen Identität sowie der zwischenmenschlichen Kommunikation und Beziehungen allgemein herstellen (vgl. Abschnitt 3.2.1). Andererseits bestehen Antwortverzerrungen (wie beispielsweise *Extreme Response Style – ERS*), die eher mit den besonderen Eigenschaften der Personen und deren individuell unterschiedlicher Wahrnehmung der vorgegebenen Antwortskala der Items in Beziehung gesetzt werden (vgl. Abschnitt 3.2.4).

In Kapitel 6 *Untersuchungen zum Dominanz-Antwortprozess* werden, basierend auf der Annahme psychometrischer Modelle zur Abbildung eines Dominanz-Antwortprozesses, drei Untersuchungen zu Antworttendenzen und idiosynkratischem Antwortverhalten durchgeführt. Studie 6.1 *Extreme und mittlere Antworttendenz im Bereich beruflicher Interessenorientierungen und Musikgeschmack* widmet sich übergreifend der Frage nach der universellen Gültigkeit eines eindimensionalen Skalierungsmodells (für einen Dominanz-Antwortprozess) für die Konstrukte berufliche Interessenorientierungen und Präferenzen des Musikge-

schmacks und untersucht auf der Ebene der sechs Skalen des AIST–R und der vier Skalen des STOMP die Personenhomogenität der vorliegenden Stichproben mit entsprechenden Modellen aus der IRT. Zur Skalierung wird einerseits das *Partial Credit Model* (Masters, 1982) als eindimensionales Modell für mehrstufige Antwortformate sowie das mixed-Rasch-Modell (vgl. Rost, 1990, 1991) als dessen Erweiterung zur Entdeckung einer, hinsichtlich der geschätzten Modellparameter, heterogenen Personenstichprobe eingesetzt. Der konzeptionell gleiche Ansatz wird in Studie 6.2 *Untersuchung zur Skalierbarkeit des BFI–K* im Sinne einer Replikation von bestehenden Befunden aus der Literatur für das Konstrukt Persönlichkeit nach dem Big-Five-Modell im Rahmen des Eigenschaftsparadigmas verfolgt. Aufbauend auf die Befunde aus den beiden Studien 6.1 *Extreme und mittlere Antworttendenz im Bereich beruflicher Interessenorientierungen und Musikgeschmack* und 6.2 *Untersuchung zur Skalierbarkeit des BFI–K*, werden in Studie 6.3 *Auswirkungen von Antworttendenzen auf empirische Befunde zum Zusammenhang zwischen Dimensionen der Persönlichkeit und beruflichen Interessenorientierungen* die empirische Bedeutung der beiden identifizierten Antworttendenzen untersucht. Zunächst werden die Auswirkungen der Antworttendenzen auf die hexagonale Struktur des Modells der beruflichen Interessenorientierung nach Holland (1997) untersucht. Darüber hinaus wird der Frage nachgegangen, ob die beiden unterschiedlichen Antworttendenzen, auf der Ebene der einzelnen Skalen beider Konstrukte, differentielle Effekte auf die empirischen Zusammenhänge verursachen (vgl. Studie 6.3 *Auswirkungen von Antworttendenzen auf empirische Befunde zum Zusammenhang zwischen Dimensionen der Persönlichkeit und beruflichen Interessenorientierungen*).

In Kapitel 7 *Untersuchungen zum Nähe–Distanz-Antwortprozess* werden Analyseansätze verfolgt, die, neben dem Dominanz-Antwortprozess, auch einen Nähe–Distanz-Antwortprozess abbilden. Aus einer eher personenzentrierten Perspektive wird der Frage nachgegangen, ob im Hinblick auf die beiden Antwortprozesse (*Dominanz* vs. *Nähe–Distanz*) aufgrund einer entsprechenden Rezeption der Items in den einzelnen Skalen unterschiedliche Personengruppen identifizierbar sind (vgl. Studie 7.1 *Seriation und Multidimensionale Skalierung zur Klassifikation der Personenstichprobe nach impliziten Antwortmodellen*).

In Studie 7.1 wird dabei untersucht, inwieweit sich innerhalb einer Stichpro-

be Personen oder Personengruppen identifizieren lassen, welche sich in ihrer Rezeption der Items in den einzelnen Skalen im Hinblick auf ihre impliziten Antwortmodelle mit unterschiedlichen Antwortprozessen unterscheiden. Es wird der Frage nachgegangen (vgl. Studie7.1), inwieweit bezüglich unterschiedlicher Antwortprozesse auch unterschiedliche Personengruppen innerhalb der analysierten Daten vorliegen. Die Skalen der drei Konstrukte werden dazu einerseits mittels der Multidimensionalen Skalierung zur Abbildung des Nähe–Distanz-Antwortprozesses und andererseits mit einem eindimensionalen Modell zur Abbildung des Dominanz-Antwortprozesses skaliert. Über Indizes für lokale Modellabweichungen zum jeweiligen psychometrischen Antwortmodell werden dabei die jeweils zum Antwortmodell passenden Personen identifiziert.

Einige Autoren (z. B. Drasgow et al., 2010b; Klinkenberg, 2001; Stark et al., 2006) weisen darauf hin, dass die eindeutige Trennung von Modellen für einerseits Dominanz- und andererseits Nähe–Distanz-Antwortprozesse durch die Polarität, aber insbesondere durch die Extremität der Itemformulierungen und deren daraus resultierenden Schwierigkeiten in empirischen Datensätzen problematisch sein kann. In Abschnitt 4.7 in Kapitel 4 *Psychometrische Modellierung* wird gezeigt, dass sich im Bereich einer mittleren Merkmalsausprägung der Personen der Verlauf der Funktion der Zustimmungswahrscheinlichkeit von Modellen mit eingipfliger ICC derjenigen von Modellen mit monoton steigender ICC für diejenigen Items angleicht, welche eher eine extreme Merkmalsausprägung repräsentieren. Die Extremität eines Urteils (nach einem Dominanz-Antwortprozess) bzw. die Zustimmung zu einem Item, welches eine entsprechend hohe Merkmalsausprägung repräsentiert, kann daher assoziiert sein mit der Unfoldingstruktur in den analysierten Daten. Basierend auf diesen theoretischen Überlegungen wird in Studie 7.2 die Konsistenz und der Zusammenhang zwischen der Zuordnung zu einem der beiden Antwortprozesse (Dominanz- vs. Nähe–Distanz-Antwortprozess) und der Antworttendenz (mittel vs. extrem) nach der Skalierung mit einem mixture-Modell für den Dominanz-Antwortprozess (Rost, 1990, 1991) untersucht.

8.2 Diskussion der empirischen Befunde

In Studie 6.1 *Extreme und mittlere Antworttendenz im Bereich beruflicher Interessenorientierungen und Musikgeschmack* können bei vier Interessenskalen (R,I,S,C; R–*Realistic*, I–*Investigative*, S–*Social*, C–*Conventional*) des AIST–R unterschiedliche Antworttendenzen zu entweder mittleren oder extremen Antwortkategorien (ERS vs. MRS) nachgewiesen werden. Mit wenigen Einschränkungen kann die Konsistenz dieser Antworttendenz der Personen für die vier Interessenskalen (R,I,S,C) über eine 3-Klassen-Lösung für die LCA zweiter Ordnung belegt werden. Die Kreuztabellierung des Klassifikationsergebnisses aus der LCA mit den 16 (2^4) möglichen Mustern (pattern) der Indikatorvariablen für die vier Dimensionen das AIST–R (R,I,S,C) qualifiziert die drei latenten Klassen zweiter Ordnung als Personenklassen mit konsistent eher mittlerer oder extremer Antworttendenz (vgl. Tabellen 6.4 und 6.5). Dies erscheint aus zweierlei Gründen bemerkenswert. Einerseits weil es sich um unterschiedliche und im hexagonalen Modell von Holland (1997) teils entgegengesetzte Interessenorientierungen handelt und andererseits, weil die Skalen des AIST–R als jeweils unipolare Merkmalsdimensionen zur Erfassung von Präferenzorientierungen angelegt sind. Im Gegensatz zu vielen anderen psychometrischen Skalen, wie beispielsweise diejenigen zur Erfassung von Dimensionen der Persönlichkeit, werden dabei lediglich inhaltlich positiv formulierte Items mit mehr oder weniger neutral formulierten Verhaltens- bzw. Handlungsalternativen für die Operationalisierung eingesetzt. Einerseits kann damit geschlussfolgert werden, dass sich das Phänomen einer unterschiedlichen Antworttendenz bei einem psychometrischen Modell zur Abbildung des Dominanz-Antwortprozesses (ERS vs. MRS), zumindest bei den hier untersuchten Stichproben eher auf die übergeordnete, individuell unterschiedliche Interpretation der vorgegebenen Antwortskala, und nicht etwa auf die Polarität oder die Inhalte der einzelnen Items zurück zu führen ist. Andererseits können die Ergebnisse aus der Studie 6.1 auch dahingehend interpretiert werden, dass die unterschiedlichen Inhalte der Items und Skalen die Unterscheidbarkeit der Antworttendenzen beeinflussen, da die beiden Antworttendenzen ERS und MRS nicht für alle sechs Skalen des AIST–R nachgewiesen werden können. Zur inhaltlichen Interpretation der gefundenen Antworttendenzen auf vier der sechs AIST–R-Skalen

bieten sich unterschiedliche Erklärungsmodelle an. Die (konsistente) Tendenz zu eher extremen Antwortkategorien (ERS) kann mit Konzepten der sozialen Erwünschtheit oder vorgetäuschtem Antwortverhalten (vgl. Abschnitt 3.2.1) erklärt werden, oder aber auch auf ein (stereotypes) unaufmerksames Antwortverhalten (vgl. Abschnitt 3.2.3) zurückgeführt werden. Die Tendenz zu eher mittleren Antwortkategorien (MRS) kann einerseits durch ein vorsichtiges Antwortverhalten (Hurley, 1998) erklärt werden, wobei die von den Items der Skalen angesprochenen Inhalte von bestimmten Personen als mehr oder weniger relevant eingestuft werden (Warr & Coffman, 1970). Andererseits tendieren Personen dazu, Fragen zu ihnen eher unbekannten Inhalten, im Sinne einer kognitiven Erleichterung (vgl. Krosnick, 1991) eher mit der „neutralen“ Mittelkategorie zu beantworten. Vor dem Hintergrund der analysierten Stichprobe mit Studierenden vornehmlich sozialwissenschaftlicher Fächer an der Universität der Bundeswehr, sowie der Inhalte der *nicht* vom Phänomen ERS und MRS betroffenen Skalen (*Artistic*: „Künstlerische Orientierung“ und *Enterprising*: „Unternehmerische Orientierung“), erscheint für die unterschiedlichen Antworttendenzen auf den anderen Skalen (R,I,S,C) die unterschiedliche Relevanz der Skaleninhalte (Hurley, 1998; Warr & Coffman, 1970, z. B.) in Verbindung mit Effekten der sozialen Erwünschtheit und vorgetäuschtem Antwortverhalten (vgl. Abschnitt 3.2.1) als Erklärung naheliegend zu sein.

In Studie 6.2 *Untersuchung zur Skalierbarkeit des BFI–K* werden die einzelnen Skalen des BFI–K zu fünf Dimensionen der Persönlichkeit im Hinblick auf deren Skalierbarkeit nach einem eindimensionalen, summativen Skalierungsmodell zur Abbildung eines *Dominanz*-Antwortprozesses untersucht. Neben dem übergreifend positiven Befund der eindimensionalen Skalierbarkeit weisen die Befunde aus Studie 6.2 auf eine methodische Problematik bei der Überprüfung einer personenhomogenen Skalierbarkeit von psychometrischen Skalen mittels mixed-Rasch-Modellen hin. So ergeben sich bei der Skalierung der Dimension Extraversion für die Stichprobe II zunächst zwei latente Personenklassen, welche auf eine Personenheterogenität hinweisen. Die genauere Untersuchung dieser beiden latenten Klassen auf der Ebene der jeweils vorherrschenden Antwortmuster indiziert aber, dass die scheinbare Personenheterogenität auch auf ein methodisches Problem der Parameterschätzung (mit der CML-Methode) vor dem Hintergrund geringer Antwortkategoriehäufigkeiten

zurückgeführt werden kann. Der ergänzende Befund aus der 2-Gruppen-KFA, dass die in einer der beiden latenten Klassen identifizierten, diskriminierenden Antwortmuster (Typen) in der jeweils andern Klasse Häufigkeiten von null aufweisen, legt nahe, dass das an sich probabilistische formulierte Modell latenter Klassen im mixed-Rasch-Modells hier deterministisch zwei (artifizielle) Klassen produziert. Diese Interpretation der Befunde wird zusätzlich dadurch unterstützt, dass die ergänzende Skalierung der fünf Dimensionen des BFI–K, basierend auf der Gesamtstichprobe, mit der gegenüber geringen Antwortkategoriehäufigkeiten robusten *PAIR*-Methode die Eindimensionalität der Skalen des BFI–K bestätigt.

Die Untersuchungen in Studie 6.3 *Auswirkungen von Antworttendenzen auf empirische Befunde zum Zusammenhang zwischen Dimensionen der Persönlichkeit und beruflichen Interessenorientierungen* widmen sich der übergeordneten Frage nach den Auswirkungen unterschiedlich ausgeprägter Antworttendenzen im Rahmen der Annahme eines summativen Skalierungsmodells zur Abbildung eines Dominanz-Antwortprozesses. Diese übergeordnete Fragestellung wird spezifisch einerseits auf empirisch gefundene Zusammenhänge zwischen Konstrukten und deren einzelnen Dimensionen bezogen und wird andererseits auf die empirische Bestätigung von konstruktspezifischen, theoretischen Annahmen, bezogen. Die Ergebnisse der Analysen zu den Zusammenhängen einzelner Dimension in den beiden Konstrukten *Persönlichkeit* und *berufliche Interessenorientierungen* deuten darauf hin, dass die unterschiedlichen Antworttendenzen (vgl. Studie 6.1) die Interkorrelationen zwischen den Skalen in nicht unerheblichem Ausmaß beeinflussen. Als übergreifendes Gesamtfazit der in der Diskussion von Studie 6.3 detailliert behandelten Befunde lässt sich feststellen, dass unterschiedliche Antworttendenzen dazu führen, dass die korrelativen Zusammenhänge zwischen den einzelnen Dimensionen der beiden Konstrukte entweder verstärkt, abgeschwächt oder sogar in der empirischen Befundlage gänzlich maskiert werden. In vergleichbarer Weise wirken sich die unterschiedlichen Antworttendenzen auch auf die empirisch zu überprüfende Gültigkeit der im Modell beruflicher Interessenorientierungen von Holland (1997) formulierten *Calculus Hypothese* (vgl. Abschnitt 2.2) aus. So indizieren die Befunde aus Studie 6.3 eine je nach Antworttendenz unterschiedlich ausgeprägte Passung der empirischen Daten an die circumplexe Struktur.

Die anzupassende circumplexe Struktur kann dabei einerseits durch die in der Calculus Hypothese theoretisch formulierten Beziehungen zwischen den sechs Dimensionen der beruflichen Interessenorientierungen vorgegeben werden. Andererseits kann sich die Überprüfung der Passung der empirischen Daten an der durch die Normstichprobe des AIST–R vorgegebenen circumplexen Struktur orientieren. Es zeigt sich in den Ergebnissen in Studie 6.3, dass die Anpassung der empirischen Daten an die durch die Normstichprobe vorgegebene circumplexe Struktur für die Teilgruppe der Personen mit mittlerer Antworttendenz am besten ausfällt. Die Anpassung der Daten an die theoretische, aus der Calculus Hypothese abgeleitete, ideale circumplexe Struktur fällt dagegen für die Teilgruppe der Personen mit extremer Antworttendenz am besten aus (vgl. Tabelle 6.11 in Studie 6.3).

Die Analysen zur Skalierung in Studie 7.1 *Seriation und Multidimensionale Skalierung zur Klassifikation der Personenstichprobe nach impliziten Antwortmodellen* überprüfen aus einer personenzentrierten Perspektive die universelle Gültigkeit des für die Skalen der drei Konstrukte angenommenen, summativen Skalierungsmodells, welches mit einem impliziten Antwortmodell zur Abbildung eines *Dominanz*-Antwortprozesses korrespondiert. Spezifisch wird untersucht, ob sich in den hier vorliegenden Fragebogendaten hinsichtlich der theoretisch denkbaren beiden alternativen Antwortmodelle (vgl. Abschnitt 1.3) unterscheidbare Personengruppen finden lassen; Personen oder Personengruppen also, die auf *dieselben* Items der jeweiligen psychometrischen Skala implizit entweder nach einem *Dominanz*-Antwortprozess oder nach einem *Nähe–Distanz*-Antwortprozess antworten. Zur Skalierung der Daten werden jeweils robuste Verfahren eingesetzt, die dazu geeignet sind, den jeweiligen Antwortprozess (*Dominanz* vs. *Nähe–Distanz*) abzubilden, ohne dabei strenge Annahmen bezüglich der Gültigkeit eines restriktiven psychometrischen Antwortmodells zu machen. Übergreifend zeigen die Befunde aus Studie 7.1, dass sich innerhalb der untersuchten Stichprobe für alle drei Konstrukte und deren Dimensionen Personen mit unterschiedlichen impliziten Antwortmodellen – einerseits nach einem Modell für den *Dominanz*-Antwortprozess und andererseits nach einem Modell für den *Nähe–Distanz*-Antwortprozess – nachweisen lassen. Allerdings weisen die Ergebnisse der Personenklassifikation nach den beiden Antwortprozessen auch darauf hin, dass sich ein nicht unerheblicher Anteil von Personen

in den Daten beiden Antwortprozessen zuordnen lässt. Diese Uneindeutigkeit in der Klassifikation ist zwar zunächst unbefriedigend, erscheint aber aus zweierlei Gründen nachvollziehbar. Einerseits werden bei den Analysen zwei jeweils allgemein formulierte, wenig restriktive psychometrische Antwortmodelle zur Modellierung der beiden Antwortprozesse angewendet. Zweitens besteht die zunächst über theoretische Betrachtungen gestützte Erklärung, dass sich bei Items, welche eine extreme Merkmalsausprägung repräsentieren die Kurvenverläufe der ICCs für den Dominanz- und den Nähe–Distanz-Antwortprozess im Bereich einer *mittleren* Merkmalsausprägung der Personen auf dem latenten Kontinuum weitgehend angleichen, was zu der Uneindeutigkeit in der Klassifikation führt.

Die Untersuchungen in Studie 7.2 *Konsistenz impliziter Antwortmodelle und Zusammenhänge mit Antworttendenzen* verfolgen zwei Hauptfragestellungen. Einerseits soll analysiert werden, ob das Antwortverhalten nach einem der beiden in Studie 7.1 identifizierten Antwortprozesse als eine konsistente, skalen- und konstruktübergreifende Reaktionstendenz der Personen aufgefasst werden kann. Die Konsistenzhypothese zu dieser Reaktionstendenz wird über eine Latent-Class-Analysis (LCA, zweiter Ordnung) und die Konfigurationsfrequenzanalyse (KFA) überprüft. Analysiert werden die Indikatorvariablen aus der Personenklassifikation in Studie 7.1. Andererseits soll analysiert werden, ob ein Zusammenhang zwischen der Antworttendenz zu entweder mittleren oder extremen Antwortkategorien und der Klassifikation der Personen nach den beiden Antwortprozessen, besteht. Mit einer KFA werden dazu signifikante Reaktionsmuster (Typen) auf den Indikatorvariablen identifiziert, die sich als Ergebnis der Personenklassifikation in den Studien 6.1 und 7.1, ergeben haben. Die Ergebnisse aus Studie 7.2 deuten darauf hin, dass die Beantwortung der Skalen aus den drei Fragebögen AIST–R, BFI–K und STOMP durch die Personen in konsistenter Weise nach einem der beiden Antwortmodelle erfolgt. Die konsistente Rezeption und Bearbeitung der einzelnen Fragen in den Skalen, entweder nach dem Dominanz- oder nach dem Nähe–Distanz-Prinzip, kann daher als übergeordnete Personeneigenschaft im Sinne einer *Metaeigenschaft* [*metatrait*] interpretiert werden (vgl. Abschnitte 3.2.3,3.3 und 4.4.2). Ferner zeigen die Befunde aus Studie 7.2, dass ein typischer Zusammenhang zwischen dem Klassifikationsergebnis zur Antworttendenz und dem Klassifika-

tionsergebnis zum impliziten Antwortmodell besteht. Dieser systematische Zusammenhang zwischen den beiden Arten der Personenklassifikation wirft einen neuen Aspekt auf das in der Literatur (vgl. Kapitel 3 und Abschnitt 3.2.4) immer wieder berichtete Phänomen unterschiedlicher Antworttendenzen (ERS. vs. MRS) nach einem summativen Skalierungsmodell. Nach den Befunden aus Studie 7.2 geht die Tendenz zu eher mittleren Antwortkategorien (MRS) mit einer uneindeutigen Klassifikation der Personen sowohl zu einem Dominanz-Antwortprozess als auch zu einem Nähe–Distanz-Antwortprozess einher. Die Befunde bestätigen insofern auch eine beispielsweise von Stark et al. (2006) angestellte theoretische Betrachtung und Hypothese (vgl. auch Abschnitt 4.7), nach der aufgrund konvergierender Itemcharakteristiken (die ICCs der Items) im Bereich einer eher mittleren Merkmalsausprägung der Personen die Unterscheidbarkeit der beiden Antwortprozesse schwierig und uneindeutig sein kann. Ferner bietet diese hier empirisch gestützte theoretische Betrachtung eine schlüssige Erklärung für die an den Ergebnissen aus Studie 7.1 zunächst zu bemängelnde, uneindeutige, doppelte Zuordnung der Personen zu *beiden* impliziten Antwortmodellen mit unterschiedlichen Antwortprozessen.

8.3 Abschließende Betrachtungen und Ausblick

Die Analysen aus den Studien 6.1, 6.2 und 6.3 in Kapitel 6 *Untersuchungen zum Dominanz-Antwortprozess* orientieren sich, mit unterschiedlichen Fragestellungen und Schwerpunktsetzungen, an dem in der Praxis der psychologischen Diagnostik üblichen und verbreiteten Skalierungsprinzip einer summativen Indexbildung. Übergeordnet betrachtet, unterstreichen die Befunde aus den einzelnen Studien und Analysen in Kapitel 6 die Relevanz individuell unterschiedlich ausgeprägter Antworttendenzen, welche einen, die empirischen Befunde verzerrenden, Einfluss auf die psychometrische Messung nehmen können.

Demgegenüber befassen sich die Studien 7.1 und 7.2 zum Antwortverhalten in Kapitel 7 *Untersuchungen zum Nähe–Distanz-Antwortprozess* mit einem (alternativen) Skalierungsprinzip zur Indexbildung gemäß einem Antwortmodell das einen *Nähe–Distanz*-Antwortprozess abbildet. Die Ergebnisse der Klassifikation nach beiden (impliziten) Antwortmodellen stützen dabei die Vermutung, dass bei der Bearbeitung der Skalen in den drei Konstrukten von unterschiedlichen Personengruppen jeweils eines der beiden Antwortmodelle (implizit) zugrunde gelegt wird. Dieser Befund indiziert entsprechend die Anwendung unterschiedlicher Skalierungsprinzipien zur Indexbildung für die jeweilige Personengruppe gemäß deren implizitem Antwortmodell. Insbesondere die Befunde aus Studie 7.1 *Seriation und Multidimensionale Skalierung zur Klassifikation der Personenstichprobe nach impliziten Antwortmodellen* werfen damit die Frage auf, inwieweit die Skalierungsergebnisse nach den unterschiedlichen Modellen auf der Ebene der Personen miteinander in Bezug gesetzt werden können. Es stellt sich die Frage der quantitativen Vergleichbarkeit der Personen anhand der durch die jeweils unterschiedliche Skalierung erzielte Rangreihung der Personen auf der gemessenen Merkmalsausprägung.

Ein solcher einfacher, direkter Vergleich der zwei Personengruppen mit den beiden unterschiedlichen Antwortmodellen dürfte sich insofern als problematisch erweisen, da die zwei bei der Skalierung erzielten Rangreihen der Personengruppen zunächst als jeweils unabhängige Verteilungen aufgefasst werden müssen. So muss eine Person, die beispielsweise in der Mitte der Verteilung derjenigen Personen liegt, welche nach einem *Dominanz-Antwortprozess*

antworten nicht notwendigerweise dasselbe Ausmaß der Merkmalsausprägung aufweisen wie eine Person, die ebenfalls in der Mitte der Verteilung ihrer Personengruppe liegt, welche aber nach einem *Nähe–Distanz-Antwortprozess* antworten. In der vorliegenden Arbeit wird daher zunächst das grundsätzliche Phänomen unterschiedlicher impliziter Antwortmodelle anhand unterschiedlicher Konstrukte und deren Dimensionen dargestellt (Studie 7.1) und dessen Zusammenhang mit dem bereits auch aus anderen Arbeiten bekannten Phänomen unterschiedlicher Antworttendenzen (*mittlere* vs. *extreme* Antworttendenz) in Studie 7.2 untersucht.

Die Frage nach einer Möglichkeit zum Vergleich der Skalierungsergebnisse nach den beiden unterschiedlichen Antwortprozessen muss in dieser Arbeit allerdings noch offenbleiben. Ein mögliches Vorgehen zum Vergleich der Personenrangreihung aus unterschiedlichen Skalierungsverfahren soll aber im Rahmen dieses Ausblicks hier kurz skizziert werden. So könnte die Vergleichbarkeit der beiden Personengruppen mit unterschiedlichen impliziten Antwortmodellen (unter bestimmten axiomatischen Annahmen) beispielsweise über (verteilungsbasierte) *Linking*-Prozeduren zum *Test-Equating* erreicht werden (vgl. z. B. Kolen & Brennan, 2014). Ein vergleichsweise einfaches Linking-Verfahren stützt sich unter der Annahme einer Normalverteilung der jeweiligen Merkmalsausprägung in den beiden zu vergleichenden Teilstichproben (bei ausreichend großen und für die Population repräsentativen Teilstichproben) auf das Prinzip der Gleichsetzung von einzelnen Abschnitten (z. B. Perzentile) der empirischen Merkmalsverteilungen. Wie bereits weiter oben schon angedeutet, muss diese Annahme einer Äquivalenz einzelner Perzentilabschnitte in den Verteilungen der beiden Personengruppen nicht unbedingt erfüllt sein, was die Anwendung dieses Linking-Prinzips mindestens fragwürdig erscheinen lässt.

Eine weitere Möglichkeit des Linkings der beiden Personengruppen mit unterschiedlichen Antwortmodellen könnte sich auf die Betrachtung der Lösungs- oder Itemkategoriewahrscheinlichkeiten zu einzelnen Items und der daraus abzuleitenden relativen Rangreihung der Itemparameter der eingesetzten psychometrischen Skalen stützen. Die a priori zu treffende Annahme müsste bei einem solchen Vorgehen darin bestehen, dass sich die relativen Rangreihen der *Items* und damit deren Itemparameter, basierend auf den Itemkategoriewahrscheinlichkeiten, invariant gegenüber den beiden unterschiedlichen Antwortmodellen

der Personen erweisen. Unter dieser Voraussetzung ergeben sich dann lediglich für die vergleichende Bestimmung der Personenparameter (der individuellen Ausprägung auf dem zu messenden Merkmal) für die beiden Personengruppen unterschiedliche Vorgehensweisen. Für diejenigen Personen, welche nach einem *Dominanz*-Antwortprozess antworten ergibt sich die individuelle Merkmalsausprägung dabei direkt aus den entsprechenden Personenparametern im Rahmen der Anwendung eines psychometrischen Modells zur Modellierung des summativen Skalierungsprinzips (vgl. dazu die in Abschnitt 4.2 in Kapitel 4 *Psychometrische Modellierung* beschriebenen Modelle). Für diejenigen Personen, welche nach einem *Nähe–Distanz*-Antwortprozess antworten, müsste dagegen die individuelle Merkmalsausprägung aus den Itemkategorieparametern (z. B. aus deren Mittelwert) derjenigen Itemkategorien bestimmt werden, welchen die betreffende Person zugestimmt hat. Zur konditionalen Bestimmung der Itemparameter (für beide Personengruppen) – also ohne Einbezug der Personenparameter – bietet sich hierbei beispielsweise der in dieser Arbeit bereits eingesetzte und in den Anhängen A und B abgeleitete und beschriebene *PAIR*-Algorithmus an. Neben dem konditionalen Prinzip des Algorithmus bietet sich dieser auch deswegen besonders an, weil das darin umgesetzte Prinzip der Itemparameterbestimmung eine starke Analogie zu dem im Kapitel 1 in Abschnitt 1.3.3 beschriebenen Skalierungsprinzip von Thurstone (1927b) zur Bestimmung der relativen Itemschwierigkeiten hat. Dieses Skalierungsprinzip wurde von Thurstone (1927b) wiederum als Voraussetzung für die Bestimmung der Merkmalsausprägung der Personen nach dem Nähe–Distanz-Antwortmodell entwickelt. Der Messwert einer Person auf der zu erfassenden Merkmalsdimension ergibt sich danach unmittelbar aus den zuvor bestimmten Schwierigkeiten derjenigen Items, welchen die Person zugestimmt hat. Die relativen Itemkategorieschwierigkeiten von Fragebogen-Items lassen sich nach diesem Prinzip empirisch über einen vollständigen Paarvergleich aller Itemkategoriehäufigkeiten bestimmen (vgl. Thurstone, 1927b, 1927c, 1929). So ein vollständiger Paarvergleich aller Itemkategoriehäufigkeiten stellt auch den Kern des *PAIR*-Algorithmus dar.

Als ein übergreifendes Ergebnis der vorliegenden Arbeit ist der Befund anzusehen, dass bei der Beantwortung von psychodiagnostischen Fragebogenverfahren zwei fundamental unterschiedliche (implizite) Antwortmodelle bei den ant-

wortenden Personen wirksam sind. Neben anderen Formen der Antwortverzerrung durch idiosynkratische Antwortmuster zeigen sich diese beiden impliziten Antwortmodelle in konsistenter Weise als skalen- und konstruktübergreifendes Phänomen. Die beiden impliziten Antwortmodelle der antwortenden Personen, die einem *Nähe–Distanz*-Antwortprozess oder den *Dominanz*-Antwortprozess folgen, korrespondieren mit zwei unterschiedlichen Gruppen von psychometrischen Modellen zur Erklärung und Modellierung von Fragebogendaten. Die fundamentale Unterschiedlichkeit der beiden Antwortmodelle indiziert für die Auswertung von Fragebogendaten die Anwendung von zwei personenspezifisch unterschiedlichen Prinzipien zur Indexbildung im Rahmen der psychologischen Diagnostik mit Fragebogenverfahren.

Anhang A

Ableitung des *PAIR*-Algorithmus aus der Modellgleichung des Rasch-Modells

Die Modellgleichung des von Georg Rasch (1960), formulierten logistischen Testmodells für den Dominanz-Antwortprozess formalisiert die Wahrscheinlichkeit der Antwort einer Person zu einer der beiden Antwortkategorien $p\left(x_{vi}|\theta_v;\sigma_i\right)$, mit $x \in \{0,1\}$, als eine logistische Funktion in Abhängigkeit der Merkmalsausprägung (Fähigkeit) θ_v einer Person v bei der Beantwortung eines Items i mit der Schwierigkeit σ_i (vgl. Gleichung A.1).

$$p\left(x_{vi}\right) = \frac{e^{x_{vi}(\theta_v-\sigma_i)}}{1+e^{(\theta_v-\sigma_i)}} \tag{A.1}$$

Die Gleichung A.1 kann in zwei Gleichungen aufgelöst werden, um jeweils die Antwortwahrscheinlichkeit in den beiden Antwortkategorien zu formalisieren. So ist die Wahrscheinlichkeit für die Wahl der Kategorie „1“ (z. B. richtige Antwort oder Zustimmung) über Gleichung A.2 definiert und die Wahrscheinlichkeit für die Kategorie „0“ (z. B. falsche Antwort oder Ablehnung) über die Gleichung A.3.

$$p\left(x_{vi}=1|\theta_v;\sigma_i\right) = \frac{e^{(\theta_v-\sigma_i)}}{1+e^{(\theta_v-\sigma_i)}} \tag{A.2}$$

$$p\left(x_{vi}=0|\theta_v;\sigma_i\right)=\frac{1}{1+e^{(\theta_v-\sigma_i)}} \tag{A.3}$$

Für die weitere Herleitung soll nun angenommen werden, dass eine Stichprobe von n Personen v (mit $v=1...n$) zwei Items i und j beantwortet. Auf der Basis der Grundannahme des Rasch-Modells der bedingten lokalen stochastischen Unabhängigkeit der Antworten zu den Items i und j, gegeben die Merkmalsausprägung θ_v der Personen, lassen sich die Wahrscheinlichkeiten der vier möglichen Antwortmuster bei der Beantwortung der beiden Items durch die multiplikative Verknüpfung der einzelnen Antwortwahrscheinlichkeiten berechnen. Über die beiden Gleichungen A.2 und A.3 lassen sich so die Wahrscheinlichkeiten der drei möglichen Summenwerte von entweder 0, 1 oder 2 oder die damit verbundenen Antwortmuster („00“, „01“, „10“, „11“), über die Gleichungen A.4 bis A.7, darstellen.

$$p\left(x_{vi}=0,x_{vj}=0|\theta_v;\sigma_i\right)=\frac{1}{1+e^{(\theta_v-\sigma_i)}}\times\frac{1}{1+e^{(\theta_v-\sigma_j)}} \tag{A.4}$$

$$p\left(x_{vi}=1,x_{vj}=0|\theta_v;\sigma_i\right)=\frac{e^{(\theta_v-\sigma_i)}}{1+e^{(\theta_v-\sigma_i)}}\times\frac{1}{1+e^{(\theta_v-\sigma_j)}} \tag{A.5}$$

$$p\left(x_{vi}=0,x_{vj}=1|\theta_v;\sigma_i\right)=\frac{1}{1+e^{(\theta_v-\sigma_i)}}\times\frac{e^{(\theta_v-\sigma_j)}}{1+e^{(\theta_v-\sigma_j)}} \tag{A.6}$$

$$p\left(x_{vi}=1,x_{vj}=1|\theta_v;\sigma_i\right)=\frac{e^{(\theta_v-\sigma_i)}}{1+e^{(\theta_v-\sigma_i)}}\times\frac{e^{(\theta_v-\sigma_j)}}{1+e^{(\theta_v-\sigma_j)}} \tag{A.7}$$

Zur weiteren Bestimmung der Differenzen der Schwierigkeiten der Items i und j beschränken sich nun die weiteren Betrachtungen auf Personen, die Item i richtig beantwortet haben und gleichzeitig Item j falsch (vgl. Gleichung A.5) und umgekehrt (vgl. Gleichung A.6) – also die Antwortmuster „01“ und „10“. Diese Beschränkung ist insofern plausibel und berechtigt, als dass sich bezüglich der Differenz der Itemschwierigkeit der beiden Items keine Informationen aus einer konstant „richtigen“ oder „falschen“ Beantwortung beider Items ableiten lassen.

Die verbundene Wahrscheinlichkeit zum Erreichen eines Summenwertes von 1 bei der Beantwortung von beiden Items kann daher durch Gleichung A.8

ausgedrückt werden.

$$p\left(x_{vi}+x_{vj}=1\right)=\frac{e^{(\theta_v-\sigma_i)}}{(1+e^{(\theta_v-\sigma_i)})\times\left(1+e^{(\theta_v-\sigma_j)}\right)}+\frac{e^{(\theta_v-\sigma_j)}}{(1+e^{(\theta_v-\sigma_i)})\times\left(1+e^{(\theta_v-\sigma_j)}\right)} \tag{A.8}$$

Die bedingte Wahrscheinlichkeit zur richtigen Beantwortung des Items i (Score $x_{vi}=1$), unter der Bedingung, dass der Summenwert für die Beantwortung beider Items $x_{vi}+x_{vj}=1$ beträgt, formal also $p\left(x_{vi}=1|x_{vi}+x_{vj}=1\right)$, kann als Verhältnis der Ergebnisse der beiden Gleichungen A.8 und A.5 ausgedrückt werden. Zum Ausdruck dieses Verhältnisses der beiden Wahrscheinlichkeiten wird Gleichung A.5 als Zähler und Gleichung A.8 als Nenner eingesetzt, was zu Gleichung A.9 führt.

$$p\left(x_{vi}=1|x_{vi}+x_{vj}=1\right)=\frac{\frac{e^{(\theta_v-\sigma_i)}}{1+e^{(\theta_v-\sigma_i)}}\times\frac{1}{1+e^{(\theta_v-\sigma_j)}}}{\frac{e^{(\theta_v-\sigma_i)}}{\left(1+e^{(\theta_v-\sigma_i)}\right)\times\left(1+e^{(\theta_v-\sigma_j)}\right)}+\frac{e^{(\theta_v-\sigma_j)}}{\left(1+e^{(\theta_v-\sigma_i)}\right)\times\left(1+e^{(\theta_v-\sigma_j)}\right)}} \tag{A.9}$$

In Analogie zu der von Kubinger (1988, S. 40) dargestellten Möglichkeit zur Eliminierung der Itemparameter zur Bestimmung der Personenparameter im Rasch-Modell (vgl. auch Fischer, 1974) lässt sich die Gleichung A.9 durch einige mathematische Umformungen zu Gleichung A.10 reduzieren, wobei dabei die Personenparameter θ_v heraus gekürzt werden.

$$p\left(x_{vi}=1|x_{vi}+x_{vj}=1\right)=\frac{e^{\sigma_i}}{e^{\sigma_i}+e^{\sigma_j}} \tag{A.10}$$

Die Gleichung A.10 führt so die bedingte Wahrscheinlichkeit für die Lösung des Items i (Score $x_{vi}=1$) auf das Verhältnis der beiden Itemschwierigkeiten σ_i und σ_j unter der Bedingung eines Summenwertes von $x_{vi}+x_{vj}=1$, zurück. Analog dazu kann die bedingte Wahrscheinlichkeit für die Lösung des Items j (Score $x_{vj}=1$) unter der Bedingung eines Summenwertes von $x_{vi}+x_{vj}=1$ durch Gleichung A.11 formalisiert werden.

$$p\left(x_{vj}=1|x_{vi}+x_{vj}=1\right)=\frac{e^{\sigma_j}}{e^{\sigma_i}+e^{\sigma_j}} \tag{A.11}$$

Der entscheidende Punkt besteht nun darin, dass die beiden Wahrscheinlichkeiten aus den Gleichungen A.10 und A.11 aus dem empirischen Daten

der Stichprobe für die Population geschätzt, oder genauer, für die Stichprobe *explizit* berechnet werden können. Diese Berechnung erfolgt, bezogen auf Gleichung A.10 durch die Bestimmung der Häufigkeit $f_{i,j}$ derjenigen Personen, welche Item i richtig beantwortet haben, unter der Bedingung, dass sie Item j falsch beantwortet haben, bezogen auf diejenigen Personen $n_{i,j} = f_{i,j} + f_{j,i}$, welche beide Items beantwortet haben und dabei gleichzeitig einen Summenwert von $x_{vi} + x_{vj} = 1$ erzielt haben. In gleicher Weise lässt sich, bezogen auf Gleichung A.11, das Verhältnis derjenigen Personen $f_{j,i}$, welche Item j richtig beantwortet haben (unter der Bedingung einer falschen Antwort auf Item i) zur Anzahl der Personen $n_{i,j} = f_{i,j} + f_{j,i}$, die beide Items beantwortet haben und dabei einen Summenwert von $x_{vi} + x_{vj} = 1$ erzielt haben, bilden.

Diese beiden Verhältnisse lassen sich durch die Gleichungen A.12 und A.13 darstellen.

$$\frac{f_{i,j}}{n_{i,j}} \simeq \frac{e^{\sigma_i}}{e^{\sigma_i} + e^{\sigma_j}} \tag{A.12}$$

$$\frac{f_{j,i}}{n_{i,j}} \simeq \frac{e^{\sigma_j}}{e^{\sigma_i} + e^{\sigma_j}} \tag{A.13}$$

Die beiden Verhältnisse in den Gleichungen A.12 und A.13 können unter der Bedingung $n_{i,j} = f_{i,j} + f_{j,i}$ für eine Stichprobe vom Umfang n umformuliert werden. Danach kann das Verhältnis der bedingten Itemkategoriehäufigkeiten in der Stichprobe zur Schätzung der Verhältnisse der Itemschwierigkeiten für die Population herangezogen werden, wie es in den Gleichungen A.14 und A.15 dargestellt ist.

$$\frac{f_{i,j}}{f_{j,i}} = \frac{\widehat{e^{\sigma_i}}}{e^{\sigma_j}} \tag{A.14}$$

$$\frac{f_{j,i}}{f_{i,j}} = \frac{\widehat{e^{\sigma_j}}}{e^{\sigma_i}} \tag{A.15}$$

Durch Logarithmieren der beiden Gleichungen A.14 und (A.15 lassen sich die Differenzen der Itemschwierigkeiten als Funktion der logarithmierten Verhältnisse der bedingten Itemkategoriehäufigkeiten $f_{i,j}$ und $f_{j,i}$ darstellen.

$$\ln\left(\frac{f_{i,j}}{f_{j,i}}\right) = \widehat{\sigma_j - \sigma_i} \tag{A.16}$$

$$\ln\left(\frac{f_{j,i}}{f_{i,j}}\right) = \widehat{\sigma_i - \sigma_j} \tag{A.17}$$

Analog zu der Anwendung der Paarvergleich-Methode in anderen Kontexten (vgl. Heine & Tarnai, 2015, sowie auch Kapitel 1, Abschnitt 1.3.3 und Kapitel 4, Abschnitt 4.5) müssen $\binom{k}{2}$ paarweise (Item-)Beurteilungen oder Itemvergleiche $f_{j,i}$ und $f_{i,j}$ (mit $i = 1, 2, ..., k$, $j = 1, 2, ..., k$ und $i \neq j$) für k Items durchgeführt werden, um eine vollständige metrische Ordnung der k Items zu ermitteln. Zur eindeutigen Identifikation der Itemparameter aus den in den Gleichungen (A.16) und (A.17) gegebenen Verhältnissen der Itemschwierigkeiten wird die im Rasch-Modell übliche Restriktion eingeführt, dass die Summe der Itemparameter gleich null ist.

Im Hinblick auf möglicherweise unvollständige Datenmatrizen, welche zur Bestimmung der Itemparameter nach der oben abgeleiteten Methode herangezogen werden, muss betont werden, dass solche fehlenden Werte keinerlei Probleme bei der Bestimmung der bedingten Itemkategoriehäufigkeiten $f_{j,i}$ und $f_{i,j}$ bereiten. Auf diesen Umstand haben bereits Choppin (1983) und auch Wright und Masters (1982) hingewiesen.

> Note that the matrix of observations need not be complete. An individual who is exposed to items i and j gets an opportunity to contribute to b_{ij} or b_{ji} and thus to the estimation of σ_i and σ_j. It is not necessary for this individual to attempt all (or indeed any) of the other items in the set. This is the algorithm's great strength in practical applications. (Choppin, 1983, S. 11)

In diesem Sinne tragen nur Personen aus einer Stichprobe zur Schätzung der bedingten Itemkategoriehäufigkeiten $f_{j,i}$ und $f_{i,j}$ (für die Population) bei, welche auf jeweils zwei Items eine gültige Antwort gegeben haben. Personen aus der Stichprobe, welche die beiden Items nicht beantwortet haben oder denen diese möglicherweise gar nicht erst vorgelegt wurden, tragen dementsprechend nicht zur Schätzung der Itemparameter bei. Die Bestimmung der Itemparameter stützt sich so nur auf die relativen Verhältnisse der bedingten Itemkategoriehäufigkeiten $f_{j,i}$ and $f_{i,j}$, unabhängig davon, durch welche Personenantworten diese zustande gekommen sind (vgl. auch Rasch, 1977).

Im Hinblick auf unvollständige Analysedaten lässt sich eine minimale Voraussetzung zu Schätzbarkeit der Itemparameter für die Population anhand

einer Stichprobe wie folgt definieren. Zur Bestimmung der verbundenen, bedingten Itemkategoriehäufigkeiten $f_{j,i}$ and $f_{i,j}$ auf einer gemeinsamen metrischen Skala, muss sichergestellt sein, dass sich die Datenmatrix nicht durch einfaches Umsortieren der Items und Personen in zwei getrennte, unverbundene Teildatensätze aufteilen lässt. Dies wäre nur dann der Fall, wenn eine Teilstichprobe der Personen n_{1v} nur eine Teilmenge der Items k_{1i} bearbeitet, während eine andere Teilstichprobe der Personen n_{2v} nur eine Teilmenge der Items k_{2i} bearbeitet, wobei gleichzeitig gelten muss $k_{1i} \notin k_{2i}$ und $n_{1v} \notin n_{2v}$. Diese minimale Voraussetzung zur Bestimmung der Itemparameter nach dem *PAIR*-Algorithmus, entspricht der (grundlegenden) *Existenzbedingung* von Modellparametern im logistischen Testmodell wie sie bereits von Fischer (1981), formuliert wurde (vgl. auch Rost, 2004).

Anhang B

Praktische Implementierung des *PAIR*-Algorithmus

Für ein einfaches Beispiel zur Darstellung der numerischen Operationen zur Berechnung der Itemparameter mit dem *PAIR*-Algorithmus sei eine einfache Datenmatrix M (vgl. Abbildung B.1) zur Analyse vorgegeben.

	$Item\ 1$	$Item\ 2$	$Item\ 3$	$Item\ 4$
$Person\ 1$	1	1	1	0
$Person\ 2$	0	1	0	1
$Person\ 3$	1	0	0	1
$Person\ 4$	1	0	0	0
$Person\ 5$	NA	1	1	NA
$Person\ 6$	1	1	NA	1
$Person\ 7$	NA	0	0	NA
$Person\ 8$	1	NA	0	0

Abbildung B.1 Datenmatrix M für $n = 8$ Personen und $k = 4$ dichotome Items

Die Datenmatrix enthält die Antworten von acht Personen zu vier dichotomen Items. Die Daten liegen zunächst im *Code*-Format vor, wobei die dichotomen Antworten mit $1 \equiv richtig$ oder Zustimmung und $0 \equiv falsch$ oder Ablehnung kodiert sind (vgl. auch Tabelle 4.1 (rechte Seite) in Abschnitt 4.2.1; fehlende Werte in der Datenmatrix sind durch den Ausdruck „*NA*“ repräsen-

tiert. Der Umfang fehlender Werte entspricht hier einem Anteil von 18.75 %.

Im ersten Schritt werden die Antwortdaten aus Matrix M als *Indikatormatrix* Z_{Daten_M} für *beide* Antwortkategorien dargestellt (vgl. Abbildung B.2), was nach Zysno (1993) auch als *Reaktions*-Format der Daten bezeichnet wird (vgl. Abschnitt 4.2.1).

	$I1.0$	$I1.1$	$I2.0$	$I2.1$	$I3.0$	$I3.1$	$I4.0$	$I4.1$
$P1$	0	1	0	1	0	1	1	0
$P2$	1	0	0	1	1	0	0	1
$P3$	0	1	1	0	1	0	0	1
$Z_{Daten_M} = P4$	0	1	1	0	1	0	1	0
$P5$	**0**	**0**	0	1	0	1	**0**	**0**
$P6$	0	1	0	1	**0**	**0**	0	1
$P7$	**0**	**0**	1	0	1	0	**0**	**0**
$P8$	0	1	**0**	**0**	1	0	1	0

Abbildung B.2 Indikatormatrix Z_{Daten_M} für beide Antwortkategorien der Items aus der Datenmatrix M.

Die in Matrix Z in Abbildung B.2 fett gedruckten Datenpunkte repräsentieren die fehlenden Werte in Matrix M. Während in der Matrix M die fehlenden Werte keinerlei Information bezüglich der Wahl einer der beiden Itemkategorien enthalten, wird diese (fehlende) Information im Reaktionsformat (Matrix Z) präzisiert. Die Repräsentation der Antwortdaten im *Reaktions*–Format in Matrix Z präzisiert die Informationen zum Antwortverhalten der Personen insofern, als dass hier der Code „0“ (für beide Kategorien eines Items) anzeigt, dass keine der beiden dargebotenen Kategorien von der betreffenden Person für das jeweilige Item ausgewählt wurde.

Der erste Schritt zur Berechnung der Itemparameter besteht nun darin, die bedingten Itemkategoriehäufigkeiten zu zählen. Für die hier vorliegenden dichotomen Items geht es darum, für jedes Item i die Häufigkeit der Lösung bzw. Zustimmung zu zählen, unter der Bedingung, dass das Item j nicht gelöst wurde, bzw. diesem nicht zugestimmt wurde – wobei aus logischen Gründen die Nebenbedingung $i \neq j$ gelten muss.

$$C_{f_{i,j};f_{j,i}} = \begin{bmatrix} 0 & 2 & 3 & 3 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 2 & 0 \end{bmatrix}$$

Abbildung B.3 Paarweise Vergleichsmatrix C mit den Einträgen $f_{i,j}$ und $f_{j,i}$ – für alle Items $k(i \neq j)$.

Für die einfache Datenmatrix M resultiert dies in einer symmetrischen paarweisen Vergleichsmatrix C (vgl. Abbildung B.3). Diese enthält die bedingten Kategoriehäufigkeiten $f_{i,j}$ und $f_{j,i}$. Die Matrix C weist Analogien zu der so genannten *Burt-Matrix* auf (Burt, 1950), wie sie im Rahmen der numerischen Operationen für die Korrespondenzanalyse (Blasius, 2001; Greenacre, 1984), gebildet wird (vgl. Abbildung B.4). Eine Erweiterung der Korrespondenzanalyse [*Correpondence Analysis* –CA] stellt die [*Multiple Correpondence Analysis* – MCA] dar (vgl. Greenacre, 2010; Greenacre & Blasius, 1994), welche eine der optimalen Skalierung oder Multidimensionalen Skalierung (MDS) äquivalente Methode darstellt, die auch unterer englischsprachigen Begriffen wie *appropriate scoring*, *dual scaling*, *homogeneity analysis*, *scalogram analysis*, oder *quantification method* bekannt ist (Abdi & Valentin, 2007, vgl. auch Abschnitt 4.5.4).

$$B_{f_{i,j};f_{j,i}} = \left[\begin{array}{cccccccc} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 5 & 2 & 2 & 3 & 1 & 3 & 2 \\ 0 & 2 & 3 & 0 & 3 & 0 & 1 & 1 \\ 1 & 2 & 0 & 4 & 1 & 2 & 1 & 2 \\ 1 & 3 & 3 & 1 & 5 & 0 & 2 & 2 \\ 0 & 1 & 0 & 2 & 0 & 2 & 1 & 0 \\ 0 & 3 & 1 & 1 & 2 & 1 & 3 & 0 \\ 1 & 2 & 1 & 2 & 2 & 0 & 0 & 3 \end{array}\right]$$

Abbildung B.4 Burt-Matrix B berechnet aus der Indikatormatrix Z der Antwort-Daten in M.

Die *Burt-Matrix* B ergibt sich dabei auf der Basis der Indikatormatrix Z

(vgl. Abbildung B.2) durch einfache Matrixmultiplikation der transponierten Indikatormatrix Z mit sich selbst – also $B = Z^T \times Z$. Der Vergleich der beiden Matrizen B und C (in Abbildung B.3) zeigt, dass in der Matrix der Paarvergleiche C im Gegensatz zur Burt-Matrix B einige Zeilen und Spalten fehlen – dies sind in diesem Beispiel die Zeilen 1, 3, 5 und 7 sowie die Spalten 2, 4, 6 und 8. Die in der Matrix C „fehlenden“ Zellen (in Abbildung B.4 grau unterlegt) beziehen sich auf die *unkonditionalen* Vergleiche der Items, welche zur Bestimmung der Itemparameter nach der *PAIR*-Methode nicht berücksichtigt werden. Die Diagonale der Matrix C enthält Einträge mit dem Wert $f_{i,j} = f_{j,i} = 0$, da sich diese Einträge auf den Vergleich des jeweiligen Items mit sich selbst beziehen, für den sich hier keine inhaltlich sinnvolle Bedeutung ergibt. Allerdings zeigt sich in diesem Beispiel, dass sich auch außerhalb der Diagonalen der paarweisen Vergleichsmatrix C Einträge mit dem Wert $f_{i,j} = f_{j,i} = 0$ befinden (vgl. Abbildung B.3). Im zweiten Rechenschritt muss aus den Einträgen in der Matrix C eine sogenannte *positiv reziproke* Matrix D mit den Einträgen $f_{i,j}/f_{j,i}$ and $f_{j,i}/f_{i,j}$ gebildet werden. Da dies allerdings im Falle von Einträgen mit dem Wert von $f_{i,j} = f_{j,i} = 0$ außerhalb der Diagonalen zu numerischen Problemen (Divisionen durch null) in diesem weiteren Rechenschritten führen würde, muss die Matrix C zuvor angemessen transformiert werden. Zur Lösung dieses numerischen Problems konnte Choppin (1983) algebraisch zeigen, dass anstatt der direkten Häufigkeiten aus den Paarvergleichen in Matrix C, äquivalent auch das Quadrat (zweite Potenz) der Matrix C (oder höhere Potenzen) eingesetzt werden kann.

Im Zusammenhang mit dem Einsatz der Methode des Paarvergleichs zur Bestimmung der Rangreihe bei paarweise Sportwettbewerben (vgl. Abschnitt 4.5) konnte M. G. Kendall (1955) zeigen, dass die resultierende Rangreihe der Wettkämpfer durch die zusätzliche Berücksichtigung der Anzahl der bereits gewonnenen Zweikämpfe der zu vergleichenden Wettbewerber stabilisiert wird. M. G. Kendall (1955) konnte weiter zeigen, dass dieses Prinzip numerisch der (wiederholten) Potenzierung der paarweisen Vergleichsmatrix C entspricht (vgl. auch David, 1971). Garner und Engelhard (2000) sowie Saaty (2008) weisen in diesem Zusammenhang darauf hin, dass die mit höher werdenden Potenzen der Matrix C resultierende Rangordnung mit derjenigen Rangordnung konvergiert, welche sich aus dem Eigenvektor des größten Eigenwertes

der Matrix C ergibt. Höhere Potenzen den Matrix C resultieren allerdings relativ schnell in numerisch großen Einträgen in der potenzierten Matrix, sodass bei den folgenden Schritten (Bildung der logarithmierten Wettquotienten) mit rechnerischen Ungenauigkeiten aufgrund von Rundungsfehlern gerechnet werden muss. M. G. Kendall (1955) stellt in diesem Zusammenhang pointiert die berechtigte Frage: *„how far one would wish to go on practical grounds“* (M. G. Kendall, 1955, S. 50) – also die Frage nach der praktisch sinnvollen Grenze der Potenzierung der Matrix C. Gesucht ist also ein sinnvolles Kriterium bis zu welcher Potenz die Matrix C in diesem Verfahren transformiert werden sollte. Zur Vermeidung der oben dargestellten numerischen Problematik erscheint eine sinnvolle Grenze darin zu liegen, die Matrix C soweit zu potenzieren, bis in der resultierenden Matrix keine Werte $f_{i,j} = f_{j,i} = 0$ außerhalb der Diagonalen mehr vorliegen.

$$C^3_{f_{i,j};f_{j,i}} = \begin{bmatrix} 8 & 15 & 28 & 26 \\ 7 & 6 & 14 & 11 \\ 1 & 2 & 4 & 6 \\ 8 & 8 & 17 & 9 \end{bmatrix}$$

Abbildung B.5 Dritte Potenz der Paarvergleich Matrix C.

In diesem Sinne zeigt sich bei der praktischer Anwendung, dass sich unter der grundsätzlichen Annahme einer homogenen, eindimensionalen Skala mit einer mittleren Anzahl von Items, die Problematik von Werten $f_{i,j} = f_{j,i} = 0$ außerhalb der Diagonalen bereits ab der dritten Potenz der Matrix C auflöst (vgl. Abbildung B.5).

Die im folgenden Schritt aus der potenzierten Matrix C gebildete *positiv reziproke* Matrix D mit den Einträgen $f_{i,j}/f_{j,i}$ and $f_{j,i}/f_{i,j}$ ist in Abbildung B.6 dargestellt.

Zur Berechnung der Itemschwierigkeitsparameter für die vier Items aus diesem Beispiel werden die Wettquotienten $f_{i,j}/f_{j,i}$ und $f_{j,i}/f_{i,j}$ in Matrix D logarithmiert und die Zeilenmittelwerte der Matrix bestimmt (vgl. Abbildung B.7). Dieses Vorgehen impliziert die im Rahmen der Modellidentifikation übliche Restriktion einer Normierung der Itemparameter auf einen Summenwert von null.

$$D = \begin{bmatrix} {}^{8}/_{8} & {}^{7}/_{15} & {}^{1}/_{28} & {}^{8}/_{26} \\ {}^{15}/_{7} & {}^{6}/_{6} & {}^{2}/_{14} & {}^{8}/_{11} \\ {}^{28}/_{1} & {}^{14}/_{2} & {}^{4}/_{4} & {}^{17}/_{6} \\ {}^{26}/_{8} & {}^{11}/_{8} & {}^{6}/_{17} & {}^{9}/_{9} \end{bmatrix}$$

Abbildung B.6 Positiv reziproke Matrix D aus Matrix C^3.

$$\ln D = \begin{bmatrix} 0.000 & -0.762 & -3.332 & -1.179 \\ 0.762 & 0.000 & -1.946 & -0.319 \\ 3.332 & 1.946 & 0.000 & 1.042 \\ 1.179 & 0.319 & -1.042 & 0.000 \end{bmatrix} \quad \rightrightarrows \quad \begin{bmatrix} -1.318 \\ -0.376 \\ 1.580 \\ 0.114 \end{bmatrix}$$

Abbildung B.7 Logarithmierte Matrix D (links) und deren Zeilenmittelwerte (rechts).

Der resultierende Vektor der Zeilenmittelwerte aus der Matrix $ln(D)$ (vgl. Abbildung B.7, rechts) enthält die Schwierigkeitsparameter σ der items i; mit $i = 1, 2, ..., k$.

Um eine Datenmatrix zu analysieren, welche polytome Items mit mehr als zwei Antwortkategorien umfasst, folgt die Prozedur des paarweisen Vergleichs grundsätzlich demselben Prinzip. Zur Bestimmung der Itemkategorieschwierigkeiten wird jedes Item in der Datenmatrix dummy-codiert, wobei eine „$0, 1$“ codierte Itemkategorie-(Super)-Matrix gebildet wird. Diese Matrix wird dann wiederum in der im obigen Beispiel beschriebenen Weise analysiert. Der erste Schritt beim paarweisen Algorithmus besteht dabei wiederum darin, die symmetrische paarweise Vergleichsmatrix C zu bilden.

Bei der Analyse der dummy-codierten Itemkategoriematrix wird daher der Algorithmus auf den Vergleich der Antworthäufigkeiten für jede Kategorie eines jeden Items erweitert. In diesem Fall repräsentiert die paarweise Vergleichsmatrix C_{fi,fj_c} mit den Einträgen f_{i_c,j_c} die Anzahl derjenigen Personen, welche bei Item i in der Kategorie c und gleichzeitig bei Item j in der Kategorie $c-1$ geantwortet haben, wodurch Choppins bedingter paarweiser Algorithmus auf polytome Antwortformate erweitert wird. Da es keinen sinnvollen Ver-

$$C_{f_{i_c j_c}} = \begin{bmatrix} 0 & 0 & 0 & 104 & 279 & 105 & 87 & 355 & 59 \\ 0 & 0 & 0 & 36 & 135 & 128 & 32 & 177 & 95 \\ 0 & 0 & 0 & 13 & 37 & 26 & 6 & 52 & 32 \\ 20 & 279 & 135 & 0 & 0 & 0 & 60 & 336 & 66 \\ 7 & 105 & 128 & 0 & 0 & 0 & 16 & 152 & 84 \\ 1 & 22 & 22 & 0 & 0 & 0 & 3 & 23 & 32 \\ 21 & 355 & 177 & 94 & 336 & 152 & 0 & 0 & 0 \\ 6 & 59 & 95 & 10 & 66 & 84 & 0 & 0 & 0 \\ 2 & 9 & 17 & 4 & 9 & 14 & 0 & 0 & 0 \end{bmatrix}$$

Abbildung B.8 Paarweise Vergleichsmatrix C für drei Items mit $m = 4$ Antwortkategorien (kodiert von 0 bis 3) mit bedingten Kategorie Häufigkeiten f_{i_c,j_c}.

gleich zwischen den Antwortkategoriehäufigkeiten innerhalb eines Items mit sich selbst gibt, bestehen symmetrisch, quadratische Bereiche entlang der Diagonale der resultierenden paarweisen Vergleichsmatrix C, welche den Wert null enthalten. Diese weisen eine Größe $(m-1) \times (m-1)$ auf, wobei m der Anzahl der Antwortkategorien entspricht. Abbildung B.8 veranschaulicht am Beispiel, wie eine solche Matrix C für drei Items mit $m = 4$ Antwortkategorien aussehen kann. Jede Gruppe von drei Spalten und Zeilen (von links nach rechts und oben nach unten) enthält die Antwortkategorie Häufigkeiten f_{i_c,j_c} für jedes Item i und j als Ergebnis des Vergleichs der Antwortkategorien i_c (mit $i_c = 1$ bis $i_c = m - 1$) mit den Antwortkategorien j_c (mit $j_c = 0$ bis $j_c = m - 2$), wobei gelten muss $i \neq j$.

Anhang C

Online Fragebögen aus dem ESF-Projekt der Universität der Bundeswehr

In diesem Anhang sind die Instrumente zur Erfassung der in dieser Arbeit analysierten Konstrukte aufgeführt. Es handelt sich um „Screenshots“ der Online Fragebögen aus dem ESF-Projekt, wie sie für die Studienjahrgänge 2007 bis 2011 an der Universität der Bundeswehr München eingesetzt wurden. Abschnitt C.1 zeigt die Fragebogenversion wie sie bis zum Jahrgang 2008 eingesetzt wurde, Abschnitt C.2 die Version für den Jahrgang 2009 und Abschnitt C.3 zeigt die Version für die Jahrgänge 2010 und 2011. Die Inhalte der Items des AIST–R unterliegen urheberrechtlichen und copyright-bezogenen Bestimmungen und sind daher in den „Screenshots“ maskiert. Der genaue Wortlaut dieser Items kann bei Bergmann und Eder (2005) eingesehen werden.

C.1 ESF-Projekt Jahrgang 2008

ESF-Projekt Online-Fragebogen

ESF-Projekt Fragebogen

1. Kennwort

Diese Erhebung ist anonym. Damit Sie später Ihre persönliche Testauswertung identifizieren können, brauchen Sie ein unverwechselbares, nur Ihnen bekanntes Kennwort. Bilden Sie Ihr persönliches Kennwort bitte nach folgender Anweisung:

1-2: 1. und 2. Buchstabe des Geburtsnamens Ihrer Mutter **3-4:** Geburts**tag** Ihrer Mutter **5-6:** Geburts**monat** Ihrer Mutter **7-8:** Die beiden **letzten** Buchstaben Ihres eigenen Geburtsortes	**Kennwort**

2. Studienfach an der UniBwM

BAU BW EIT ETTI GEO INF LRT MB PÄD SPO SWI WINF WOW

Studiere nicht an der UniBwM, sondern ... *(z.B. "studiere BWL an der LMU"; oder "berufstätig als Bankkauffrau")*

3. Studienjahrgang

2003 2004 2005 2006 2007 Studiere nicht an der UniBwM oder früherer Jahrgang

4. TSK - welche Uniform tragen Sie?

Heer Luftwaffe Marine
nicht zutreffend / kein Soldat

5. Soldent-Studat:

Geben Sie bitte auf einer Skala von 0 bis 100 % an, ob Sie sich während Ihrer 3+ Jahre an der UniBw 100%ig als Student/in ... oder mehr als zum Studium abkommandierter Soldat (also zu 0% Student/in) sehen.
Lassen Sie diese Frage unbeantwortet, wenn Sie nicht an der UniBw studieren!

%ig Student(in)

6. Alter

unter 21 Jahre 23 Jahre 26 Jahre
21 Jahre 24 Jahre über 26 Jahre
22 Jahre 25 Jahre

7. Geschlecht

männlich weiblich

Die folgenden Fragen zur Studienfachwahl richten sich nur an Studierende und Hochschulabsolventen

1. Welche der folgenden Fächer haben Sie in der Phase der Berufs- und Studienwahl in Erwägung gezogen?

		1 überhaupt nicht erwogen	2	3	4 sehr ernsthaft erwogen bzw. gewählt
01	Architektur				
02	Bauingenieurwesen				
03	Biologie				
04	Betriebswirtschaftslehre				
05	Elektrotechnik				
06	Erziehungswissenschaft / Pädagogik				
07	Geoinformatik				
08	Germanistik				
09	Geschichte				
10	Informatik				
11	Jura				
12	Maschinenbau				
13	Mathematik				
14	Medizin				
15	Physik				
16	Politikwissenschaften				
17	Psychologie				

18	Publizistik				
19	Soziologie				
20	Sportwissenschaft				
21	Wirtschafts- und Organisationswissenschaften				
22	anderes Fach, und zwar				
23	anderes Fach, und zwar				

2. Ist das von Ihnen gewählte Studium Ihr Wunschstudium gewesen?

- Ja
- Nein, mein Wunschstudium wäre _____ gewesen

BFI

Machen Sie diesen psychologischen Test, um etwas über Ihre Persönlichkeit zu erfahren! Dieser Test misst, was nach Ansicht vieler Psychologen die fünf grundlegenden Dimensionen der Persönlichkeit sind.

Bei den folgenden Aussagen geht es darum, wie Sie sich selbst sehen. Ein Beispiel: Sind Sie der Ansicht, daß Sie gerne Zeit mit anderen verbringen? Bitte entscheiden Sie, inwieweit die jeweilige Aussage für Sie zutrifft. Es gibt keine richtigen oder falschen Antworten, allerdings werden Sie keine zutreffenden Ergebnisse erhalten, wenn Sie die Fragen nicht ernsthaft und wahrheitsgemäß beantworten.

Ihre Aufgabe ist es, für jede Aussage auf einer Skala von -2 bis +2 anzuklicken, wie sehr die Aussage für Sie zutrifft oder nicht zutrifft.

-2	-1	0	+1	+2
Sehr unzutreffend	Unzutreffend	teils/teils	Zutreffend	Sehr zutreffend

	Ich sehe mich selbst als jemand, der ...	-2	-1	0	+1	+2
1.	...zuverlässig und gewissenhaft arbeitet					
2.	...gerne Überlegungen anstellt, mit Ideen spielt					
3.	...aus sich herausgeht, gesellig ist					
4.	...leicht nervös und unsicher wird					
5.	...nicht den selben Regeln und Gesetzen unterworfen sein sollte, wie die meisten Menschen					
6.	...praktische Lösungen lieber mag als theoretische Diskussionen					
7.	...sich viele Sorgen macht					
8.	...nur wenig künstlerische Interessen hat					
9.	...Aufgaben gründlich erledigt					
10.	...dem eine hohe gesellschaftliche Stellung etwas bedeutet					
11.	...begeisterungsfähig ist, andere mitreißen kann					
12.	...oft Krach mit anderen hat.					
13.	...ruhig bleibt, selbst in Stresssituationen					

	Ich sehe mich selbst als jemand, der ...	-2	-1	0	+1	+2
14.	...bequem ist und zur Faulheit neigt					
15.	...eher zurückhaltend und reserviert ist					
16.	...künstlerische und ästhetische Eindrücke schätzt					
17.	...eher still und wortkarg ist					
18.	...rücksichtsvoll und einfühlsam zu anderen ist					
19.	...dazu neigt, unordentlich zu sein					
20.	...den Luxus liebt und das auch zeigt					
21.	...lieber kooperiert als konkurriert					
22.	...emotional ausgeglichen und nicht leicht aus der Fassung zu bringen ist					
23.	...mehr Respekt verdient als der durchschnittliche Mensch					
24.	...eine aktive Vorstellungskraft hat, phantasievoll ist					
25.	...schroff und abweisend zu anderen sein kann					
26.	...ziemlich durchschnittlich ist					

STOMP
Short Test of Music Preferences

Sag mir was du hörst, und ich sag dir, wer du bist. Psychologen der Universität von Texas fanden heraus, dass der Musikgeschmack Rückschlüsse auf die Persönlichkeit zulässt. "Musik durchdringt viele Bereiche unseres täglichen Lebens, wir hören sie im Auto am Weg zur Arbeit, entspannt zu Hause oder mit Freunden an der Bar", meint Sam Gosling. "Nahezu jeder hört auf irgendeine Weise Musik. Unsere Untersuchungen ergaben, dass die Persönlichkeit eine wesentliche Rolle bei der Wahl der Musik spielt".

Geben Sie für jedes der unten aufgeführten Musikgenre an, wie gerne oder ungern Sie diese Art von Musik hören. Verwenden Sie für Ihre Antworten die folgende 7-stufige Skala:

-3	-2	-1	0	+1	+2	+3
Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr

		-3	-2	-1	0	+1	+2	+3
1.	Klassik							
2.	Blues							
3.	Country							
4.	Electronica							

5.	Folk							
6.	Rap/hip-hop							
7.	Soul/R&B							
8.	Populäre Volksmusik							
9.	NDW							
10.	Alternative							
11.	Jazz							
12.	Rock							
13.	Pop							
14.	Heavy Metal							
15.	Filmmusik/Titelmelodien							

Hörproben für vier Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.
(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

		-3	-2	-1	0	+1	+2	+3
1.	MusicDimension1.mp3							
2.	MusicDimension2.mp3							
3.	MusicDimension3.mp3							
4.	MusicDimension4.mp3							

SDO
Soziale Dominanzorientierung

Im Folgenden finden Sie eine Reihe von Aussagen darüber, in welcher Beziehung gesellschaftliche Gruppen zueinander stehen sollten. Gesellschaftliche Gruppen können dabei z.B. ethnische Gruppen, politische Gruppen, religiöse Gruppen, Berufsgruppen oder auch die beiden Geschlechter sein. Bitte geben Sie an, wie stark Sie persönlich den Aussagen zustimmen.

Verwenden Sie für Ihre Antworten die folgende 7-stufige Skala:

-3 -2 -1 0 +1 +2 +3
Stimme ich überhaupt nicht zu — Teils/teils — Stimme voll und ganz zu

		-3	-2	-1	0	+1	+2	+3
1.	Wir würden mehr Probleme schaffen als lösen, wenn wir alle Gruppen gleich behandeln würden.							
2.	Wir sollten unser Möglichstes tun, um die Bedingungen für die unterschiedlichen Gruppen anzugleichen.							
3.	Um im Leben vorwärts zu kommen, ist es manchmal notwendig, auf anderen Gruppen herum zu treten.							
4.	Keine Gruppe von Menschen ist mehr wert als irgendeine andere Gruppe.							
5.	Es macht keinen Sinn, die Einkommen so gleich wie möglich zu gestalten.							
6.	Wenn bestimmte Gruppen unter sich bleiben würden, hätten wir weniger Probleme.							
7.	Es ist ein echtes Problem, daß bestimmte Gruppen oben sind und andere unten.							
8.	Keine einzelne Gruppe sollte in der Gesellschaft dominieren.							

RWA

Es folgen einige Aussagen, die sich auf verschiedene politische und gesellschaftliche Themen beziehen. Wie ist Ihre persönliche Meinung dazu?
Verwenden Sie für Ihre Antworten die folgende 5-stufige Skala:

-2 -1 0 +1 +2
Stimme überhaupt nicht zu — Teils/teils — Stimme voll und ganz zu

		-2	-1	0	+1	+2
1.	Es ist gut, dass die jungen Leute heutzutage größere Freiheiten haben, ihr eigenes Ding zu machen und gegen Dinge zu protestieren, die sie nicht mögen.					
2.	Die wahren Schlüssel zum guten Leben sind Gehorsam, Disziplin und Tugend.					
3.	Man sollte alten, traditionellen Glaubenssätzen weniger Beachtung schenken und stattdessen seine persönlichen Moralvorstellungen über Gut und Böse entwickeln.					
4.	Es gibt kein Verbrechen, das die Todesstrafe rechtfertigen würde.					
5.	Gehorsam und Achtung vor der Autorität sind die wichtigsten Tugenden, die Kinder lernen sollten.					
6.	Tugendhaftigkeit und Gesetzestreue bringen uns auf lange Sicht weiter als das ständige Infragestellen der Grundfesten unserer Gesellschaft.					
7.	Gleichgeschlechtliche Lebensgemeinschaften sollten der Ehe in jeder Hinsicht gleichgestellt werden.					
8.	Die Zeiten, in denen Frauen sich unterordnen mussten, sollten ein für alle Mal vorbei sein. Der Platz einer Frau in der Gesellschaft sollte sein, wo immer sie möchte.					
9.	Was unser Land wirklich braucht, ist ein starker, entschlossener Kanzler, der uns wieder auf unseren richtigen Weg bringt.					
10.	Es ist wichtig, die Rechte von Radikalen und Abweichlern in jeder Hinsicht zu wahren.					
11.	Was wir in unserem Land anstelle von mehr Bürgerrechten wirklich brauchen, ist eine anständige Portion Recht und Ordnung.					
12.	Die Abkehr von der Tradition wird sich eines Tages als fataler Fehler herausstellen.					

PARTEIEN

Wie sympathisch sind Ihnen die folgenden politischen Parteien?

Verwenden Sie für Ihre Antworten die folgende 7-stufige Skala:

-3 -2 -1 0 +1 +2 +3

Sehr unsympathisch — Neutral — Sehr sympathisch

		-3	-2	-1	0	+1	+2	+3
1.	CDU							
2.	CSU							
3.	SPD							
4.	Grüne							
5.	F.D.P.							
6.	Die Linke / PDS							
7.	REP							
8.	NPD/DVU							

LIRE
Links-Rechts Schema

Menschen haben eine bestimmte politische Grundhaltung, die man üblicherweise als eher "links", eher "Mitte" oder eher "rechts" bezeichnet.

Wie würden Sie Ihre persönliche politische Grundhaltung auf der folgenden Skala einordnen?

Mitte

links ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ **rechts**

AIST

Der Allgemeine-Interessen-Strukturtest (AIST) besteht aus einer Liste mit verschiedensten Tätigkeiten. Geben Sie bitte für jede einzelne davon an, wie sehr diese Sie interessiert bzw. interessieren würde.

Interessieren heißt: etwas gerne tun, etwas wegen der Sache selbst tun.

Sie können für jede Tätigkeit bis zu 5 Punkten vergeben, je nachdem, wie groß Ihr Interesse ist. Verschieden viele Punkte sollen bedeuten:

1 — Das interessiert mich gar nicht; das tue ich **nicht** gerne

2 — Das interessiert mich wenig

3 — Das interessiert mich **etwas**

4 — Das interessiert mich ziemlich

5 — Das interessiert mich sehr; das tue ich **sehr** gerne

		1	2	3	4	5
1.	.....					
2.	.....					
3.	.....					
4.	.....					
5.	.....					
6.	.....					
7.	.....					
8.	.....					
9.	.....					
10.	.....					
11.	.....					
12.	.....					
13.	.....					
14.	.....					
15.	.....					
16.	.....					
17.	.....					
18.	.....					
19.	.....					
20.	.....					

		1	2	3	4	5
21.	.....					
22.	.....					
23.	.....					
24.	.....					
25.	.....					
26.	.....					
27.	.....					
28.	.....					
29.	.....					
30.	.....					
31.	.....					
32.	.....					

33.	.....					
34.	.....					
35.	.....					
36.	.....					
37.	.....					
38.	.....					
39.	.....					
40.	.....					

		1	2	3	4	5
41.	.....					
42.	.....					
43.	.....					
44.	.....					
45.	.....					
46.	.....					
47.	.....					
48.	.....					
49.	.....					
50.	.....					
51.	.....					
52.	.....					
53.	.....					
54.	.....					
55.	.....					
56.	.....					
57.	.....					
58.	.....					
59.	.....					
60.	.....					

Verhalten im Alltag

1. Wie oft sind Sie **im letzten Monat** mit Freunden oder Ihrem Partner / Ihrer Partnerin **ausgegangen**?

[ ] mal

2. Wie oft haben Sie sich **im vergangenen Trimester (Oktober bis Dezember) mit Verwandten oder Freunden aus Ihrer Schulzeit** getroffen?

[ ] mal

3. Anzahl Freunde:
Bitte nehmen Sie sich für diese Frage etwas Zeit. Gemeint sind alle Personen (ohne Familienangehörige), die Sie gut kennen und mit denen Sie gegenseitige Sympathie, Vertrauen und Hilfsbereitschaft verbindet.

keine	1 bis 5	6 bis 10	11 bis 15	16 bis 20	21 bis 25	26 bis 30	über 30
○	○	○	○	○	○	○	○

4. Wie zufrieden Sind Sie mit Ihren sozialen Beziehungen (Kommilitonen, Freunde und Bekannte)?

Unzufrieden	1	2	3	4	5	Sehr zufrieden
	○	○	○	○	○	

5. Haben Sie in den letzten 3 Monaten ein Museum oder eine Ausstellung besucht? Wie oft?

kein mal	1 mal	2 mal	3 mal	4 mal	5 mal	öfter als 5 mal
○	○	○	○	○	○	○

6. Wann haben Sie Ihre Steuererklärung für **2006** abgegeben?

- ○ im Mai oder früher
- ○ Juni bis September
- ○ Oktober bis Dezember
- ○ später oder noch immer nicht
- ○ Frage unzutreffend / kann nicht beantwortet werden

7. Wenn Sie Autofahrer sind:Wieviel PS (ca.) hat Ihr Auto? Wenn Sie nur die kW-Zahl wissen: 1kW = 1.36 PS.

[ ] PS (0 für "habe kein Auto")

Die folgenden 6 Fragen richten sich nur an Studierende (aller Fächer)

8. Wie oft haben Sie im vergangenen Trimester von Kommilitonen Vorlesungsmitschriften etc. oder Kopien bekommen?

☐ mal (ca.)

9. Wie oft haben Sie selbst im vergangenen Trimester Vorlesungsmitschriften / Kopien für andere zur Verfügung gestellt?

☐ mal (ca.)

10. Wie oft kommen Sie unvorbereitet zu einer Veranstaltung?

○ ○ ○ ○ ○ ○ ○

nie manchmal sehr häufig

11. Kommt es vor, daß Sie ein Studieninhalt so fesselt, daß Sie über das Pflichtpensum hinausgehende Informationen und Literatur suchen und lesen?

○ ○ ○ ○ ○ ○ ○

nie manchmal sehr häufig

12. Verglichen mit meinen Kommilitonen/innen komme ich in einer Prüfungssituation

○ ○ ○ ○ ○

Viel weniger als andere in Stress | Nicht mehr und nicht weniger | Viel mehr als andere in Stress

13. Wenn bei mir der Adrenalin-Pegel vor einer Klausur steigt, fördert das nur meine Konzentration und Leistung.

Stimmt überhaupt nicht 1 2 3 4 5 Stimmt voll und ganz

○ ○ ○ ○ ○

---

14. Fitness / Sport:
Wie oft gehen Sie normalerweise ins Fitness-Studio oder treiben sonstigen Sport um sich fit zu halten (jew. 1 Stunde oder mehr):

○ ○ ○ ○

mehrmals/Woche | 1mal / Woche | seltener | nie

Ernährung: Wie oft essen Sie ...

15. eine **fleischlose** Sauce zu Spaghetti und anderen Nudeln

○ ○ ○ ○

normalerweise | oft | manchmal | selten oder nie

16. Fisch oder Geflügel statt **rotem** Fleisch

○ ○ ○ ○

normalerweise | oft | manchmal | selten oder nie

17. ein vegetarisches Gericht

○ ○ ○ ○

normalerweise | oft | manchmal | selten oder nie

18. Gemüse

○ ○ ○ ○

7+ mal / Woche | 4-6 mal / Woche | 2-3 mal / Woche | seltener oder nie

19. grünen Salat oder Tomatensalat

○ ○ ○ ○

7+ mal / Woche | 4-6 mal / Woche | 2-3 mal / Woche | seltener oder nie

20. Obst

○ ○ ○ ○

7+ mal / Woche | 4-6 mal / Woche | 2-3 mal / Woche | seltener oder nie

21. Müsli und andere ballaststoffreiche Zerealien

○ ○ ○ ○

7+ mal / Woche | 4-6 mal / Woche | 2-3 mal / Woche | seltener oder nie

Bitte beachten Sie, daß Sie Ihre Antworten nur dann erfolgreich absenden können, wenn Sie (fast) alle Fragen beantwortet haben!

Hier ist noch Raum für Kommentare:

Wenn Sie das möchten, können Sie mir hier Ihre Email-Adresse mitteilen:

Vielen Dank für Ihre Mitarbeit!

Abschicken

Nach dem Absenden folgt eine Eingangsbestätigung - bitte etwas Geduld ....

C.2 ESF-Projekt Jahrgang 2009

ESF-Projekt Online-Fragebogen

UniBwM Fak Päd - Schmolck - März 2009 *ESF-Projekt FT09*

ESF-Projekt Fragebogen

1. Kennwort

Diese Erhebung ist anonym. Damit Sie später Ihre persönliche Testauswertung identifizieren können, brauchen Sie ein unverwechselbares, nur Ihnen bekanntes Kennwort. Bilden Sie Ihr persönliches Kennwort bitte nach folgender Anweisung:

1-2:	1. und 2. Buchstabe des Geburtsnamens Ihrer Mutter	**Kennwort**
3-4:	Geburts**tag** Ihrer Mutter	
5-6:	Geburts**monat** Ihrer Mutter	
7-8:	Die beiden **letzten** Buchstaben Ihres eigenen Geburtsortes	

2. Studienfach an der UniBwM

BAU BW EIT ETTI GEO INF LRT MB MBA ME PÄD SPO SWI WINF WOW

Studiere nicht an der UniBwM, sondern ... (z.B. *"studiere BWL an der LMU"*; oder *"berufstätig als Bankkauffrau"*)

3. Studienjahrgang

2004 2005 2006 2007 2008 Studiere nicht an der UniBwM oder früherer Jahrgang

4. TSK - welche Uniform tragen Sie?

Heer (alter Ausbildungsgang) Heer (neuer Ausbildungsgang) Truppengattung Heer

Lufwaffe Marine

nicht zutreffend / kein Soldat

5. Soldent-Studat:

Geben Sie bitte auf einer Skala von 0 bis 100 % an, ob Sie sich während Ihrer 3+ Jahre an der UniBw 100%ig als Student/in ... oder mehr als zum Studium abkommandierter Soldat (also zu 0% Student/in) sehen.
Lassen Sie diese Frage unbeantwortet, wenn Sie nicht an der UniBw studieren!

%ig Student(in)

6. Alter

unter 20 Jahre 22 Jahre 25 Jahre
20 Jahre 23 Jahre 26 Jahre
21 Jahre 24 Jahre über 26 Jahre

7. Geschlecht

männlich weiblich

Die folgenden Fragen zur Studienfachwahl richten sich nur an Studierende und Hochschulabsolventen

1. Welche der folgenden Fächer haben Sie in der Phase der Berufs- und Studienwahl in Erwägung gezogen?

		1 überhaupt nicht erwogen	2	3	4 sehr ernsthaft erwogen bzw. gewählt
01	Architektur	○	○	○	○
02	Bauingenieurwesen	○	○	○	○
03	Biologie	○	○	○	○
04	Betriebswirtschaftslehre	○	○	○	○
05	Elektrotechnik	○	○	○	○
06	Erziehungswissenschaft / Pädagogik	○	○	○	○
07	Geoinformatik	○	○	○	○
08	Germanistik	○	○	○	○
09	Geschichte	○	○	○	○
10	Informatik	○	○	○	○
11	Jura	○	○	○	○
12	Maschinenbau	○	○	○	○
13	Mathematik	○	○	○	○
14	Medizin	○	○	○	○
15	Physik	○	○	○	○
16	Politikwissenschaften	○	○	○	○
17	Psychologie	○	○	○	○
18	Publizistik	○	○	○	○
19	Soziologie	○	○	○	○
20	Sportwissenschaft	○	○	○	○
21	Wirtschafts- und Organisationswissenschaften	○	○	○	○
22	anderes Fach, und zwar	○	○	○	○
23	anderes Fach, und zwar	○	○	○	○

2. Ist das von Ihnen gewählte Studium Ihr Wunschstudium gewesen?

○ Ja

○ Nein, mein Wunschstudium wäre _____ gewesen

BFI

Machen Sie diesen psychologischen Test, um etwas über Ihre Persönlichkeit zu erfahren! Dieser Test misst, was nach Ansicht vieler Psychologen die fünf grundlegenden Dimensionen der Persönlichkeit sind.

Bei den folgenden Aussagen geht es darum, wie Sie sich selbst sehen. Ein Beispiel: Sind Sie der Ansicht, daß Sie gerne Zeit mit anderen verbringen? Bitte entscheiden Sie, inwieweit die jeweilige Aussage für Sie zutrifft. Es gibt keine richtigen oder falschen Antworten, allerdings werden Sie keine zutreffenden Ergebnisse erhalten, wenn Sie die Fragen nicht ernsthaft und wahrheitsgemäß beantworten.

Ihre Aufgabe ist es, für jede Aussage auf einer Skala von -2 bis +2 anzuklicken, wie sehr die Aussage für Sie zutrifft oder nicht zutrifft.

-2	-1	0	+1	+2
Sehr unzutreffend	Unzutreffend	teils/teils	Zutreffend	Sehr zutreffend

	Ich sehe mich selbst als jemand, der ...	-2	-1	0	+1	+2
1.	...zuverlässig und gewissenhaft arbeitet	○	○	○	○	○
2.	...gerne Überlegungen anstellt, mit Ideen spielt	○	○	○	○	○
3.	...aus sich herausgeht, gesellig ist	○	○	○	○	○
4.	...leicht nervös und unsicher wird	○	○	○	○	○
5.	...nicht den selben Regeln und Gesetzen unterworfen sein sollte, wie die meisten Menschen	○	○	○	○	○
6.	...praktische Lösungen lieber mag als theoretische Diskussionen	○	○	○	○	○
7.	...sich viele Sorgen macht	○	○	○	○	○

8.	...nur wenig künstlerische Interessen hat					
9.	...Aufgaben gründlich erledigt					
10.	...dem eine hohe gesellschaftliche Stellung etwas bedeutet					
11.	...begeisterungsfähig ist, andere mitreißen kann					
12.	...oft Krach mit anderen hat.					
13.	...ruhig bleibt, selbst in Stresssituationen					
14.	...bequem ist und zur Faulheit neigt					

	Ich sehe mich selbst als jemand, der ...	-2	-1	0	+1	+2
15.	...eher zurückhaltend und reserviert ist					
16.	...künstlerische und ästhetische Eindrücke schätzt					
17.	...eher still und wortkarg ist					
18.	...rücksichtsvoll und einfühlsam zu anderen ist					
19.	...dazu neigt, unordentlich zu sein					
20.	...den Luxus liebt und das auch zeigt					
21.	...lieber kooperiert als konkurriert					
22.	...emotional ausgeglichen und nicht leicht aus der Fassung zu bringen ist					
23.	...mehr Respekt verdient als der durchschnittliche Mensch					
24.	...eine aktive Vorstellungskraft hat, phantasievoll ist					
25.	...schroff und abweisend zu anderen sein kann					
26.	...nicht zu wilden Spekulationen neigt					
27.	...gern ohne Umschweife an die Sache geht					

STOMP
Short Test of Music Preferences

Sag mir was du hörst, und ich sag dir, wer du bist. Psychologen der Universität von Texas fanden heraus, dass der Musikgeschmack Rückschlüsse auf die Persönlichkeit zulässt. "Musik durchdringt viele Bereiche unseres täglichen Lebens, wir hören sie im Auto am Weg zur Arbeit, entspannt zu Hause oder mit Freunden an der Bar", meint Sam Gosling. "Nahezu jeder hört auf irgendeine Weise Musik. Unsere Untersuchungen ergaben, dass die Persönlichkeit eine wesentliche Rolle bei der Wahl der Musik spielt".

Geben Sie für jedes der unten aufgeführten Musikgenre an, wie gerne oder ungern Sie diese Art von Musik hören. Verwenden Sie für Ihre Antworten die folgende 7-stufige Skala:

-3 -2 -1 0 +1 +2 +3

Mag ich überhaupt nicht — Neutral / keine Meinung — Mag ich sehr

		-3	-2	-1	0	+1	+2	+3
1.	Klassik							
2.	Blues							
3.	Country							
4.	Folk							
5.	Reggae							
6.	Religiöse Musik							
7.	Rap/hip-hop							
8.	Heavy Metal							
9.	Soul/R&B							
10.	Elektronische Musik							
11.	Traditionelle Volksmusik							

		-3	-2	-1	0	+1	+2	+3
12.	Populäre Volksmusik							
13.	Schlager							
14.	NDW							
15.	Alternative							
16.	Techno							
17.	Punk							
18.	Jazz							

19.	Rock							
20.	Pop							
21.	Oldies							
22.	Filmmusik/Titelmelodien							

Hörproben für vier Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.
(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

		-3	-2	-1	0	+1	+2	+3
1.	MusicDimension1.mp3							
2.	MusicDimension2.mp3							
3.	MusicDimension3.mp3							
4.	MusicDimension4.mp3							

SDO
Soziale Dominanzorientierung

Im Folgenden finden Sie eine Reihe von Aussagen darüber, in welcher Beziehung gesellschaftliche Gruppen zueinander stehen sollten. Gesellschaftliche Gruppen können dabei z.B. ethnische Gruppen, politische Gruppen, religiöse Gruppen, Berufsgruppen oder auch die beiden Geschlechter sein. Bitte geben Sie an, wie stark Sie persönlich den Aussagen zustimmen.

Verwenden Sie für Ihre Antworten die folgende 7-stufige Skala:

-3 -2 -1 0 +1 +2 +3

Stimme ich überhaupt nicht zu — Teils/teils — Stimme voll und ganz zu

		-3	-2	-1	0	+1	+2	+3
1.	Wir würden mehr Probleme schaffen als lösen, wenn wir alle Gruppen gleich behandeln würden.							
2.	Wir sollten unser Möglichstes tun, um die Bedingungen für die unterschiedlichen Gruppen anzugleichen.							
3.	Um im Leben vorwärts zu kommen, ist es manchmal notwendig, auf anderen Gruppen herum zu treten.							
4.	Keine Gruppe von Menschen ist mehr wert als irgendeine andere Gruppe.							
5.	Es macht keinen Sinn, die Einkommen so gleich wie möglich zu gestalten.							
6.	Wenn bestimmte Gruppen unter sich bleiben würden, hätten wir weniger Probleme.							
7.	Es ist ein echtes Problem, daß bestimmte Gruppen oben sind und andere unten.							
8.	Keine einzelne Gruppe sollte in der Gesellschaft dominieren.							

RWA

Es folgen einige Aussagen, die sich auf verschiedene politische und gesellschaftliche Themen beziehen. Wie ist Ihre persönliche Meinung dazu?
Verwenden Sie für Ihre Antworten die folgende 5-stufige Skala:

-2 -1 0 +1 +2

Stimme überhaupt nicht zu — Teils/teils — Stimme voll und ganz zu

		-2	-1	0	+1	+2
1.	Es ist gut, dass die jungen Leute heutzutage größere Freiheiten haben, ihr eigenes Ding zu machen und gegen Dinge zu protestieren, die sie nicht mögen.					
2.	Die wahren Schlüssel zum guten Leben sind Gehorsam, Disziplin und Tugend.					
3.	Man sollte alten, traditionellen Glaubenssätzen weniger Beachtung schenken und stattdessen seine persönlichen Moralvorstellungen über Gut und Böse entwickeln.					

4.	Es gibt kein Verbrechen, das die Todesstrafe rechtfertigen würde.	○	○	○	○	○
5.	Gehorsam und Achtung vor der Autorität sind die wichtigsten Tugenden, die Kinder lernen sollten.	○	○	○	○	○
6.	Tugendhaftigkeit und Gesetzestreue bringen uns auf lange Sicht weiter als das ständige Infragestellen der Grundfesten unserer Gesellschaft.	○	○	○	○	○
7.	Gleichgeschlechtliche Lebensgemeinschaften sollten der Ehe in jeder Hinsicht gleichgestellt werden.	○	○	○	○	○
8.	Die Zeiten, in denen Frauen sich unterordnen mussten, sollten ein für alle Mal vorbei sein. Der Platz einer Frau in der Gesellschaft sollte sein, wo immer sie möchte.	○	○	○	○	○
9.	Was unser Land wirklich braucht, ist ein starker, entschlossener Kanzler, der uns wieder auf unseren richtigen Weg bringt.	○	○	○	○	○
10.	Es ist wichtig, die Rechte von Radikalen und Abweichlern in jeder Hinsicht zu wahren.	○	○	○	○	○
11.	Was wir in unserem Land anstelle von mehr Bürgerrechten wirklich brauchen, ist eine anständige Portion Recht und Ordnung.	○	○	○	○	○
12.	Die Abkehr von der Tradition wird sich eines Tages als fataler Fehler herausstellen.	○	○	○	○	○

PARTEIEN

Wie sympathisch sind Ihnen die folgenden politischen Parteien?

Verwenden Sie für Ihre Antworten die folgende 7-stufige Skala:

-3 -2 -1 0 +1 +2 +3

Sehr unsympathisch — Neutral — Sehr sympathisch

		-3	-2	-1	0	+1	+2	+3
1.	CDU	○	○	○	○	○	○	○
2.	CSU	○	○	○	○	○	○	○
3.	SPD	○	○	○	○	○	○	○
4.	Grüne	○	○	○	○	○	○	○
5.	F.D.P.	○	○	○	○	○	○	○
6.	Die Linke	○	○	○	○	○	○	○
7.	REP	○	○	○	○	○	○	○
8.	NPD/DVU	○	○	○	○	○	○	○

LIRE
Links-Rechts Schema

Menschen haben eine bestimmte politische Grundhaltung, die man üblicherweise als eher "links", eher "Mitte" oder eher "rechts" bezeichnet.

Wie würden Sie Ihre persönliche politische Grundhaltung auf der folgenden Skala einordnen?

Mitte

links ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ **rechts**

AIST

Der Allgemeine-Interessen-Strukturtest (AIST) besteht aus einer Liste mit verschiedensten Tätigkeiten. Geben Sie bitte für jede einzelne davon an, wie sehr diese Sie interessiert bzw. interessieren würde.

Interessieren heißt: etwas gerne tun, etwas wegen der Sache selbst tun.

Sie können für jede Tätigkeit bis zu 5 Punkten vergeben, je nachdem, wie groß Ihr Interesse ist. Verschieden viele Punkte sollen bedeuten:

1	2	3	4	5
Das interessiert mich gar nicht; das tue ich **nicht** gerne	Das interessiert mich wenig	Das interessiert mich **etwas**	Das interessiert mich ziemlich	Das interessiert mich sehr; das tue ich **sehr** gerne

		1	2	3	4	5

1.	.....					
2.	.....					
3.	.....					
4.	.....					
5.	.....					
6.	.....					
7.	.....					
8.	.....					
9.	.....					
10.	.....					
11.	.....					
12.	.....					
13.	.....					
14.	.....					
15.	.....					
16.	.....					
17.	.....					
18.	.....					
19.	.....					
20.	.....					

		1	2	3	4	5
21.	.....					
22.	.....					
23.	.....					
24.	.....					
25.	.....					
26.	.....					
27.	.....					
28.	.....					
29.	.....					
30.	.....					
31.	.....					
32.	.....					
33.	.....					
34.	.....					
35.	.....					
36.	.....					
37.	.....					
38.	.....					
39.	.....					
40.	.....					

		1	2	3	4	5
41.	.....					
42.	.....					
43.	.....					
44.	.....					
45.	.....					
46.	.....					

47.	.....	○	○	○	○	○
48.	.....	○	○	○	○	○
49.	.....	○	○	○	○	○
50.	.....	○	○	○	○	○
51.	.....	○	○	○	○	○
52.	.....	○	○	○	○	○
53.	.....	○	○	○	○	○
54.	.....	○	○	○	○	○
55.	.....	○	○	○	○	○
56.	.....	○	○	○	○	○
57.	.....	○	○	○	○	○
58.	.....	○	○	○	○	○
59.	.....	○	○	○	○	○
60.	.....	○	○	○	○	○

Verhalten im Alltag

Unternehmungen - *Wie oft besuchten Sie im **letzten halben Jahr** ? ..*

1. Kino	☐ mal
2. Disco oder **Party**	☐ mal
3. Rock- oder **Popkonzerte**	☐ mal
4. Theater oder klassische Musikveranstaltungen	☐ mal
5. Museum oder **Ausstellung**	☐ mal
6. Kneipe oder **Cafe**	☐ mal
7. Sportstadion	☐ mal

8. Wie viele Leute laden Sie üblicherweise zu Ihrer Geburtstagsfeier ein?

☐ Personen

9. Wie zufrieden Sind Sie mit Ihren sozialen Beziehungen (Kommilitonen, Kameraden, Freunde und Bekannte)?

Unzufrieden 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Sehr zufrieden

10. Pünktlichkeit: In meinem Freundeskreis bin ich bekannt als jemand, der

○	○	○	○	○
*über*pünktlich	pünktlich	nahezu pünktlich	mal so mal so	selten pünktlich ist

11. Rauchen: Sind Sie ?

○	○	○	○	○
Nichtraucher	Gelegenheits-Raucher	leichter Raucher	mäßiger Raucher	starker Raucher

12. Alkohol: trinken Sie ?

○	○	○	○	○
gar nicht	selten	gelegentlich	häufig	sehr häufig

13. Fitness / Sport:

Wie oft gehen Sie normalerweise ins Fitness-Studio oder treiben sonstigen Sport um, sich fit zu halten (jew. 1 Stunde oder mehr):

täglich	mehrmals / Woche	1mal / Woche	seltener	nie
○	○	○	○	○

14. Gehen Sie zum **Blutspenden?**

regelmäßig	gelegentlich	selten	nie	aus gesundheitlichen Gründen nicht möglich
○	○	○	○	○

15. Ihre **Lieblingsfarbe**:

- ○ Blau
- ○ Rot
- ○ Gelb
- ○ Grün
- ○ Andere, nämlich: [ ]

16. Wann haben Sie Ihre Steuererklärung für **2007** abgegeben?

- ○ April oder früher
- ○ im Mai
- ○ Juni bis September
- ○ Oktober bis Dezember
- ○ später oder noch immer nicht
- ○ Frage unzutreffend / kann nicht beantwortet werden

17. Wenn Sie Autofahrer sind:Wieviel PS (ca.) hat Ihr Auto? Wenn Sie nur die kW-Zahl wissen: 1kW = 1.36 PS.

[ ] PS (0 für "habe kein Auto")

Die folgenden 6 Fragen richten sich nur an Studierende (aller Fächer)

18. Wie oft haben Sie im vergangenen Trimester von Kommilitonen Vorlesungsmitschriften etc. oder Kopien bekommen?

[ ] mal (ca.)

19. Wie oft haben Sie selbst im vergangenen Trimester Vorlesungsmitschriften / Kopien für andere zur Verfügung gestellt?

[ ] mal (ca.)

20. Wie oft kommen Sie unvorbereitet zu einer Veranstaltung?

nie			manchmal			sehr häufig
○	○	○	○	○	○	○

21. Kommt es vor, daß Sie ein Studieninhalt so fesselt, daß Sie über das Pflichtpensum hinausgehende Informationen und Literatur suchen und lesen?

nie			manchmal			sehr häufig
○	○	○	○	○	○	○

22. Verglichen mit meinen Kommilitonen/innen komme ich in einer Prüfungssituation

Viel weniger als andere in Stress		Nicht mehr und nicht weniger		Viel mehr als andere in Stress
○	○	○	○	○

Bitte beachten Sie, daß Sie Ihre Antworten nur dann erfolgreich absenden können, wenn Sie (fast) alle Fragen beantwortet haben!

Hier ist noch Raum für Kommentare:

Wenn Sie das möchten, können Sie mir hier Ihre Email-Adresse mitteilen:

Vielen Dank für Ihre Mitarbeit!

Abschicken

Nach dem Absenden folgt eine Eingangsbestätigung - bitte etwas Geduld ....

C.3 ESF-Projekt Jahrgang 2011

Anzeigeoptionen
Info: Hier können Sie optional die Anzeigeoptionen verändern. Wenn Sie eine Sprache auswählen, die keine eigenen Textelemente hat, werden die Textelemente der Standardsprache angezeigt.

Anzeigeoptionen einstellen:
- [ ] Filter anzeigen
- [ ] Pretest-Kommentare anzeigen
- [ ] Todos anzeigen
- [ ] Trigger anzeigen
- [ ] Plausichecks anzeigen
- [ ] Randomisierung abschalten
- [ ] Interne Verlinkungen ausblenden
- [ ] Nur den Fragebogen ausdrucken

Sprache: Deutsch

Einstellungen speichern

Informationen zur Umfrage SWM_2011_kurz

Umfrage-Nr.	208226
Autor	Jörg-Henrik Heine
Mitarbeiter	
Start	2011-02-25 00:00:00
Ende	2011-12-31 00:00:00

Fragebogen

1 [Seiten-ID: 1097576] [L]
Eingangsseite
UniBwM Fak Päd - Tarnai *SWM-Projekt WT11*

SWM-Projekt Fragebogen

Projekt im WT 2011

Dieses Projekt ist Bestandteil der Lehrveranstaltungen der *Professur Sozialwissenschaftliche Methodenlehre*, das regelmäßig in den *Einführungen* durchgeführt und ausgewertet wird. Die Studierenden lernen dabei die Anwendung von Erhebungsinstrumenten, Tests und Skalen zunächst aus der Befragtenperspektive kennen.

Im Projekt wollen wir gemeinsam mit den Studierenden klären, wie man die Genauigkeit und Aussagekraft der Untersuchungsinstrumente kritisch überprüfen kann. Die Bedeutung der sog. Reliabilität oder formalen Messgenauigkeit wird demonstriert. Interessanter, aber auch viel schwieriger zu prüfen ist die Frage, was ein Test tatsächlich inhaltlich misst, und welche Schlussfolgerungen aufgrund der Testergebnisse möglich sind.

Der SWM-Projekt Fragebogen enthält neben Kurzformen psychologischer Tests zu Persönlichkeit, Interessen, Einzelfragen zum Alltagsverhalten und als Schwerpunkt der Befragung **Hörproben und Fragen zum Musikgeschmack**.

Wie schon letztes Jahr wollen wir ausdrücklich auch Freunde/innen der Studierenden und andere Interessenten an unserer Testerhebung einladen mitzumachen.

Achtung: Manche Fragen müssen beantwortet werden, damit der Fragebogen verschickt werden kann. Es empfiehlt sich daher, von Anfang an wirklich alle Fragen zu beantworten!

2 [Seiten-ID: 1097577] [L]
Frage 1
UniBwM Fak Päd - Tarnai *SWM-Projekt WT11*

SWM-Projekt Fragebogen

Kennwort

Diese Erhebung ist anonym. Damit die Testauswertung eindeutig ist und mit künftigen oder früheren Erhebungen verglichen werden kann, geben Sie ein unverwechselbares, nur Ihnen bekanntes Kennwort an. Bilden Sie Ihr persönliches Kennwort bitte nach folgender Anweisung:

1-2: 1. und 2. Buchstabe des Geburtsnamens Ihrer Mutter
3-4: Geburts**tag** Ihrer Mutter (z.B. 03)
5-6: Geburts**monat** Ihrer Mutter (z.B. 06)
7-8: Die beiden **letzten** Buchstaben Ihres eigenen Geburtsortes
Kennwort

3 [Seiten-ID: 1097578] [L]
Frage 2

Studienfach bzw. Fakultät an der UniBwM

Studienfach bzw. Fakultät an der UniBwM
- ○ BAU
- ○ BW
- ○ EIT
- ○ ETTI
- ○ GEO
- ○ INF
- ○ LRT
- ○ MB
- ○ MBA
- ○ ME
- ○ PÄD
- ○ SPO
- ○ SWI
- ○ WINF
- ○ WOW
- ○ Studiere nicht an der UniBwM, sondern ... (z.B. "studiere BWL an der LMU"; oder "berufstätig als Bankkauffrau")

4 [Seiten-ID: 1097579] [L]

Frage 3

Studienjahrgang

2006 | 2007 | 2008 | 2009 | 2010 | Studiere nicht an der UniBwM oder früherer Jahrgang

5 [Seiten-ID: 1097580] [L]

Frage 5

Soldent-Studat:

Geben Sie bitte auf einer Skala von 0 bis 100 % an, ob Sie sich während Ihrer 3+ Jahre an der UniBw 100%ig als Student/in ... oder mehr als zum Studium abkommandierter Soldat (also zu 0% Student/in) sehen.

Lassen Sie diese Frage unbeantwortet, wenn Sie nicht an der UniBw studieren!

%ig Student(in)

6 [Seiten-ID: 1097581] [L]

Frage 6

Alter

unter 20 Jahre	22 Jahre	25 Jahre
20 Jahre	23 Jahre	26 Jahre
21 Jahre	24 Jahre	über 26 Jahre

7 [Seiten-ID: 1097582] [L]

Frage 7

Geschlecht

männlich | weiblich

8 [Seiten-ID: 1097583] [L]

Frage 8

Die folgenden Fragen zur Studienfachwahl richten sich nur an Studierende und Hochschulabsolventen

1. Welche der folgenden Fächer haben Sie in der Phase der Berufs- und Studienwahl in Erwägung gezogen?

	1 überhaupt nicht erwogen	2	3	4 sehr ernsthaft erwogen bzw. gewählt
Architektur				
Bauingenieurwesen				
Biologie				
Betriebswirtschaftslehre				
Elektrotechnik				
Erziehungswissenschaft / Pädagogik				
Geoinformatik				
Germanistik				
Geschichte				
Informatik				
Jura				
Maschinenbau				
Mathematik				
Medizin				
Physik				
Politikwissenschaften				
Psychologie				
Publizistik				
Soziologie				
Sportwissenschaft				
Wirtschafts- und Organisationswissenschaften				
anderes Fach, und zwar				
anderes Fach, und zwar				

2. Ist das von Ihnen gewählte Studium Ihr Wunschstudium gewesen?

Ja

Nein, mein Wunschstudium wäre folgendes gewesen

9 [Seiten-ID: 1097584] [L]

Frage 10

BFI

Machen Sie diesen psychologischen Test, um etwas über Ihre Persönlichkeit zu erfahren! Dieser Test misst, was nach Ansicht vieler Psychologen die fünf grundlegenden Dimensionen der Persönlichkeit sind.

Bei den folgenden Aussagen geht es darum, wie Sie sich selbst sehen. Ein Beispiel: Sind Sie der Ansicht, daß Sie gerne Zeit mit anderen verbringen? Bitte entscheiden Sie, inwieweit die jeweilige Aussage für Sie zutrifft. Es gibt keine richtigen oder falschen Antworten, allerdings werden keine zutreffenden Ergebnisse erhalten, wenn Sie die Fragen nicht ernsthaft und wahrheitsgemäß beantworten.

Ihre Aufgabe ist es, für jede Aussage auf einer Skala von -2 bis +2 anzuklicken, wie sehr die Aussage für Sie zutrifft oder nicht zutrifft.

Ich sehe mich selbst als jemand, der ...

	Sehr unzutreffend	Unzutreffend	teils/teils	Zutreffend	Sehr zutreffend
...zuverlässig und gewissenhaft arbeitet					
...gerne Überlegungen anstellt, mit Ideen spielt					
...aus sich herausgeht, gesellig ist					
...leicht nervös und unsicher wird					
...nicht den selben Regeln und Gesetzen unterworfen sein sollte, wie die meisten Menschen					
...praktische Lösungen lieber mag als theoretische Diskussionen					
...sich viele Sorgen macht					
...nur wenig künstlerische Interessen hat					
...Aufgaben gründlich erledigt					
...dem eine hohe gesellschaftliche Stellung etwas bedeutet					
...begeisterungsfähig ist, andere mitreißen kann					
...oft Krach mit anderen hat.					
...ruhig bleibt, selbst in Stresssituationen					
...bequem ist und zur Faulheit neigt					
...eher zurückhaltend und reserviert ist					
...künstlerische und ästhetische Eindrücke schätzt					
...eher still und wortkarg ist					
...rücksichtsvoll und einfühlsam zu anderen ist					
...dazu neigt, unordentlich zu sein					
...den Luxus liebt und das auch zeigt					
...lieber kooperiert als konkurriert					
...emotional ausgeglichen und nicht leicht aus der Fassung zu bringen ist					
...mehr Respekt verdient als der durchschnittliche Mensch					
...eine aktive Vorstellungskraft hat, phantasievoll ist					
...schroff und abweisend zu anderen sein kann					
...nicht zu wilden Spekulationen neigt					
...gern ohne Umschweife an die Sache geht					

10 [Seiten-ID: 1097585] [L]
Frage 11

STOMP
Short Test of Music Preferences

Sag mir was du hörst, und ich sag dir, wer du bist. Psychologen der Universität von Texas fanden heraus, dass der Musikgeschmack Rückschlüsse auf die Persönlichkeit zulässt. "Musik durchdringt viele Bereiche unseres täglichen Lebens, wir hören sie im Auto am Weg zur Arbeit, entspannt zu Hause oder mit Freunden an der Bar", meint Sam Gosling. "Nahezu jeder hört auf irgendeine Weise Musik. Unsere Untersuchungen ergaben, dass die Persönlichkeit eine wesentliche Rolle bei der Wahl der Musik spielt".

Geben Sie für jedes der unten aufgeführten Musikgenre an, wie gerne oder ungern Sie diese Art von Musik hören. Verwenden Sie für Ihre Antworten die folgende 7-stufige Skala:

	Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
Klassik							
Blues							
Country							
Folk							
Reggae							
Religiöse Musik							
Rap/hip-hop							
Heavy Metal							
Soul/R&B							
Elektronische Musik							
Traditionelle Volksmusik							
Populäre Volksmusik							
Schlager							
NDW - Neue Deutsche Welle							
Alternative							
Techno							
Punk							
Jazz							
Rock							
Pop							
Oldies							
Filmmusik/Titelmelodien							

11 [Seiten-ID: 1097586] [L]
Frage 11_MB1

Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

12 [Seiten-ID: 1097587] [L]
Frage 11_MB2
Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

13 [Seiten-ID: 1097588] [L]
Frage 11_MB3
Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

14 [Seiten-ID: 1097589] [L]
Frage 11_MB4
Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

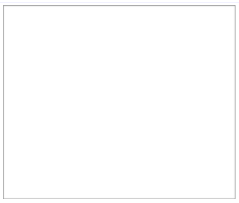
(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

15 [Seiten-ID: 1097590] [L]
Frage 11_MB5
Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○



16 [Seiten-ID: 1097591] [L]

Frage 11_MB6

Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

17 [Seiten-ID: 1097592] [L]

Frage 11_MB7

Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

18 [Seiten-ID: 1097593] [L]

Frage 11_MB8

Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

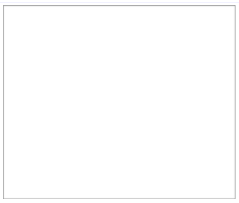
19 [Seiten-ID: 1097594] [L]

Frage 11_MB9

Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○



20 [Seiten-ID: 1097595] [L]
Frage 11_MB10
Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

21 [Seiten-ID: 1097596] [L]
Frage 11_MB11
Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

22 [Seiten-ID: 1097597] [L]
Frage 11_MB12
Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

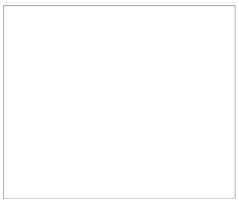
(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

23 [Seiten-ID: 1097598] [L]
Frage 11_MB13
Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○



24 [Seiten-ID: 1097599] [L]

Frage 11_MB14

Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

25 [Seiten-ID: 1097600] [L]

Frage 11_MB15

Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

26 [Seiten-ID: 1097601] [L]

Frage 11_MB16

Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

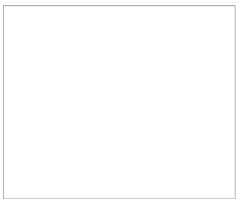
27 [Seiten-ID: 1097602] [L]

Frage 11_MB17

Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○



28 [Seiten-ID: 1097603] [L]
Frage 11_MB18
Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

29 [Seiten-ID: 1097604] [L]
Frage 11_MB19
Hörproben für verschiedene Dimensionen des Musikgeschmacks: Anklicken, Anhören und einstufen, wie Ihnen diese Art von Musik gefällt.

(Bitte freilassen, falls das Anhören auf Ihrem Computer nicht funktioniert.)

Mag ich überhaupt nicht			Neutral / keine Meinung			Mag ich sehr
○	○	○	○	○	○	○

30 [Seiten-ID: 1097605] [L]
Frage 15

LIRE
Links-Rechts-Schema

Menschen haben eine bestimmte politische Grundhaltung, die man üblicherweise als eher "links", eher "Mitte" oder eher "rechts" bezeichnet.

Wie würden Sie Ihre persönliche politische Grundhaltung auf der folgenden Skala einordnen?

links				eher	Mitte				rechts
○	○	○	○	○	○	○	○	○	○

31 [Seiten-ID: 1097606] [L]
Frage 16

AIST

Der Allgemeine-Interessen-Strukturtest (AIST) besteht aus einer Liste mit verschiedensten Tätigkeiten. Geben Sie bitte für jede einzelne davon an, wie sehr diese Sie interessiert bzw. interessieren würde.

Interessieren heißt: etwas gerne tun, etwas wegen der Sache selbst tun. Sie können für jede Tätigkeit bis zu 5 Punkten vergeben, je nachdem, wie groß Ihr Interesse ist. Verschieden viele Punkte sollen bedeuten:

	1 Das interessiert mich gar nicht; das tue ich nicht gerne	2 Das interessiert mich wenig	3 Das interessiert mich etwas	4 Das interessiert mich ziemlich	5 Das interessiert mich sehr; das tue ich sehr gerne
.....	○	○	○	○	○
.....	○	○	○	○	○
.....	○	○	○	○	○
.....	○	○	○	○	○
.....	○	○	○	○	○
.....	○	○	○	○	○
.....	○	○	○	○	○
.....	○	○	○	○	○
.....	○	○	○	○	○
.....	○	○	○	○	○

32 [Seiten-ID: 1097607] [L]
Frage 17

Verhalten im Alltag

Unternehmungen - Wie oft besuchten Sie im **letzten halben Jahr** ? .. Anzahl der Besuche eintragen!

1.Kino
2. Disco oder **Party**
3. Rock- oder **Popkonzerte**

4. Theater oder **klassische Musikveranstaltungen**
5. Museum oder **Ausstellung**
6. Kneipe oder **Cafe**
. Sportstadion

33 [Seiten-ID: 1097608] [L]

Frage 18

Wie viele Leute laden Sie üblicherweise zu Ihrer Geburtstagsfeier ein?

Personen

34 [Seiten-ID: 1097609] [L]

Frage 19

Wie zufrieden Sind Sie mit Ihren sozialen Beziehungen (Kommilitonen, Kameraden, Freunde und Bekannte)?

Unzufrieden 1	2	3	4	Sehr zufrieden 5

35 [Seiten-ID: 1097610] [L]

Frage 20

Pünktlichkeit:

In meinem Freundeskreis bin ich bekannt als jemand, der

überpünktlich	pünktlich	nahezu pünktlich	mal so mal so	selten pünktlich ist

36 [Seiten-ID: 1097611] [L]

Frage 21

Rauchen: Sind Sie ?

Nichtraucher	Gelegenheits-raucher	leichter Raucher	mäßiger Raucher	starker Raucher

37 [Seiten-ID: 1097612] [L]

Frage 22

Alkohol:

Trinken Sie ?

gar nicht	selten	gelegentlich	häufig	sehr häufig

38 [Seiten-ID: 1097613] [L]

Frage 23

Fitness / Sport:

Wie oft gehen Sie normalerweise ins Fitness-Studio oder treiben sonstigen Sport um, sich fit zu halten (jew. 1 Stunde oder mehr):

täglich	mehrmals / Woche	1mal / Woche	seltener	nie

39 [Seiten-ID: 1097614] [L]

Frage 24

Gehen Sie zum Blutspenden?

regelmäßig	gelegentlich	selten	nie	aus gesundheitlichen Gründen nicht möglich

40 [Seiten-ID: 1097615] [L]

Frage 25

Ihre Lieblingsfarbe:

- Blau
- Rot
- Gelb
- Grün
- Andere, nämlich:

41 [Seiten-ID: 1097616] [L]

Frage 26

Wann haben Sie Ihre Steuererklärung für 2009 abgegeben?

- April oder früher
- im Mai
- Juni bis September
- Oktober bis Dezember
- später oder noch immer nicht
- Frage unzutreffend / kann nicht beantwortet werden

42 [Seiten-ID: 1097617] [L]

Frage 27

Wenn Sie Autofahrer sind: Wieviel PS (ca.) hat Ihr Auto?

Wenn Sie nur die kW-Zahl wissen: 1kW = 1.36 PS.

PS (0 für "habe kein Auto")

43 [Seiten-ID: 1097618] [L]

Frage 28

Wie oft haben Sie im vergangenen Trimester von Kommilitonen Vorlesungsmitschriften etc. oder Kopien bekommen?

mal (ca.)

44 [Seiten-ID: 1097619] [L]

Frage 29

Wie oft haben Sie selbst im vergangenen Trimester Vorlesungsmitschriften / Kopien für andere zur Verfügung gestellt?

mal (ca.)

45 [Seiten-ID: 1097620] [L]

Frage 30

Wie oft kommen Sie unvorbereitet zu einer Veranstaltung?

nie			manchmal			sehr häufig

46 [Seiten-ID: 1097621] [L]

Frage 31

Kommt es vor, daß Sie ein Studieninhalt so fesselt, daß Sie über das Pflichtpensum hinausgehende Informationen und Literatur suchen und lesen?

nie **manchmal** **sehr häufig**

○ ○ ○ ○ ○ ○ ○

47 [Seiten-ID: 1097622] [L]

Frage 32

Verglichen mit meinen Kommilitonen/innen komme ich in einer Prüfungssituation

Viel weniger als andere in Stress **Nicht mehr und nicht weniger** **Viel mehr als andere in Stress**

○ ○ ○ ○ ○

48 [Seiten-ID: 1097623] [L]

Frage 33

Hier ist noch Raum für Kommentare:

49 [Seiten-ID: 1097624] [L]

Endseite

Sie haben nun den Fragebogen zu Ende ausgefüllt!

Vielen Dank für Ihre Teilnahme!

Anhang D

Ergänzende Abbildungen zur Untersuchung 7.1 in Kapitel 7

Dieser Anhang enthält die zusätzlichen Abbildungen zu der Untersuchung 7.1. Wiedergeben sind die grafischen Darstellungen der reorganisierten Teildatenmatrizen aus der Gesamtstichprobe jeweils für die Skalen der drei Konstrukte zu beiden impliziten Antwortmodellen bzw. Antwortprozessen. Das methodische Vorgehen zur Skalierung und Klassifikation der Personen nach den beiden impliziten Antwortmodellen ist in Abschnitt 7.1 beschrieben.

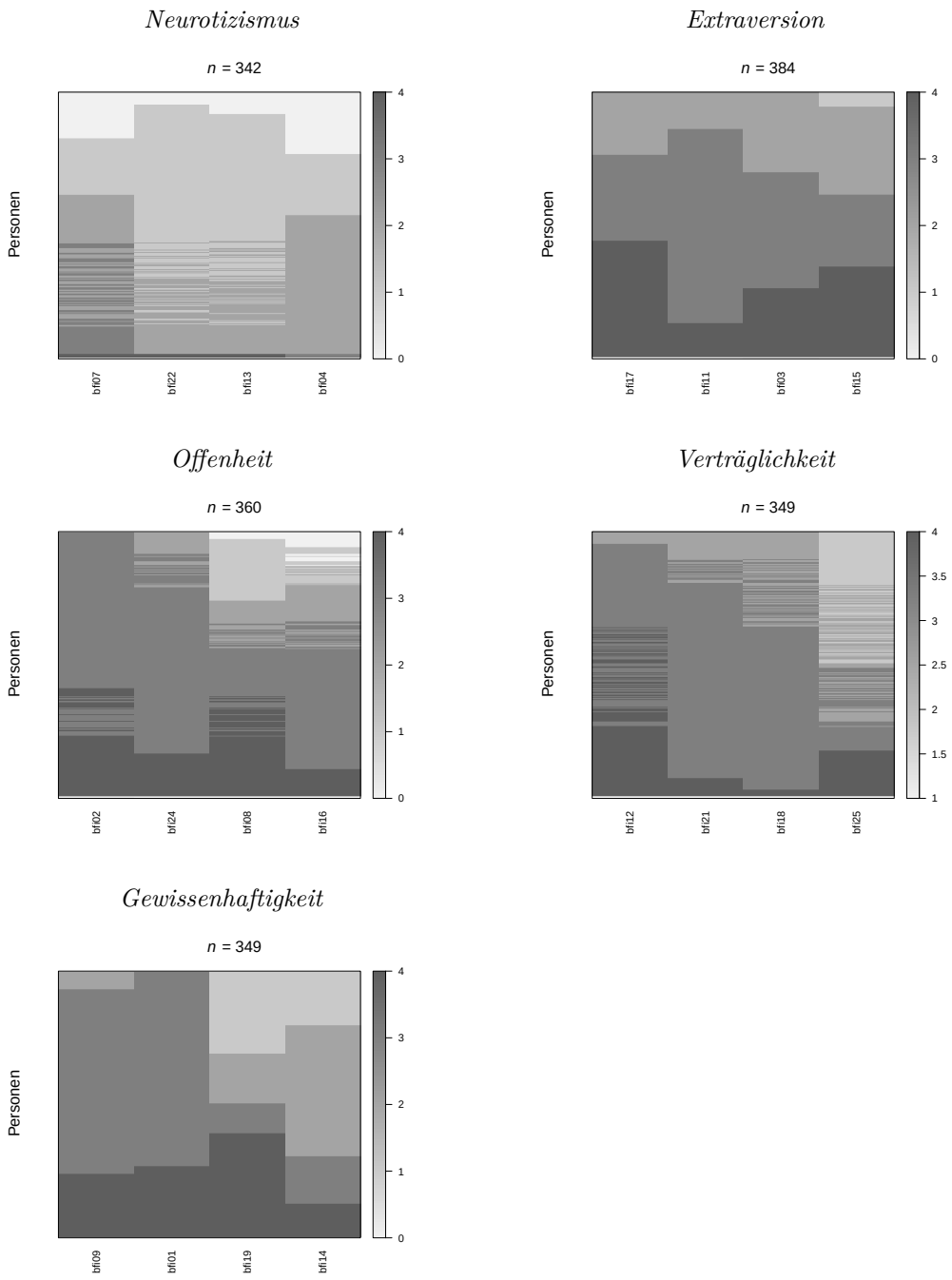


Abbildung D.1 Stichprobe I und II: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Dominanz*-Antwortprozess (PCM) für fünf Skalen des BFI–K; Graustufen entsprechen den Antwortkategorien: Dunkel ≡ 4 = *„Sehr zutreffend“* – hell ≡ 0 = *„Sehr unzutreffend“*.

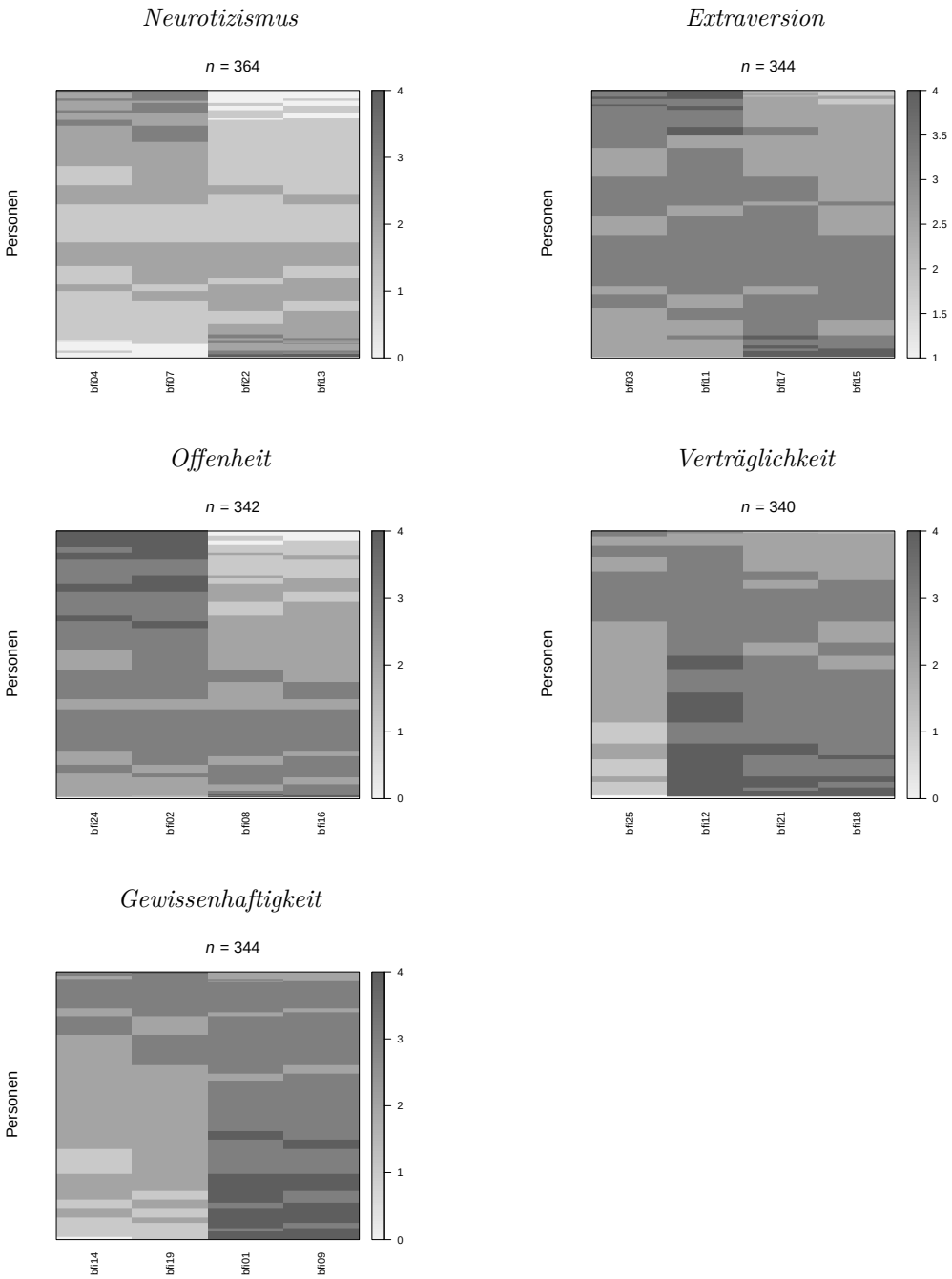


Abbildung D.2 Stichprobe I: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Nähe–Distanz*-Antwortprozess (MDS) für fünf Skalen des BFI–K; Graustufen entsprechen den Antwortkategorien: Dunkel ≡ 4 = *„Sehr zutreffend“* – hell ≡ 0 = *„Sehr unzutreffend“*.

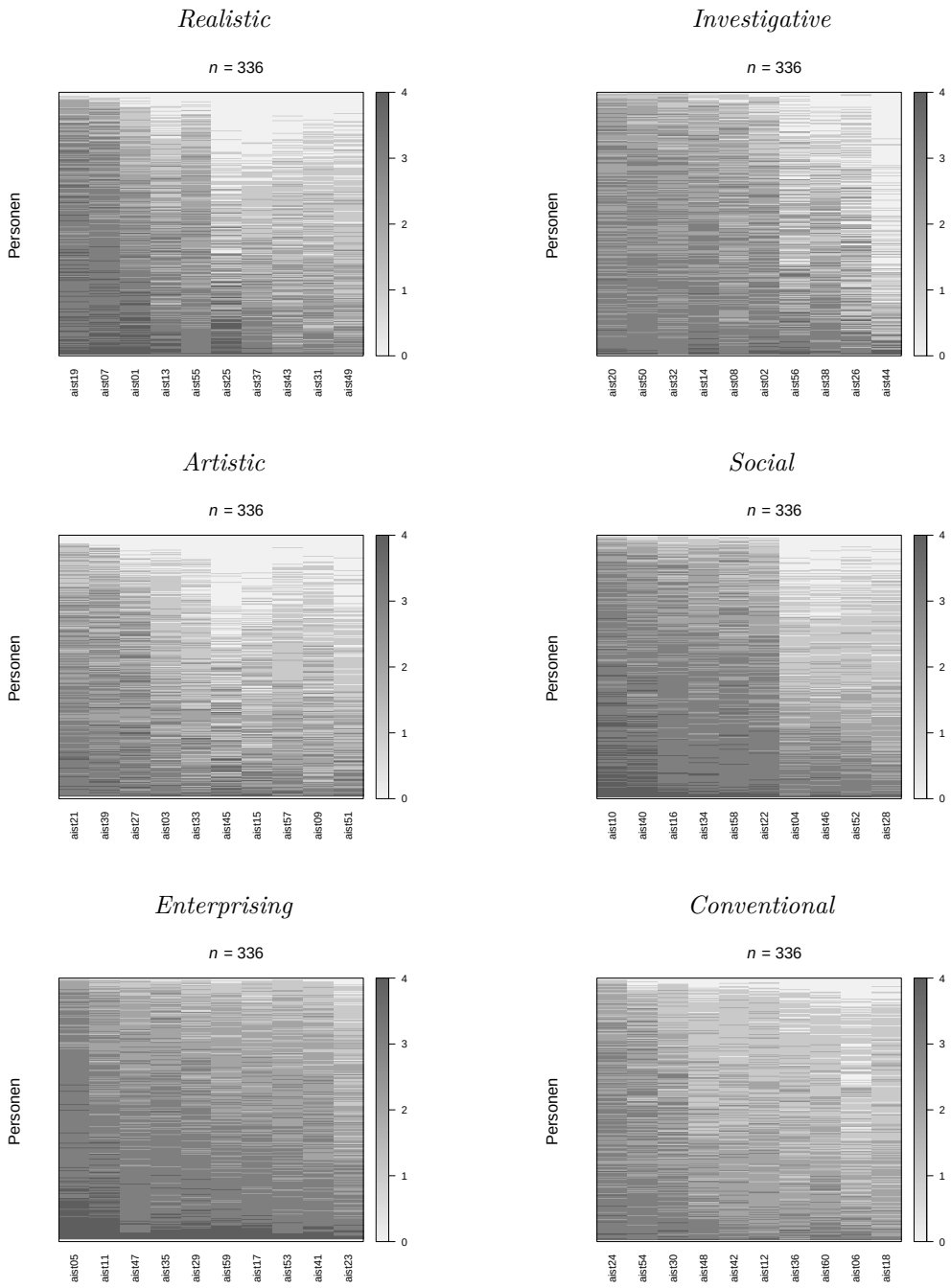


Abbildung D.3 Stichprobe I und II: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Dominanz*-Antwortprozess (PCM) für sechs Skalen des AIST-R; Graustufen entsprechen den Antwortkategorien: Dunkel $\equiv 4 =$ *„Das interessiert mich sehr; das tue ich sehr gerne“* – hell $\equiv 0 =$ *„Das interessiert mich gar nicht; das tue ich nicht gerne“*.

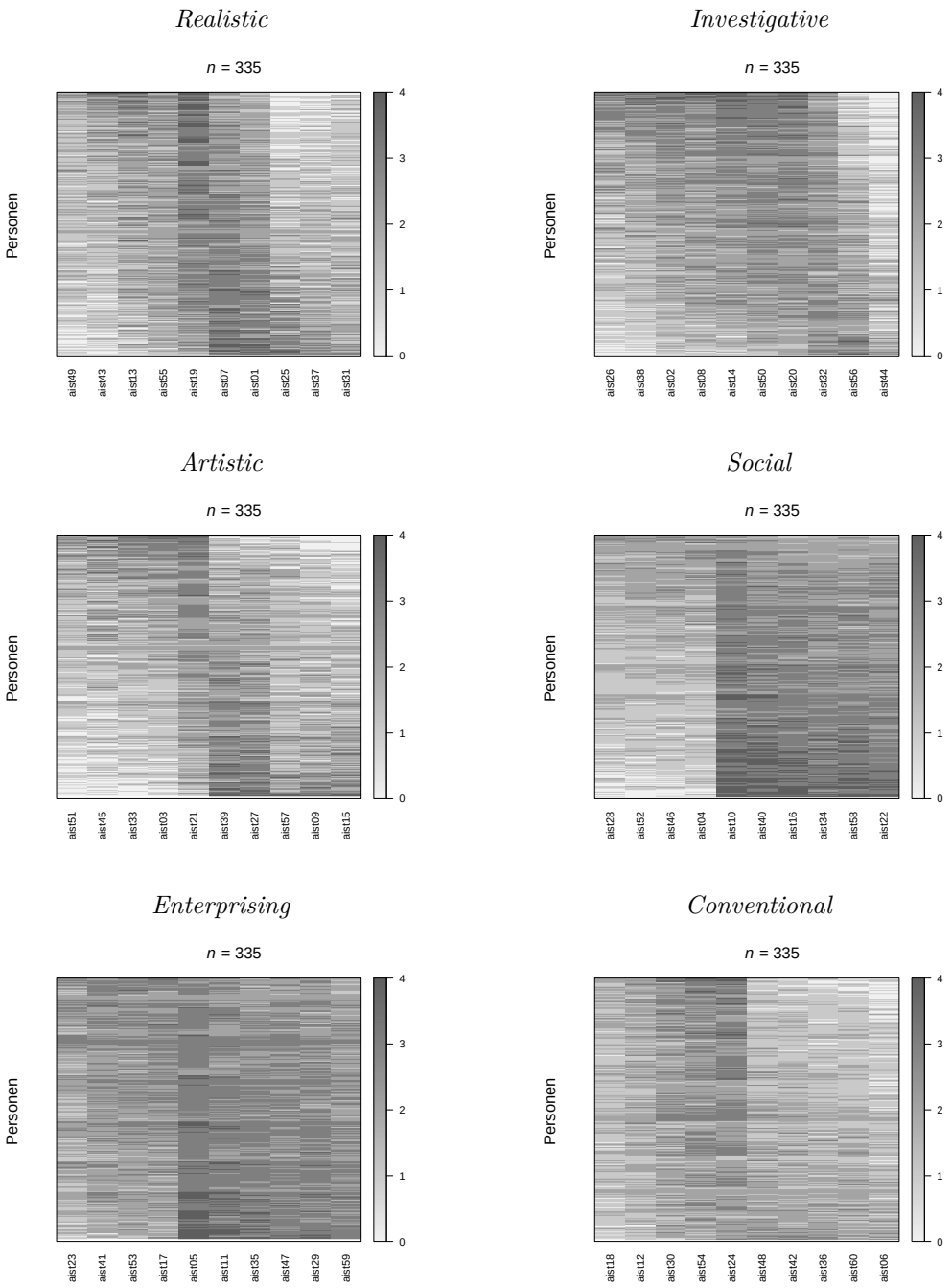


Abbildung D.4 Stichprobe I und II: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Nähe–Distanz*-Antwortprozess (MDS) für sechs Skalen des AIST-R; Graustufen entsprechen den Antwortkategorien: Dunkel ≡ 4 = *„Das interessiert mich sehr; das tue ich sehr gerne“* – hell ≡ 0 = *„Das interessiert mich gar nicht; das tue ich nicht gerne“*.

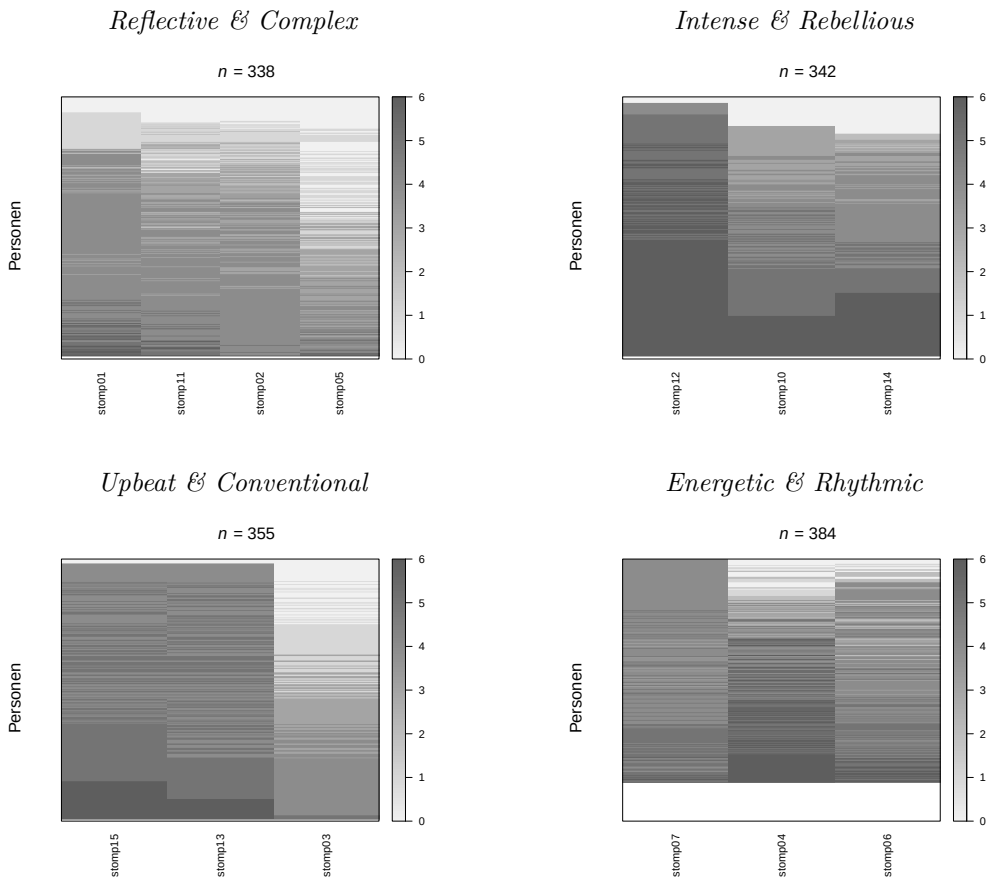


Abbildung D.5 Stichprobe I: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Dominanz*-Antwortprozess (PCM) für vier Skalen des STOMP; Graustufen entsprechen den Antwortkategorien: Dunkel $\equiv 6$ *„Mag ich sehr“* – hell $\equiv 0 =$ *„Mag ich überhaupt nicht“*.

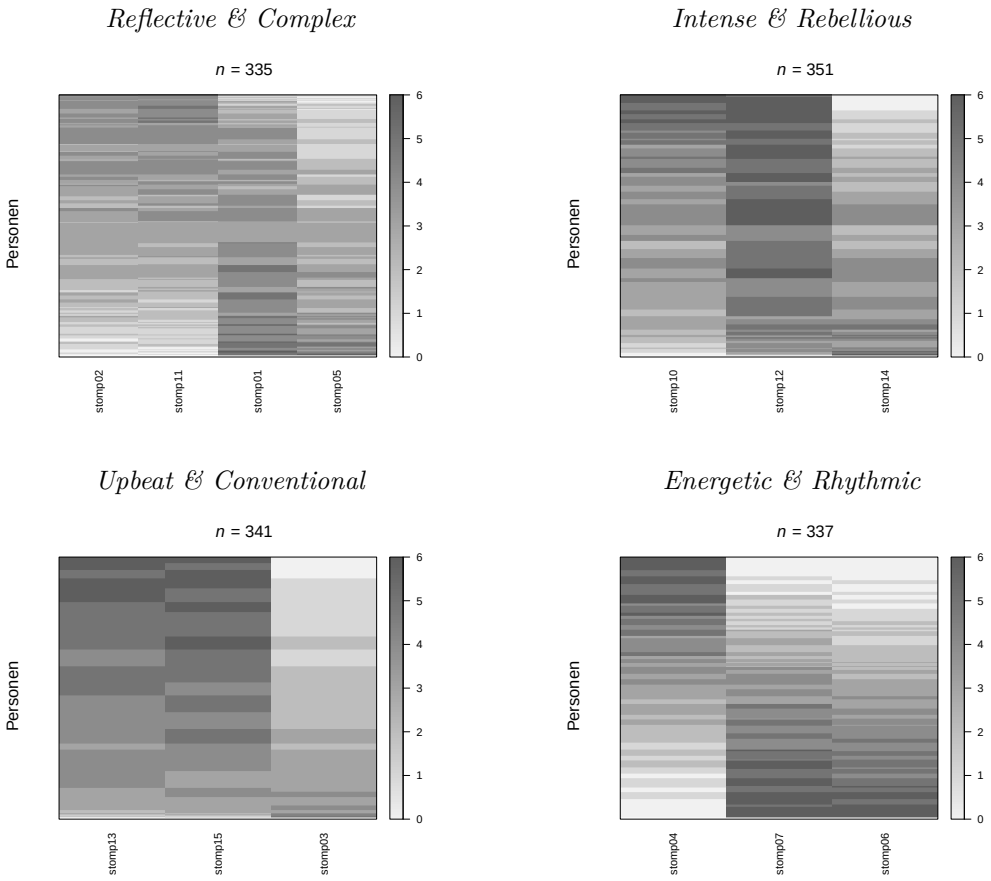


Abbildung D.6 Stichprobe I: Grafische Darstellung der reorganisierten Datenmatrizen für klassifizierte Personen nach dem *Nähe–Distanz*-Antwortprozess (MDS) für vier Skalen des STOMP; Graustufen entsprechen den Antwortkategorien: Dunkel ≡ 6 *„Mag ich sehr“* – hell ≡ 0 = *„Mag ich überhaupt nicht“*.

Literaturverzeichnis

Abdi, H. & Valentin, D. (2007). Multiple correspondence analysis. In N. J. Salkind & K. Rasmussen (Hrsg.), *Encyclopedia of measurement and statistics* (Bd. II, S. 651–657). Thousand Oaks: SAGE Publications.

Abe, C., Holland, J. L., Lutz, S. W. & Richards, J. M. (1965). *A description of american college freshmen.* (Research Report Nr. ACT-RR-1-MAR-65). Iowa City: U.S. Department of Health, Education & Welfare Office of Education.

Adams, E. & Messick, S. (1958). An axiomatic formulation and generalization of successive intervals scaling. *Psychometrika*, *23* (4), 355–368. doi: 10.1007/BF02289784

Adams-Webber, J. & Benjafield, J. (1973). The relation between lexical marking and rating extremity in interpersonal judgment. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *5* (3), 234.

Adorno, T. W. (1950). *The authoritarian personality.* New York: Harper.

Aghajani, M., Veer, I. M., Tol, M.-J. v., Aleman, A., Buchem, M. A. v., Veltman, D. J., ... Wee, N. J. v. d. (2014). Neuroticism and extraversion are associated with amygdala resting-state functional connectivity. *Cognitive, Affective, & Behavioral Neuroscience*, *14* (2), 836–848. doi: 10.3758/s13415-013-0224-0

Ahrens, W. (1901). Zur relativen Bewertung von Turnierpartien. *Wiener Schachzeitung*, *IV* (1), 181–192.

Aitkin, I. & Aitkin, M. (2011). *Statistical modeling of the national assessment of educational progress.* New York: Springer New York.

Ajzen, I. (1991). Theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50* (2), 179–211.

Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology*, *52* (1), 27–58. doi: 10.1146/annurev.psych.52.1.27

Ajzen, I. & Fishbein, M. (1972). Attitudes and opinions. *Annual Review of Psychology*, *23* (1), 487–544. doi: 10.1146/annurev.ps.23.020172.002415

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Hrsg.), *2nd International Symposium in Information Theory* (S. 267–281). Budapest: Akademiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatik Control*, *AC-19* (6), 716–723.

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, *49* (6), 1621–1630. doi: 10.1037/0022-3514.49.6.1621

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J. & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology*, *68* (5), 804–825. doi: 10.1037/0022-3514.68.5.804

Allport, G. W. (1927). Concepts of trait and personality. *Psychological Bulletin*, *24* (5), 284–293. doi: 10.1037/h0073629

Allport, G. W. & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47* (1), 1–171. doi: 10.1037/h0093360

Allport, G. W. & Vernon, P. E. (1930). The field of personality. *Psychological Bulletin*, *27* (10), 677–730. doi: 10.1037/h0072589

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34* (1), 42–54.

Andersen, E. B. (1973a). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26* (1), 31–44. doi: 10.1111/j.2044-8317.1973.tb00504.x

Andersen, E. B. (1973b). A goodness of fit test for the Rasch model. *Psychometrika*, *38* (1), 123–140. doi: 10.1007/BF02291180

Andersen, E. B. (1995). Residualanalysis in the polytomous Rasch model. *Psychometrika*, *60* (3), 375–393. doi: 10.1007/BF02294382

Anderson, D. R., Burnham, K. P. & White, G. C. (1998). Comparison of akai-

ke information criterion and consistent akaike information criterion for model selection and statistical inference from capture-recapture studies. *Journal of Applied Statistics*, *25* (2), 263–282.

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2* (4), 581–594. doi: 10.1177/014662167800200413

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika*, *43* (4), 561–573.

Andrich, D. (1978c). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, *38* (3), 665–680. doi: 10.1177/001316447803800308

Andrich, D. (1982). An extension of the rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, *47* (1), 105–113.

Andrich, D. (1985). An elaboration of guttman scaling with rasch models for measurement. *Sociological Methodology*, *15*, 33–80.

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement*, *12* (1), 33–51.

Andrich, D. (1989). A probabilistic IRT model for unfolding preference data. *Applied Psychological Measurement*, *13* (2), 193–216.

Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement*, *19* (3), 269–290. doi: 10.1177/014662169501900306

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: reconciling Thurstone and Likert methodologies. *British Journal of Mathematical & Statistical Psychology*, *49*, 347–365.

Andrich, D. (2004)jan. Controversy and the Rasch Model: A characteristic of incompatible paradigms? *Medical Care*, *42* (Supplement), I-7–I-16. doi: 10.1097/01.mlr.0000103528.48582.7c

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous rasch model. *Psychometrika*, *75* (2), 292–308. doi: 10.1007/s11336-010-9154-8

Andrich, D. & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological*

Measurement, *17* (3), 253–276. doi: 10.1177/014662169301700307

Andrich, D. & Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement*, *4* (3), 205–221.

Arbeitskreis OPD. (1996). *Operationalisierte psychodynamische Diagnostik Grundlagen und Manual* (1. Aufl.). Bern: Huber.

Arce-Ferrer, A. J. (2006). An investigation into the factors influencing extreme-response style: improving meaning of translated and culturally adapted rating scales. *Educational and Psychological Measurement*, *66* (3), 374–392. doi: 10.1177/0013164405278575

Arias, V. B. & Arias, B. (2017). The negative wording factor of core self-evaluations scale (CSES): methodological artifact, or substantive specific variance? *Personality and Individual Differences*, *109*, 28–34. doi: 10 .1016/j.paid.2016.12.038

Aschenbrenner, K. (1981). Efficient sets, decision heuristics, and single-peaked preferences. *Journal of Mathematical Psychology*, *23* (3), 227–256. doi: 10.1016/0022-2496(81)90061-4

Asendorpf, J. B. & Neyer, F. J. (2012). *Psychologie der Persönlichkeit* (5. Aufl.). Berlin: Springer. doi: 10.1007/978-3-642-30264-0

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., ... De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology*, *86* (2), 356–366.

Austin, E., Deary, I., Gibson, G., McGregor, M. & Dent, J. (1998). Individual response spread in self-report scales: Personality correlations and consequences. *Personality and individual differences*, *24* (3), 421–438.

Austin, E. J., Deary, I. J. & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences*, *40* (6), 1235–1245. doi: 10.1016/j.paid.2005.10.018

Ayala, R. (2009). *The theory and practice of item response theory.* New York: Guilford Press.

Babkin, B. P. (1949). *Pavlov a biography.* Chicago: University of Chicago Press.

Bäckström, M. (2007). Higher-order factors in a Five-Factor personality inventory and its relation to social desirability. *European Journal of Psychological Assessment*, *23* (2), 63–70.

Bäckström, M., Björklund, F. & Larsson, M. R. (2009)jun. Five-Factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality*, *43* (3), 335–344. doi: 10.1016/j.jrp.2008.12.013

Baer, R. A., Wetter, M. W. & Berry, D. T. R. (1992). Detection of underreporting of psychopathology on the MMPI: A meta-analysis. *Clinical Psychology Review*, *12* (5), 509–525. doi: 10.1016/0272-7358(92)90069-K

Baghaei, P., Yanagida, T. & Heene, M. (2017). Development of a descriptive fit statistic for the Rasch model. *North American Journal of Psychology*, *19* (1), 155–168.

Baker, F. B. & Harwell, M. R. (1996). Computing elementary symmetric functions and their derivatives: A didactic. *Applied Psychological Measurement*, *20* (2), 169–192. doi: 10.1177/014662169602000206

Baker, F. B. & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. Boca Raton: CRC Press.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60* (3), 361–370.

Barrick, M. R. & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology*, *81* (3), 261–272. doi: 10.1037/0021-9010.81.3.261

Barrick, M. R., Mount, M. K. & Gupta, R. (2003). Meta-analysis of the relationship between the Five-Factor model of personality and Holland's occupational types. *Personnel Psychology*, *56* (1), 45–74. doi: 10.1111/j.1744-6570.2003.tb00143.x

Barton, M. A. & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, *1981* (1), i–8. doi: 10.1002/j.2333-8504.1981.tb01255.x

Bass, B. M. (1955). Authoritarianism or acquiescence? *The Journal of Abnormal and Social Psychology*, *51* (3), 616.

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin*, *91* (1), 3–26. doi: 10.1037/0033-2909.91.1.3

Baumeister, R. F., Masicampo, E. J. & Vohs, K. D. (2011). Do conscious thoughts cause behavior? *Annual Review of Psychology*, *62* (1), 331–361. doi: 10.1146/annurev.psych.093008.131126

Baumeister, R. F. & Tice, D. M. (1988). Metatraits. *Journal of Personality*, *56* (3), 571–598.

Baumgarten, F. (1933). *Die Charaktereigenschaften*. Bern: A. Francke.

Baumgartner, H. & Steenkamp, J. B. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38* (2), 143–156.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M. & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, *27* (5), 319–334. doi: 10.1177/0146621603257518

Bechtel, G. G. (1968). Folded and unfolded scaling from preferential paired comparisons. *Journal of Mathematical Psychology*, *5* (2), 333–357. doi: 10.1016/0022-2496(68)90079-5

Beins, B. C. (1994). Barnum Effect. In R. J. Corsini (Hrsg.), *Encyclopedia of Psychology* (S. 130–131). New York: John Wiley & Sons, Inc.

Bejar, I. I. (1977). An application of the continuous response level model to personality measurement. *Applied Psychological Measurement*, *1* (4), 509–521. doi: 10.1177/014662167700100407

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74* (3), 183–200. doi: 10.1037/h0024835

Bem, D. J. (1977). Predicting more of the people more of the time: some thoughts on the allen-potkay studies of intraindividual variability. *Journal of Personality*, *45* (3), 327–333.

Bem, D. J. (1983). Further déjà vu in the search for cross-situational consistency: A response to mischel and peake. *Psychological Review*, *90* (4), 390–393.

Bem, D. J. & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior.

Psychological Review, *81* (6), 506.

Bem, D. J. & Funder, D. C. (1978). Predicting more of the people more of the time: Assessing the personality of situations. *Psychological Review*, *85* (6), 485.

Bensch, D., Paulhus, D. L., Stankov, L. & Ziegler, M. (2017). Teasing apart overclaiming, overconfidence, and socially desirable responding. *Assessment*. doi: 10.1177/1073191117700268

Bentler, P. M. (1969). Semantic space is (approximately) bipolar. *The Journal of Psychology*, *71* (1), 33–40. doi: 10.1080/00223980.1969.10543067

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107* (2), 238–246.

Bentler, P. M., Jackson, D. N. & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin*, *76* (3), 186–204. doi: 10.1037/h0031474

Berg, I. A. (1957). Deviant responses and deviant people: The formulation of the deviation hypothesis. *Journal of Counseling Psychology*, *4* (2), 154–161. doi: 10.1037/h0048027

Berg, I. A. & Collier, J. S. (1953). Personality and group differences in extreme response sets. *Educational and Psychological Measurement*, *13* (2), 164–169. doi: 10.1177/001316445301300202

Bergmann, C. (2001). Personality and vocational interests: Evaluation of the correspondence between the Five-Factor model of personality and Holland's six vocational orientations. In K. W. Kallus, N. Posthumus & P. Jimenez (Hrsg.), *Current psychological research in austria.* Graz: Akademische Druck- und Verlagsanstalt.

Bergmann, C. & Eder, F. (1999). *Allgemeiner Interessen-Struktur-Test (AIST), Umwelt-Struktur-Test (UST). Testmanual* (2. Aufl.). Weinheim: Beltz-Test.

Bergmann, C. & Eder, F. (2005). *AIST-R Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-R) - Revision.* Göttingen: Beltz Test.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C. & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4* (3), 340–345. doi:

10.1037/1040-3590.4.3.340

Bertea, P. E. & Zait, A. (2014). Response styles in cross-cultural research–evidence from historical regions. *CrossCultural Management Journal*, *XVI* (30), 19–27.

Bertin, J. (1977). *La graphique et le traitement graphique de l'information*. Paris: Flammarion.

Bever, T. G. (1988). A cognitive theory of emotion and aesthetics in music. *Psychomusicology: A Journal of Research in Music Cognition*, *7* (2), 165–175. doi: 10.1037/h0094171

Biderman, M., Nguyen, N., Cunningham, C. & Ghorbani, N. (2011). The ubiquity of common method variance: The case of the Big Five. *Journal of Research in Personality*, *45*, 417–429.

Bing, M. N., Kluemper, D., Kristl Davison, H., Taylor, S. & Novicevic, M. (2011). Overclaiming as a measure of faking. *Organizational Behavior and Human Decision Processes*, *116* (1), 148–162. doi: 10.1016/j.obhdp.2011.05.006

Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. *Educational and Psychological Measurement*, *45*, 523–534.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T. & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14* (4), 317–335.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Hrsg.), *Statistical theories of mental test scores* (S. 395–479). Reading: Addison-Wesley.

Blasius, J. (2001). *Korrespondenzanalyse*. München: Oldenbourg Wissenschaftsverlag.

Blasius, J. & Lautsch, E. (1990). Die komplementäre Anwendung zweier Verfahren : Korrespondenzanalyse und Konfigurationsfrequenzanalyse. *ZA-Information / Zentralarchiv für Empirische Sozialforschung* (27), 110–133.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*

(4), 443–459. doi: 10.1007/BF02293801

Bock, R. D., Gibbons, R. & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12* (3), 261–280. doi: 10 .1177/014662168801200305

Boeije, H. R. (2004). And then there were three: Self-presentational styles and the presence of the partner as a third person in the interview. *Field Methods*, *16* (1), 3 –22. doi: 10.1177/1525822X03259228

Boerner, R. J. (2015). Körperbau und Temperament. In R. J. Boerner (Hrsg.), *Temperament: Theorie, Forschung, Klinik* (S. 121–159). Berlin: Springer. doi: 10.1007/978-3-642-39505-5_5

Bolt, D. M. (2005). Limited- and full-information estimation of item response theory models. In R. P. McDonald, A. Maydeu-Olivares & J. J. McArdle (Hrsg.), *Contemporary psychometrics: a Festschrift for roderick p. mcdonald.* Mahwah, N.J: Lawrence Erlbaum Associates.

Boltzmann, L. (1877). Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht. In *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften* (Bde. LXXVI, Heft III, S. 373–435). Wien: Karl Gerolds Sohn.

Bond, J. A. (1987). The process of responding to personality items: Inconsistent responses to repeated presentation of identical items. *Personality and Individual Differences*, *8* (3), 409–417. doi: 10.1016/0191-8869(87) 90042-0

Bond, T. G. & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3. Aufl.). New York & London: Routledge.

Boon, J., Gozna, L. & Hall, S. (2008). Detecting 'faking bad' on the Gudjonsson Suggestibility Scales. *Personality and Individual Differences*, *44* (1), 263–272. doi: 10.1016/j.paid.2007.08.005

Borg, I. (1992). *Grundlagen und Ergebnisse der Facettentheorie.* Bern: Verlag Hans Huber.

Borg, I. & Mohler, P. P. (1993). Zur Indexbildung in der Facettentheorie. *ZUMA-Nachrichten*, *17* (33), 10–24.

Borg, I. & Staufenbiel, T. (1993). Facet theory and design for attitude mea-

surement and its application. In *New directions in attitude measurement* (S. 206–237). Berlin, New York: Walter de Gruyter.

Borg, I. & Staufenbiel, T. (2007). *Lehrbuch - Theorien und Methoden der Skalierung*. Bern: Huber.

Borkenau, P. & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae*. Göttingen: Hogrefe, Verl. für Psychologie.

Borkenau, P. & Ostendorf, F. (2008). *NEO-Fünf-Faktoren-Inventar nach Costa und McCrae ; NEO-FFI* (2., neu normierte und vollständig überarb. Aufl.). Göttingen: Hogrefe.

Borkenau, P., Riemann, R., Spinath, F. M. & Angleitner, A. (2006). Genetic and environmental influences on person × situation profiles. *Journal of Personality*, *74* (5), 1451–1480. doi: 10.1111/j.1467-6494.2006.00416.x

Borkenau, P., Zaltauskas, K. & Leising, D. (2009). More may be better but there may be too much: Optimal trait level and self-enhancement bias. *Journal of Personality*, *77* (3), 825–858. doi: 10.1111/j.1467-6494.2009.00566.x

Borsboom, D. (2008). Latent Variable Theory. *Measurement: Interdisciplinary Research and Perspectives*, *6* (1-2), 25–53. doi: 10.1080/15366360802035497

Borsboom, D. & Mellenbergh, G. J. (2004)jan. Why Psychometrics is Not Pathological A Comment on Michell. *Theory & Psychology*, *14* (1), 105–120. doi: 10.1177/0959354304040200

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.). Berlin: Springer. doi: 10.1007/978-3-642-12770-0

Bossuyt, P. M. (1990). *A comparison of probabilistic unfolding theories for paired comparisons data*. Berlin: Springer.

Bossuyt, P. M. & Roskam, E. E. (1989). Maximum likelihood unidimensional unfolding in a probabilistic model without parametric assumptions. *Advances in Psychology*, *60*, 77–98. doi: 10.1016/S0166-4115(08)60231-9

Box, G. (1979). Robustness in the Strategy of Scientific Model Building. In R. L. Launer & G. N. Wilkinson (Hrsg.), *Robustness in statistics: proceedings of a workshop* (S. 201–236). New York: Academic Press.

Boyce, A. C. (1915). *Methods for measuring teachers' efficiency* (Bd. 2). Chicago: University of Chicago Press.

Brady, H. E. (1985). Statistical consistency and hypothesis testing for nonmetric multidimensional scaling. *Psychometrika*, *50* (4), 509–537. doi: 10.1007/BF02296267

Brady, H. E. (1989). Factor and ideal point analysis for interpersonally incomparable data. *Psychometrika*, *54* (2), 181–202. doi: 10.1007/BF02294514

Brady, H. E. (1990). Traits versus issues: Factor versus ideal-point analysis of candidate thermometer ratings. *Political Analysis*, *2*, 97–129.

Brehm, M. & Feger, H. (2001). Feature pattern analysis in the context of other models and methods. *Psychologische Beiträge*, *43* (2), 444–457.

Brenner, C. (1994). The mind as conflict and compromise formation. *Journal of Clinical Psychoanalysis*, *3*, 473–563.

Britt, T. W. (1993). Metatraits: Evidence relevant to the validity of the construct and its implications. *Journal of Personality and Social Psychology*, *65* (3), 554.

Britt, T. W. & Shepperd, J. A. (1999). Trait relevance and trait assessment. *Personality & Social Psychology Review (Lawrence Erlbaum Associates)*, *3* (2), 108.

Brown, A. & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Industrial and Organizational Psychology*, *3* (4), 489–493.

Brown, C., Templin, J. & Cohen, A. (2015). Comparing the Two- and Three-Parameter Logistic Models via Likelihood Ratio Tests: A Commonly Misunderstood Problem. *Applied Psychological Measurement*, *39* (5), 335–348. doi: 10.1177/0146621614563326

Browne, M. W. (2000). Psychometrics. *Journal of the American Statistical Association*, *95* (450), 661–665. doi: 10.2307/2669413

Brusco, M. & Steinley, D. (2006). Inducing a blockmodel structure of two-mode binary data using seriation procedures. *Journal of Mathematical Psychology*, *50* (5), 468–477. doi: 10.1016/j.jmp.2006.05.005

Bryce, T. (1981). Rasch-Fitting. *British Educational Research Journal*, *7* (2), 137–153. doi: 10.1080/0141192810070203

Buckland, S. T., Burnham, K. P. & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, *53* (2), 603–618. doi: 10.2307/2533961

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2., aktualisierte Aufl.). München: Pearson.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. aktualisierte und erweiterte Aufl.). München: Pearson.

Burnham, K. P. & Anderson, D. R. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, *28*, 111–119.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33* (2), 261–304. doi: 10.1177/0049124104268644

Burnham, K. P., Anderson, D. R. & Burnham, K. P. (2002). *Model selection and multimodel inference: a practical information-theoretic approach* (2. Aufl.). New York: Springer.

Burt, C. (1939). The factorial analysis of emotional traits. *Journal of Personality*, *7* (3), 238–254.

Burt, C. (1949). Alternative methods of factor analysis and their relations to pearson's method of'principal axes'. *British Journal of Statistical Psychology*, *2* (2), 98–121. doi: 10.1111/j.2044-8317.1949.tb00271.x

Burt, C. (1950). The factorial analysis of qualitative data. *British Journal of Statistical Psychology*, *3* (3), 166–185. doi: 10.1111/j.2044-8317.1950.tb00296.x

Buss, A. H. & Finn, S. E. (1987). Classification of personality traits. *Journal of Personality and Social Psychology*, *52* (2), 432–444. doi: 10.1037/0022-3514.52.2.432

Buss, D. M. & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review*, *90* (2), 105–126. doi: 10.1037/0033-295X.90.2.105

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75* (4), 581–612. doi: 10.1007/s11336-010-9178-0

Cai, L. & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and*

Statistical Psychology, *66* (2), 245–276. doi: 10.1111/j.2044-8317.2012.02050.x

Cai, L., Maydeu-Olivares, A., Coffman, D. L. & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical & Statistical Psychology*, *59*, 173–194. doi: 10.1348/000711005X66419

Cai, L., Yang, J. S. & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16* (3), 221–248. doi: 10.1037/a0023350

Campbell, D. (2001). *The mozart effect*. New York: HarperCollins.

Campbell, D. P., Borgen, F. H., Eastes, S. H., Johansson, C. B. & Peterson, R. A. (1968). A set of basic interest scales for the Strong Vocational Interest Blank for men. *Journal of Applied Psychology; Journal of Applied Psychology*, *52* (6p2), 1.

Campbell, D. P. & Holland, J. L. (1972). A merger in vocational interest research: Applying Holland's theory to Strong's data. *Journal of Vocational Behavior*, *2* (4), 353–376. doi: 10.1016/0001-8791(72)90012-7

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56* (2), 81–105. doi: 10.1037/h0046016

Canli, T., Sivers, H., Whitfield, S. L., Gotlib, I. H. & Gabrieli, J. D. E. (2002). Amygdala response to happy faces as a function of extraversion. *Science (New York, N.Y.)*, *296* (5576), 2191. doi: 10.1126/science.1068749

Carifio, J. & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, *3* (3), 106–116.

Carlson, R. (1975). Personality. *Annual Review of Psychology*, *26* (1), 393–414. doi: 10.1146/annurev.ps.26.020175.002141

Carroll, J. D. & Arabie, P. (1980). Multidimensional scaling. *Annual review of psychology*, *31* (1), 607–649.

Carter, N. T., Dalal, D. K., Lake, C. J., Lin, B. C. & Zickar, M. J. (2011). Using mixed-model item response theory to analyze organizational survey responses: An illustration using the job descriptive index. *Organizational*

Research Methods, *14* (1), 116–146.

Carter, N. T., Lake, C. J. & Zickar, M. J. (2010). Toward understanding the psychology of unfolding. *Industrial and Organizational Psychology*, *3* (4), 511–514.

Carter, N. T. & Zickar, M. J. (2011). The influence of dimensionality on parameter estimation accuracy in the generalized graded unfolding model. *Educational and Psychological Measurement*, *71* (5), 765–788. doi: 10.1177/0013164410387594

Cattell, R. B. (1944). Interpretation of the twelve primary personality factors. *Personality*, *13* (1), 55–91.

Cattell, R. B. (1945). The principal trait clusters for describing personality. *Psychological Bulletin*, *42* (3), 129–161. doi: 10.1037/h0060679

Cattell, R. B. (1946). *The description and measurement of personality.* Oxford England: World Book Company.

Cattell, R. B. (1968). Trait-view theory of perturbations in ratings and self ratings (L(BR)- and Q-data): Its application to obtaining pure trait score estimates in questionnaires. *Psychological Review*, *75* (2), 96–113. doi: 10.1037/h0025604

Cattell, R. B. (1980). Two basic models for personality and environment interaction and the need for their substantive investigation. *Multivariate Behavioral Research*, *15* (3), 243–247.

Cattell, R. B. (1988a). The data box its ordering of total resources in terms of possible relational systems. In J. R. Nesselroade & R. B. Cattell (Hrsg.), *Handbook of Multivariate Experimental Psychology* (2. Aufl., S. 69–130). New York: Plenum Press.

Cattell, R. B. (1988b). The meaning and strategic use of factor analysis. In J. R. Nesselroade & R. B. Cattell (Hrsg.), *Handbook of Multivariate Experimental Psychology* (2. Aufl., S. 131–203). New York: Plenum Press.

Cattell, R. B. & Anderson, J. C. (1953). The measurement of personality and behavior disorders by the I. P. A. T. Music Preference Test. *Journal of Applied Psychology*, *37* (6), 446–454.

Cattell, R. B. & Saunders, D. R. (1954a). Beiträge zur Faktoren-Analyse der Persönlichkeit. *Zeitschrift für Experimentelle und Angewandte Psycho-*

logie, *2*, 325–357.

Cattell, R. B. & Saunders, D. R. (1954b). Musical prefferences and personality diagnosis: I. A factorization of one hundred and twenty themes. *Journal of Social Psychology*, *39* (1), 3–24.

Chambers, J. R. & Windschitl, P. D. (2004). Biases in social comparative judgments: The role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin*, *130* (5), 813–838. doi: 10.1037/0033-2909.130.5.813

Chamorro-Premuzic, T., Swami, V., Furnham, A. & Maakip, I. (2009). The Big Five personality traits and uses of music. *Journal of Individual Differences*, *30* (1), 20–27. doi: 10.1027/1614-0001.30.1.20

Chang, L. (1995). Connotatively consistent and reversed connotatively inconsistent items are not fully equivalent: Generalizability study. *Educational and Psychological Measurement*, *55* (6), 991–997. doi: 10.1177/0013164416667978

Chapman, L. J. & Campbell, D. T. (1957). Response set in the F Scale. *The Journal of Abnormal and Social Psychology*, *54* (1), 129–132.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14* (3), 464–504. doi: 10.1080/10705510701301834

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F. & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36* (4), 523–562.

Chernyshenko, O. S., Stark, S., Drasgow, F. & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19* (1), 88–105.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9* (2), 233–255. doi: 10.1207/S15328007SEM0902_5

Choppin, B. H. (1968). Item bank using sample-free calibration. *Nature*, *219* (5156), 870–872. doi: 10.1038/219870a0

Choppin, B. H. (1982). The use of latent trait models in the measurement of

cognitive abilities and skills. In D. Spearritt (Hrsg.), *The Improvement of Measurement in Education and Psychology: Contributions of Latent Trait Theories* (S. 41–63). Melbourne: The Australian Council for Educational Research Ltd.

Choppin, B. H. (1983). *A fully conditional estimation procedure for rasch model parameters* (CSE Report Nr. 196). Los Angeles: University of California, Graduate School of Education Center for the Study of Evaluation.

Christensen, K. B., Makransky, G. & Horton, M. C. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41* (3), 178–194.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40* (1), 5–32. doi: 10.1007/BF02291477

Churchyard, J. S., Pine, K. J., Sharma, S. & Fletcher, B. C. (2014). Same traits, different variance. *SAGE Open*, *4* (1), 1–11. doi: 10.1177/2158244014522634

Clark, S. S. & Giacomantonio, S. G. (2013). Music preferences and empathy: Toward predicting prosocial behavior. *Psychomusicology: Music, Mind, and Brain*, *23* (3), 177–186. doi: 10.1037/a0034882

Cliff, N., Collins, L. M., Zatkin, J., Gallipeau, D. & McCormick, D. J. (1988). An ordinal scaling method for questionnaire and other ordinal data. *Applied Psychological Measurement*, *12* (1), 83 –97. doi: 10.1177/014662168801200108

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7* (3), 249–253. doi: 10.1177/014662168300700301

Cohen, L. (1979). Approximate expressions for parameter estimates in the rasch model. *British Journal of Mathematical and Statistical Psychology*, *32* (1), 113–120. doi: 10.1111/j.2044-8317.1979.tb00756.x

Cole, J. S. & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education*, *57* (2), 119–130. doi: 10.1353/jge.0.0018

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, *57* (3), 145–158. doi: 10.1037/h0060984

Coombs, C. H. (1951). Mathematical models in psychological scaling. *Journal*

of the American Statistical Association, *46* (256), 480–489. doi: 10.2307/2280397

Coombs, C. H. (1952). *A theory of psychological scaling* (Nr. 34). Ann Arbor, MI: University of Michigan Press.

Coombs, C. H. (1956). The Scale Grid: Some interrelations of data models. *Psychometrika*, *21* (4), 313–329. doi: 10.1007/BF02296299

Coombs, C. H. (1967). *A theory of data* (2. Aufl.). New York: Wiley.

Coombs, C. H. & Avrunin, G. S. (1977a). Single-Peaked Functions and the Theory of Preference. *Psychological Review*, *84* (2), 216–230. doi: 10.1037/0033-295X.84.2.216

Coombs, C. H. & Avrunin, G. S. (1977b). A theorem on single-peaked preference functions in one dimension. *Journal of Mathematical Psychology*, *16* (3), 261–266. doi: 10.1016/0022-2496(77)90056-6

Coombs, C. H. & Coombs, L. C. (1976). 'Don't know'. Item ambiguity or respondent uncertainty? *Public Opinion Quarterly*, *40* (4), 497.

Coombs, C. H., Raiffa, H. & Thrall, R. M. (1954). Some views on mathematical models and measurement theory. *Psychological Review*, *61* (2), 132–144. doi: 10.1037/h0063044

Costa, P. T. & McCrae, R. R. (1985). *The NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T. & McCrae, R. R. (1992a). Four ways five factors are basic. *Personality and Individual Differences*, *13* (6), 653–665. doi: 10.1016/0191-8869(92)90236-I

Costa, P. T. & McCrae, R. R. (1992b). Reply to Eysenck. *Personality and Individual Differences*, *13* (8), 861–865. doi: 10.1016/0191-8869(92)90002-7

Costa, P. T. & McCrae, R. R. (1992c). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional Manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., McCrae, R. R. & Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences*, *12* (9), 887–898. doi: 10.1016/0191-8869(91)90177-D

Costa, P. T., McCrae, R. R. & Holland, J. L. (1984). Personality and vocational

interests in an adult sample. *Journal of Applied Psychology*, *69* (3), 390–400. doi: 10.1037/0021-9010.69.3.390

Couch, A. & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology*, *60* (2), 151.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research (JMR)*, *17* (4), 407–422.

Cressie, N. & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, *46* (3), 440 – 464.

Cronbach, L. J. (1941). An experimental comparison of the multiple true-false and multiple multiple-choice tests. *Journal of Educational Psychology*, *32* (7), 533–543. doi: 10.1037/h0058518

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology*, *33* (6), 401–415. doi: 10.1037/h0054677

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6* (4), 475–494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10* (1), 3–31. doi: 10.1177/001316445001000101

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16* (3), 297–334.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12* (11), 671–684. doi: 10.1037/h0043943

Crowne, D. P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24* (4), 349–354. doi: 10.1037/h0047358

Dahlbender, R. W. & Tritt, K. (2011). Einführung in die OPD. *Psychotherapie*, *16* (1), 28–39.

Dalal, D. K. & Carter, N. T. (2015a). Consequences of ignoring ideal point items for applied decisions and criterion-related validity estimates. *Journal of Business and Psychology*, *30* (3), 483–498. doi: 10.1007/s10869-014-9377-2

Dalal, D. K. & Carter, N. T. (2015b). Negatively worded items negatively impact survey research. In C. E. Lance & R. J. Vandenberg (Hrsg.), *More statistical and methodological myths and urban legends.* Abingdon: Routledge.

Damarin, F. & Messick, S. (1965). Response styles as personality variables: A theoretical integration of multivariate research. *ETS Research Bulletin Series*, *1965* (1), i–116.

Darwin, C. R. (1859). *On the origin of species - by means of natural selection* (1. Aufl.). London: John Murray.

Darwin, C. R. (1872). *The expression of the emotions in man and animals.* London: John Murray.

David, H. A. (1971). Ranking the players in a round robin tournament. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *39* (2), 137. doi: 10.2307/1402170

David, H. A. (1988). *The method of paired comparisons.* London: Griffin.

Davies, S. E., Connelly, B. S., Ones, D. S. & Birkland, A. S. (2015). The general factor of personality: the 'Big One' a self-evaluative trait, or a methodological gnat that won't go away? *Personality and Individual Differences*, *81*, 13–22. doi: 10.1016/j.paid.2015.01.006

Davis, C. G., Thake, J. & Weekes, J. R. (2012). Impression managers: Nice guys or serious criminals? *Journal of Research in Personality*, *46* (1), 26–31. doi: 10.1016/j.jrp.2011.11.001

Dayton, C. M. & Macready, G. B. (1980). A scaling model with response errors and intrinsically unscalable respondents. *Psychometrika*, *45* (3), 343–356.

de Falguerolles, A., Friedrich, F. & Sawitzki, G. (1997). A tribute to J. Bertin's graphical data analysis. *SoftStat*, *97*, 11–20.

de Fruyt, F. & Mervielde, I. (1997). The Five-Factor model of personality and Holland's RIASEC interest types. *Personality and Individual Differences*, *23* (1), 87–103. doi: DOI:10.1016/S0191-8869(97)00004-4

de Jong, M. G., Pieters, R. & Fox, J.-P. (2010). Reducing Social Desirability Bias Through Item Randomized Response: An Application to Measure Underreported Desires. *Journal of Marketing Research (JMR)*, *47* (1), 14–27.

de la Torre, J. (2006). Markov chain monte carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement*, *30* (3), 216–232. doi: 10.1177/0146621605282772

de Leeuw, J. & Bettonvil, B. (1986). An upper bound for sstress. *Psychometrika*, *51* (1), 149–153. doi: 10.1007/BF02294008

de Leeuw, J. & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, *31* (3), 1–30.

Delsing, M. J. M. H., ter Bogt, T. F. M., Engels, R. C. M. E. & Meeus, W. H. J. (2008). Adolescents' music preferences and personality characteristics. *European Journal of Personality*, *22* (2), 109–130. doi: 10.1002/per.665

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39* (1), 1–38.

de Vries, R. E., Zettler, I. & Hilbig, B. E. (2014). Rethinking trait conceptions of social desirability scales: Impression management as an expression of honesty-humility. *Assessment*, *21* (3), 286–299. doi: 10.1177/1073191113504619

Dicken, C. F. (1967). Acquiescence in the MMPI a method variance artifact? *Psychological Reports*, *20*, 927–933.

Dickson, D. H. & Kelly, I. W. (1985). The 'Barnum Effect' in personality assessment: a review of the literature. *Psychological Reports*, *57* (2), 367–382. doi: 10.2466/pr0.1985.57.2.367

Diekmann, A. (2014). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen* (9. Aufl.). Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.

Diener, E. (1999). Introduction to the special section on the structure of emotion. *Journal of Personality and Social Psychology*, *76* (5), 803–804. doi: 10.1037/0022-3514.76.5.803

Diers, C. J. (1964). Social desirability and acquiescence in response to personality items. *Journal of consulting psychology*, *28* (1), 71.

Digman, J. M. (1990). Personality structure: Emergence of the Five-Factor model. *Annual Review of Psychology*, *41* (1), 417–440. doi: 10.1146/annurev.ps.41.020190.002221

DiStefano, C. & Motl, R. W. (2009). Personality correlates of method ef-

fects due to negatively worded items on the rosenberg self-esteem scale. *Personality and Individual Differences*, *46* (3), 309–313. doi: 10.1016/j.paid.2008.10.020

Divgi, D. R. (1986). Does the rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, *23* (4), 283–298.

Dollinger, S. J. (1993). Research note: Personality and music preference: extraversion and excitement seeking or openness to experience? *Psychology of Music*, *21* (1), 73 –77. doi: 10.1177/030573569302100105

Dolnicar, S. & Grün, B. (2007). Cross-cultural differences in survey response patterns. *International Marketing Review*, *24* (2), 127–143.

Donlon, T. F. & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, *28* (1), 105–113. doi: 10.1177/001316446802800110

Donoghue, J. (1994). An empirical-examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31* (4), 295–311. doi: 10 .1111/j.1745-3984.1994.tb00448.x

Douglas, R. J. (1967). The hippocampus and behavior. *Psychological Bulletin*, *67* (6), 416–442. doi: 10.1037/h0024599

Drasgow, F., Chernyshenko, O. S. & Stark, S. (2010a). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology*, *3* (4), 465–476. doi: 10.1111/j.1754-9434.2010.01273.x

Drasgow, F., Chernyshenko, O. S. & Stark, S. (2010b). Improving the measurement of psychological variables: Ideal point models rock! *Industrial and Organizational Psychology*, *3* (4), 515–520.

Drasgow, F., Levine, M. V. & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38* (1), 67–86. doi: 10.1111/j.2044-8317.1985.tb00817.x

Drobny, F. (1900). Zum Begriffe 'Turnierstärke'. *Wiener Schachzeitung*, *III* (1), 171–176.

Drobny, F. (1901). Ueber eine neu Art der Preisvertheilung. *Wiener Schachzeitung*, *IV* (1), 2–4.

Dunlop, P. D., Telford, A. D. & Morrison, D. L. (2012). Not too little, but not too much: The perceived desirability of responses to personality items. *Journal of Research in Personality*, *46* (1), 8–18. doi: 10.1016/j.jrp.2011.10.004

Dunning, D. & Cohen, G. L. (1992). Egocentric definitions of traits and abilities in social judgment. *Journal of Personality and Social Psychology*, *63* (3), 341–355.

Dunning, D. & McElwee, R. O. (1995). Idiosyncratic trait definitions: Implications for self-description and social judgment. *Journal of Personality and Social Psychology*, *68* (5), 936–946.

Dwight, S. A., Porter Wolf, P. & Golden, J. H. (2002). Metatraits: Enhancing criterion-related validity through the assessment of traitedness. *Journal of Applied Social Psychology*, *32* (10), 2202–2212.

Eagle, M. N. (2007). Psychoanalysis and its critics. *Psychoanalytic Psychology*, *24* (1), 10–24. doi: 10.1037/0736-9735.24.1.10

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: The Dryden Press Inc.

Edwards, A. L. (1961). Social desirability or acquiescence in the MMPI? A case study with the SD scale. *The Journal of Abnormal and Social Psychology*, *63* (2), 351.

Edwards, A. L. (1983). *Techniques of attitude scale construction*. Ardent Media.

Edwards, A. L. & Abbott, R. D. (1973). Measurement of personality traits: Theory and technique. *Annual review of psychology*, *24* (1), 241–278.

Eid, M. & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16* (1), 20–30.

Eid, M. & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications* (S. 255–270). New York: Springer.

Ellingson, J. E., Sackett, P. R. & Connelly, B. S. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology*, *92* (2), 386–395. doi:

10.1037/0021-9010.92.2.386

Ellingson, J. E., Sackett, P. R. & Hough, L. M. (1999). Social desirability corrections in personality measurement: issues of applicant comparison and construct validity. *Journal of Applied Psychology*, *84* (2), 155.

Ellingson, J. E., Smith, D. B. & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology*, *86* (1), 122–133. doi: 10.1037/0021-9010.86.1.122

Ellson, D. G. & Ellson, E. C. (1953). Historical note on the rating scale. *Psychological Bulletin*, *50* (5), 383–384. doi: 10.1037/h0054149

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Routledge.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32* (3), 224–247. doi: 10.1177/0146621607302479

Emons, W. H. M. (2009). Detection and diagnosis of person misfit from patterns of summed polytomous item scores. *Applied Psychological Measurement*, *33* (8), 599–619. doi: 10.1177/0146621609334378

Endler, N. S. & Magnusson, D. (1976). Toward an interactional psychology of personality. *Psychological Bulletin*, *83* (5), 956–974. doi: 10.1037/0033-2909.83.5.956

Eysenck, H. J. (1992a). Four ways five factors are not basic. *Personality and Individual Differences*, *13* (6), 667–673. doi: 10.1016/0191-8869(92)90237-J

Eysenck, H. J. (1992b). A reply to Costa and McCrae. P or A and C—the role of theory. *Personality and Individual Differences*, *13* (8), 867–868. doi: 10.1016/0191-8869(92)90003-8

Eysenck, H. J. (1993). 'The structure of phenoytypic personality traits': Comment. *American Psychologist*, *48* (12), 1299–1300. doi: 10.1037/0003-066X.48.12.1299.b

Fechner, G. T. (1860a). *Elemente der Psychophysik I.* Leipzig: Breitkopf und Härtel.

Fechner, G. T. (1860b). *Elemente der Psychophysik II.* Leipzig: Breitkopf und Härtel.

Federn, E. (2005). Sigmund Freud. In H. E. Lück & R. Miller (Hrsg.),

Illustrierte Geschichte der Psychologie (S. 141–144). Weinheim: Beltz.

Ferguson, G. A. (1942). Item selection by the constant process. *Psychometrika*, *7* (1), 19–29. doi: 10.1007/BF02288601

Ferrando, P. J. (2004). Person reliability in personality measurement: An item response theory analysis. *Applied Psychological Measurement*, *28* (2), 126–140. doi: 10.1177/0146621603260917

Ferrando, P. J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research*, *42* (3), 481–507.

Ferrando, P. J. (2012). Assessing inconsistent responding in E and N measures: An application of person-fit analysis in personality. *Personality and Individual Differences*, *52* (6), 718–722. doi: 10.1016/j.paid.2011.12.036

Ferrando, P. J., Condon, L. & Chico, E. (2004). The convergent validity of acquiescence: an empirical study relating balanced scales and separate acquiescence scales. *Personality and Individual Differences*, *37* (7), 1331–1340. doi: 10.1016/j.paid.2004.01.003

Ferrando, P. J. & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, *63* (2), 427–448.

Festinger, L. (1957). *Theory of cognitive dissonance.* Stanford: Stanford University Press.

Festinger, L. & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, *58* (2), 203–210. doi: 10.1037/h0041593

Fichten, C. S. & Sunerton, B. (1983). Popular horoscopes and the 'Barnum Effect.'. *Journal of Psychology*, *114* (1), 123.

Filter, R. O. (1921). An experimental study of character traits. *Journal of Applied Psychology*, *5* (4), 297–317. doi: 10.1037/h0070544

Finn, J. A., Ben-Porath, Y. S. & Tellegen, A. (2015). Dichotomous versus polytomous response options in psychopathology assessment: method or meaningful variance? *Psychological Assessment*, *27* (1), 184–193. doi: 10.1037/pas0000044

Finney, S. J. & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Hrsg.), *Structural equation modeling: A second course.* Greenwich, Conn: IAP -

Information Age Publishing Inc.

Fischer, G. H. (1970). *A further note on estimation in Rasch's measurement model with two categories of answers* (Research Bulletin Nr. 3). Vienna: University of Vienna.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *46* (1), 59–77.

Fischer, G. H. & Scheiblechner, H. (1970a). Algorithmen und Programme fuer das probabilistische Testmodell von Rasch. *Psychologische Beitrage* (12), 23–51.

Fischer, G. H. & Scheiblechner, H. H. (1970b). *Two simple methods for asymptotically unbiased estimation in Rasch's measurement model with two categories of answers* (Research Bulletin Nr. 1). Wien: Psychologisches Institut der Universität Wien.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *222* (594-604), 309–368. doi: 10.1098/rsta.1922.0009

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700–725.

Fiske, D. W. & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin*, *52* (3), 217–250. doi: 10.1037/h0045276

Fisseni, H.-J. (1998). *Persönlichkeitspsychologie auf der Suche nach einer Wissenschaft ; ein Theorienüberblick* (4., überarb. und erw. Aufl.). Göttingen: Hogrefe.

Fleeson, W. & Noftle, E. (2008). The end of the person–situation debate: an emerging synthesis in the answer to the consistency question. *Social and Personality Psychology Compass*, *2* (4), 1667–1684. doi: 10.1111/j.1751-9004.2008.00122.x

Fonagy, P. (2015). The effectiveness of psychodynamic psychotherapies: An update. *World Psychiatry*, *14* (2), 137–150. doi: 10.1002/wps.20235

Fonagy, P., Rost, F., Carlyle, J.-a., McPherson, S., Thomas, R., Pasco Fearon, R. M., ... Taylor, D. (2015). Pragmatic randomized controlled trial of

long-term psychoanalytic psychotherapy for treatment-resistant depression: the Tavistock Adult Depression Study (TADS). *World Psychiatry*, *14* (3), 312–321.

Forer, B. R. (1949). The fallacy of personal validation: a classroom demonstration of gullibility. *The Journal of Abnormal and Social Psychology*, *44* (1), 118–123. doi: 10.1037/h0059240

Forero, C. G. & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full Information methods. *Psychological Methods*, *14* (3), 275–299. doi: 10.1037/a0015825

Formann, A. K. (1984). *Die Latent-Class-Analyse: Einführung in Theorie und Anwendung*. Beltz.

Formann, A. K. (1986). A note on the computation of the second-order derivatives of the elementary symmetric functions in the Rasch model. *Psychometrika*, *51* (2), 335–339.

Formann, A. K. (2002). Identifying types, response errors, and unscalable respondents from personality questionnaires. *Psychologische Beiträge*, *44*, 78–93.

Fraley, C. & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97* (458), 611–631. doi: 10.1198/016214502760047131

Franke, G. H. (2002). Faking bad in personality inventories: Consequences for the clinical context. *Psychologische Beiträge*, *44*, 17–23.

Freud, S. (1911). *Die Traumdeutung* (3. vermehrte Aufl.). Leipzig & Wien: Franz Deuticke.

Freud, S. (1923). *Das Ich und das Es*. Wien - Leipzig - Zürich: Internationaler Psychoanalytischer Verlag.

Freud, S. (1933). Die Zerlegung der psychischen Persönlichkeit. In *Neue Folge der Vorlesungen zur Einführung in die Psychoanalyse.*

Frey, A., Hartig, J. & Rupp, A. A. (2009). Booklet designs in large-scale assessments of student achievement: theory and practice. *Educational Measurement: Issues and Practice*, *28* (3), 39–53.

Freyd, M. (1924). Introverts and Extroverts. *Psychological Review*, *31* (1), 74–87. doi: 10.1037/h0075875

Friedman, L. (2011). Charles Brenner A Practitioner's Theorist. *Journal of*

the American Psychoanalytic Association, *59* (4), 679–700. doi: 10.1177/0003065111414891

Fromm, E. (1932). Die Psychoanalytische Charakterologie und ihre Bedeutung für die Sozialpsychologie. *Zeitschrift für Sozialforschung*, *1* (3), 253–.

Funder, D. C. (2001). Personality. *Annual Review of Psychology*, *52* (1), 197–221. doi: 10.1146/annurev.psych.52.1.197

Funder, D. C. & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology*, *60* (5), 773.

Furnham, A. & Schofield, S. (1987). Accepting personality test feedback: A review of the Barnum effect. *Current Psychology*, *6* (2), 162–178. doi: 10.1007/BF02686623

Förstl, H. & Förstl-Hautzinger-Roth (Hrsg.). (2006). *Neurobiologie psychischer Störungen*. Heidelberg: Springer Medizin.

Gaier, E. L. & Lee, M. C. (1953). Pattern analysis: the configural approach to predictive measurement. *Psychological Bulletin*, *50* (2), 140–148. doi: 10.1037/h0057283

Gaier, E. L., Lee, M. C. & McQuitty, L. L. (1953). Response patterns in a test of logical inference. *Educational and Psychological Measurement*, *13* (4), 550–567. doi: 10.1177/001316445301300402

Galton, F. (1884). Measurement of character. *Fortnightly Review*, *36*, 179 – 185.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, *45* (273-279), 135–145.

Garner, M. & Engelhard, G. (2000). Rasch measurement teory, the method of paired comparisons and graph theory. In M. Wilson & G. Engelhard (Hrsg.), *Objective measurement: theory into practice* (Bd. 5, S. 259–286). Stamford, Conneticut: Ablex Publishing Corporation.

Garner, M. & Engelhard Jr, G. (2002). An eigenvector method for estimating item parameters of the dichotomous and polytomous rasch models. *Journal of Applied Measurement*, *1* (3), 107–128.

Gebhardt, M., Heine, J.-H. & Sälzer, C. (2015). Schulische Kompetenzen von Schülerinnen und Schülern ohne sonderpädagogischen Förderbedarf im

gemeinsamen Unterricht. *Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete*, *84* (3), 246. doi: 10.2378/vhn2015.art28d

Gebhardt, M., Heine, J.-H., Zeuch, N. & Förster, N. (2015). Lernverlaufsdiagnostik im Mathematikunterricht der zweiten Klasse: Raschanalysen und Empfehlungen zur Adaptation eines Testverfahrens für den Einsatz in inklusiven Klassen. *Empirische Sonderpädagogik*, *7* (3), 206–222.

Gediga, G. (1998). *Skalierung: Eine Einführung in die Methodik zur Entwicklung von Test- und Messinstrumenten in den Verhaltenswissenschaften*. LIT.

Geiger, M., Sauter, R., Olderbak, S. & Wilhelm, O. (2016). Faking ability: Measurement and validity. *Personality and Individual Differences*, *101*, 480. doi: 10.1016/j.paid.2016.05.147

Geisser, S. (1992). Introduction to Fisher (1922) on the mathematical foundations of theoretical statistics. In S. Kotz & N. L. Johnson (Hrsg.), *Breakthroughs in Statistics* (S. 1–10). Springer New York. doi: 10.1007/978-1-4612-0919-5_1

Ghiselin, M. T. (1973). Darwin and evolutionary Psychology darwin initiated a radically new way of studying behavior. *Science*, *179* (4077), 964–968. doi: 10.1126/science.179.4077.964

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., ... Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31* (1), 4–19. doi: 10.1177/0146621606289485

Gifi, A. (1991). *Nonlinear multivariate analysis* (korrigierte Aufl.). Chichester [u.a.]: Wiley.

Gigerenzer, G. (1981). *Messung und Modellbildung in der Psychologie*. München: UTB Reinhardt.

Gignac, G. E. (2013). Modeling the balanced inventory of desirable responding: Evidence in favor of a revised model of socially desirable responding. *Journal of Personality Assessment*, *95* (6), 645–656. doi: 10.1080/00223891.2013.816717

Gilbert, D. T. (1991). How mental systems believe. *American psychologist*, *46* (2), 107.

Glas, C. A. W. (1988). The Rasch model and multistage testing. *Journal of*

Educational Statistics, *13* (1), 45–52. doi: 10.2307/1164950

Glas, C. A. W. (2009). What IRT can and cannot do. *Measurement: Interdisciplinary Research and Perspectives*, *7* (2), 91–93. doi: 10.1080/15366360903117020

Glas, C. A. W. & Verhelst, N. D. (1995). Tests of fit for polytomous rasch models. In G. Fischer & I. Molenaar (Hrsg.), *Rasch models: Foundations, recent developments, and applications.* New York: Springer.

Glickman, M. E. (1995). A comprehensive guide to chess ratings. *The American Chess Journal*, *3*, 59–102.

Glickman, M. E. (2005). Adaptive paired comparison design. *Journal of Statistical Planning and Inference*, *127*, 279–293.

Globalpark AG. (2010). *UNIPARK - EFS-Survey.* Köln: Questback GmbH; vormals Globalpark AG. `https://www.unipark.com/`

Goldberg, L. R. (1990). An alternative description of personality: the big-five factor structure. *Journal of Personality and Social Psychology*, *59* (6), 1216–1229.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4* (1), 26–42. doi: http://dx.doi.org/10.1037/1040-3590.4.1.26

Goldberg, L. R. (1993). The structure of phenotypic personality-traits. *American Psychologist*, *48* (1), 26–34. doi: 10.1037//0003-066X.48.1.26

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, *33* (2), 234–246. doi: 10.1111/j.2044-8317.1980.tb00610.x

Goldstein, H. (2015). Rasch measurement: a response to Payanides, Robinson and Tymms. *British Educational Research Journal*, *41* (1), 176–179. doi: 10.1002/berj.3170

Goldstein, H. & Blinkhorn, S. (1982). The Rasch model still does not fit. *British Educational Research Journal*, *8* (2), 167–170. doi: 10.1080/0141192820080207

Gollwitzer, M., Eid, M. & Jürgensen, R. (2005). Response styles in the assessment of anger expression. *Psychological Assessment*, *17* (1), 56–69. doi: 10.1037/1040-3590.17.1.56

González-Romá, V. & Espejo, B. (2003). Testing the middle response categories «not sure»,«in between» and «?» in polytomous items. *Psicothema*, *15*, 278–284.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association*, *70* (352), 755–768.

Goodyer, I. M., Tsancheva, S., Byford, S., Dubicka, B., Hill, J., Kelvin, R., ... Fonagy, P. (2011). Improving mood with psychoanalytic and cognitive therapies (IMPACT): a pragmatic effectiveness superiority trial to investigate whether specialised psychological treatment reduces the risk for relapse in adolescents with moderate to severe unipolar depression: study protocol for a randomised controlled trial. *Trials*, *175* (12), 1–12. doi: 10.1186/1745-6215-12-175

Gottfredson, G. D. (1999). John L. Holland's contributions to vocational psychology: A review and evaluation. *Journal of Vocational Behavior*, *55* (1), 15–40. doi: doi:DOI:10.1006/jvbe.1999.1695

Gottfredson, G. D., Jones, E. M. & Holland, J. L. (1993). Personality and vocational interests: The relation of Holland's six interest dimensions to five robust dimensions of personality. *Journal of Counseling Psychology*, *40* (4), 518.

Graf-Nold, A. (2005). Carl Gustaf Jung - der Individualisierungsprozeß und die Gegensatznatur der Psyche. In H. E. Lück & R. Miller (Hrsg.), *Illustrierte Geschichte der Psychologie* (S. 151–157). Weinheim: Beltz.

Gray, J. A. (1970). The psychophysiological basis of introversion-extraversion. *Behaviour Research and Therapy*, *8* (3), 249–266.

Gray, S. H. (2002). Evidence-based psychotherapeutics. *Psychodynamic Psychiatry*, *30* (1), 3-16.

Graziano, W. G. & Tobin, R. M. (2002). Agreeableness: Dimension of personality or social desirability artifact? *Journal of Personality*, *70* (5), 695–728. doi: 10.1111/1467-6494.05021

Green, D. P., Goldman, S. L. & Salovey, P. (1993). Measurement error masks bipolarity in affect ratings. *Journal of Personality and Social Psychology*, *64* (6), 1029–1041. doi: 10.1037/0022-3514.64.6.1029

Green, D. P., Salovey, P. & Truax, K. M. (1999). Static, dynamic, and causative

bipolarity of affect. *Journal of Personality and Social Psychology*, *76* (5), 856–867. doi: 10.1037/0022-3514.76.5.856

Greenacre, M. J. (Hrsg.). (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Greenacre, M. J. (2010). Correspondence Analysis. In P. P. B. McGaw (Hrsg.), *International encyclopedia of education* (3. Aufl., S. 103–111). Oxford: Elsevier.

Greenacre, M. J. & Blasius, J. (1994). *Correspondence analysis in the social sciences: recent developments and applications*. Academic Press.

Greenleaf, E. A. (1992a). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research (JMR)*, *29* (2), 176–188.

Greenleaf, E. A. (1992b). Measuring extreme response style. *Public Opinion Quarterly*, *56* (3), 328–351.

Greving, B. (2007). Messen und Skalieren von Sachverhalten. In S. Albers (Hrsg.), *Methodik der empirischen Forschung* (S. 65–78). Wiesbaden: Gabler.

Grieve, R. & McSwiggan, C. (2014). Predicting intentions to fake in psychological testing: Which normative beliefs are important? *Revista de Psicología del Trabajo y de las Organizaciones*, *30* (1), 23–28. doi: 10.5093/tr2014a3

Griffith, R. L., Lee, L. M., Peterson, M. H. & Zickar, M. J. (2011). First dates and little white lies: A trait contract classification theory of applicant faking behavior. *Human Performance*, *24* (4), 338–357. doi: 10.1080/08959285.2011.597475

Griffith, R. L. & Peterson, M. H. (2011). One piece at a time: The puzzle of applicant faking and a call for theory. *Human Performance*, *24* (4), 291–301. doi: 10.1080/08959285.2011.597474

Gross, O. (1902). *Die cerebrale Sekundärfunction*. Leipzig: F.C.W. Vogel.

Grünbaum, A. (1988). *Die Grundlagen der Psychoanalyse: Eine philosophische Kritik*. Stuttgart: Reclam.

Gu, H., Wen, Z. & Fan, X. (2015). The impact of wording effect on reliability and validity of the Core Self-Evaluation Scale (CSES): A bi-factor perspective. *Personality and Individual Differences*, *83*, 142–147. doi:

10.1016/j.paid.2015.04.006

Guilford, J. P. (1934). Introversion-extroversion. *Psychological Bulletin*, *31* (5), 331–354. doi: 10.1037/h0072741

Guilford, J. P. (1975). Factors and factors of personality. *Psychological Bulletin*, *82* (5), 802–814. doi: 10.1037/h0077101

Guilford, J. P. & Braly, K. W. (1930). Extroversion and introversion. *Psychological Bulletin*, *27* (2), 96–107. doi: 10.1037/h0073968

Guilford, J. P. & Guilford, R. B. (1939a). Personality factors D, R, T, and A. *The Journal of Abnormal and Social Psychology*, *34* (1), 21–36. doi: 10.1037/h0056344

Guilford, J. P. & Guilford, R. B. (1939b). Personality factors N and GD. *The Journal of Abnormal and Social Psychology*, *34* (2), 239–248. doi: 10.1037/h0063296

Gulliksen, H. (1946). Paired comparisons and the logic of measurement. *Psychological Review*, *53* (4), 199–213. doi: 10.1037/h0061673

Gulliksen, H. (1950). *Theory of Mental Test*. New York: Wiley.

Gur, R. C. & Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology*, *37* (2), 147–169. doi: 10.1037/0022-3514.37.2.147

Gustafsson, J.-E. (1980a). A solution of the conditional estimation problem for long tests in the rasch model for dichotomous items. *Educational and Psychological Measurement*, *40* (2), 377 –385. doi: 10.1177/001316448004000214

Gustafsson, J.-E. (1980b). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *33* (2), 205–233. doi: 10.1111/j.2044-8317.1980.tb00609.x

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150.

Guttman, L. (1947). The cornell technique for scale and intensity analysis. *Educational and Psychological Measurement*, *7* (2), 247–279. doi: 10.1177/001316447007002004

Guttman, L. (1950). The basis of scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. Star & J. A. Clausen (Hrsg.), *Studies in social psychology in World War II* (Bd. IV Measurement and

prediction, S. 362–412). Princeton: Princeton University Press.

Haaga, D. A. F., Ahrens, A. H., Schulman, P., Seligman, M. E. P., DeRubeis, R. J. & Minarik, M. L. (1995). Metatraits and cognitive assessment: Application to attributional style and depressive symptoms. *Cognitive Therapy and Research*, *19* (1), 121–142. doi: 10.1007/BF02229680

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, *5* (5), 815–841.

Haberman, S. J., Sinharay, S. & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, *78* (3), 417–440. doi: 10.1007/s11336-012-9305-1

Häcker, H. O. (2014). Response set. In M. A. A. Wirtz (Hrsg.), *Dorsch Lexikon der Psychologie* (18. Aufl., S. 1443). Bern: Verlag Hans Huber.

Haladyna, T. M. & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53* (4), 999–1010. doi: 10.1177/0013164493053004013

Hamamura, T., Heine, S. & Paulhus, D. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences*, *44* (4), 932–942.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston : Hingham, MA, U.S.A: Kluwer-Nijhoff Pub. ; Distributors for North America, Kluwer Boston.

Hardy, B. & Ford, L. R. (2014). It's not me, it's you miscomprehension in surveys. *Organizational Research Methods*, *17* (2), 138–162. doi: 10.1177/1094428113520185

Harnisch, D. L. & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, *18* (3), 133–146.

Hart, C. M., Ritchie, T. D., Hepper, E. G. & Gebauer, J. E. (2015). The Balanced Inventory of Desirable Responding short form (BIDR-16). *SAGE Open*, *5* (4), 2158244015621113. doi: 10.1177/2158244015621113

Harvey, R. J. (2016). Improving measurement via item response theory great idea, but hold the Rasch. *The Counseling Psychologist*, 0011000015615427. doi: 10.1177/0011000015615427

Hauenstein, N. M. A., Bradley, K. M., O'Shea, P. G., Shah, Y. J. & Magill, D. P. (2017). Interactions between motivation to fake and personality item characteristics: Clarifying the process. *Organizational Behavior and Human Decision Processes*, *138*, 74–92. doi: 10.1016/j.obhdp.2016.11.002

Hayes, A. F. & Dunning, D. (1997). Construal processes and trait ambiguity: Implications for self-peer agreement in personality judgment. *Journal of Personality and Social Psychology*, *72* (3), 664–677. doi: 10.1037/0022-3514.72.3.664

He, J. & van de Vijver, F. J. R. (2015). Self-presentation styles in self-reports: Linking the general factors of response styles, personality traits, and values in a longitudinal study. *Personality and Individual Differences*, *81*, 129–134. doi: 10.1016/j.paid.2014.09.009

He, J., van de Vijver, F. J. R., Espinosa, A. D. & Mui, P. H. (2014). Toward a unification of acquiescent, extreme, and midpoint response styles: A multilevel study. *International Journal of Cross Cultural Management*, *14* (3), 306–322. doi: 10.1177/1470595814541424

Healey, B. J. (1973). Pilot study on the applicability of the Music Preference Test of Personality. *Journal of Music Therapy*, *10* (1), 36–45. doi: 10.1093/jmt/10.1.36

Heilbrun, A. B. (1962). Social desirability and the relative validities of achievement scales. *Journal of Consulting Psychology*, *26* (4), 383.

Heine, J.-H. (2010). *Extremer Antwortstil als Determinante bei der Beantwortung des NEO-PI-R* (Unveröffentlichte Diplomarbeit). München: Ludwig-Maximilians-Universität.

Heine, J.-H. (2019). *pairwise: Rasch Model Parameters by Pairwise Algorithm* (R package version 0.4.4-5.2). `https://CRAN.R-project.org/package=pairwise`

Heine, J.-H., Alexandrowicz, R. W. & Stemmler, M. (2019). *confreq: Configural Frequencies Analysis Using Log-Linear Modeling* (R package version 1.5.4-5). `https://CRAN.R-project.org/package=confreq`

Heine, J.-H., Gebhardt, M., Schwab, S., Neumann, P., Gorges, J. & Wild, E.

(2018). Testing psychometric properties of the CFT 1-R for students with special educational needs. *Psychological Test and Assessment Modeling*, *60* (1), 3–27.

Heine, J.-H., Mang, J., Borchert, L., Gomolka, J., Kröhme, U., Goldhammer, F. & Sälzer, C. (2016). Kompetenzmessung in PISA 2015. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme & O. Köller (Hrsg.), *PISA 2015 Eine Studie zwischen Kontinuität und Innovation* (S. 383–430). Münster: Waxmann Verlag.

Heine, J.-H., Sälzer, C., Borchert, L., Siberns, H. & Mang, J. (2013). Technische Grundlagen des fünften internationalen Vergleichs. In M. Prenzel, C. Sälzer, E. Klieme & O. Köller (Hrsg.), *PISA 2012 - Fortschritte und Herausforderungen in Deutschland.* Münster: Waxmann.

Heine, J.-H. & Tarnai, C. (2015). Pairwise rasch model item parameter recovery under sparse data conditions. *Psychological Test and Assessment Modeling*, *57* (1), 3–36.

Heiser, W. J. & Meulman, J. (1983). Analyzing rectangular tables by joint and constrained multidimensional scaling. *Journal of Econometrics*, *22* (1-2), 139–167. doi: 10.1016/0304-4076(83)90097-0

Helmes, E., Holden, R. R. & Ziegler, M. (2015). Response bias, malingering, and impression management. In G. J. Boyle, D. H. Saklofske & G. Matthews (Hrsg.), *Measures of personality and social psychological constructs.* Amsterdam: Academic Press.

Hemker, B. T., Sijtsma, K., Molenaar, I. W. & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, *61* (4), 679–693. doi: 10.1007/BF02294042

Hemker, B. T., Sijtsma, K., Molenaar, I. W. & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62* (3), 331–347. doi: 10.1007/BF02294555

Henson, J. M., Reise, S. P. & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling: A Multidisciplinary Journal*, *14* (2), 202–226. doi: 10.1080/10705510709336744

Hernández, A., Drasgow, F. & González-Romá, V. (2004). Investigating the

functioning of a middle category by means of a mixed-measurement model. *Journal of applied psychology*, *89* (4), 687.

Hernández, A., Espejo, B. & González-Romá, V. (2006). The functioning of central categories middle level and sometimes in graded response scales: does the label matter? *PSICOTHEMA-OVIEDO-*, *18* (2), 300.

Herzberg, P. Y. (2002). Zur psychometrischen Optimierung einer Reaktanzskala mittels klassischer und IRT-basierter Analysemethoden. *Diagnostica*, *48* (4), 163–171. doi: 10.1026//0012-1924.48.4.163

Higgins, N. C., Zumbo, B. D. & Hay, J. L. (1999). Construct validity of attributional style: Modeling context-dependent item sets in the attributional style questionnaire. *Educational and Psychological Measurement*, *59* (5), 804–820. doi: 10.1177/00131649921970152

Hill, M. O. (1974). Correspondence analysis: A neglected multivariate method. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *23* (3), 340–354. doi: 10.2307/2347127

Hofstee, W. K. B., Berge, J. M. F. T. & Hendriks, A. A. J. (1998). How to score questionnaires. *Personality and Individual Differences*, *25* (5), 897–909. doi: 10.1016/S0191-8869(98)00086-5

Hoijtink, H. (1990). A latent trait model for dichotomous choice data. *Psychometrika*, *55* (4), 641–656. doi: 10.1007/BF02294613

Hoijtink, H. (1991). The measurement of latent traits by proximity items. *Applied Psychological Measurement*, *15* (2), 153–169. doi: 10.1177/014662169101500205

Holden, R. R. & Book, A. S. (2009). Using hybrid Rasch-latent class modeling to improve the detection of fakers on a personality inventory. *Personality and Individual Differences*, *47* (3), 185–190. doi: 10.1016/j.paid.2009.02.024

Holden, R. R. & Passey, J. (2010). Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and Individual Differences*, *49* (5), 446–450. doi: 10.1016/j.paid.2010.04.015

Holland, J. L. (1958). A personality inventory employing occupational titles. *Journal of Applied Psychology*, *42* (5), 336–342. doi: 10.1037/h0047330

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling*

Psychology, *6* (1), 35–45. doi: doi:10.1037/h0040767

Holland, J. L. (1963). A theory of vocational choice. I. Vocational images and choice. *Vocational Guidance Quarterly*, *11* (4), 232–239.

Holland, J. L. (1965). *Manual for the Vocational Preference Inventory* (5. Aufl.). Iowa City: Educational Research Associates.

Holland, J. L. (1966). A psychological classification scheme for vocations and major fields. *Journal of Counseling Psychology*, *13* (3), 278–288. doi: 10.1037/h0023725

Holland, J. L. (1971). *A counselor's guide: For use with the self directed search.* Consulting Psychologists Press.

Holland, J. L. (1973). *Making vocational choices;.*

Holland, J. L. (1975). *Manual for the vocational preference inventory.* Palo Alto, Ca.: Consulting Psychologists Press.

Holland, J. L. (1979). *The self-directed search professional manual.* Palo Alto, California: Consulting Psychologists Press.

Holland, J. L. (1985). *Making vocational choices. A theory of vocational personalities and work environments.* Englewood-Cliffs, NJ: Prentice-Hall.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments.* Odessa: Psychological Assessment Resources.

Holland, J. L. (1999). Why interest inventories are also personality inventories. In M. L. Savickas & A. R. Spokane (Hrsg.), *Vocational Interests: Meaning, Measurement, and Counseling Use* (illustrierte Aufl., S. 87–133). Davies-Black Publishing.

Holland, J. L., Gottfredson, D. C. & Power, P. G. (1980). Some diagnostic scales for research in decision making and personality: Identity, information, and barriers. *Journal of Personality and Social Psychology*, *39* (6), 1191–1200. `http://content.apa.org/journals/psp/39/6/1191` doi: 10.1037/h0077731

Holland, J. L., Johnston, J. A. & Asama, F. N. (1994). More Evidence for the Relationship Between Holland's Personality Types and Personality Variables. *Journal of Career Assessment*, *2* (4), 331–340. doi: 10.1177/106907279400200401

Holland, J. L., Johnston, J. A. & Asama, N. F. (1993). The Vocational Identity

Scale: A Diagnostic and Treatment Tool. *Journal of Career Assessment*, *1* (1), 1–12. doi: 10.1177/106907279300100102

Holland, J. L., Krause, A. H., Nixon, M. E. & Trembath, M. F. (1953). The classification of occupations by means of Kuder Interest profiles: I. The development of interest groups. *Journal of Applied Psychology*, *37* (4), 263–269. doi: 10.1037/h0057095

Holland, J. L., Whitney, D., Cole, N. & Richards, J. (1969). *An empirical occupational classification derived from a theory of personality and intended for practice and research.* (Bericht). Iowa City, IA: American Coll. Testing Program.

Hong, S. & Min, S.-Y. (2007). Mixed Rasch modeling of the Self-Rating Depression Scale incorporating latent class and Rasch rating scale models. *Educational and Psychological Measurement*, *67* (2), 280–299. doi: 10.1177/0013164406292072

Hooper, D., Coughlan, J. & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6* (1), 53–60.

Horan, P. M., DiStefano, C. & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling*, *10* (3), 435–455.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24* (6), 417–441. doi: 10.1037/h0071325

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D. & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, *75* (5), 581–595. doi: 10.1037/0021-9010.75.5.581

Hoyt, C. (1945). The principle of likelihood as a basis for tests of significance. *The Journal of Experimental Education*, *13* (3).

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6* (1), 1–55. doi: 10.1080/10705519909540118

Huang, J. & Mead, A. D. (2014). Effect of personality item writing on psy-

chometric properties of ideal-point and Likert scales. *Psychological Assessment*, *26* (4), 1162–1172. doi: 10.1037/a0037273

Hubert, L. (1974). Problems of seriation using a subject by item response matrix. *Psychological Bulletin*, *81* (12), 976–983. doi: 10.1037/h0037348

Hubert, L. (1976). Seriation using asymmetric proximity matrices. *British Journal of Mathematical and Statistical Psychology*, *29* (1), 32–52. doi: 10.1111/j.2044-8317.1976.tb00701.x

Hubert, L. & Arabie, P. (1987). Evaluating order hypotheses within proximity matrices. *Psychological Bulletin*, *102* (1), 172–178.

Hui, C. H. & Triandis, H. C. (1989)sep. Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20* (3), 296–309. doi: 10.1177/0022022189203004

Humphry, S. M. (2011)mar. The Role of the Unit in Physics and Psychometrics. *Measurement: Interdisciplinary Research and Perspectives*, *9* (1), 1–24. doi: 10.1080/15366367.2011.558442

Humphry, S. M. (2013)dec. A middle path between abandoning measurement and measurement theory. *Theory & Psychology*, *23* (6), 770–785. doi: 10.1177/0959354313499638

Hurley, J. R. (1998). Timidity as a response style to psychological questionaires. *The Journal of Psychology*, *132* (2), 201 – 210.

Hurvich, C. M. & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76* (2), 297.

Iachan, R. (1984). A measure of agreement for use with the holland classification system. *Journal of Vocational Behavior*, *24* (2), 133–141.

Idaszak, J. R. & Drasgow, F. (1987). A revision of the Job Diagnostic Survey: Elimination of a measurement artifact. *Journal of Applied Psychology*, *72* (1), 69–74. doi: 10.1037/0021-9010.72.1.69

Imbasciati, A. (2003). Cognitive sciences and psychoanalysis: A possible convergence. *Psychodynamic Psychiatry*, *31* (4), 627.

Ivanov, V. K. & Geake, J. G. (2003). The Mozart Effect and primary school children. *Psychology of Music*, *31* (4), 405 –413. doi: 10.1177/03057356030314005

Jackson, D. & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, *55* (4), 218.

Jacoby, W. G. (1991). *Data theory and dimensional analysis*. Thousand Oaks: SAGE Publications, Inc.

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education*, *38* (12), 1217–1218. doi: 10.1111/j.1365-2929.2004.02012.x

Jansen, P. G. W. (1981). Spezifisch objektive Messung im Falle monotoner Einstellungsitems. *Zeitschrift für Sozialpsychologie*, *12* (1), 24–41.

Jansen, P. G. W. (1983). *Rasch analysis of attitudinal data* (Monographie). Radboud Universiteit, Nijmegen.

Jerrim, J., Micklewright, J., Heine, J.-H., Sälzer, C. & McKeown, C. (2018)feb. PISA 2015: how big is the 'mode effect' and what has been done about it? *Oxford Review of Education*, *44* (4), 476–493.

Jo, M.-S. (2000). Controlling social- desirability bias via method factors of direct and indirect questioning in structural equation models. *Psychology & Marketing*, *17* (2), 137–148.

Jo, M.-S., Nelson, J. & Kiecker, P. (1997). A model for controlling social desirability bias by direct and indirect questioning. *Marketing Letters*, *8* (4), 429–437. doi: 10.1023/A:1007951313872

Joe, H. & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75* (3), 393–419. doi: 10.1007/s11336-010-9165-5

Joerin Fux, S. (2003). *Persönlichkeit und Berufstätigkeit. Theorie und Instrumente von John Holland im deutschsprachigen Raum, unter Adaptation und Weiterentwicklung von Self-directed Search (SDS) und Position Classification Inventory (PCI).* Göttingen: Cuvillier Verlag.

Joerin Fux, S., Stoll, F., Bergmann, C. & Eder, F. (2003). *EXPLORIX – Das Werkzeug zur Berufswahl und Laufbahnplanung.* Bern: Huber.

Johanson, G. & Alsmadi, A. (2002). Differential person functioning. *Educational and Psychological Measurement*, *62* (3), 435–443. doi: 10.1177/00164402062003003

Johanson, G. A. & Osborn, C. J. (2004). Acquiescence as differential person functioning. *Assessment & Evaluation in Higher Education*, *29* (5), 535–548.

John, O. P., Angleitner, A. & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European*

journal of Personality, *2* (3), 171–203.

John, O. P., Donahue, E. M. & Kentle, R. L. (1991). *The Big Five Inventory-Versions 4a and 54* (Bericht). Berkeley: University of California, Berkeley.

John, O. P. & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Hrsg.), *Handbook of personality: Theory and research* (2. Aufl., S. 102–138). New York: Guilford.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39* (1), 103–129. doi: 10.1016/j.jrp.2004.09.009

Johnson, M. S. (2006). Nonparametric estimation of item and respondent locations from unfolding-type items. *Psychometrika*, *71* (2), 257–279. doi: 10.1007/s11336-003-1098-9

Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software*, *20* (10), 1–24.

Johnson, M. S. & Junker, B. W. (2003). Using data augmentation and markov chain monte carlo for the estimation of unfolding response models. *Journal of Educational and Behavioral Statistics*, *28* (3), 195–230. doi: 10.3102/10769986028003195

Jones, E. E., Gergen, K. J. & Jones, R. G. (1963). Tactics of ingratiation among leaders and subordinates in a status hierarchy. *Psychological Monographs: General and Applied*, *77* (3), 1–20. doi: 10.1037/h0093832

Jones, L. V. & Thissen, D. (2006). A History and Overview of Psychometrics. In C. R. Rao & S. Sinharay (Hrsg.), *Psychometrics* (Bde. Handbook of Statistics, 26). Amsterdam: Elsevier.

Judd, C. M. & Kulik, J. A. (1980). Schematic effects of social attitudes on information processing and recall. *Journal of Personality and Social Psychology*, *38* (4), 569–578. doi: 10.1037/0022-3514.38.4.569

Jung, C. G. (1921). *Psychologische Typen*. Zürich: Rascher.

Kam, C. C. S. & Meyer, J. P. (2015). Implications of item keying and item valence for the investigation of construct dimensionality. *Multivariate Behavioral Research*, *50* (4), 457–469. doi: 10.1080/00273171.2015.1022640

Kamakura, W. A. & Balasubramanian, S. K. (1989). Tailored interviewing: An

application of item response theory for personality measurement. *Journal of Personality Assessment*, *53* (3), 502.

Kandel, E., Schwartz, J. & Jessel, T. (2000). *Principles of neural science* (4. Aufl.). New York: McGraw-Hill.

Kandel, E. R. (2005). *Psychiatry, psychoanalysis, and the new biology of mind* (1. Aufl.). Washington, DC: American Psychiatric Pub.

Kaplan, K. J. (1972). On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychological Bulletin*, *77* (5), 361–372. doi: 10.1037/h0032590

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of applied measurement*, *1* (2), 152–176.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, *2* (4), 389–423.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16* (4), 277–298.

Karabatsos, G. (2006). Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology*, *50* (2), 123–148.

Keats, J. A. (1974). Applications of projective transformations to test theory. *Psychometrika*, *39* (3), 359–360. doi: 10.1007/BF02291709

Keller, F. (2012). Latent-Class und Mixed-Rasch-Modelle zur Identifizierung skalierbarer und unskalierbarer Personengruppen in der Allgemeinen Depressionsskala. In W. Kempf & R. Langeheine (Hrsg.), *Item-Response-Modelle in der sozialwissenschaftlichen Forschung.* Berlin: Regener.

Keller, F. & Kempf, W. (1997). Some latent trait and latent class analyses of the Beck-Depression-Inventory (BDI). In J. Rost & R. Langeheine (Hrsg.), *Applications of latent trait and latent class models in the social sciences* (S. 314–323). Münster: Waxmann.

Kendall, D. G. (2004). Seriation. In *Encyclopedia of statistical sciences.* John Wiley & Sons, Inc.

Kendall, M. G. (1955). Further contributions to the theory of paired comparisons. *Biometrics*, *11* (1), 43–62. doi: 10.2307/3001479

Kenrick, D. T. & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist*, *43* (1), 23–34. doi: 10.1037/0003-066X.43.1.23

Kernberg, O. F. (1993). The current status of psychoanalysis. *Journal of the American Psychoanalytic Association*, *41* (1), 45–62. doi: 10.1177/000306519304100102

Khalid, M. N. & Glas, C. A. W. (2014). A scale purification procedure for evaluation of differential item functioning. *Measurement*, *50*, 186–197. doi: 10.1016/j.measurement.2013.12.019

Khorramdel, L. (2014). The influence of different rating scales on impression management in high stakes assessment. *Psychological Test and Assessment Modeling*, *56* (2), 154–167.

Kiefer, C. & Benit, N. (2016). What is applicant faking behavior? A review on the current state of theory and modeling techniques. *Journal of European Psychology Students*, *7* (1), 9–19. doi: 10.5334/jeps.345

Kieser, M. & Victor, N. (1999). Configural frequency analysis (CFA) revisited - a new look at an old approach. *Biometrical journal*, *41* (8), 967–983.

Klages, L. (1910). *Prinzipien der Charakterologie*. Leipzig: Barth.

Klages, L. (1926). *Die Grundlagen der Charakterkunde*. Leipzig: Barth.

Klauer, K. C. (1995). The assessment of person fit. In *Rasch models. Foundations, recent developments, and applications* (S. 97–110). New York: Springer.

Klinkenberg, E. (2001). A logistic IRT model for decreasing and increasing item characteristic curves. In A. Boomsma, M. A. J. Duijn & T. A. B. Snijders (Hrsg.), *Essays on Item Response Theory* (Bde. Lecture notes in statistics, 157, S. 173–192). New York: Springer.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, *55* (2), 312–320. doi: 10.1037/0022-3514.55.2.312

Knowles, E. S. & Byers, B. (1996). Reliability shifts in measurement reactivity. Driven by content engagement or self-engagement? *Journal of Personality and Social Psychology*, *70* (5), 1080–1090. doi: 10.1037/0022-3514.70.5.1080

Knowles, E. S. & Condon, C. A. (1999). Why people say yes. A dual-process

theory of acquiescence. *Journal of Personality and Social Psychology*, *77* (2), 379–386. doi: 10.1037/0022-3514.77.2.379

Knowles, E. S. & Nathan, K. T. (1997). Acquiescent responding in self-reports: Cognitive style or social concern? *Journal of Research in Personality*, *31*, 293–301.

Knuth, D. E. (1992). Two notes on notation. *The American Mathematical Monthly*, *99* (5), 403–422. doi: 10.2307/2325085

Knutson, B., Wolkowitz, O. M., Cole, S. W., Chan, T., Moore, E. A., Johnson, R. C., ... Reus, V. I. (1998). Selective alteration of personality and social behavior by serotonergic intervention. *American Journal of Psychiatry*, *155* (3), 373–379.

Kohut, H. (1971). *The Analysis of the Self. A Systematic Approach to the Psychoanalytic Treatment of Narcissistic Personality Disorders.* New York: International Universities Press.

Kohut, H. (1977). *The Restoration of the Self.* Madison CO: International Universities Press.

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking.* New York: Springer.

Konstabel, K., Aavik, T. & Allik, J. (2006). Social desirability and consensual validity of personality traits. *European Journal of Personality*, *20* (7), 549–566. doi: 10.1002/per.593

Koretz, D. (2005). *Alignment, high stakes, and the inflation of test scores* (CSE Report Nr. 655). Harvard: Graduate School of Education.

Kraepelin, E. (1983). *Lebenserinnerungen* (H. Hippius, G. Peters & D. Ploog, Hrsg.). Berlin: Springer.

Krapp, A. & Lewalter, D. (2001). Development of interests and interest-based motivational orientations. A longitudinal study in vocational school and work settings. In S. Volet & S. Järvelä (Hrsg.), *Motivation in learning contexts: theoretical and methodological implications* (S. 209–232). London: Elsevier.

Kraut, J. (1995). *Testkonstruktion und Testtheorie.* Weinheim: Beltz, Psychologie Verlags Union.

Krauth, J. & Lienert, G. A. (1973). *Die Konfigurationsfrequenzanalyse (KFA) und ihre Anwendung in Psychologie und Medizin: ein multivariates nicht-*

parametrisches Verfahren zur Aufdeckung von Typen und Syndromen ; mit 70 Tabellen. Freiburg: Alber Karl.

Kretschmer, E. (1977). *Körperbau und Charakter: Untersuchungen zum Konstitutionsproblem und zur Lehre von den Temperamenten* (26., neubearb. u. erw. Aufl.). Berlin: Springer.

Kromrey, H. (1994). *Empirische Sozialforschung* (6. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5* (3), 213–236. doi: 10.1002/acp.2350050305

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50* (1), 537.

Krosnick, J. A. & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51* (2), 201.

Kruger, J. (1999). Lake Wobegon be gone! The 'below-average effect' and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology*, *77* (2), 221–232. doi: 10.1037/0022-3514.77.2.221

Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77* (6), 1121.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29* (1), 1–27. doi: 10.1007/BF02289565

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, *29* (2), 115–129.

Kubinger, K. D. (1988). Aktueller Stand und kritsche Würdigung der Probabilistischen Testtheorie. In K. D. Kubinger (Hrsg.), *Moderne Test Theorie*. Weinheim: Psychologie Verlags Union.

Kubinger, K. D. (2000). Replik auf Jürgen Rost 'Was ist aus dem Rasch-Modell geworden?': Und für die Psychologische Diagnostik hat es doch revolutionäre Bedeutung. *Psycologische Rundschau*, *51* (1), 33–34.

Kubinger, K. D. (2002). On faking personality inventories. *Psychologische*

Beiträge, *44* , 10–16.

Kubinger, K. D. (2005). Psychological test calibration using the rasch model - some critical suggestions on traditional approaches. *International Journal of Testing*, *5* (4), 377–394. doi: 10.1207/s15327574ijt0504_3

Kubinger, K. D. (2017). Adaptive testing. In *Principles and methods of test construction: standards and recent advances* (Bde. Psychological Assessment - Science and Practice, 3, S. 104–119). Göttingen: Hogrefe.

Kubinger, K. D. & Draxler, C. (2007a). A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In M. von Davier & C. H. Carstensen (Hrsg.), *Multivariate and mixture distribution Rasch models: Extensions and applications.* (S. 293–309). New York, NY US: Springer Science + Business Media.

Kubinger, K. D. & Draxler, C. (2007b). Probleme bei der Testkonstruktion nach dem Rasch-Modell. *Diagnostica*, *53* (3), 131–143. doi: 10.1026/0012-1924.53.3.131

Kubinger, K. D., Steinfeld, J., Reif, M. & Yanagida, T. (2012). Biased (conditional) parameter estimation of a Rasch model calibrated item pool administered according to a branched testing design. *Psychological Test and Assessment Modeling*, *54* (4), 450–460.

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2* (3), 151–160. doi: 10.1007/BF02288391

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, *33* (2), 188–229. doi: 10.1177/0049124103262065

Kulas, J. T., Klahr, R. & Knights, L. (2018). Confound it! Social desirability and the 'reverse-scoring' method effect. *European Journal of Psychological Assessment*. doi: 10.1027/1015-5759/a000459

Kulas, J. T. & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, *43* (3), 489–493.

Kulas, J. T. & Stachowski, A. A. (2013). Respondent rationale for neither agreeing nor disagreeing: Person and item contributors to middle category endorsement intent on Likert personality indicators. *Journal of Research*

in Personality, *47* (4), 254–262. doi: 10.1016/j.jrp.2013.01.014

Kulas, J. T., Stachowski, A. A. & Haynes, B. A. (2008). Middle Response Functioning in Likert-responses to Personality Items. *Journal of Business and Psychology*, *22* (3), 251–259. doi: 10.1007/s10869-008-9064-2

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22* (1), 79–86. doi: 10.2307/2236703

Kuncel, N. R. & Borneman, M. J. (2007). Toward a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment*, *15* (2), 220–231. doi: 10.1111/j.1468-2389.2007.00383.x

Kuncel, N. R., Borneman, M. J. & Kiger, T. (2012). Innovative item response process and bayesian faking detection methods. In M. Ziegler, C. MacCann & R. Roberts (Hrsg.), *New perspectives on faking in personality assessment* (S. 3–16). Oxford: Oxford University Press.

Kuncel, N. R. & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items. Implications for detecting desirable response style and scale development. *Personnel Psychology*, *62* (2), 201–228.

Kwan, V. S. Y., John, O. P., Kenny, D. A., Bond, M. H. & Robins, R. W. (2004). Reconceptualizing individual differences in self-enhancement bias: An interpersonal approach. *Psychological Review*, *111* (1), 94–110. doi: 10.1037/0033-295X.111.1.94

Kwan, V. S. Y., John, O. P., Robins, R. W. & Kuang, L. L. (2008). Conceptualizing and assessing self-enhancement bias: A componential approach. *Journal of Personality and Social Psychology*, *94* (6), 1062.

Kyngdon, A. (2006). An empirical study into the theory of unidimensional unfolding. *Journal of Applied Measurement*, *7* (4), 369–393.

Lam, T. C. & Stevens, J. J. (1994). Effects of content polarization, item wording, and rating scale width on rating response. *Applied Measurement in Education*, *7* (2), 141–158.

Lamiell, J. T. (2006). Ursprungsmythos. William Stern (1871-1938) und der 'Ursprungsmythos' der differentiellen Psychologie. *Journal für Psychologie*.

Lance, C. E., Cornwell, J. M. & Mulaik, S. A. (1988). Limited information

parameter estimates for latent or mixed manifest and latent variable models. *Multivariate Behavioral Research*, *23* (2), 171–187. doi: 10.1207/s15327906mbr2302_3

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. *Diagnostica*, *47* (3), 111–121. doi: 10.1026//0012-1924.47.3.111

Langeheine, R. (1980). *Log-lineare Modelle zur multivariaten Analyse qualitativer Daten. Eine Einführung*. München: Oldenbourg Verlag.

Langmeyer, A., Guglhör-Rudan, A. & Tarnai, C. (2012). What do music preferences reveal about personality? A cross-cultural replication using self-ratings and ratings of music samples. *Journal of Individual Differences*, *33* (2), 119–130. doi: 10.1027/1614-0001/a000082

Lanning, K. (1991). *Consistency, scalability, and personality measurement*. New York: Springer. doi: 10.1007/978-1-4612-3072-4

Larson, L. M. & Borgen, F. H. (2002). Convergence of vocational interests and personality: Examples in an adolescent gifted sample. *Journal of Vocational Behavior*, *60* (1), 91–112. doi: 10.1006/jvbe.2001.1821

Larson, L. M., Rottinghaus, P. J. & Borgen, F. H. (2002). Meta-analyses of Big Six interests and Big Five personality factors. *Journal of Vocational Behavior*, *61* (2), 217–239.

Latcheva & Davidov, E. (2014). Skalen und Indizes. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S. 745–756). Wiesbaden: Springer VS.

Laux, H. (2005). *Entscheidungstheorie* (6., durchges. Aufl.). Berlin: Springer.

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. Star & J. A. Clausen (Hrsg.), *Studies in social psychology in World War II* (Bd. IV Measurement and prediction, S. 362–412). Princeton: Princeton University Press.

Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Hrsg.), *Psychology: A Study of a Science* (Bd. 3. Formulations of the person and the social context, S. 476–543). New York: McGraw-Hill Book Company, Inc.

Lee, H. & Geisinger, K. F. (2015). The matching criterion purification for differential item functioning analyses in a large-scale assessment. *Educational and Psychological Measurement*, 1–23. doi: 10.1177/0013164415585166

Lee, K. & Ashton, M. C. (2004). Psychometric properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research*, *39*, 329–358. doi: 10.1207/s15327906mbr3902_8

Leenen, I. & Van Mechelen, I. (2004). A conjunctive parallelogram model for pick any/n data. *Psychometrika*, *69* (3), 401–420.

Lentz, T. F. (1938). Acquiescence as a factor in the measurement of personality. In *Proceedings of forty-sixth annual meeting of the American Psychological Association* (Bde. Psychological Bulletin, 35(9), S. 659). Columbus, Ohio: Psychological Review Company. doi: 10.1037/h0055433

Lenzner, T., Kaczmirek, L. & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*, *24* (7), 1003–1020. doi: 10.1002/acp.1602

Leunbach, G. (1961). On quantitative models for qualitative data. *Acta Sociologica*, *5* (3), 144–156.

Levashina, J., Morgeson, F. P. & Campion, M. A. (2009). They don't do It often, but they do it well: Exploring the relationship between applicant mental abilities and faking. *International Journal of Selection and Assessment*, *17* (3), 271–281. doi: 10.1111/j.1468-2389.2009.00469.x

Levin, R. A. & Zickar, M. J. (2002). Investigating self-presentation, lies, and bullshit. Understanding faking and its effects on selection decisions using theory, field research, and simulation. In C. L. Hulin, J. M. Brett & F. Drasgow (Hrsg.), *The psychology of work: Theoretically based empirical research* (S. 253–276). Mahwah, NJ: Lawrence Erlbaum.

Levine, M. & Drasgow, F. (1979). Approprietness Measurement. Basic Principles and Validating Studies. In D. J. Weiss (Hrsg.), *Proceedings of the 1979 Conference on Computerized Adaptive Testing* (S. 285–347). Washington, D.C.: Distributed by ERIC Clearinghouse.

Levine, M. V. & Drasgow, F. (1982). Appropriateness measurement. Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, *35*, 42–56.

Levine, M. V. & Drasgow, F. (1983). The relation between incorrect option

choice and estimated ability. *Educational and Psychological Measurement*, *43* (3), 675–685. doi: 10.1177/001316448304300301

Levine, M. V. & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, *53* (2), 161–176. doi: 10.1007/BF02294130

Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4* (4), 269–290. doi: 10.2307/1164595

Levitin, D. J., Grahn, J. & London, J. (2017). The psychology of music: Rhythm and movement. *Annual Review of Psychology*. doi: 10.1146/annurev-psych-122216-011740

Lienert, G. A. (1971). Die Konfigurationsfrequenzanalyse: I. Ein neuer Weg zu Typen und Syndromen. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *19* (2), 99–115.

Lienert, G. A. & Krauth, J. (1975). Configural frequency analysis as a statistical tool for defining types. *Educational and Psychological Measurement*, *35* (2), 231–238. doi: 10.1177/001316447503500201

Liiv, I. (2010). Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining*, *3* (2), 70–91. doi: 10.1002/sam.10071

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22 140*, 55.

Likert, R., Roslow, S. & Murphy, G. (1934). A simple and reliable method of scoring the Thurstone attitude scales. *Journal of Social Psychology*, *5* (2).

Linacre, J. M. (1999). Understanding rasch measurement: Estimation methods for rasch measures. *Journal of Outcome Measurement*, *3* (4), 382–405.

Linacre, J. M. (2002). What do INFIT and OUTFIT, mean-square and standardized mean? *Rasch Measurement Transactions*, *16* (2), 878.

Linacre, J. M. (2004a). Discrimination, guessing and carelessness asymptotes: estimating irt parameters with rasch. *Rasch Measurement Transactions*, *18* (1), 959–960.

Linacre, J. M. (2004b). Rasch model estimation: Further topics. *Journal of Applied Measurement*, *5* (1), 95–110.

Linacre, J. M. & Fisher Jr, W. P. (2012). Harvey Goldstein's Objections

to Rasch Measurement: A Response from Linacre and Fisher. *Rasch Measurement Transactions*, *26* (3), 1383–9.

Linzer, D. A. & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42* (10), 1–29.

Litle, P. & Zuckerman, M. (1986). Sensation seeking and music preferences. *Personality and Individual Differences*, *7* (4), 575–578. doi: 10.1016/0191-8869(86)90136-4

Liu, T., Lan, T. & Xin, T. (2016). Detecting random responses in a personality scale using IRT-based person-fit indices. *European Journal of Psychological Assessment*, 1–11. doi: 10.1027/1015-5759/a000369

Liu, W.-Y., Weber, B., Reuter, M., Markett, S., Chu, W.-C. & Montag, C. (2013). The Big Five of personality and structural imaging revisited: a VBM - DARTEL study. *NeuroReport*, *24* (7), 375–380. doi: 10.1097/WNR.0b013e328360dad7

Lombardi, L. & Pastore, M. (2015). Robust evaluation of fit indices to fake-good perturbation of ordinal data. *Quality & Quantity*, 1–25. doi: 10.1007/s11135-015-0282-1

Lönnqvist, J.-E., Paunonen, S., Tuulio-Henriksson, A., Lönnqvist, J. & Verkasalo, M. (2007). Substance and style in socially desirable responding. *Journal of Personality*, *75* (2), 291–322. doi: 10.1111/j.1467-6494.2006.00440.x

Lord, F. M. (1944). Reliability of multiple-choice tests as a function of number of choices per item. *Journal of Educational Psychology*, *35*, 175–180.

Lorge, I. (1937). Gen-like: Halo or reality. In H. Rogers (Hrsg.), *Proceedings of the eighth spring meeting, Eastern Branch, American Psychological Association* (Bde. Psychological Bulletin, 34(8), S. 545–546). Poughkeepsie, New York: Psychological Review Company. doi: 10.1037/h0059154

Lozano, L. M., García-Cueto, E. & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *4* (2), 73–79. doi: 10.1027/1614-2241.4.2.73

Luce, R. D. (1977a). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15* (3), 215–233.

Luce, R. D. (1977b). Thurstone's discriminal processes fifty years later. *Psychometrika*, *42* (4), 461–489.

Luo, G. (1998a). A general formulation for unidimensional unfolding and pairwise preference models: making explicit the latitude of acceptance. *Journal of Mathematical Psychology*, *42* (4), 400–417. doi: 10.1006/jmps.1998.1206

Luo, G. (1998b). A general formulation of multidimensional unfolding models involving the latitude of acceptance. In A. Rizzi, M. Vichi & H. H. Bock (Hrsg.), *Advances in data science and classification: Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98), Università 'La Sapienza', Rome, 21-24 July, 1998.* New York: Springer.

Luo, G. (2001). A class of probabilistic unfolding models for polytomous responses. *Journal of Mathematical Psychology*, *45* (2), 224–248. doi: 10.1006/jmps.2000.1310

Luo, G. & Andrich, D. (2005). Estimating parameters in the Rasch model in the presence of null categories. *Journal of Applied Measurement*, *6* (2), 128–146.

Lykken, D. T., Bouchard, T. J., McGue, M. & Tellegen, A. (1990). The Minnesota Twin Family Registry: Some initial findings. *Acta Geneticae Medicae Et Gemellologiae*, *39* (1), 35–70.

MacCann, C., Ziegler, M. & Roberts, R. (2012). Faking in personality assessment. In M. Ziegler, C. MacCann & R. Roberts (Hrsg.), *New perspectives on faking in personality assessment* (S. 309–329). Oxford: Oxford University Press.

MacCann, R. G. (2004). Reliability as a function of the number of item options derived from the "Knowledge or Random Guessing" model. *Psychometrika*, *69* (1), 147–157. doi: 10.1007/BF02295844

Maij-de Meij, A. M., Kelderman, H. & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, *32* (8), 611–631. doi: 10.1177/0146621607312613

Mair, P., Hatzinger, R., Maier, M. J. & Rusch, T. (2018). *eRm: Extended*

Rasch Modeling (R package version 0.16-2).

Maniaci, M. R. & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. doi: 10.1016/j.jrp.2013.09.008

Maraun, M. D. & Rossi, N. T. (2001). The extra-factor phenomenon revisited: Unidimensional unfolding as quadratic factor analysis. *Applied Psychological Measurement*, *25* (1), 77–87. doi: 10.1177/01466216010251006

Marcus, B. (2009). 'Faking' from the applicant's perspective: A theory of self-presentation in personnel selection settings. *International Journal of Selection and Assessment*, *17* (4), 417–430. doi: 10.1111/j.1468-2389.2009.00483.x

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, *60* (4), 523–547. doi: 10.1007/BF02294327

Maris, G. & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement: Interdisciplinary Research and Perspectives*, *7* (2), 75–88. doi: 10.1080/15366360903070385

Maris, G. & Maris, E. (2002). *Are attitude items monotone or single-peaked? An analysis using Bayesian methods.* (Bericht). Arnhem: Citogroep.

Markett, S., Montag, C., Melchers, M., Weber, B. & Reuter, M. (2016). Anxious personality and functional efficiency of the insular-opercular network: A graph-analytic approach to resting-state fMRI. *Cognitive, Affective, & Behavioral Neuroscience*, *16* (6), 1039–1049. doi: 10.3758/s13415-016-0451-2

Markett, S., Weber, B., Voigt, G., Montag, C., Felten, A., Elger, C. & Reuter, M. (2013). Intrinsic connectivity networks and personality: the temperament dimension harm avoidance moderates functional connectivity in the resting brain. *Neuroscience*, *240*, 98–105.

Marsh, H. W. (1996). Positive and negative global self-esteem: a substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70* (4), 810–819.

Martin, B. A., Bowen, C. C. & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences*, *32* (2), 247–256. doi: 10.1016/S0191-8869(01)00021-6

Martin, S. B., Covell, D. J., Joseph, J. E., Chebrolu, H., Smith, C. D., Kelly,

T. H., ... Gold, B. T. (2007). Human experience seeking correlates with hippocampus volume: convergent evidence from manual tracing and voxel-based morphometry. *Neuropsychologia*, *45* (12), 2874–2881. doi: 10.1016/j.neuropsychologia.2007.05.009

Martin-Löf, P. (1973). *Statistiska Modeller* (Anteckningar från seminarier läsåret 1969-70, utarbetade av Rolf Sundberg.[Notizen aus Seminaren des Studienjahres 1969-70, überarbeitet von Rolf Sundberg.]). Stockholm: Institutet för försäkringsmatematik och matematisk statistik vid Stockholms universitet.[Institut für Versicherungsmathematik und Mathematische Statistik an der Universität Stockholm.].

Maslow, A. H. (1962). *Toward a psychology of being*. Princeton: Van Nostrand.

Maslow, A. H. (1969). Toward a humanistic biology. *American Psychologist*, *24* (8), 724–735. doi: 10.1037/h0027859

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47* (2), 149–174. doi: 10.1007/BF02296272

Masters, G. N. (1988). Measurement models for ordered response categories. In R. Langeheine & J. Rost (Hrsg.), *Latent trait and latent class models* (S. 11–29). New York: Springer.

Masters, G. N. & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 101–121). New York: Springer.

Matell, M. S. & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: reliability and validity. *Educational and Psychological Measurement*, *31* (3), 657–674. doi: 10.1177/001316447103100307

Matschinger, H. & Krebs, D. (1998). Zum Problem der Abbildung eindimensional konzipierter Konstrukte bei entgegengesetzter Itempolung. *ZA-Information / Zentralarchiv für Empirische Sozialforschung*, *43*, 81–110.

Matz, B. W. (2002). *Die Konstitutionstypologie von Ernst Kretschmer* (Unveröffentlichte Dissertation). Freie Universität Berlin, Germany.

Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika*, *66* (2), 209–227.

Maydeu-Olivares, A. (2005). Linear item response theory, nonlinear item re-

sponse theory, and factor analysis: a unified framework. In R. P. McDonald, A. Maydeu-Olivares & J. J. McArdle (Hrsg.), *Contemporary Psychometrics: a Festschrift for Roderick P. McDonald* (S. 73–100). Mahwah, NJ: Lawrence Erlbaum Associates.

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika*, *71* (1), 57–77. doi: 10.1007/s11336-005-0773-4

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11* (3), 71–101. doi: 10.1080/15366367.2013.831680

Maydeu-Olivares, A. (2015). Evaluating the fit of IRT models. In S. P. Reise & D. A. Revicki (Hrsg.), *Handbook of item response theory modeling: Applications to typical performance assessment* (S. 111–127). New York: Routledge.

Maydeu-Olivares, A., Hernández, A. & McDonald, R. P. (2006). A multidimensional ideal point item response theory model for binary data. *Multivariate Behavioral Research*, *41* (4), 445–472. doi: 10.1207/s15327906mbr4104_2

Maydeu-Olivares, A. & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2(n) contingency tables: A unified framework. *Journal of the American Statistical Association*, *100* (471), 1009–1020. doi: 10.1198/016214504000002069

Maydeu-Olivares, A. & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71* (4), 713–732. doi: 10.1007/s11336-005-1295-9

Maydeu-Olivares, A., Kramp, U., Garcia-Forero, C., Gallardo-Pujol, D. & Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods*, *41* (2), 295–308. doi: 10.3758/BRM.41.2.295

McCrae, R. R. & Costa, P. T. (1983a). Joint factors in self-reports and ratings: Neuroticism, extraversion and openness to experience. *Personality and Individual Differences*, *4* (3), 245–255. doi: 10.1016/0191-8869(83)90146-0

McCrae, R. R. & Costa, P. T. (1983b). Social desirability scales: More sub-

stance than style. *Journal of Consulting and Clinical Psychology*, *51* (6), 882–888. doi: 10.1037/0022-006X.51.6.882

McCrae, R. R. & Costa, P. T. (1985). Updating Norman's 'adequacy taxonomy': Intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology*, *49* (3), 710–721. doi: 10.1037/0022-3514.49.3.710

McCrae, R. R. & Costa, P. T. (1987). Validation of the Five-Factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52* (1), 81–90. doi: 10.1037/0022-3514.52.1.81

McCrae, R. R. & John, O. P. (1992). An introduction to the Five-Factor model and its applications. *Journal of Personality*, *60* (2), 175–215. doi: 10.1111/j.1467-6494.1992.tb00970.x

McDonald, R. P. & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30* (1), 23–40. doi: 10.1207/s15327906mbr3001_2

McFarland, L. A. & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, *85* (5), 812–821. doi: 10.1037/0021-9010.85.5.812

McFarland, L. A. & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology*, *36* (4), 979–1016. doi: 10.1111/j.0021-9029.2006.00052.x

McGee, R. K. (1962). The relationship between response style and personality variables: I. The measurements of response acquiescence. *The Journal of Abnormal and Social Psychology*, *64* (3), 229–233. doi: 10.1037/h0043076

McIntyre, H. H. (2011). Investigating response styles in self-report personality data via a joint structural equation mixture modeling of item responses and response times. *Personality and Individual Differences*, *50* (5), 597–602. doi: 10.1016/j.paid.2010.12.001

McNamara, L. & Ballard, M. E. (1999). Resting arousal, sensation seeking, and music preference. *Genetic, Social, and General Psychology Monographs*, *125* (3), 229–250.

McPherson, J. & Mohr, P. (2005). The role of item extremity in the emergence of keying-related factors: an exploration with the life orientation test.

Psychological Methods, *10* (1), 120–131. doi: 10.1037/1082-989X.10.1.120

McReynolds, P. & Ludwig, K. (1987). On the history of rating scales. *Personality and Individual Differences*, *8* (2), 281–283. doi: 10.1016/0191-8869(87)90188-7

Meade, A. W. & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17* (3), 437–455. doi: 10.1037/a0028085

Meehl, P. E. (1956). Wanted - a good cook-book. *American Psychologist*, *11* (6), 263–272. doi: 10.1037/h0044164

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18* (4), 311–314. doi: 10.1177/014662169401800402

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9* (1), 3–8. doi: 10.1207/s15324818ame0901_2

Meijer, R. R., Molenaar, I. W. & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, *18* (2), 111–120. doi: 10.1177/014662169401800202

Meijer, R. R., Muijtjens, A. M. M. & Vleuten, C. P. M. v. d. (1996). Non parametric person-fit research: some theoretical issues and an empirical example. *Applied Measurement in Education*, *9* (1), 77–89. doi: 10.1207/s15324818ame0901_7

Meijer, R. R., Niessen, A. S. M. & Tendeiro, J. N. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*, *23* (1), 52–62. doi: 10.1177/1073191115577800

Meijer, R. R. & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, *8* (3), 261–272. doi: 10.1207/s15324818ame0803_5

Meijer, R. R. & Sijtsma, K. (2001a). Methodology review: evaluating person fit. *Applied Psychological Measurement*, *25* (2), 107–135. doi: 10.1177/01466210122031957

Meijer, R. R. & Sijtsma, K. (2001b). Person fit statistics: What is their purpose? *Rasch Measurement Transactions*, *15* (2), 823–823.

Meisenberg, G. & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, *44* (7), 1539–1550.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7* (2), 105–118. doi: 10.2307/1164960

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19* (1), 91–100.

Mellenbergh, G. J. (2001). Outline of a faceted theory of item response data. In A. Boomsma, M. A. J. van Duijn & T. A. B. Snijders (Hrsg.), *Essays on Item Response Theory* (Bde. Lecture notes in statistics, 157, S. 415–432). New York: Springer.

Menictas, C., Wang, P. & Fine, B. (2011). Assessing flat-lining response style bias in online research. *Australasian Journal of Market & Social Research*, *19* (2), 34–44.

Merritt, S. M. (2012). The two-factor solution to Allen and Meyer's (1990) Affective Commitment Scale: Effects of negatively worded items. *Journal of Business and Psychology*, *27* (4), 421–436. doi: 10.1007/s10869-011-9252-3

Mersman, J. L. & Shultz, K. S. (1998). Individual differences in the ability to fake on personality measures. *Personality and Individual Differences*, *24* (2), 217–227. doi: 10.1016/S0191-8869(97)00160-8

Messick, S. & Jackson, D. N. (1957). Authoritarianism or acquiescence in Bass's data. *The Journal of Abnormal and Social Psychology*, *54* (3), 424–426. doi: 10.1037/h0041682

Meulman, J., Hubert, L. & Heiser, W. J. (1998). The data theory scaling system. In A. Rizzi, M. Vichi & H. H. Bock (Hrsg.), *Advances in data science and classification: proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98), Università 'La Sapienza', Rome, 21-24 July, 1998.* New York: Springer.

Michaelides, M. P., Zenger, M., Koutsogiorgi, C., Brähler, E., Stöbel-Richter, Y. & Berth, H. (2016). Personality correlates and gender invariance of wording effects in the German version of the Rosenberg Self-Esteem Scale. *Personality and Individual Differences*, *97*, 13–18. doi: 10.1016/j.paid.2016.03.011

Michell, J. (2000)jan. Normal Science, Pathological Science and Psychometrics. *Theory & Psychology*, *10* (5), 639–667. doi: 10.1177/0959354300105004

Michell, J. (2004). Item Response Models, Pathological Science and the Shape of Error Reply to Borsboom and Mellenbergh. *Theory & Psychology*, *14* (1), 121–129. doi: 10.1177/0959354304040201

Michell, J. (2008)may. Is Psychometrics Pathological Science? *Measurement: Interdisciplinary Research & Perspective*, *6* (1-2), 7–24. doi: 10.1080/15366360802035489

Millsap, R. E. & Maydeu-Olivares, A. (Hrsg.). (2009). *The SAGE handbook of quantitative methods in psychology*. Los Angeles: SAGE Publications Inc.

Minkov, M. (2017). Middle responding: An unobtrusive measure of national cognitive ability and personality. *Personality and Individual Differences*, *113*, 187–192. doi: 10.1016/j.paid.2017.03.041

Mischel, W. (1968). *Personality and assessment*. Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, *80* (4), 252–283. doi: 10.1037/h0035002

Mischel, W. (2004). Toward an Integrative Science of the Person. *Annual Review of Psychology*, *55* (1), 1–22. `http://www.annualreviews.org/doi/10.1146/annurev.psych.55.042902.130709` doi: 10.1146/annurev.psych.55.042902.130709

Mischel, W. (2009). From personality and assessment (1968) to personality science, 2009. *Journal of Research in Personality*, *43* (2), 282–290.

Mischel, W. & Peake, P. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review*, *89* (6), 730.

Mischel, W. & Shoda, Y. (1995). A cognitive–affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102* (2), 246–268.

Mischel, W., Shoda, Y. & Mendoza-Denton, R. (2002). Situation-behavior profiles as a locus of consistency in personality. *Current Directions in Psychological Science*, *11*, 50–54. doi: 10.1111/1467-8721.00166

Mohler, P. P. (2006). Sampling from a universe of items and the de-

machiavellization of questionnaire design. In M. Braun & P. P. Mohler (Hrsg.), *Beyond the horizon of measurement Festschrift in honor of Ingwer Borg* (Bde. ZUMA Nachrichten - Spezial, 10, S. 9–14). Mannheim: ZUMA.

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. The Hague: Mouton & Co.

Mokken, R. J. & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6* (4), 417–430. doi: 10.1177/014662168200600404

Mokken, R. J., Lewis, C. & Sijtsma, K. (1986). Rejoinder to 'The Mokken Scale: A critical discussion'. *Applied Psychological Measurement*, *10* (3), 279–285. doi: 10.1177/014662168601000306

Molenaar, I. W. (1995). Estimation of item parameters. In G. Fischer & I. Molenaar (Hrsg.), *Rasch models: Foundations, recent developments, and applications* (S. 39–51). New York: Springer.

Molenaar, I. W. (1997a). Lenient or strict application of IRT with an eye on practical consequences. In J. Rost & R. Langeheine (Hrsg.), *Applications of latent trait and latent class models in the social sciences* (S. 38–49). Münster: Waxmann.

Molenaar, I. W. (1997b). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of Modern Item Response Theory* (S. 369–380). New York: Springer.

Molenaar, I. W. & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55* (1), 75–106.

Molenaar, I. W. & Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education*, *9* (1), 27–45. doi: 10.1207/s15324818ame0901_4

Montag, C., Reuter, M., Weber, B., Markett, S. & Schoene-Bake, J.-C. (2012). Individual differences in trait anxiety are associated with white matter tract integrity in the left temporal lobe in healthy males but not females. *Neuroscience*, *217*, 77–83.

Moore, D. A. & Small, D. A. (2007). Error and bias in comparative judgment: On being both better and worse than we think we are. *Journal of Personality and Social Psychology*, *92* (6), 972–989. doi: 10.1037/

0022-3514.92.6.972

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity*, *42* (6), 779–794. doi: 10.1007/s11135-006-9067-x

Moosbrugger, H. (2012). 05 Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 103–117). Berlin: Springer.

Moosbrugger, H. & Kelava, A. (Hrsg.). (2012). *Testtheorie und Fragebogenkonstruktion*. Berlin: Springer. doi: 10.1007/978-3-642-20072-4

Morf, M. E. & Jackson, D. N. (1972). An analysis of two response styles: True responding and item endorsement. *Educational and Psychological Measurement*, *32* (2), 329–353. doi: 10.1177/001316447203200210

Mosteller, F. (1951). Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika*, *16* (2), 207–218. doi: 10.1007/BF02289116

Mount, M. K., Barrick, M. R., Scullen, S. M., Rounds, J. & Sackett, P. (2005). Higher-order dimensions of the Big Five personality traits and the Big Six vocational interest types. *Personnel Psychology*, *58* (2), 447–478.

Muck, P. M. (2007). AIST-R - Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-R) - Revision. *Zeitschrift für Arbeits- und Organisationspsychologie*, *51* (1), 26–31.

Mummendey, H. D. (2015). Selbstdarstellungstheorie. In D. Frey & M. Irle (Hrsg.), *Motivations-, Selbst- und Informationsverarbeitungstheorien* (2. Aufl., Bd. 3, S. 212–233). Bern: Huber.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16* (2), 159 –176. doi: 10.1177/014662169201600206

Murtha, T. C., Kanfer, R. & Ackerman, P. L. (1996). Toward an interactionist taxonomy of personality and situations: An integrative situational—dispositional representation of personality traits. *Journal of Personality and Social Psychology*, *71* (1), 193–207.

Naemi, B. D., Beal, D. J. & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, *77* (1), 261–286. doi:

10.1111/j.1467-6494.2008.00545.x

Nagy, G. (2007). *Berufliche Interessen, kognitive und fachgebundene Kompetenzen Ihre Bedeutung für die Studienfachwahl und die Bewährung im Studium* (Monographie, Freie Universität Berlin, Berlin). `https://refubium.fu-berlin.de/handle/fub188/10012`

Nagy, G., Marsh, H. W., Lüdtke, O. & Trautwein, U. (2009). Representing the circles in our minds: Confirmatory factor analysis of circumplex stuctures and profiles. In T. Teo & M. S. Khine (Hrsg.), *Structural Equation Modeling in Educational Research - Concepts and Applications* (S. 287 – 316). Rotterdam Boston Taipei: Sense Publishers.

Narens, L. (1981). On the scales of measurement. *Journal of Mathematical Psychology*, *24* (3), 249–275. doi: 10.1016/0022-2496(81)90045-6

Narens, L. & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99* (2), 166–180. doi: 10.1037/0033-2909.99.2.166

Nauta, M. M. (2010). The development, evolution, and status of Holland's theory of vocational personalities: Reflections and future directions for counseling psychology. *Journal of Counseling Psychology*, *57* (1), 11–22. doi: 10.1037/a0018213

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15* (3), 263–280. doi: 10.1002/ejsp.2420150303

Nettle, D. (2006). The evolution of personality variation in humans and other animals. *The American Psychologist*, *61* (6), 622–631. doi: 10.1037/0003-066X.61.6.622

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16* (1), 1–32. doi: 10.2307/1914288

Nichols, D. S., Greene, R. L. & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, *45* (2), 239–250. doi: 10.1002/1097-4679(198903)45:2<239::AID-JCLP2270450210>3.0.CO;2-1

Niessen, A. S. M., Meijer, R. R. & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal*

of Research in Personality, *63*, 1–11. doi: 10.1016/j.jrp.2016.04.010

Nishisato, S. (1978a). Errata to: Optimal scaling of paired comparison and rank order data: An alternative to Guttman's formulation. *Psychometrika*, *43* (4), 587–587. https://doi.org/10.1007/BF02293818 doi: 10.1007/BF02293818

Nishisato, S. (1978b). Optimal scaling of paired comparison and rank order data: An alternative to guttman's formulation. *Psychometrika*, *43* (2), 263–271. https://doi.org/10.1007/BF02293868 doi: 10.1007/BF02293868

North, A. C. & Hargreaves, D. J. (1996). Situational influences on reported musical preference. *Psychomusicology: A Journal of Research in Music Cognition*, *15* (1-2), 30–45. doi: 10.1037/h0094081

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Nye, C. D., Do, B. R., Drasgow, F. & Fine, S. (2008). Two-step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment*, *16* (2), 112–120.

Nye, C. D., Newman, D. A. & Joseph, D. L. (2010). Never say 'always'? Extreme item wording effects on scalar invariance and item response curves. *Organizational Research Methods*, *13* (4), 806–830. doi: 10.1177/1094428109349512

O'Brien, E. & LaHuis, D. M. (2011). Do applicants and incumbents respond to personality items similarly? A comparison of dominance and ideal point response models. *International Journal of Selection and Assessment*, *19* (2), 109–118. doi: 10.1111/j.1468-2389.2011.00539.x

O'Connell, M. S., Kung, M.-C. & Tristan, E. (2011). Beyond impression management: Evaluating three measures of response distortion and their relationship to job performance. *International Journal of Selection and Assessment*, *19* (4), 340–351. doi: 10.1111/j.1468-2389.2011.00563.x

OECD. (2014). *PISA 2012 technical report*. OECD Publishing.

OECD. (2019). *PISA - Programme for International Student Assessment*. http://www.oecd.org/pisa/

Ones, D. S., Viswesvaran, C. & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81* (6), 660–679. doi: 10.1037/0021-9010.81.6.660

Oppenheimer, D. M., Meyvis, T. & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45* (4), 867–872. doi: 10.1016/j.jesp.2009.03.009

Orth, B. (1989). On the axiomatic foundations of unfolding: With applications to political party preferences of german voters. *Advances in Psychology*, *60*, 221–235. doi: 10.1016/S0166-4115(08)60238-1

Ostendorf, F. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae - Revidierte Fassung*. Göttingen: Hogrefe.

Osterlind, S. J. (1990). Toward a uniform definition of a test item. *Educational Research Quarterly*, *14* (4), 2–5.

Osterlind, S. J. (2002). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (2. Aufl.). Dordrecht: Kluwer Academic Publishers.

Panayides, P., Robinson, C. & Tymms, P. (2015)feb. Rasch measurement: a response to Goldstein. *British Educational Research Journal*, *41* (1), 180–182. doi: 10.1002/berj.3182

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46* (3), 598–609. doi: 10.1037/0022-3514.46.3.598

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. Shaver & L. S. Wrightsmann (Hrsg.), *Measures of social psychological attitudes* (Bd. 1. Measures of personality and social psychological attitudes, S. 17–59). San Diego: Academic Press.

Paulhus, D. L. (1998a). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, *66* (6), 1025–1060.

Paulhus, D. L. (1998b). *Manual for Balanced Inventory of Desirable Responding (BIDR-7)*. Toronto: Multi-Health Systems.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. Braun, D. Jackson & D. Wiley (Hrsg.), *The role of constructs in psychological and educational measurement* (S. 49–69). Mahwah, N.J.: Lawrence Erlbaum Associates.

Paulhus, D. L. (2012). Overclaiming on personality questionnaires. In M. Zieg-

ler, C. MacCann & R. Roberts (Hrsg.), *New perspectives on faking in personality assessment* (S. 151–164). Oxford: Oxford University Press.

Paulhus, D. L. & Reid, D. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, *60* (2), 307.

Pauls, C. A. & Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. *Personality and Individual Differences*, *37* (6), 1137–1151. doi: 10.1016/j.paid.2003.11.018

Paunonen, S. V. & LeBel, E. P. (2012). Socially desirable responding and its elusive effects on the validity of personality assessments. *Journal of Personality & Social Psychology*, *103* (1), 158–175.

Pawlik, K. (1971). *Dimensionen des Verhaltens* (2. Aufl.). Bern: Hans Huber.

Payne, E. (1967). Musical taste and personality. *British Journal of Psychology*, *58* (1).

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *Series 6* (2), 559–572.

Peer, E. & Gamliel, E. (2011). Too reliable to be true? Response bias as a potential source of inflation in paper-and-pencil questionnaire reliability. *Practical Assessment, Research & Evaluation*, *16* (9), 2.

Pfanzagl, J. (1959). A general theory of measurement applications to utility. *Naval Research Logistics Quarterly*, *6* (4), 283–294. doi: 10.1002/nav.3800060404

Pfanzagl, J. (1971). *Theory of measurement* (2. Aufl.). Würzburg-Wien: Physica-Verlag.

Piedmont, R. L. & Aycock, W. (2007). An historical analysis of the lexical emergence of the Big Five personality adjective descriptors. *Personality and Individual Differences*, *42* (6), 1059–1068.

Piedmont, R. L., McCrae, R. R., Riemann, R. & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology*, *78* (3), 582–593. doi: 10.1037/0022-3514.78.3.582

Pilotte, W. J. & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement*, *50* (3), 603–610. doi: 10.1177/0013164490503016

Plieninger, H. (2017). Mountain or Molehill? A Simulation Study on the Impact of Response Styles. *Educational and Psychological Measurement*, *77* (1), 32–53. doi: 10.1177/0013164416636655

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *The Journal of Applied Psychology*, *88* (5), 879–903. doi: 10.1037/0021-9010.88.5.879

Ponocny, I. & Klauer, K. C. (2002). Towards identification of unscalable personality questionnaire respondents: The use of person fit indices. *Psychologische Beiträge*, *44* (1), 22–40.

Popper, K. (1935). *Logik der Forschung.* Wien: Springer Verlag.

Porst, R. (2000). *Question Wording - Zur Formulierung von Fragebogen-Fragen* (Bericht Nr. 2). Mannheim Germany: Zentrum für Umfragen, Methoden und Analysen.

Post, W. J. (1992). *Nonparametric unfolding models. A latent structure approach.* (Unveröffentlichte Dissertation). University of Groningen, Groningen.

Post, W. J. & Snijders, T. A. B. (1993). Nonparametric unfolding models for dichotomous data. *Methodika*, *VII*, 130–156.

Post, W. J., van Duijn, M. A. & van Baarsen, B. (2001). Single-peaked or monotone tracelines? On the choice of an IRT model for scaling data. In A. Boomsma, M. A. J. van Duijn & T. A. B. Snijders (Hrsg.), *Essays on Item Response Theory* (Bde. Lecture notes in statistics, 157, S. 391–413). New York: Springer.

Pratkanis, A. R. (1989). The cognitive representation of attitudes. In A. R. Pratkanis, S. Breckler & A. G. Greenwald (Hrsg.), *Attitudes structure and function* (S. 71–98). Hillsdale: Erlbaum.

Pratkanis, A. R. & Greenwald, A. G. (1989). A sociocognitive model of attitude structure and function. In L. Berkowitz (Hrsg.), *Advances in Experimental Social Psychology* (Bd. 22, S. 245–285). San Diego: Academic Press.

Preinerstorfer, D. & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, *65* (2), 251–262. doi: 10.1111/j.2044-8317.2011

.02020.x

Prenzel, M. (1988). *Die Wirkweise von Interesse. Ein pädagogisch-psychologisches Erklärungsmodell.* Opladen: Westdeutscher Verlag.

Puchhammer, M. (1988). Die Berücksichtigung von Rateparametern im Modell von Rasch. In K. D. Kubinger (Hrsg.), *Moderne Test Theorie* (S. 271–280). Weinheim ; München: Psychologie Verlags Union.

R Core Team. (2018). R: A language and environment for statistical computing [Software-Handbuch]. Vienna, Austria. `https://www.R-project.org/`

Radocy, R. E. & Boyle, D. J. (2003). *Psychological foundations of musical behaviour* (4. Aufl.). Springfield, Illinois: Charles C Thomas Publisher Ltd.

Rammstedt, B. (1997). *Die deutsche Version des Big Five Inventory (BFI): Übersetzung und Validierung eines Fragebogens zur Erfassung des Fünf-Faktoren-Modells der Persönlichkeit.* [Unveroeffentlichte Diplomarbeit]. Bielefeld.

Rammstedt, B. & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K):. *Diagnostica*, *51* (4), 195–206. doi: 10.1026/0012-1924.51.4.195

Randall, W. M. & Rickard, N. S. (2017). Personal music listening. *Music Perception: An Interdisciplinary Journal*, *34* (5), 501–514. doi: 10.1525/mp.2017.34.5.501

Ranger, J. & Kuhn, J.-T. (2012). Assessing fit of item response models using the information matrix test. *Journal of Educational Measurement*, *49* (3), 247–268. doi: 10.1111/j.1745-3984.2012.00174.x

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Kopenhagen: Danmarks pædagogiske Institut.

Rasch, G. (1966a). An informal report on the present state of a theory of objectivity in comparisons. In *Proceedings of the NUFFIC International Summer Session in Science at 'Het Oude Hof'.* The Hague.

Rasch, G. (1966b). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, *19* (1), 49–57. doi: 10.1111/j.2044-8317.1966.tb00354.x

Rasch, G. (1977). *On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements.*

Rauscher, F. H. & Shaw, G. L. (2016). Key components of the mozart effect.

Perceptual and Motor Skills, *86* (3), 835–841. doi: 10.2466/pms.1998.86.3.835

Rauscher, F. H., Shaw, G. L. & Ky, C. N. (1993). Music and spatial task performance. *Nature*, *365* (6447), 611. doi: 10.1038/365611a0

Rauscher, F. H., Shaw, G. L. & Ky, K. N. (1995). Listening to Mozart enhances spatial-temporal reasoning: towards a neurophysiological basis. *Neuroscience Letters*, *185* (1), 44–47.

Rauthmann, J. (2014). Person-situation debatte. In M. A. Wirtz (Hrsg.), *Dorsch Lexikon der Psychologie*. Göttingen: Hogrefe.

Rawlings, D., Barrantes i Vidal, N. & Furnham, A. (2000). Personality and aesthetic preference in Spain and England: Two studies relating sensation seeking and openness to experience to liking for paintings and music. *European Journal of Personality*, *14* (6), 553–576.

Rawlings, D. & Ciancarelli, V. (1997). Music preference and the Five-Factor model of the NEO Personality Inventory. *Psychology of Music*, *25* (2), 120 –132. doi: 10.1177/0305735697252003

Rawlings, D., Hodge, M., Sherr, D. & Dempsey, A. (1995). Toughmindedness and preference for musical excerpts, categories and triads. *Psychology of Music*, *23* (1), 63 –80. doi: 10.1177/0305735695231005

Ray, J. V., Hall, J., Rivera-Hudson, N., Poythress, N. G., Lilienfeld, S. O. & Morano, M. (2013). The relation between self-reported psychopathic traits and distorted response styles: A meta-analytic review. *Personality Disorders: Theory, Research, and Treatment*, *4* (1), 1–14. doi: 10.1037/a0026482

Regan, P. C., Snyder, M. & Kassin, S. M. (1995). Unrealistic optimism: Self-enhancement or person positivity? *Personality and Social Psychology Bulletin*, *21* (10), 1073 –1082. doi: 10.1177/01461672952110008

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14* (2), 127–137.

Reise, S. P. & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65* (1), 143–151. doi: 10.1037/0022-3514.65.1.143

Reise, S. P. & Waller, N. G. (2003). How many IRT parameters does it take to

model psychopathology items? *Psychological Methods*, *8* (2), 164–184. doi: 10.1037/1082-989X.8.2.164

Reise, S. P. & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5* (1), 27–48. doi: 10.1146/annurev.clinpsy.032408.153553

Renner, K.-H., Heydasch, T. & Ströhlein, G. (2012). *Forschungsmethoden der Psychologie*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Rentfrow, P. J., Goldberg, L. R. & Levitin, D. J. (2011). The structure of musical preferences: A Five-Factor model. *Journal of Personality & Social Psychology*, *100* (6), 1139–1157.

Rentfrow, P. J., Goldberg, L. R., Stillwell, D. J., Kosinski, M., Gosling, S. D. & Levitin, D. J. (2012). The song remains the same. a replication and extension of the music model. *Music Perception*, *30* (2), 161–185. doi: 10.1525/MP.2012.30.2.161

Rentfrow, P. J. & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of personality and social psychology*, *84* (6), 1236–1256.

Rey, J. J., Abad, F. J., Barrada, J. R., Garrido, L. E. & Ponsoda, V. (2014). The impact of ambiguous response categories on the factor structure of the GHQ–12. *Psychological Assessment*, *26* (3), 1021–1030. doi: 10.1037/a0036468

Rigg, M. (1937). Musical expression: an investigation of the theories of Erich Sorantin. *Journal of Experimental Psychology*, *21* (4), 442–455. doi: 10.1037/h0056388

Roberts, J. S. (1995). *Item response theory approaches to attitude measurement* (Unpublished doctoral dissertation). University of South Carolina.

Roberts, J. S., Donoghue, J. R. & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24* (1), 3–32. doi: 10.1177/01466216000241001

Roberts, J. S. & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement*, *20* (3), 231–255. doi: 10.1177/014662169602000305

Roberts, J. S., Laughlin, J. E. & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement*, *59* (2), 211–233. doi: 10.1177/00131649921969811

Robitzsch, A., Kiefer, T. & Wu, M. (2018). *TAM: Test Analysis Modules* (R package version 3.0-21). `https://CRAN.R-project.org/package=TAM`

Rocereto, J. F., Puzakova, M., Anderson, R. E. & Kwak, H. (2011). The role of response formats on extreme response style: A case of Likert-type vs. semantic differential scales. *Advances in International Marketing*, *10* (22), 53–71.

Rodebaugh, T. L., Woods, C. M. & Heimberg, R. G. (2007). The reverse of social anxiety is not always the opposite: The reverse-scored items of the Social Interaction Anxiety Scale do not belong. *Behavior Therapy*, *38* (2), 192–206. doi: 10.1016/j.beth.2006.08.001

Rodebaugh, T. L., Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L. & Rapee, R. M. (2004). More information from fewer questions: The factor structure and item properties of the original and brief Fear of Negative Evaluation Scale. *Psychological Assessment*, *16* (2), 169–181. doi: 10.1037/1040-3590.16.2.169

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24* (2), 3–13. doi: 10.1111/j.1745-3992.2005.00006.x

Rogers, C. R. (1946). Significant aspects of client-centered therapy. *American Psychologist*, *1* (10), 415–422. doi: 10.1037/h0060866

Rogers, C. R. (1947). Some observations on the organization of personality. *American Psychologist*, *2* (9), 358–368. doi: 10.1037/h0060883

Rogers, C. R. (1959). A theory of therapy, personality, and interpersonal relationships, as developed in the client-centered framework. In S. Koch (Hrsg.), *Psychology: A Study of a Science* (Bd. 3. Formulations of the person and the social context, S. 184–256). New York: McGraw-Hill Book Company, Inc.

Rogers, C. R. (1961). *On becoming a person.* Boston, MA: Houghton-Mifflin.

Rogers, H. J. & Hattie, J. A. (1987). A monte carlo investigation of several

person and item fit statistics for item response models. *Applied Psychological Measurement*, *11* (1), 47–57. doi: 10.1177/014662168701100103

Rohner, S. J. (1985). Cognitive-emotional response to music as a function of music and cognitive complexity. *Psychomusicology: A Journal of Research in Music Cognition*, *5* (1-2), 25–38. doi: 10.1037/h0094202

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, *1978* (9), 222–245.

Romer, M. (1975). Notes sur l'application du traitement graphique de l'information en sociologie empirique. Comparaison avec l'analyse factorielle des correspondances. *Revue Française de Sociologie; Paris*, *16* (1), 79–94. https://search.proquest.com/docview/1303246134/citation/A1D588851CBA4231PQ/7

Rorer, L. G. (1965). The great response style myth. *Psychological Bulletin*, *63* (3), 129–156.

Rosenzweig, S. (1951). Idiodynamics in personality theory with special reference to projective methods. *Psychological Review*, *58* (3), 213–223. doi: 10.1037/h0053907

Rosenzweig, S. (1985). Freud and experimental psychology: The emergence of idiodynamics. In S. Koch & D. E. Leary (Hrsg.), *A Century of psychology as science* (S. 135–207). McGraw-Hill.

Roskam, E. E. (1985). Current issues in item response theory: Beyond psychometrics. In *Measurement Personality Assess* (1. Aufl., S. 3–19). Amsterdam: Elsevier Science Publishing Company.

Ross, A. O. (1987). *Personality : the scientific study of complex human behavior*. New York: Holt, Rhinehart, and Winston.

Roßmann, J. (2017). *Satisficing in Befragungen*. Wiesbaden: Springer Fachmedien Wiesbaden. doi: 10.1007/978-3-658-16668-7

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement*, *12* (4), 397–409. doi: 10.1177/014662168801200408

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14* (3), 271–282. doi: 10.1177/014662169001400305

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44* (1), 75–92. doi: 10.1111/j.2044-8317.1991.tb00951.x

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau*, *50* (3), 140–156.

Rost, J. (2000). Haben ordinale Rasch-Modelle variierende Trennschärfen? Eine Antwort auf die Wiener Repliken. *Psychologische Rundschau*, *51* (1), 36–37.

Rost, J. (2002). When personality questionnaires fail to be unidimensional. *Psychologische Beiträge*, *44*, 108–125.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2. Aufl.). Bern: Huber.

Rost, J., Carstensen, C. & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Hrsg.), *Applications of latent trait and latent class models in the social sciences* (S. 324–332). Münster: Waxmann.

Rost, J., Carstensen, C. H. & von Davier, M. (1999). Sind die Big Five Rasch-skalierbar? *Diagnostica*, *45* (3), 119–127. doi: 10.1026//0012-1924.45.3.119

Rost, J. & Georg, W. (1991). Alternative Skalierungsmöglichkeiten zur klassischen Testtheorie am Beispiel der Skala 'Jugendzentrismus'. *ZA-Information / Zentralarchiv für Empirische Sozialforschung*, *28*, 52–74.

Rost, J. & Langeheine, R. (1997). A guide through latent structure models for categorical data. In J. Rost & R. Langeheine (Hrsg.), *Applications of latent trait and latent class models in the social sciences* (S. 13–37). Münster: Waxmann.

Rost, J. & Luo, G. (1997). An application of a Rasch-based unfolding model to a questionnaire on adolescent centrism. In J. Rost & R. Langeheine (Hrsg.), *Applications of latent trait and latent class models in the social sciences* (S. 278–286). Münster: Waxmann.

Rost, J. & von Davier, M. (1994). A conditional item-fit index for Rasch models. *Applied Psychological Measurement*, *18* (2), 171–182. doi: 10.1177/014662169401800206

Roster, C. A., Albaum, G. & Rogers, B. (2006). Can cross-national/cultural

studies presume etic equivalency in respondents' use of extreme categories of Likert rating scales? *International Journal of Market Research*, *48* (6), 741–759.

Roszkowski, M. J. & Soven, M. (2010). Shifting gears: consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35* (1), 117–134. doi: 10.1080/02602930802618344

Roth, G. (2001). *Fühlen, Denken, Handeln Wie das Gehirn unser Verhalten steuert* (5. Aufl.). Frankfurt am Main: Suhrkamp.

Rundquist, E. A. (1966). Item and response characteristics in attitude and personality measurement: A reaction to L. G. Rorer's The Great Response-Style Myth. *Psychological Bulletin*, *66* (3), 166–177. doi: 10.1037/h0023709

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55* (1), 3–38.

Rushton, J. P. & Endler, N. S. (1977). Person by situation interactions in academic achievement. *Journal of Personality*, *45* (2), 297–309. doi: 10.1111/j.1467-6494.1977.tb00153.x

Russell, J. A. & Carroll, J. M. (1999a). On the bipolarity of positive and negative affect. *Psychological Bulletin*, *125* (1), 3–30. doi: 10.1037/0033-2909.125.1.3

Russell, J. A. & Carroll, J. M. (1999b). The phoenix of bipolarity: Reply to Watson and Tellegen (1999). *Psychological Bulletin*, *125* (5), 611–617. doi: 10.1037/0033-2909.125.5.611

Saal, F. E., Downey, R. G. & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88* (2), 413–428. doi: 10.1037/0033-2909.88.2.413

Saaty, T. (2008). Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, *102* (2), 251–318.

Sackeim, H. A. & Gur, R. C. (1979). Self-deception, other-deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology*, *47* (1), 213–215. doi: 10.1037/0022-006X.47.1.213

Sadler, M., Hunger, J. & Miller, C. (2010). Personality and impression management: Mapping the Multidimensional Personality Questionnaire onto 12 self-presentation tactics. *Personality and Individual Differences*, *48* (5), 623–628.

Saint-Mont, U. (2012)aug. What measurement is all about. *Theory & Psychology*, *22* (4), 467–485. doi: 10.1177/0959354311429997

Sälzer, C. & Heine, J.-H. (2016). Students' skipping behavior on truancy items and (school) subjects and its relation to test performance in PISA 2012. *International Journal of Educational Development*, *46*, 103–113. doi: https://doi.org/10.1016/j.ijedudev.2015.10.009

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *35* (1), 139–139. doi: 10.1007/BF02290599

Samejima, F. (1999). General Graded Response Model..

Saucier, G. & Goldberg, L. R. (2001). Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of Personality*, *69* (6), 847–879.

Schäfer, T. (2008). *Determinants of music preference* (Unveröffentlichte Dissertation). Chemnitz University of Technology, Chemnitz.

Schäfer, T. & Mehlhorn, C. (2017). Can personality traits predict musical style preferences? A meta-analysis. *Personality and Individual Differences*, *116*, 265–273. doi: 10.1016/j.paid.2017.04.061

Schinka, J. A., Dye, D. A. & Curtiss, G. (1997). Correspondence between Five-Factor and RIASEC models of personality. *Journal of Personality Assessment*, *68* (2), 355.

Schlenker, B. R. (2003). Self-Presentation. In *Handbook of self and identity* (S. 492–518). New York: Guilford.

Schlenker, B. R. & Weigold, a. M. F. (1992). Interpersonal processes involving impression regulation and management. *Annual Review of Psychology*, *43* (1), 133–168. doi: 10.1146/annurev.ps.43.020192.001025

Schmit, M. J. & Ryan, A. M. (1993). The Big Five in Personnel Selection:

Factor Structure in Applicant and Nonapplicant Populations. *Journal of Applied Psychology*, *78* (6), 966–974.

Schmitt, N. & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9* (4), 367–373. doi: 10.1177/014662168500900405

Schmolck, P. (2003). *ESF-Projekt WT03: Beschreibung der eingesetzten Instrumente und individuelle Testergebnisse* (Unveröffentlichter Projektbericht). Neubiberg: Universität der Bundeswehr München.

Schmolck, P. (2004). *ESF-Projekt WT04 Individuelle Testergebnisse* (unveröffentlichter Projektbericht Nr. 1). Neubiberg: Universität der Bundeswehr München.

Schmolck, P. (2005). *ESF-Projekt WT05 Individuelle Testergebnisse* (unveröffentlichter Projektbericht Nr. 2). Neubiberg: Universität der Bundeswehr München.

Schmolck, P. (2006a). *ESF-Projekt FT06: Beschreibung der eingesetzten Instrumente und individuelle Testergebnisse* (Unveröffentlichter Projektbericht). Neubiberg: Universität der Bundeswehr München.

Schmolck, P. (2006b). *ESF-Projekt WT06 Individuelle Testergebnisse* (unveröffentlichter Projektbericht Nr. 3). Neubiberg: Universität der Bundeswehr München.

Schneewind, K. A. (1982). *Persönlichkeitstheorien I* (Bde. Erträge der Forschung, 168). Darmstadt: Wissenschaftliche Buchgesellschaft.

Scholz, J., Klein, M. C., Behrens, T. E. J. & Johansen-Berg, H. (2009). Training induces changes in white-matter architecture. *Nature Neuroscience*, *12* (11), 1370. doi: 10.1038/nn.2412

Schriesheim, C. A. & Eisenbach, R. J. (1995). An exploratory and confirmatory factor-analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *Journal of Management*, *21* (6), 1177–1193.

Schriesheim, C. A., Eisenbach, R. J. & Bailey, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement*, *51* (1), 67–78. doi: 10.2466/pr0.1999.85.1.213

Schriesheim, C. A. & Hill, K. D. (1981). Controlling acquiescence response

bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement*, *41* (4), 1101–1114. doi: 10.1177/0013164481041004 20

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6* (2), 461–464.

Schönemann, P. H. (1970). On metric multidimensional unfolding. *Psychometrika*, *35* (3), 349–366.

Seashore, C. E. (1938). *Psychology of Music*. New York: Mcgraw-Hill Book Company.

Segura, S. L. & González-Romá, V. (2003). How Do Respondents Construe Ambiguous Response Formats of Affect Items? *Journal of Personality & Social Psychology*, *85* (5), 956–968.

Seiwald, B. B. (2002). Replicability and generalizability of Kubinger's results: Some more studies on faking personality inventories. *Psychologische Beiträge*, *44* , 17–23.

Shadel, W. G. & Cervone, D. (1993). The Big Five versus nobody? *American Psychologist*, *48* (12), 1300–1302. doi: 10.1037/0003-066X.48.12.1300

Sheldon, W. H. (1963). *The Varieties of Human Physique: An Introduction to Constitutional Psychology*. Hafner Pub.

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27* (3), 219–246. doi: 10.1007/BF02289621

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27* (3), 219–246. doi: 10.1007/BF02289621

Shoda, Y. (1999). A unified framework for the study of behavioral consistency: Bridging person x situation interaction and the consistency paradox. *European Journal of Personality*, *13*, 361–387.

Shoda, Y. & Mischel, W. (2000). Reconciling contextualism with the core assumptions of personality psychology. *European Journal of Personality*, *14* (5), 407–428. doi: 10.1002/1099-0984(200009/10)14:5<407::AID-PER391>3.0.CO;2-3

Shoda, Y., Mischel, W. & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: incorporating psychologi-

cal situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology*, *67* (4), 674–687.

Siem, F. M. (1998). Metatraits and self-schemata: Same or different? *Journal of Personality*, *66* (5), 783–803. doi: 10.1111/1467-6494.00032

Sijtsma, K. (1986). A coefficient of deviance of response Patterns. *Kwantitatieve Methoden*, *7* (22), 131–145.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology*, *22* (6), 786–809. doi: 10.1177/0959354312454353

Sijtsma, K. & Emons, W. H. M. (2013)dec. Separating models, ideas, and data to avoid a paradox: Rejoinder to Humphry. *Theory & Psychology*, *23* (6), 786–796. doi: 10.1177/0959354313503724

Sijtsma, K. & Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, *25* (4), 391–415. doi: 10.3102/10769986025004391

Sijtsma, K. & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, *16* (2), 149–157. doi: 10.1177/014662169201600204

Sijtsma, K. & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks: SAGE.

Sikkema, P. (1999). *Jongeren 890?990: Een generatie waar om gevochten wordt [Youth 890?990: A generation that is being fought for]* (Bericht). Amsterdam: Interview-NSS.

Silvia, P. J., Fayn, K., Nusbaum, E. C. & Beaty, R. E. (2015). Openness to experience and awe in response to nature and music: Personality and profound aesthetic experiences. *Psychology of Aesthetics, Creativity, and the Arts*. doi: 10.1037/aca0000028

Singer, S., Decker, O. & Glaesmer, H. (2007). Testinformation-EXPLORIX. *Diagnostica*, *53* (1), 53–55. doi: 10.1026/0012-1924.53.1.53

Sixtl, F. (1973). Probabilistic unfolding. *Psychometrika*, *38* (2), 235–248. doi: 10.1007/BF02291116

Skinner, B. F. (1963). Operant behavior. *American Psychologist*, *18* (8), 503–515. doi: 10.1037/h0045185

Skinner, B. F. (1965). *Science and human behavior*. New York, NY: Free

Press.

Sliter, K. A. & Zickar, M. J. (2014). An IRT examination of the psychometric functioning of negatively worded personality items. *Educational and Psychological Measurement*, *74* (2), 214–226. doi: 10.1177/0013164413504584

Sloboda, J. (1986). *The Musical Mind: the cognitive psychology of music* (Nr. no. 5). New York : Oxford: Oxford University Press ; Clarendon Press.

Sloboda, J. A. (1991). Music structure and emotional response: Some empirical findings. *Psychology of Music*, *19* (2), 110 –120. doi: 10.1177/0305735691192002

Smith, D. B. & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology*, *87* (2), 211–219. doi: 10.1037/0021-9010.87.2.211

Spada, H. (Hrsg.). (1992). *Lehrbuch allgemeine Psychologie* (2. Aufl.). Bern: Huber.

Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology*, *15* (2), 201–292. doi: 10.2307/1412107

Spector, P. E. (1992). *Summated rating scale construction: An introduction* (M. S. Lewis-Beck, Hrsg.). Thousand Oaks: SAGE Publications Inc.

Spector, P. E. (2006). Method variance in organizational research truth or urban legend? *Organizational Research Methods*, *9* (2), 221–232. doi: 10.1177/1094428105284955

Spector, P. E. & Brannick, M. T. (2010). If Thurstone Was Right, What Happens When We Factor Analyze Likert Scales? *Industrial and Organizational Psychology*, *3* (4), 502–503. doi: 10.1111/j.1754-9434.2010.01280.x

Spector, P. E., Rosen, C. C., Richardson, H. A., Williams, L. J. & Johnson, R. E. (2017). A new perspective on method variance: A measure-centric approach. *Journal of Management*, 0149206316687295. doi: 10.1177/0149206316687295

Spector, P. E., Van Katwyk, P. T., Brannick, M. T. & Chen, P. Y. (1997). When two factors don't reflect two constructs: how item characteristics can produce artifactual factors. *Journal of Management*, *23* (5), 659–677. doi: 10.1016/S0149-2063(97)90020-9

Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Wien; New York: Springer.

Stachowiak, H. (1983). *Modelle, Konstruktion der Wirklichkeit*. München: Fink.

Stark, S., Chernyshenko, O. S., Drasgow, F. & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91* (1), 25–39.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25* (2), 173–180.

Steiger, J. H. & Lind, J. C. (1980). Statistically based tests for the number of common factors. Iowa City, IA.

Stemmler, M. (2014). *Person-centered methods: Configural Frequency Analysis (CFA) and other methods for the analysis of contingency tables*. New York: Springer Publishing Company.

Stemmler, M. & Bingham, C. R. (2003). Log-linear modeling and two-sample CFA in the search of discrimination types. *Psychology Science*, *45* (2), 421–429.

Stemmler, M. & Heine, J.-H. (2017). Using Configural Frequency Analysis as a Person-centered Analytic Approach with Categorical Data. *International Journal of Behavioral Development*, *41* (5), 632–646. doi: 10.1177/0165025416647524

Stern, W. (1900). *Über Psychologie der individuellen Differenzen: Ideen zu einer differentiellen Psychologie* (Nr. 12). Leipzig: Barth.

Stern, W. (1911). *Die Differentielle Psychologie in ihren methodischen Grundlagen*. Leipzig: Verlag von Johann Ambrosius Barth.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103* (2684), 677–680. doi: 10.2307/1671815

Stewart, T. J. & Frye, A. W. (2004). Investigating the use of negatively phrased survey items in medical education settings: Common wisdom or common mistake? *Academic Medicine*, *79* (10), 18–20.

Steyer, R. & Eid, M. (2000). *Messen und Testen*. Berlin: Springer.

Stokman, F. & van Schuur, W. H. (1980). Basic scaling. *Quality and Quantity*, *14* (1), 5–30. doi: 10.1007/BF00154792

Strong, E. K. (1943). *Vocational interests of men and women*. Stanford:

Stanford University Press.

Sugiura, N. (1978). Further analysis of the data by Akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, *7* (1), 13–26. doi: 10.1080/03610927808827599

Summers, F. (2008). Theoretical insularity and the crisis of psychoanalysis. *Psychoanalytic Psychology*, *25* (3), 413–424. doi: 10.1037/0736-9735.25.3.413

Suszek, H., Holas, P., Wyrzykowski, T., Lorentzen, S. & Kokoszka, A. (2015). Short-term intensive psychodynamic group therapy versus cognitive-behavioral group therapy in day treatment of anxiety disorders and comorbid depressive or personality disorders: study protocol for a randomized controlled trial. *Trials*, *16*. doi: 10.1186/s13063-015-0827-6

Suárez-Falcón, J. C. & Glas, C. A. W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *56* (1), 127–143. doi: 10.1348/000711003321645395

Swain, S. D., Weathers, D. & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research (JMR)*, *45* (1), 116–131.

Taber, B. J. (2006). Vocational Preference Inventory (VPI). In J. Greenhaus & G. Callanan (Hrsg.), *Encyclopedia of Career Development.* Thousand Oaks: SAGE Publications, Inc.

Tan, S.-L., Pfordresher, P. & Harré, R. (2010). *Psychology of music: from sound to significance*. Hove, East Sussex [England] ; New York, NY: Psychology Press.

Tarnai, C. & Rost, J. (1990). *Identifying aberrant response patterns in the Rasch model: The Q index*. Münster: Institut für sozialwissenschaftliche Forschung (ISF).

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, *49* (1), 95–110. doi: 10.1007/BF02294208

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational and Behavioral Statistics*, *10* (1), 55–73. doi: 10.3102/10769986010001055

Tatsuoka, K. K. & Linn, R. L. (1983). Indices for detecting unusual pat-

terns: Links between two general approaches and potential applications. *Applied Psychological Measurement*, *7* (1), 81–96.

Tay, L., Drasgow, F., Rounds, J. & Williams, B. A. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal? *Journal of Applied Psychology; Journal of Applied Psychology*, *94* (5), 1287.

Taylor, D., Carlyle, J.-a., McPherson, S., Rost, F., Thomas, R. & Fonagy, P. (2012). Tavistock Adult Depression Study (TADS): a randomised controlled trial of psychoanalytic psychotherapy for treatment-resistant/treatment-refractory forms of depression. *BMC Psychiatry*, *12*, 60. doi: 10.1186/1471-244X-12-60

Tedeschi, J. T. (Hrsg.). (1981). *Impression management theory and social psychological research*. New York: Academic Press.

Tedeschi, J. T., Schlenker, B. R. & Bonoma, T. V. (1971). Cognitive dissonance: Private ratiocination or public spectacle? *American Psychologist*, *26* (8), 685–695. doi: 10.1037/h0032110

Tekman, H. G. & Hortaçsu, N. (2002). Music and social identity: Stylistic identification as a response to musical style. *International Journal of Psychology*, *37* (5), 277–285.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality*, *56* (3), 621–663. doi: 10.1111/j.1467-6494.1988.tb00905.x

Tendeiro, J. N. & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement*, *51* (3), 239–259. doi: 10.1111/jedm.12046

Tendeiro, J. N., Meijer, R. R. & Niessen, A. S. M. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74* (5), 1–27. doi: 10.18637/jss.v074.i05

Tett, R. P., Freund, K. A., Christiansen, N. D., Fox, K. E. & Coaster, J. (2012). Faking on self-report emotional intelligence and personality tests: Effects of faking opportunity, cognitive ability, and job type. *Personality and Individual Differences*, *52* (2), 195–201. doi: 10.1016/j.paid.2011.10.017

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47* (2), 175–186. doi: 10.1007/

BF02296273

Thissen, D. (2013). The meaning of goodness-of-fit tests: Commentary on "Goodness-of-fit Assessment of Item Response Theory Models". *Measurement: Interdisciplinary Research and Perspectives*, *11* (3), 123–126. doi: 10.1080/15366367.2013.835205

Thissen, D., Reeve, B. B., Bjorner, J. B. & Chang, C.-H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research*, *16* (1), 109–119. doi: 10.1007/s11136-007-9169-5

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51* (4), 567–577. doi: 10.1007/BF02295596

Thorndike, R. L., Angoff, W. H. & American Council on Education. (1971). *Educational measurement*. Washington: American Council on Education.

Thurstone, L. L. (1927a). A law of comparative judgement. *Psychological Review*, *34*, 273–286.

Thurstone, L. L. (1927b). The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, *21* (4), 384–400. doi: 10.1037/h0065439

Thurstone, L. L. (1927c). Psychophysical analysis. *The American journal of psychology*, *38* (3), 368–389.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 529–554.

Thurstone, L. L. (1929). Fechner's law and the method of equal appearing intervals. *Journal of Experimental Psychology*, *12* (3), 214.

Thurstone, L. L. (1934). The vectors of mind. *Psychological Review*, *41* (1), 1–32. doi: 10.1037/h0075959

Thurstone, L. L. (1935). *The Vectors Of Mind Multiple Factor Analysis For The Isolation Of Primary Traits*. The University Of Chicago Press.

Thurstone, L. L. (1938). *Primary Mental Abilities*. University of Chicago Press.

Thurstone, L. L. & Chave, E. (1929). *Measurement of attitude: a psychophysical method and some experiments with a scale for measuring attitude toward the church*. Chicago & London: The University Of Chicago Press.

Tietz, V. (1900a). Ueber eine neu Art der Preisvertheilung. *Wiener Schach-*

zeitung, *III* (1), 223–230.

Tietz, V. (1900b). Zum Capitel: Turnierstärken. *Wiener Schachzeitung*, *III* (1), 1–4.

Todt, E. (1978). *Das Interesse. Empirische Untersuchungen zu einem Motivationskonzept*. Bern u.a.: Huber.

Todt, E. & Schreiber, S. (1998). Development of Interests. In A. Hoffmann, A. Krapp, K. A. Renninger & J. Baumert (Hrsg.), *Interest and learning: Proceedings of the Seeon conference on interest and gender* (S. 25–40). Kiel: Institute for Science Education at the University of Kiel (IPN).

Tokar, D. M. & Swanson, J. L. (1995). Evaluation of the Correspondence between Holland's Vocational Personality Typology and the Five-Factor Model of Personality. *Journal of Vocational Behavior*, *46* (1), 89–108. doi: 10.1006/jvbe.1995.1006

Torgerson, W. S. (1961). Scaling and test theory. *Annual Review of Psychology*, *12* (1), 51–70. doi: 10.1146/annurev.ps.12.020161.000411

Torgerson, W. S. (1967). *Theory and methods of scaling* (7. Aufl.). New York: Wiley.

Tourangeau, R. & Rasinski, K. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103* (3), 299.

Tourangeau, R., Rips, L. J. & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

Tracey, T. J. G. (2003). Interest traitedness as a moderator of interest–occupation congruence. *Journal of Vocational Behavior*, *62* (1), 1–10. doi: 10.1016/S0001-8791(02)00011-8

Tracey, T. J. G. (2016). A note on socially desirable responding. *Journal of Counseling Psychology*, *63* (2), 224–232. doi: 10.1037/cou0000135

Tucker, L. R. & Messick, S. (1963). An individual differences model for multidimensional scaling. *Psychometrika*, *28* (4), 333–367. doi: 10.1007/BF02289557

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39–55.

Überla, K. (1977). *Faktorenanalyse: Eine systematische Einführung für Psychologen, Mediziner, Wirtschafts- und Sozialwissenschaftler*. Springer-Verlag.

Uziel, L. (2010). Rethinking social desirability scales from impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, *5* (3), 243–262.

van den Wittenboer, G. L. H., Hox, J. J. & De Leeuw, E. D. (1997). Aberrant response patterns in elderly respondents: Latent class analysis of respondent scalability. In J. Rost & R. Langeheine (Hrsg.), *Applications of latent trait and latent class models in the social sciences* (S. 155–162). Münster: Waxmann.

van den Wollenberg, A. (1982). Two new test statistics for the rasch model. *Psychometrika*, *47* (2), 123–140. doi: 10.1007/BF02296270

van der Ark, L. A. (2001). Relationships and Properties of Polytomous Item Response Theory Models. *Applied Psychological Measurement*, *25* (3), 273–282. doi: 10.1177/01466210122032073

van der Flier, H. (1982). Deviant Response Patterns and Comparability of Test Scores. *Journal of Cross-Cultural Psychology*, *13* (3), 267–298. doi: 10.1177/0022002182013003001

van der Kloot, W., Kroonenberg, P. M. & Bakker, D. (1985). Implicit Theories of Personality: Further Evidence of Extreme Response Style. *Multiariate Behavioral Research*, *20*, 369–387.

van Schuur, W. H. (1984). *Structure in Political Beliefs: A New Model for Stochastic Unfolding with Application to European Party Activists*. Amsterdam: CT Press.

van Schuur, W. H. (1992). Nonparametric unidimensional unfolding for multicategory data. *Political Analysis*, *4* (1), 41–74.

van Schuur, W. H. (1995). The use of the unfolding model in survey measurement. Survey Measurement and Process Quality. In *Proceedings of the International Conference on Survey Measurement and Process Quality.* Bristol.

van Schuur, W. H. (1997). Nonparametric IRT Models for Dominance and Proximity Data. In M. Wilson, G. Engelhard Jr & K. Draney (Hrsg.), *Objective Measurement: Theory into Practice* (Bd. 4, S. 313–331). Greenwich: Ablex Publishing Corporation.

van Schuur, W. H. (2006). The Unfolding Fallacy Unveiled: Visualizing Structures of Dichotomous Unidimensional Item-Response-Theory Da-

ta by Multiple Correspondence Analysis. In M. Greenacre & J. Blasius (Hrsg.), *Multiple correspondence analysis and related methods.* Chapman & Hall/CRC.

van Schuur, W. H. (2011). *Ordinal Item Response Theory: Mokken Scale Analysis.* SAGE.

van Schuur, W. H. & Kiers, H. A. L. (1994). Why Factor Analysis Often Is the Incorrect Model for Analyzing Bipolar Concepts, and What Model to Use Instead. *Applied Psychological Measurement*, *18* (2), 97–110. doi: 10.1177/014662169401800201

van Schuur, W. H. & Kruijtbosch, M. (1995). Measuring subjective well-being: Unfolding the Bradburn Affect Balance Scale. *Social Indicators Research*, *36* (1), 49–74. doi: 10.1007/BF01079396

van Schuur, W. H. & Molenaar, I. W. (1982). MUDFOLD: Multiple Stochastic Unidimensional Unfolding. In H. Caussinus, P. Ettinger & R. Tomassone (Hrsg.), *COMPSTAT 1982 5th Symposium held at Toulouse 1982* (S. 419–424). Physica-Verlag HD. doi: 10.1007/978-3-642-51461-6_64

van Sonderen, E., Sanderman, R. & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *Plos One*, *8* (7), e68967. doi: 10.1371/journal.pone.0068967

Vautier, S., Veldhuis, M., Lacot, E. & Matton, N. (2012)dec. The ambiguous utility of psychometrics for the interpretative foundation of socially relevant avatars. *Theory & Psychology*, *22* (6), 810–822. doi: 10.1177/0959354312450093

Verhelst, N. D. & Glas, C. A. W. (1995). The one parameter logistic model. In G. Fischer & I. Molenaar (Hrsg.), *Rasch models: Foundations, recent developments, and applications* (S. 215–238). New York: Springer.

Verhelst, N. D. & Verstralen, H. H. F. M. (1993). A stochastic unfolding model derived from the partial credit model. *Kwantitative Methoden*, *42*, 73–92.

Victor, N. (1983). A note on contingency tables with one structural zero. *Biometrical Journal*, *25*, 283–289.

Victor, N. (1989). An alternativ approach to Configural Frequency Analysis. *Methodika*, *3*, 61–73.

Victor, N. & Kieser, M. (1991). A test procedure for an alternative approach

to Configural Frequency Analysis. *Methodika*, *5*, 87–97.

Viswanathan, M. (2005). *Measurement error and research design.* Thousand Oaks: SAGE Publications.

Viswesvaran, C. & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59* (2), 197–210. doi: 10.1177/00131649921969802

Vittersø, J., Biswas-Diener, R. & Diener, E. (2005). The divergent meanings of life satisfaction: Item response modeling of the Satisfaction With Life scale in greenland and norway. *Social Indicators Research*, *74* (2), 327–348.

von Davier, M. (2001). *WINMIRA 2001.* Groningen, The Netherlands: ASC-Assessment Systems Corporation USA and Science Plus Group.

von Davier, M. (2009). Is there need for the 3pl Model? Guess what? *Measurement: Interdisciplinary Research and Perspectives*, *7* (2), 110–114. doi: 10.1080/15366360903117079

von Davier, M. & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. v. Davier & C. H. Carstensen (Hrsg.), *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications* (1. Aufl.). Springer US.

von Eye, A. (2004). Base models for configural frequency analysis. *Psychology Science*, *46*, 150–170.

von Eye, A. & Mair, P. (2008). A functional approach to Configural Frequency Analysis. *Austrian Journal of Statistics*, *37* (2), 161–173.

Waller, N. G. & Reise, S. P. (1992). Genetic and environmental influences on item response pattern scalability. *Behavior Genetics*, *22* (2), 135–152.

Waller, N. G., Tellegen, A., McDonald, R. P. & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality*, *64* (3), 545–576. doi: 10.1111/j.1467-6494.1996.tb00521.x

Wang, W.-C., Chen, H.-F. & Jin, K.-Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement*, *75* (1), 157–178. doi: 10.1177/0013164414528209

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54* (3), 427–450. doi: 10.1007/BF02294627

Warr, P. & Coffman, T. (1970). Personality, involvement and extremity of judgement. *British Journal of Social and Clinical Psychology*, *9* (2), 108–121.

Warrens, M. J. & Heiser, W. J. (2006). Scaling unidimensional models with multiple correspondence analysis. In M. Greenacre & J. Blasius (Hrsg.), *Multiple correspondence analysis and related methods.* Chapman & Hall/CRC.

Watson, D. & Tellegen, A. (1999). Issues in dimensional structure of affect—Effects of descriptors, measurement error, and response formats: Comment on Russell and Carroll (1999). *Psychological Bulletin*, *125* (5), 601–610. doi: 10.1037/0033-2909.125.5.601

Watzlawick, P., Beavin, J. & Jackson, D. (2000). *Menschliche Kommunikation: Formen, Störungen, Paradoxien*. Huber.

Weekers, A. M. & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models: An illustration with a Dutch dominance and unfolding personality inventory. *European Journal of Psychological Assessment*, *24* (1), 65–77.

Weijters, B. & Baumgartner, H. (2012). Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research (JMR)*, *49* (5), 737–747.

Weijters, B., Geuens, M. & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, *34* (2), 105.

Weijters, B., Geuens, M. & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological methods*, *15* (1), 96.

Weinberger, A. H., Darkes, J., Del Boca, F. K., Greenbaum, P. E. & Goldman, M. S. (2006)mar. Items as context: Effects of item order and ambiguity on factor structure. *Basic and Applied Social Psychology*, *28* (1), 17–26. doi: 10.1207/s15324834basp2801_2

Wermuth, V. N. (1973). Anmerkungen zur Konfigurationsfrequenzanalyse. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *3*, 5–21.

Westlye, L. T., Bjornebekk, A., Grydeland, H., Fjell, A. M. & Walhovd, K. B. (2011). Linking an anxiety-related personality trait to brain white matter microstructure diffusion tensor imaging and harm avoidance. *Archives*

of General Psychiatry, *68* (4), 369–377.

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M. & Ostendorf, F. (2013). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences*, *34* (2), 69–81. doi: 10.1027/1614-0001/a000102

Wetzel, E., Böhnke, J. R. & Rose, N. (2016). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement*, *76* (2), 304–324. doi: 10.1177/0013164415591848

Wetzel, E., Carstensen, C. H. & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47* (2), 178–189. doi: 10.1016/j.jrp.2012.10.010

Wetzel, E. & Hell, B. (2013). Gender-related differential item functioning in vocational interest measurement: An analysis of the AIST-R. *Journal of Individual Differences*, *34* (3), 170–183. doi: 10.1027/1614-0001/a000112

Wetzel, E., Lüdtke, O., Zettler, I. & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23* (3), 279–291. doi: 10.1177/1073191115583714

Wiggins, J. S. (1962). Strategic, method, and stylistic variance in the MMPI. *Psychological Bulletin*, *59* (3), 224–242.

Wiggins, J. S. (1964). Convergences among stylistic response measures from objective personality tests. *Educational and Psychological Measurement*, *24* (3), 551–562. doi: 10.1177/001316446402400310

Wiggins, J. S. (1973). *Personality and prediction*. Reading: Addison-Wesley.

Wilhelm von Ockham. (1967). *Scriptum in Librum Primum Sententiarum Ordinatio - Prologus et Distinctio Prima* (Bd. 1 Opera Theologica). St. Bonaventure, N.Y: Cura Instituti Franciscani Universitatis S. Bonaventurae.

Williams, E. F., Dunning, D. & Kruger, J. (2013). The hobgoblin of consistency: Algorithmic judgment strategies underlie inflated self-assessments of performance. *Journal of Personality and Social Psychology*, *104* (6), 976–994. doi: 10.1037/a0032416

Willse, J. T. (2014). *mixRasch: Mixture Rasch Models with JMLE* (R package version 1.1). `https://CRAN.R-project.org/package=mixRasch`

Willson, T. D., Dunn, D. S., Kraft, D. & Lisle, D. L. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Hrsg.), *Advances in experimental social psychology* (Bd. 22, S. 287–343). San Diego: Academic Press.

Wilson, T., Diesterhoft, J. F., Gelman, S. A., Hollon, S., Murphy, K. R. & Treisman, A. (2008). 2008 Award Winners: John L. Holland, Award for distinguished scientific applications of psychology. *American Psychologist*, *63* (8), 672–674.

Wing, H. D. (1941). A factorial study of musical tests. *British Journal of Psychology. General Section*, *31* (4).

Woollett, K. & Maguire, E. A. (2011). Acquiring 'the Knowledge' of London's Layout Drives Structural Brain Changes. *Current Biology*, *21* (24-2), 2109. doi: 10.1016/j.cub.2011.11.018

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14* (2), 97–116.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Hrsg.), *The new rules of measurement: What every psychologist and educator should know* (S. 65–104). New York: Psychology Press. doi: 10.4324/9781410603593

Wright, B. D., Gaskell, G. D. & O'Muircheartaigh, C. A. (1994). How much is 'quite a bit'? Mapping between numerical values and vague quantifiers. *Applied Cognitive Psychology*, *8* (5), 479–496. doi: 10.1002/acp.2350080506

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D. & Masters, G. N. (1990). Computation of OUTFIT and INFIT statistics. *Rasch Measurement Transactions*, *3* (4), 84–5.

Wright, B. D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29* (1), 23 –48. doi: 10.1177/001316446902900102

Wu, M. & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14* (4), 339–355.

Wu, M. L., Adams, R. J., Wilson, M. & Haldane, S. A. (2007). *ACER Con-*

Quest: Generalised item response modeling software. Melbourne: ACER.

Yamamoto, K. (1989). *Hybrid model of IRT and latent class models.* (Research Report Nr. RR-89-41). Princeton: Educational Testing Service (ETS).

Yamamoto, K. & Everson, H. T. (1995). Modeling the mixture of IRT and pattern responses by a modified HYBRID model. *ETS Research Report Series*, *1995* (1), i–26. doi: 10.1002/j.2333-8504.1995.tb01651.x

Yamamoto, M., Naga, S. & Shimizu, J. (2007). Positive musical effects on two types of negative stressful conditions. *Psychology of Music*, *35* (2), 249 –275. doi: 10.1177/0305735607070375

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8* (2), 125–145. doi: 10.1177/014662168400800201

Young, F. W. (1984). Scaling. *Annual Review of Psychology*, *35* (1), 55–81. doi: 10.1146/annurev.ps.35.020184.000415

Zener, T. & Schnuelle, L. (1976). Effects of the self-directed search on high school students. *Journal of Counseling Psychology*, *23* (4), 353–359.

Zentner, M., Grandjean, D. & Scherer, K. R. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, *8* (4), 494–521. doi: 10.1037/1528-3542.8.4.494

Zickar, M. J., Gibby, R. E. & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, *7* (2), 168–190. doi: 10.1177/1094428104263674

Zickar, M. J. & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology*, *84* (4), 551–563. doi: 10.1037/0021-9010.84.4.551

Ziegler, M. (2011). Applicant faking: a look into the black box. *The Industrial-Organizational Psychologist*, *49* (1), 29–36.

Ziegler, M. (2015). "F*** you, I won't do what you told me!" - response biases as threats to psychological assessment. *European Journal of Psychological Assessment*, *31* (3), 153–158. doi: 10.1027/1015-5759/a000292

Ziegler, M. & Bühner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement*, *69* (4), 548 –565. doi: 10.1177/0013164408324469

Ziegler, M., Maaß, U., Griffith, R. & Gammon, A. (2015). What is the nature of faking? Modeling distinct response patterns and quantitative differences in faking at the same time. *Organizational Research Methods*, 1094428115574518. doi: 10.1177/1094428115574518

Ziegler, M., MacCann, C. & Roberts, R. (2012). *New perspectives on faking in personality assessment*. Oxford: Oxford University Press.

Ziegler, M., Toomela, A. & Bühner, M. (2009). A reanalysis of Toomela (2003): Spurious measurement error as cause for common variance between personality factors. *Psychology Science Quarterly*, *51* (1), 65–75.

Zubin, J. (1934). The determination of response-patterns in personality inventories. *Psychological Bulletin*, *31* (9), 713.

Zubin, J. (1937). The determination of response patterns in personality adjustment inventories. *Journal of Educational Psychology*, *28* (6), 401–413. doi: 10.1037/h0058522

Zuckerman, M., Eysenck, S. & Eysenck, H. J. (1978). Sensation seeking in England and America: Cross-cultural, age, and sex comparisons. *Journal of Consulting and Clinical Psychology*, *46* (1), 139–149.

Zuckerman, M. & Norton, J. (1961). Response set and content factors in the California F Scale and the Parental Attitude Research Instrument. *The Journal of Social Psychology*, *53* (2), 199–210.

Zweigenhaft, R. L. (2008). A Do Re Mi Encore. *Journal of Individual Differences*, *29* (1), 45–55. doi: 10.1027/1614-0001.29.1.45

Zysno, P. V. (1993). Polytome Skalogramm-Analyse. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *14* (1), 37–49.