universität wien

Diplomarbeit

Titel der Diplomarbeit

Erfassung von Antworttendenzen in optischen Ratingskalen mittels multidimensionaler Item Response Theorie (MIRT) Modelle im NEO-FFI

verfasst von

Claudia Krutis

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, 2014

Studienkennzahl: A 298

Studienrichtung: Psychologie

Betreuerin: Mag. Dr. Lale Khorramdel-Ameri

Danksagung

Ich möchte mich an dieser Stelle bei all jenen Personen bedanken, die mich bei der Erstellung dieser Diplomarbeit unterstützt haben.

An erster Stelle danke ich Mag. Dr. Lale Khorramdel-Ameri für die engagierte und verständnisvolle Betreuung, ihre Geduld und die Einführung in einen spannenden, neuen Themenbereich.

Ein besonderes Dankeschön geht an Mag. Alina Bugelnig für die tatkräftige Hilfe und moralische Unterstützung während der Datenerhebung und der Erstellung der Diplomarbeit sowie für die wundervolle Freundschaft, die sich daraus ergeben hat.

Vielen Dank auch an meine Freunde und Familie, vor allem meine Mutter Eva Mokesch und meine Großmutter Herta Germann, die immer an mich geglaubt haben. Ganz speziellen Dank auch an meinen Vater, Josef Krutis, ohne dessen moralische und finanzielle Unterstützung ich nie hätte studieren können.

Weiters möchte ich mich auch noch bei Mag. Katharina Ebenberger, Barbara Franzl und Karin Pieber für die Hilfe und Motivation bedanken, die mir beide zukommen haben lassen.

Herzlichen Dank auch an die Personen, die so kurzfristig das Korrekturlesen dieser Diplomarbeit auf sich genommen haben.

Anmerkung

Die Datenerhebung für diese Diplomarbeit wurde gemeinsam mit einer anderen Diplomandin (Alina Bugelnig) durchgeführt und beide Diplomarbeiten beruhen auf derselben Basisliteratur. Sollte es zu Überschneidungen und Ähnlichkeiten im Text kommen, sind diese unbeabsichtigt; trotz möglicher Ähnlichkeiten handelt es sich um völlig unterschiedliche Fragestellungen.

Abstract

Der verfälschende Einfluss von Antworttendenzen auf Fragebogenergebnisse in Untersuchungen und folglich deren Validität ist mittlerweile bekannt (vgl. Baumgartner & Steenkamp, 2001; De Jong, Steenkamp, Fox & Baumgartner, 2008; Dolnicar & Grun, 2009; Weijters, Schillewaert & Geuens, 2008). Um diesen Einfluss deutlicher zu untersuchen und schlussendlich auszuschalten oder zumindest zu minimieren, bedarf es allerdings weiterer Forschung, vor allem in Hinblick auf die eindimensionale Messung der Antworttendenzen (vgl. Khorramdel & von Davier, 2014; von Davier & Khorramdel (2013). In der vorliegenden Diplomarbeit wurde deshalb untersucht, ob mit Hilfe eines neuen multidimensionalen Item Response Theorie (MIRT) Ansatzes von Böckenholt (2012), erweitert durch Khorramdel und von Davier (2014) und von Davier und Khorramdel (2013), die Antworttendenz zur Mitte und Antworttendenz zu extremen Urteilen in den untersuchten Daten vorhanden bzw. eindimensional messbar sind, sowie ob die Daten von diesen bereinigt werden können. Die Datenerhebung fand im Rahmen eines *Online-Self-Assessments* statt, das Erhebungsinstrument war das NEO-Fünf-Faktoren-Inventar (NEO-FFI; Borkenau & Ostendorf, 1993). Die durch die Testung gewonnenen polytomen Daten wurden in binäre Pseudoitems zerlegt, welche die Antwortsubprozesse der Testpersonen darstellen; im Anschluss daran wurden diese Pseudoitems mit IRT Modellen über ein- oder mehrdimensionalen Faktoren für Antworttendenzen und Merkmale modelliert. Die Ergebnisse zeigen, dass die untersuchten Antworttendenzen tatsächlich in den Daten vorhanden sind und zumindest die Tendenz zur Mitte eindimensional messbar ist. Für die Tendenz zu extremen Urteilen konnte keine eindimensionale Messung erzielt werden. Weiters wurde festgestellt, dass eine bestimmte – von Antworttendenzen bereinigte – Kodierung der Items (Pseudoitems *d*; nur jene Antwortmöglichkeiten wurden kodiert, die weitgehend unbeeinflusst von Antworttendenzen schienen) eine passende Messung der fünf untersuchten Persönlichkeitsmerkmale darstellten.

The biasing influence of response styles on the results of questionnaire studies and their validity has been known for a while (cf. Baumgartner & Steenkamp, 2001; De Jong, Steenkamp, Fox, & Baumgartner, 2008; Dolnicar & Grun, 2009; Weijters,

Schillewaert & Geuens, 2008). To further investigate and eliminate response styles or at least minimizing their influence, additional research is needed, especially concerning the unidimensional measuring of response styles (cf. Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013). Therefore, this thesis examined whether midpoint and extreme response styles were present in the collected data, whether they could be measured unidimensionally and whether the data could be corrected for those response styles. For this, a new multidimensional Item Response Theory (MIRT) approach was used, based on an idea of Böckenholt (2012), expanded by Khorramdel and von Davier (2014) and von Davier and Khorramdel (2013). The data was collected through an *Online-Self-Assessment*, the instrument used was the *NEO-Fünf-Faktoren-Inventar* (NEO-FFI; Borkenau & Ostendorf, 1993). The collected polytomous data were decomposed into binary pseudo items, which represent response sub processes; these pseudo items were then modelled with uni- and multidimensional response style and trait factors using IRT models. Results show that the examined response styles are present in the data and that the midpoint response style could be measured unidimensionally while this was not possible for the extreme response style. Furthermore, it was determined that a certain item coding (the pseudo items *d*), corrected for response styles (by coding only those response categories not affected by response styles), were a suitable measure for the examined five personality traits.

Inhaltsverzeichnis

1 Einleitung

Eines der großen Themengebiete der Psychologie ist seit langer Zeit die Erfassung und Messung der menschlichen Persönlichkeit. Das hierzu am häufigsten verwendete Instrument ist der Persönlichkeitsfragebogen. Dieser ist dadurch charakterisiert, dass die befragten Personen eine Selbstauskunft geben (Rost, 2004). Trotz der Entwicklung gewisser psychologisch-diagnostischer Verfahren – sogenannter Objektiver Persönlichkeitstests – welche die Ausprägungen der interessierenden Merkmale objektiv über das tatsächlich beobachtete Verhalten in bestimmten Situationen erfassen, erfreut sich der Persönlichkeitsfragebogen ungebrochener Beliebtheit. Er wird aufgrund seiner vielfältigen Einsatzmöglichkeiten und der zeit- und kostengünstigen Anwendung nicht nur in individuellen Beratungen (z.B. Persönlichkeits-, Gesundheits- und Klinische Psychologie) verwendet, sondern beispielsweise auch in der Personalauswahl (trotz der Verfälschbarkeit der Antworten) oder für sogenannte webbasierte *Self-Assessments* (z.B. Selbsttestung zur Studieneignung).

Die oben angeführte Selbstauskunft über das eigene Erleben und Verhalten in einem Persönlichkeitsfragebogen kann einige Probleme aufwerfen: so stellt sich in Bezug auf manche Antwortformate die Frage der Zumutbarkeit (zu wenige Antwortmöglichkeiten nehmen der Testperson die Möglichkeit des Nuancieren ihrer Antworten, verbale und numerische Antwortmöglichkeiten können unklar oder nicht eindeutig sein etc.), während z.B. in Bewerbungssituationen der Problematik der Verfälschbarkeit eine hohe Bedeutung zukommt. Ebenfalls als problematisch werden sogenannte Antworttendenzen angesehen. Hierbei handelt es sich um bestimmte (vom eigentlich zu erfassenden Merkmal unabhängige) Antwortmuster, die manche Personen z.B. aufgrund des spezifischen Inhalts eines oder mehrerer Items bzw. des ganzen Persönlichkeitsfragebogens oder aufgrund dessen formaler Gestaltung in Bezug auf die Beantwortung zeigen (Arnold, Eysenck & Meili, 1996) und welche die eigentliche Messung verzerren. Weitere Gründe für das Auftreten von Antworttendenzen können eine geringe Testbearbeitungsmotivation (z.B. in sog. „*Low Stakes Assessments*“) oder grundsätzliche Verständnisprobleme bezüglich des Frageninhalts (z.B. aufgrund sprachlicher Schwierigkeiten) sein. Treten solche Antworttendenzen auf, ist eine wahre

Beurteilung bzw. eine faire Erfassung der Persönlichkeit mit Hilfe eines Persönlichkeitsfragebogens kaum möglich, weshalb auf diesem Gebiet intensiv geforscht wird um Antworttendenzen erfassen und kontrollieren zu können. Einer der Lösungsansätze dabei ist der Einsatz der Item Response Theorie (IRT). Der große Vorteil der IRT liegt in der Möglichkeit der eindimensionalen Messung der Antworttendenzen – kann eine solche Messung erzielt werden, wird tatsächlich nur das jeweils interessierende Konstrukt gemessen (in diesem Fall die Antworttendenz) (Khorramdel & von Davier, 2013). Weiters wird in der IRT sowohl die Fähigkeit der Testperson als auch die Schwierigkeit der Items berücksichtigt, was vorteilhaft ist, da beides mit Antworttendenzen interagieren kann (Bolt & Johnson, 2009; De Jong, Steenkamp, Fox & Baumgartner, 2008). Auf weitere Vorteile wird in Kapitel 2.3.2.1 (Item Response Theorie Ansätze) eingegangen. Einer der neuesten IRT Ansätze stammt von Böckenholt (2012), der die beobachteten Daten in binäre Pseudoitems zerlegt und mit diesen IRT Modelle berechnete. Dieser Ansatz wurde von Khorramdel und von Davier (2014) und von Davier und Khorramdel (2013) durch den Vergleich eindimensionaler Antworttendenzfaktoren mit mehrdimensionalen Persönlichkeitsfaktoren erweitert.

Die vorliegende Diplomarbeit soll nun den eben erwähnten Ansatz weiter ausbauen. Während sich Khorramdel und von Davier (2014) und von Davier und Khorramdel (2013) mit der Erfassung von bestimmten formalen Antworttendenzen in einer verbalen Ratingskala beschäftigt haben, liegt das Hauptaugenmerk dieser Diplomarbeit auf der Erfassung und Messung derselben Antworttendenzen in einer optischen Ratingskala, ebenfalls mit Hilfe von mehrdimensionalen IRT Modellen, aufbauend auf dem bereits erwähnten Ansatz von Böckenholt (2012).

Im ersten theoretischen Teil der vorliegenden Diplomarbeit wird in Kapitel 1 eine allgemeine Einleitung gegeben, in Kapitel 2.1 wird ausführlicher auf die Thematik des Persönlichkeitsfragebogens eingegangen. Danach werden in Kapitel 2.2 diverse Antwortformate und die damit verbundenen Vor- und Nachteile abgehandelt. Die generellen und spezifischen Problematiken in Bezug auf Persönlichkeitsfragebögen, die sich teilweise auch aus den bereits angesprochenen Antwortformaten ableiten, werden in Kapitel 2.3 näher ausgeführt. Zuletzt wird in Kapitel 2.3.1 noch ein Überblick über

Antworttendenzen im Allgemeinen und die Tendenz zur Mitte bzw. die Tendenz zu extremen Antworten im Speziellen gegeben.

Kapitel 2.3.3 befasst sich mit der Item Response Theorie und den darauf basierenden Modellen (z.B. Rasch Modell oder 2PL Modell), die unter anderem zur Berechnung der Ergebnisse verwendet wurde. Schließlich wird hier auch noch näher auf den zur Berechnung hauptsächlich verwendeten Ansatz von Böckenholt (2012) und erweiternde Ansätze (Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013) eingegangen.

Im empirischen Teil wird zunächst in Kapitel 3.1 nochmals das Ziel der vorliegenden Diplomarbeit ausgewiesen. Im Anschluss finden sich in Kapitel 3.2 Angaben zur Methode und zur verwendeten Stichprobe; weiters sind hier auch die Hypothesen zu finden, die zur Prüfung der Fragestellungen formuliert wurden. In Kapitel 3.3 erfolgt die Präsentation der Ergebnisse.

Nach einer Diskussion und Kritik an den Ergebnissen und der Methodik, folgt im selben Kapitel 3.4 noch ein Ausblick auf mögliche weiterführende Forschung auf diesem Gebiet.

Die vorliegende Arbeit wird durch eine Zusammenfassung der wesentlichen Ergebnisse und Erkenntnisse in Kapitel 3.5 abgeschlossen.

2 Theoretischer Teil

2.1 Persönlichkeitsfragebogen

Asendorpf und Neyer (2012) beschreiben die Persönlichkeit eines Menschen als die Gesamtheit seiner Persönlichkeitseigenschaften – dies betrifft nicht nur die individuellen Besonderheiten der körperlichen Erscheinung, sondern auch die Regelmäßigkeiten im Verhalten und Erleben. Die Erfassung und Messung dieses für einen Menschen „typischen" Verhaltens und Erlebens erfolgt zumeist mittels einer Selbstauskunft der befragten Person in einem Persönlichkeitsfragebogen. Mit Hilfe dieser Fragebögen kann festgestellt werden, ob das interessierende Merkmal in einer hohen Ausprägung vorliegt, was grundsätzlich für das Vorhandensein dieses Merkmals spricht (Moosbrugger & Kelava, 2012). In den meisten Fällen werden mit Persönlichkeitsfragebögen gleich mehrere Merkmale bzw. Persönlichkeitsdimensionen auf jeweils einer eigenen Skala erfasst (also mehrdimensional), während eindimensionale Verfahren zur Erfassung der Persönlichkeit eher selten sind.

Becker definiert den Persönlichkeitsfragebogen als „*ein standardisiertes Instrument zur Erhebung von Selbst- und Fremdberichtsdaten, aus denen unter Anwendung testtheoretisch begründeter Auswertungsprinzipien Testwerte abgeleitet werden, die als Indikatoren für den individuellen Ausprägungsgrad von Persönlichkeitseigenschaften dienen*" (Becker, 2003, S. 332). Obwohl in dieser Definition auch Fremdberichtsdaten erwähnt werden, erheben die meisten Persönlichkeitsfragebögen ausschließlich Selbstauskünfte. Als Selbstauskunft ist hier die (reflektierte) Selbsteinschätzung der befragten Person hinsichtlich ihrer Eigenschaften und typischen Verhaltens- und Erlebensweisen zu sehen (Kubinger, 2009). Diese „selbstbezogenen" Angaben sind von Erinnerungsvermögen, Aufmerksamkeit, Selbsterkenntnis etc. abhängig und sehr anfällig für unwillkürliche Fehler und Verzerrungen sowie absichtliche Verfälschungen (Bortz & Döring, 2006).

Dass sich der Persönlichkeitsfragebogen trotz bereits existierender, objektiver Verfahren (Objektive Persönlichkeitstests) – im Gegensatz zu den durch die Selbstauskunft subjektiven Persönlichkeitsfragebögen – immer noch größter Beliebtheit

erfreut, dürfte unter anderem auf wirtschaftliche Gründe zurückzuführen sein. Genaueres hierzu findet sich in Kapitel 2.2 zum Thema Antwortformate, in dem auch die jeweiligen Vor- und Nachteile angeführt sind.

Als Erschaffer des Vorbilds für andere Persönlichkeitsfragebögen gilt laut Kubinger (2009) Raymond B. Cattell. Er suchte mit faktorenanalytischen Methoden nach voneinander unabhängigen Eigenschaften, um so die Vielfalt an Persönlichkeiten erklären zu können. Diese faktorenanalytisch gewonnenen Persönlichkeitsdimensionen basierten – anders als in vielen aktuellen Verfahren – noch nicht auf speziellen Theorien oder der Zusammenstellung bestimmter inhaltlicher Kriterien (Perleth, 2003).

Mittlerweile umfasst die Forschung zu faktorenanalytischen Gesamtsystemen der Persönlichkeit eine Vielzahl von Modellen. Das Bekannteste ist das Fünf-Faktoren-Modell (auch Big Five genannt) (McCrae & Costa, 1983). Heutzutage beziehen sich viele Verfahren zur Erfassung von Persönlichkeit auf dieses Modell, so z.B. auch der Persönlichkeitsfragebogen in der Studie zur vorliegenden Diplomarbeit. Im Big Five Persönlichkeitsmodell wird davon ausgegangen, dass sich Menschen hinsichtlich folgender unabhängiger fünf Faktoren ihrer Persönlichkeit wesentlich unterscheiden: Neurotizismus, Extraversion, Offenheit für Erfahrungen, Verträglichkeit und Gewissenhaftigkeit (Kubinger, 2009).

Basierend auf den interessierenden Merkmalen bzw. Persönlichkeitsdimensionen, wie z.B. den soeben erwähnten Big Five, können Items für Persönlichkeitsfragebögen erstellt werden. Rost (2004) definiert Items als „*die Bestandteile eines Tests, die eine Reaktion oder Antwort hervorrufen sollen*“ (S.18). In Persönlichkeitsfragebögen stellen sich Items hauptsächlich als Fragen, Aussagen (Statements) oder Eigenschaftslisten dar, während in Leistungstests auch andere Formate wie z.B. Zahlenfolgen, Analogien oder Durchstreichaufgaben vorgegeben werden (Seiwald, 2003).

Rost (2004) gibt weiters an, dass ein Item die kleinste Beobachtungseinheit in einem Test ist und aus zwei Komponenten besteht:

- Itemstamm – egal in welcher Darstellung (Fragen, Statements, Problemstellungen etc.), der Item- oder Aufgabenstamm stellt das Reizmaterial dar, auf das die befragte Person reagieren soll.
- Antwortformat – mit Hilfe des Antwortformats wird die Antwort der befragten Person auf das Reizmaterial erfasst.

2.2 Antwortformat

Seiwald (2003) gibt an, dass gerade das Antwortformat grundsätzliche Eigenschaften eines psychologisch-diagnostischen Verfahrens bestimmt. Es gibt eine Vielzahl von Möglichkeiten, um die Reaktion einer Testperson auf das ihr präsentierte Reizmaterial festzuhalten. Einen guten Überblick liefern hier z.B. Moosbrugger & Kelava (2012).

Das für die vorliegende Diplomarbeit verwendete Antwortformat findet sich in der Kategorie der Aufgaben mit gebundenem Antwortformat, genauer unter den Beurteilungsaufgaben. In diesen geht es darum, die Ausprägung eines interessierenden Persönlichkeitsmerkmals festzustellen. Hierzu wird der Grad der Zustimmung bzw. Ablehnung in Bezug auf ein vorgegebenes Statement herangezogen (Moosbrugger & Kelava, 2012), welcher entweder mit Hilfe von Analog- oder Ratingskalen ermittelt wird. Bei Analogskalen handelt es sich laut Kubinger (2009) um ein kontinuierliches Antwortformat; die Testperson kann also auf einem Kontinuum zwischen zwei Extremwerten ihre Antwort frei wählen. Moosbrugger & Kelava (2012) sprechen deshalb auch von einer kontinuierlichen Analogskala. Im Gegensatz hierzu kann – wie bereits erwähnt – auch eine diskret gestufte Ratingskala verwendet werden. Da dies in der vorliegenden Diplomarbeit der Fall war, wird im Folgenden genauer auf dieses Antwortformat eingegangen.

2.2.1 Ratingskala

Die Ratingskala wird sehr häufig als Antwortformat in Persönlichkeitsfragebögen verwendet und gleicht technisch gesehen dem bekannten Multiple-Choice-Format insofern, als dass die befragte Person aus verschiedenen Antwortmöglichkeiten jene auswählen soll, die ihr passend erscheinen. Anders als im Leistungsbereich gibt es in Persönlichkeitsfragebögen allerdings keine richtigen und falschen Antworten; die

Antwortmöglichkeiten stellen hier mehr oder weniger fein abgestufte Ausprägungsgrade dar, mit Hilfe derer sich die Testperson in Bezug auf ein interessierendes Persönlichkeitsmerkmal selbst beschreibt. Es ist somit passender, das Antwortformat nicht als Multiple-Choice-Format, sondern als Ratingskala zu bezeichnen (Kubinger, 2009). Asendorpf und Neyer (2012) merken weiters an, dass fünfstufige verbale Skalen der Zustimmung auch Likert-Skalen genannt werden, nach dem Statistiker Rensis Likert.

Rost (2004) beschreibt die Ratingskala als ein Antwortformat mit mehreren (also mehr als zwei) Antwortkategorien, die item-unspezifisch sind (dieselbe Benennung der einzelnen Antwortmöglichkeiten gilt für mehrere oder alle Items in einem Fragebogen) und für die befragte Person eine Rangordnung darstellen (womit man von ordinalen Antwortkategorien sprechen kann). Mit Hilfe dieser Antwortkategorien kann die befragte Person nun mehr oder weniger nuanciert ihre Zustimmung oder Ablehnung zu einem bestimmten Iteminhalt ausdrücken.

Die Ratingskala, wie auch im Allgemeinen das gebundene Antwortformat, hat den Vorteil sehr ökonomisch zu sein. Sowohl Vorgabe als auch Auswertung können am PC erfolgen und haben somit zumeist einen niedrigeren Bearbeitungs- und Auswertungsaufwand als das beim freien Antwortformat der Fall ist, bei welchem keine Antwortmöglichkeiten angeboten werden (die Testperson muss die Antworten also selbst formulieren). Auch die Vorgabe als Gruppenverfahren ist möglich, ebenso wie interpersonelle Vergleiche. Durch standardisierte Vorgaben zur Auswertung liegt beim gebundenen Antwortformat bzw. der Ratingskala auch eine sehr hohe Verrechnungssicherheit vor. Rost (2004) nennt als prinzipiellen Vorteil der Ratingskala, dass das darin verwendete Antwortformat für alle Items im Test oder Fragebogen gilt. Somit ist nicht nur der Konstruktionsaufwand für den/die Testkonstrukteur/in verringert, sondern auch der Bearbeitungsaufwand für die Testperson, da sich diese auf den Antwortmodus einstellen und gleichartige Maßstäbe für alle Items benutzen kann. Bühner (2011) merkt als weitere Vorteile an, dass mit Hilfe einer Ratingskala sehr differenzierte Informationen über die Ausprägung eines Merkmals und die Angleichung der Differenziertheit der Fragen an die Differenzierungsfähigkeit der Testperson und den Untersuchungszweck möglich sind.

Während einige der Nachteile der Ratingskala bei den im Anschluss folgenden verschiedenen Varianten erläutert werden, weisen Bortz und Döring (2006) noch zusätzlich auf die Möglichkeit von systematischen Urteilsfehlern hin, die eventuell die Brauchbarkeit der Urteile einschränken – als die wichtigsten Urteilsfehler nennen sie u.a. den Haloeffekt (der Gesamteindruck oder eine hervorstechende Eigenschaft beeinflussen die Beurteilung unterschiedlicher Eigenschaften), den Milde-Härte-Fehler (Personen werden systematisch zu positiv oder zu negativ eingestuft) und den Primacy-Recency-Effekt (es kommt aufgrund der Position der zu beurteilenden Objekte zu Urteilsverzerrungen).

Wie bei den meisten Antwortformaten gibt es auch bei der Ratingskala verschiedene Varianten, die großen Einfluss auf das Antwortverhalten einer Testperson haben können. Diese werden im Folgenden – auch in Hinblick auf ihre Vor- und Nachteile – erläutert.

- Polarität des Items: uni- vs. bipolar

Von einer unipolaren Skala ist laut Rost (2004) dann die Rede, wenn sie von einem Nullpunkt nur in eine Richtung geht. Moosbrugger & Kelava (2012) definieren den erwähnten Nullpunkt als einen Bezugspunkt für das geringste Ausmaß an Zustimmung oder Ablehnung und den gegenüberliegenden positiven oder negativen Extrempol als Markierung für die stärkste Zustimmung bzw. Ablehnung. Eine Steigung der Häufigkeit, der Intensität oder des Grades der Zustimmung bzw. Ablehnung erfolgt somit wieder nur in eine Richtung. Bortz & Döring (2006) empfehlen die Verwendung einer unipolaren Skala z.B. bei Merkmalen mit einem natürlichen Nullpunkt, wie beispielsweise das Ausmaß der Belästigung durch Lärm.

Lässt sich zu einem Begriff ein passender Gegenbegriff finden, kann in einem Fragebogen auch eine bipolare Antwortskala eingesetzt werden. Rost (2004) gibt an, dass die Antwortkategorien in einer bipolaren Skala von einem negativen Pol über einen Mittelpunkt zu einem positiven Pol gehen. Eine solche Skala hat laut Rost (2004) meistens gleich viele Antwortkategorien auf jeder Seite, das heißt, sie ist symmetrisch. Bortz & Döring (2006) erklären den neutralen Mittelpunkt damit, dass so die Gegensätzlichkeit der verwendeten Begriffe, Zahlen, Symbole, etc. noch stärker betont

werden kann. Weiters geben die Autoren an, dass der Vorteil der bipolaren gegenüber der unipolaren Skala darin liegen kann, dass sich die gegensätzlichen Begriffe gegenseitig definieren und somit die Präzision der Urteile erhöhen können.

Rost (2004) beschreibt, dass die Verwendung einer uni- oder bipolaren Skala sowohl vom Iteminhalt, als auch von dem zu messenden Persönlichkeitsmerkmal abhängt; letzteres kann unipolar (z.B. Ängstlichkeit) oder bipolar (z.B. Extraversion vs. Introversion) definiert sein.

- Anzahl der Skalenstufen

Laut Rost (2004) bestimmt die Anzahl der Skalenstufen darüber, wie differenziert das abgestufte Urteil erfasst wird. Die Auswahl der Anzahl hängt unter anderem davon ab, wie genau bzw. differenziert die befragten Personen eine bestimmte Frage beantworten können und wie genau zwischen ihnen unterschieden werden soll (Bühner, 2011). Während Rost (2004) beschreibt, dass jede Anzahl zwischen 3 und 10 Antwortkategorien möglich ist, gibt Kubinger (2009) eine Anzahl zwischen 3 und 5 Skalenstufen als üblich an. Durch die Nutzung von mehr Antwortkategorien kann zwar die Validität und die Reliabilität eines Tests gesteigert werden, allerdings sind diese Steigerungen ab 7 Antwortkategorien gering (Faulbaum, Prüfer & Rexroth, 2009).

Zu viele Antwortmöglichkeiten können sogar negative Folgen haben: so stellen Bortz & Döring (2006) fest, dass bei sehr vielen Abstufungen (z. B. 100) meist Skalen ausgewählt werden, die durch 5 oder 10 teilbar sind; sie führen dies darauf zurück, dass bei zu feiner Differenzierung der Antwortkategorien das Urteilsvermögen der Testperson überfordert wird. Bühner (2011) gibt außerdem an, dass es zu Problemen kommen kann, wenn verbale Abstufungen zu nahe beieinander liegen oder nicht eindeutig sind. Als Beispiel nennt er den Begriff „gelegentlich“, wenn dieser vom Begriff „manchmal“ gefolgt wird – hier ist unklar, durch welchen dieser Begriffe eine höhere Intensität ausgedrückt werden soll. Die Anzahl von Skalenstufen kann außerdem auch das Auftreten von Antworttendenzen beeinflussen, auf welche im später folgenden Kapitel 2.3.1 noch genauer eingegangen wird.

- Neutrale Antwortkategorie: ja oder nein?

Auch die Entscheidung für die Vorgabe einer geraden oder ungeraden Anzahl von Skalenstufen ist in Bezug auf eine Ratingskala wichtig. Spart man eine mittlere,

neutrale Antwortkategorie aus, nutzt man also eine gerade Anzahl von Antwortalternativen, zwingt man die Testperson, zumindest tendenziell ein Urteil in eine Richtung der Skala abzugeben (Bortz & Döring, 2006). Dies kann allerdings dazu führen, dass Personen mit einer tatsächlich mittleren Ausprägung der zu messenden Persönlichkeitseigenschaft die passendste Antwortmöglichkeit genommen wird.

Wird eine ungerade Anzahl an Skalenstufen verwendet, wird also eine neutrale, mittlere Antwortkategorie eingefügt, kann dieses Problem umgegangen werden, dafür werden andere aufgeworfen: so zeigt sich laut Moosbrugger und Kelava (2012), dass Testpersonen die neutrale Antwortmöglichkeit teilweise auch als Ausweichoption nutzen, weil sie die Frage als unpassend empfinden oder nicht verstehen bzw. die Antwort verweigern oder diese nicht wissen. Bortz und Döring (2006) liefern hierzu das Stichwort „Ambivalenz-Indifferenz-Problem".

Das Anbieten einer „Ich weiß nicht"- oder „Das kann ich nicht beantworten"-Kategorie kann hier hilfreich sein, allerdings müssen diese Antworten in einer statistischen Auswertung als fehlende Werte berechnet werden, was wiederum problematisch sein kann (Bühner, 2011).

Auch die Motivation einer Testperson kann bei einer Testung mit einer neutralen Mittelkategorie zu Problemen führen: Rost (2004) führt an, dass motivierte Personen oft die mittlere Antwortkategorie vermeiden und diese somit seltener auftritt, als eigentlich zu erwarten wäre. Dies kann die Qualität der Messung beeinträchtigen. Moosbrugger und Kelava (2012) sprechen von einem konstruktfremden Antwortverhalten, das sowohl negativen Einfluss auf die Validität eines Tests haben kann, als auch auf die Interpretation von Befunden.

- Benennung der Kategorien

Hat man sich auf eine Anzahl der zu verwendenden Skalenstufen festgelegt, ist auch die Überlegung in Bezug auf deren Bezeichnung wichtig. Hierzu gibt es wieder mehrere Möglichkeiten:

1. numerisch: in einer numerischen Ratingskala wird die Bedeutung der Skalenpunkte mit Hilfe von Zahlen dargestellt. Der Vorteil der numerischen Bezeichnung liegt darin, dass diese eindeutig und knapp ist; allerdings kann es auch vorkommen, dass nicht alle Testpersonen diese abstrakte Darstellungsform verstehen (Bortz & Döring, 2006). Eine Benennung mit Zahlen kann außerdem

zur Annahme verleiten, dass es sich bei der vorliegenden Ratingskala um eine Intervallskala handelt und somit eine besonders präzise Messung vorliegt. Dies ist jedoch nicht immer der Fall, da die Gleichheit der Abstände zwischen den Skalenpunkten nicht automatisch den Abständen in der subjektiven Wahrnehmung der Testperson entsprechen müssen (Rost, 2004; Moosbrugger & Kelava, 2012). Beispiel für eine numerische Ratingskala: -2 / -1 / 0 / 1 / 2.

2. verbal: Die Bezeichnung der Skalenpunkte erfolgt in der verbalen Ratingskala mittels Worten. Diese sprachliche Umschreibung vereinheitlicht die Bedeutung der Antwortstufen laut Rost (2004) intersubjektiv – die Benennung mit Worten ist somit weniger abstrakt und zumeist verständlicher als jene mit Zahlen. Als Antwortmöglichkeiten stehen u.a. Angaben zu Häufigkeit, Wahrscheinlichkeit, Intensität und Bewertung zur Verfügung. Besonders positiv sind hier konkrete Häufigkeitsangaben (z.B. „mindestens zweimal täglich“), da ein verbindlicher Maßstab zum Vergleich zwischen verschiedenen Testpersonen vorliegt und Urteilsfehler und der Einfluss von Antworttendenzen minimiert werden können (Rost, 2004). Die Schwierigkeit der verbalen Ratingskala liegt vor allem darin, Worte zu finden, die tatsächlich gleichwertige Abstände zwischen den Ausprägungen eines Merkmals darstellen und vor allem, dass diese von unterschiedlichen Personen auch gleich verstanden werden. Wie schon im numerischen Antwortformat angeführt, kann bei der Benennung einer Ratingskala mit Worten die automatische Annahme des Vorliegens von Intervallskalenniveau und somit einer besonders genauen Messung zu Problemen führen. Beispiel für eine verbale Ratingskala: starke Ablehnung / Ablehnung / Neutral / Zustimmung / starke Zustimmung.

3. symbolisch: In der optischen Ratingskala werden Symbole zur Bezeichnung der Skalenpunkte verwendet. Diese können vielfältige Formen annehmen, wie z.B. Kreise, Smileys oder Plus-/Minus-Zeichen. Die Bedeutung von sprachlichen Umschreibungen kann subjektiven Schwankungen unterliegen (Rost, 2004), während die Bedeutung von Symbolen zumeist eindeutig und auf einen Blick erfassbar ist (Bortz & Döring, 2006). Deshalb werden diese gerne bei Fragebögen für Kinder eingesetzt und wirken auch für Erwachsene auflockernd.

Mit Hilfe von Symbolen kann auch der Eindruck von übertriebener mathematischer Exaktheit vermieden werden (Rost, 2004), was bei numerischen Bezeichnungen ein Problem darstellen kann. Auch beim Einsatz von Symbolen als Benennung der Skalenpunkte ist auf Äquidistanz zu achten. Wie schon beim numerischen und verbalen Antwortformat erwähnt, ist auch im optischen Antwortformat eine automatische Annahme des Intervallskalenniveaus der Ratingskala problematisch. Beispiel für eine symbolische Ratingskala: -- / - / o / + / ++.

Durch die Kombination der verschiedenen Bezeichnungen (z.B. numerisch und verbal) hofft man, die Vorteile beider Benennungsvarianten nutzen zu können. Eine weitere Möglichkeit, die Skalenpunkte zu bezeichnen, stellt eine Verankerung durch Fallbeispiele dar.

Da in der vorliegenden Diplomarbeit Smileys als optisches Antwortformat verwendet wurden und diese somit für diese Diplomarbeit von besonderer Bedeutung sind, wird im Folgenden etwas genauer darauf eingegangen.

2.2.1.1 Smileys als optisches Antwortformat

Ein besonders häufig eingesetztes optisches Antwortformat sind Smileys. Der Einsatz dieser als symbolische Antwortmöglichkeiten in Ratingskalen geht auf Kunin (1955) zurück, der sie im Rahmen einer Befragung zum Thema Arbeitszufriedenheit in einem Automobilkonzern in den USA entwickelte. Er ging davon aus, dass es vor allem Menschen mit niedrigen verbalen Fähigkeiten schwer fällt, ihre Gefühle und Einstellungen in den Worten eines anderen Menschen (im Falle einer Testung des Testkonstrukteurs/der Testkonstrukteurin) auszudrücken. Um diese sprachlichen Schwierigkeiten zu umgehen, entwickelte er eine Reihe von stilisierten Gesichtern, bei denen ausschließlich die Mundpartie variierte – somit wurde den Testpersonen ermöglicht jenes Smiley auszuwählen, dessen Ausdruck am ehesten den eigenen Gefühlen und Einstellungen entsprach (Kunin, 1998).

Gewisse Gesichtsausdrücke werden kulturübergreifend mit gewissen Gefühlen in Verbindung gebracht, wie zum Beispiel Glück oder Traurigkeit (Ekman, 1971). Dies

kann insofern auf Smileys umgelegt werden, als dass ein stilisiertes Gesicht mit einer stark nach oben gebogenen Mundpartie in den meisten Fällen als glücklich interpretiert wird, während ein solches mit einer stark nach unten gebogenen Mundpartie als traurig/unglücklich anzusehen ist. Dieses intuitive Verständnis verschiedener Gesichtsausdrücke und deren Intensität ermöglicht den vielfältigen Einsatz von Smileys als optisches Antwortformat in Fragebögen.

Neben der Verwendung zur Erfassung der Zufriedenheit in verschiedenen Bereichen (z.B. Arbeit, Leben, Wohlbefinden, Service, etc.), werden Smileys als Antwortmöglichkeiten auch im diagnostischen Bereich eingesetzt, oft auch in Kombination mit einem anderen Antwortformat (z.B. verbal). Ein Beispiel hierfür ist das Inventar zur Persönlichkeitsdiagnostik in Situationen (Schaarschmidt & Fischer, 1999) – nach der Beschreibung einer Situation soll die Testperson auf einer vierstufigen, verbalen Antwortskala (stimmt genau – stimmt eher schon – stimmt eher nicht – stimmt gar nicht) die für sie treffendste Antwort auswählen. Im Anschluss daran findet jeweils eine Beurteilung der Zufriedenheit der eigenen Reaktionen statt, wofür fünf Smileys zur Beantwortung zur Verfügung stehen. Ratingskalen mit Smileys als Antwortformat eignen sich auch für den Einsatz bei Personen, die kognitiv beeinträchtigt sind oder – wie bereits erwähnt – bei bestehenden Sprachbarrieren. Durch die intuitive Verständlichkeit können sie weiters auch Kindern vorgegeben werden – im diagnostischen Bereich geschieht dies z.B. mittels der „Smiley-Analog-Skala“ (Pothmann, 1996), mit deren Hilfe die Schmerzintensität in der Schmerztherapie erfasst wird.

Wie in der Anwendung von verbalen, numerischen und anderen optischen Antwortformaten ist auch beim Einsatz von Smileys darauf zu achten, dass die Abstände zwischen den vorgegebenen stilisierten Gesichtern als gleichwertig wahrgenommen werden. Jäger (2004) entwickelte hierzu eine Antwortskala mit fünf Smileys, die sowohl äquidistant als auch eindimensional ist. Weiters ist darauf zu achten, dass in Bezug auf die Bedeutung von Smileys auch intra- und interindividuelle Unterschiede bestehen können – Elfering und Grebner (2011) stellten hierzu in einer Studie zum Thema Arbeitszufriedenheit fest, dass sowohl der Affektzustand und das Persönlichkeitsmerkmal Neurotizismus einer Testperson als auch der Umfang der

dargebotenen Smileys einen Einfluss auf die Antworten von Testpersonen haben können.

2.3 Problematik bei Persönlichkeitsfragebögen

Einige der bei den Varianten der Ratingskala im Allgemeinen erwähnten Probleme treffen auch auf Persönlichkeitsfragebögen im Speziellen zu. So können z.B. das automatische Ausgehen vom Vorliegen eines Intervallskalenniveaus oder der Einsatz einer neutralen oder „Ich weiß nicht"-Antwortkategorie zu messtheoretischen Problemen führen; weiters kann auch die Verwendung verschiedener sprachlicher Ausdrücke und Formulierungen problematisch sein, vor allem, wenn diese für unterschiedliche Personen oder in unterschiedlichen Kulturkreisen nicht dasselbe bedeuten oder verschieden aufgefasst werden können. Becker (2003) weist außerdem darauf hin, dass es durch die Verwendung von Fremd- oder selten gebrauchten Wörtern, komplizierten Sätzen und (doppelten) Verneinungen zu Verständnisproblemen kommen kann.

Hambros (2003) führt zum Thema Zumutbarkeit von Persönlichkeitsfragebögen an, dass Fragen zum persönlichen Intimbereich kritisch sein können, ebenso wie der Einsatz des dichotomen Antwortformats (z.B. ja/nein) – dieses lässt der Testperson keine Möglichkeit, ihre Antworten entsprechend zu nuancieren und kann im schlechtesten Fall Reaktanz auslösen. Auch die Durchschaubarkeit eines Persönlichkeitsfragebogens kann hinsichtlich dessen Zumutbarkeit problematisch sein: befindet sich die Testperson in einer Auswahlsituation und die Messintention des Persönlichkeitsfragebogens ist zu leicht durchschaubar, kann es sein, dass sich die Testperson nicht ernst genommen fühlt und verärgert reagiert, was einen negativen Effekt auf die Beantwortung der Fragen haben kann.

Auch die Testperson selbst kann eine Fehlerquelle bei der Diagnostik mittels Persönlichkeitsfragebögen sein: so beschreibt Becker (2003) Validitätseinbußen durch das Nichtverfügen über die erforderlichen intellektuellen oder sprachlichen Voraussetzungen, durch ein verzerrtes Selbstbild, durch eine starke psychische Beeinträchtigung zum Zeitpunkt der Untersuchung, durch nicht ausreichend vorhandene Motivation, durch das Zeigen von Antworttendenzen und durch das absichtliche

Verfälschen ihrer Antworten. Dieser letzter Punkt hängt auch stark mit der Problematik der Untersuchungssituation zusammen – handelt es sich um eine für die Testperson persönlich wichtige Situation (wie Auswahl- und Bewerbungsverfahren), kann sie eventuell ihr Antwortverhalten z.B. in eine sozial erwünschte Richtung lenken. Die Problematik der sozialen Erwünschtheit gehört zum Überbegriff der Antworttendenzen, auf die im Folgenden näher eingegangen wird.

2.3.1 Antworttendenzen

Der Begriff Antworttendenzen beschreibt die Neigung einer Testperson zu systematischen Antwortmustern in Ratingskalen, die unabhängig vom Iteminhalt, also des zu messenden Persönlichkeitsmerkmals, sind (Paulhus, 1991; Rost, 2004). Vielfach wird davon ausgegangen, dass die Beantwortung der Items in einem Persönlichkeitsfragebogen aufgrund deren Inhalts erfolgt und somit der resultierende Wert eine Ausprägung des interessierenden Persönlichkeitsmerkmals darstellt; da Antworttendenzen jedoch inhaltsunabhängig erfolgen, darf diese Grundannahme bei deren Vorliegen nicht getätigt werden (Harzing, 2006), da sie den wahren Wert einer Testperson verfälschen.

Schon Nunnally (1967, zit. nach Khorramdel & von Davier, 2014) ging davon aus, dass Antworttendenzen bei ähnlichen Items in einem kurzen Zeitraum (wie z.B. während der Vorgabe eines Persönlichkeitsfragebogens) relativ stabile Charaktereigenschaften darstellen. Auch Javaras & Ripley (2007), deren Modell nicht nur die Erfassung der Einstellungen der Testpersonen erlaubt, sondern auch die von Antworttendenzen, gehen davon aus, dass sich individuelle Antwortmuster über einen kurzen Zeitraum nicht ändern. Weijters, Geuens & Schillewaert (2010a) untersuchten diese Stabilität auch in einer Längsschnittstudie – sie gaben im Abstand von einem Jahr denselben Testpersonen zwei Online-Persönlichkeitsfragebögen vor und stellten fest, dass auch über diesen längeren Zeitraum die Antworttendenzen relativ stabil blieben. Antworttendenzen können in zwei grundlegende Formen unterteilt werden:

2.3.1.1 Formale Antworttendenzen („response styles“)

Darunter sind all jene Tendenzen zu verstehen, die sich aufgrund der formalen Gestaltung des Antwortformats eines Persönlichkeitsfragebogens ergeben (Arnold, Eysenck & Meili, 1996). In der Literatur lassen sich sehr viele unterschiedliche formale Antworttendenzen finden. Im Folgenden werden jene zwei näher beschrieben, die für die vorliegende Diplomarbeit besonders wichtig sind:

- Tendenz zur Mitte

Bei der Tendenz zur Mitte bevorzugt die Testperson (bewusst oder unbewusst) unabhängig vom Iteminhalt die mittlere (neutrale) Antwortkategorie. Dies kann aufgrund von Unsicherheit (die Testperson möchte eine definitive Entscheidung z.B. wegen unzureichendem Wissen vermeiden), Gleichgültigkeit (die Testperson hat kein Interesse am Thema), Verweigerung (die Testperson möchte ihre wirklichen Einstellungen verbergen oder hält ein Item bzw. das Thema für unpassend) oder Verständnisproblemen (inhaltlich oder in Bezug auf die Formulierung) geschehen (Baumgartner & Steenkamp, 2001; Rost 2004). Moosbrugger & Kelava (2012) weisen weiters darauf hin, dass manche Testpersonen davon ausgehen, dass das Ankreuzen der mittleren Antwortkategorie eher dem Antwortverhalten von „normalen“ oder „typische“ Personen entspricht, und diese deshalb wählen, ohne den genauen Inhalt des Items zu beachten. Ein Beispiel für die Tendenz zur Mitte zeigt Abbildung 1.

		X		

Abbildung 1: Beispiel für Tendenz zur Mitte.

- Tendenz zu extremen Urteilen

Die Testperson wählt hier in einer Ratingskala bzw. einem Persönlichkeitsfragebogen mit mehr als drei Antwortmöglichkeiten jene aus, die sich an den äußeren Enden befinden. Dies geschieht – wie bei der Tendenz zur Mitte – wieder unabhängig vom Iteminhalt und ist somit kein Ausdruck von extremer Zustimmung oder Ablehnung, sondern stellt eine Antworttendenz dar. Baumgartner & Steenkamp (2001) führen an, dass dies häufiger bei Items mit hoher Bedeutung für die Testperson vorkommt und mit höherer Ängstlichkeit und eventuell abweichendem Verhalten assoziiert werden kann. Die Tendenz zu extremen Urteilen wird von den Autoren außerdem noch als Charakteristik von Personen mit schlecht entwickelten Schemata und wenig

ausdifferenzierten kognitiven Strukturen beschrieben und kann als Ausdruck von Unnachgiebigkeit, Rechthaberei und Intoleranz gegenüber Vieldeutigkeit angesehen werden. Ein Beispiel für die Tendenz zu extremen Urteilen zeigt Abbildung 2.

Abbildung 2: Beispiel für Tendenz zu extremen Urteilen

Weitere wichtige formale Antworttendenzen sind die Akquieszenz oder Ja-sage-Tendenz (die Testperson stimmt unabhängig vom Iteminhalt den Fragen bzw. Aussagen spontan und unreflektiert zu), Positions- und Reihenfolgeeffekte (die Position und/oder die Reihenfolge der Items kann einen Einfluss auf das Antwortverhalten einer Testperson haben), das Raten (hier wird die richtige Lösung in einem Multiple Choice-Test – wie z.B. einem Leistungs- oder Intelligenztest – geraten) und die „zufällige" Beantwortung (die Beantwortung erfolgt systematisch oder willkürlich, z.B. in „Antwort-Rhythmen" oder blockweise) (Seiwald, 2003).

2.3.1.2 Inhaltliche Antworttendenzen („response sets")

Zu den inhaltlichen Antworttendenzen zählen all jene, bei denen verfälschende Antworten aufgrund des spezifischen Inhalts einzelner Items oder des gesamten Fragebogens zu finden sind (Arnold, Eysenck & Meili, 1996). Sie sind somit inhaltsrelevant und dahingehend von den formalen Antworttendenzen (die ja unabhängig vom Inhalt bzw. dem zu messenden Konstrukt sind) abzugrenzen. Dementsprechend ist die Bezeichnung als „Antworttendenz" hier eher kritisch zu sehen, da der Unterschied zu den inhaltsunabhängigen bzw. konstruktunabhängigen Antworttendenzen nicht genug betont wird. Im Englischen wird z.B. für inhaltliche Antworttendenzen der Begriff *response sets* gebraucht (Rost, 2004), die formalen Antworttendenzen aber noch einmal deutlich mit dem Begriff *response styles* unterschieden. Die folgend beschriebenen inhaltlichen Antworttendenzen sind somit nicht als Antworttendenzen im Sinne von *response styles* zu sehen.

- Soziale Erwünschtheit

Dieser Begriff beschreibt die Tendenz einer Testperson, sich durch bewusstes Verfälschen der Antworten in einem Persönlichkeitsfragebogen in einer sozial

erwünschten Weise darzustellen; sie beschreibt sich also vielleicht vorteilhafter, als sie sich selbst sieht. Im Gegensatz zur Selbsttäuschung (*Self-deceptive Enhancement*) handelt es sich hierbei um eine Fremdtäuschung (*Impression Management*) (Moosbrugger & Kelava, 2012). Dies geschieht vor allem in Situationen, in der die Folgen der Testung für die Testperson selbst besonders wichtig sind, wie z.B. im Rahmen eines Bewerbungsverfahrens (Seiwald, 2003). Eine bewusste Verfälschung in Richtung sozialer Erwünschtheit ist vor allem bei Persönlichkeitsfragebögen mit hoher Augenscheinvalidität möglich, da hier die Messintention leicht durchschaubar ist (Kersting, 2003; Kubinger, 2003). Als Gegenmaßnahmen wurden hier u.a. sogenannte Kontroll- oder Lügenskalen (Moosbrugger & Kelava, 2012) und Objektive Persönlichkeitstests (Kubinger, 2003) eingesetzt. Auch die Zusicherung der Anonymität und die Aufklärung über den Untersuchungszweck gelten als Maßnahmen zur Verringerung des Effekts der sozialen Erwünschtheit (Moosbrugger & Kelava, 2012), ebenso wie die Instruktion zum ehrlichen und spontanen Antworten (Seiwald, 2003).

Als weitere Formen der inhaltlichen Antworttendenzen werden Simulation und Dissimulation, defensive Einstellung, Lügen und Abweichreaktionen angesehen (Arnold, Eysenck & Meili, 1998).

Wie bereits erwähnt stellen Antworttendenzen insofern ein großes Problem dar, als dass sie sowohl den wahren Wert einer Testperson in Bezug auf ein interessierendes Persönlichkeitsmerkmal, als auch die Korrelation zwischen den Skalen verfälschen, also diese erhöhen oder absenken, können (Baumgartner & Steenkamp, 2001). Dadurch wird die Validität eines Tests beeinträchtigt; so kann es beispielsweise dazu kommen, dass unterschiedliche Ergebnisse für verschiedene Untersuchungsgruppen als Unterschiede in den zu messenden Persönlichkeitsmerkmalen gedeutet werden, obwohl die Ergebnisse in Wirklichkeit durch das Vorhandensein von Antworttendenzen verzerrt sind. Dies konnte u. a. im Bereich der nationalen und internationalen Marktforschung (De Jong, Steenkamp, Fox & Baumgartner, 2008), für die Art der Befragung (Weijters, Schillewaert & Geuens, 2008) und Bewertungen von Lehrpersonal durch Studierende (Dolnicar & Grun, 2009) festgestellt werden.

Auch die Eindimensionalität einer Messung kann durch das Vorliegen von Antworttendenzen negativ beeinflusst werden – wird z.B. eine mittlere Antwortkategorie nicht dazu genutzt, die wahre Merkmalsausprägung auf einer Skala darzustellen, sondern um die Antwort zu verweigern oder damit auszudrücken, dass man das Item für unpassend hält, misst der Test bzw. der Persönlichkeitsfragebogen nicht mehr eindimensional, also nur das intendierte Merkmal, sondern zweidimensional (also das intendierte Merkmal und die Antworttendenz) (Rost, 2004). Chun, Campbell & Yoo (1974) bestätigten dies auch für die Tendenz zu extremen Urteilen.

Die Beeinflussung der Eindimensionalität und Validität durch Antworttendenzen kann dazu führen, dass Gruppenunterschiede nicht mehr vergleichbar sind (Chun, Campbell & Yoo, 1974; Harzing, 2006; Morren, Gelissen & Vermunt, 2012). Dies stellt vor allem in *large-scale assessments* ein Problem dar, wie Bolt & Newton (2011) national für Nordamerika und Buckley (2009) international in Bezug auf die PISA-Studie von 2006 bestätigen konnten. Ebenfalls betroffen sind interkulturelle Untersuchungen (Chun, Campbell & Yoo, 1974; Morren, Gelissen & Vermunt, 2012).

Die Unvergleichbarkeit der Gruppenunterschiede basiert u.a. auch darauf, dass Antworttendenzen abhängig von Variablen wie z.B. Alter, Einkommen, Bildungsgrad, verschiedene Persönlichkeitseigenschaften, den vier Kulturdimensionen von Hofstede (2001), Geschlecht und Kultur vorkommen können. So belegen beispielsweise einige Studien, dass Frauen eher eine Tendenz zu extremen Urteilen zeigen als Männer (Austin, Deary & Egan, 2006; De Jong, Steenkamp, Fox & Baumgartner, 2008; Weijters, Geuens & Schillewaert, 2010b; Khorramdel & von Davier, 2014). Interkulturell konnten ebenfalls Unterschiede in den Präferenzen zu verschiedenen Antworttendenzen gefunden werden: während lateinamerikanische (Hui & Triandis, 1989; Khorramdel & von Davier, 2014) und afro-amerikanische Testpersonen (Bachman & O'Malley, 1984) vermehrt eine Tendenz zu extremen Urteilen im Vergleich mit nordamerikanischen und Testpersonen europäischer Herkunft zeigten, bevorzugten chinesische und japanische Studierende mit einer ähnlichen Vergleichsgruppe eindeutig die Tendenz zur Mitte (Chen, Lee & Stevenson, 1995; Harzing, 2006; Hamamura, Heine & Paulhus, 2008; Khorramdel & von Davier, 2014). In Europa wurde vor allem im Mittelmeerraum verglichen mit Nordwesteuropa eine

Tendenz zu extremen Urteilen festgestellt (van Herk, Poortinga & Verhallen, 2004). Weiters wurde die Tendenz zu extremen Urteilen vor allem in Griechenland, Italien und in spanisch sprechenden Ländern aufgezeigt (Harzing, 2006). Von Studierenden wird international die Tendenz zu extremen Urteilen eher vermieden (Dolnicar & Grun, 2009).

Studien ergaben außerdem, dass die Persönlichkeitseigenschaft Extraversion eher mit einer Tendenz zu extremen Urteilen verbunden ist (Harzing, 2006); Austin, Deary & Egan (2006) bestätigen dies für die Eigenschaften Extraversion und Gewissenhaftigkeit. Tests bzw. Fragebögen in der Muttersprache werden ebenfalls eher mit einer Tendenz zu extremen Urteilen beantwortet; werden diese auf Englisch vorgegeben (falls dies nicht die Muttersprache ist), scheint eher eine Tendenz zur Mitte vorzuherrschen (Harzing, 2006).

2.3.2 Erfassung und Kontrolle von Antworttendenzen

Die einfachste Möglichkeit, Antworttendenzen zu vermeiden, ist die Vorgabe eines dichotomen Antwortformats (z.B. ich stimme zu/ich stimme nicht zu). Diese Vorgehensweise hat jedoch mehrere Nachteile: so können einerseits wichtige Informationen zur Intensität der Einstellung der Testperson verloren gehen (Böckenholt, 2012); andererseits kann – wie bereits erwähnt – durch die begrenzte Möglichkeit des Nuancierens Reaktanz entstehen. Manche Testpersonen antworten dann willkürlich bzw. untypisch und eventuell sogar entgegen ihrer wahren Eigenschaften (Kubinger, 2009); weiters kann es dazu kommen, dass die Testpersonen mit höheren „Faking"-Tendenzen bzw. einem höheren *Impression Management* reagieren, also eher sozial erwünscht antworten (Khorramdel & Kubinger, 2006).

Da normalerweise in psychologischen Testungen mehrkategorielle Antwortformate vorgezogen werden, wurden in Bezug auf diese einige Ansätze entwickelt, um Antworttendenzen messen und kontrollieren zu können; diese bergen allerdings spezifische Probleme. Eher simplere Methoden nutzen z.B. Häufigkeiten der gewählten Antwortkategorien (Bachman & O'Malley, 1984; Baumgartner & Steenkamp, 2001; van Herk, Poortinga & Verhallen, 2004; Harzing, 2006; Buckley, 2009; Reynolds & Smith, 2010) oder die Standardabweichung der Summe aller Items einer Testperson

(Baumgartner & Steenkamp, 2001). Diese Ansätze sind zwar einfach in der Anwendung, allerdings werden heterogene Items benötigt, die möglichst nicht miteinander korrelieren – solche Items sind nicht immer leicht zu finden und können die Validität der Messung negativ beeinflussen (Bolt & Newton, 2011). Ähnliches gilt auch für die Methode umgepolte Items in den Fragebogen mitaufzunehmen und die doppelten Zustimmungen auszuzählen (Hox, de Leeuw & Kreft, 1991; Johnson, Kulesa, Cho & Shavitt, 2005); hier ist es oft schwierig umgepolte Items passend zu formulieren. Andere Ansätze, wie z.B. MMTMs (*multi-trait-multi-method models*) (Saris & Aalberts, 2003; Saris, Satorra & Coenders, 2004) und die Spezifikation eines Methodenfaktors in einer konfirmatorischen Faktorenanalyse (Billiet & McClendon, 2000; Welkenhuysen-Gybels, Billiet & Cambré, 2003) sind rechnerisch bereits komplexer, erfassen jedoch weder die Tendenz zur Mitte, noch die Tendenz zu extremen Urteilen. Bei Ansätzen, die zur Erfassung und Kontrolle der Antworttendenzen latente Klassenanalysen mittels Regression (Moors, 2009) oder konfirmatorischer Faktorenanalyse (Moors, 2003; Kieruj & Moors, 2010; Moors, 2012; Kieruj & Moors, 2012) nutzen, wird jeweils eine spezifische Software benötigt, um die komplexen Rechenvorgänge vorzunehmen. Der Einsatz der RIRS (*representative indicators for response styles*) (Baumgartner & Steenkamp, 2001) und der RIRMACS Methode (*representative indicators response styles means and covariance structure*) (Weijters, Geuens & Schillewaert, 2008) erfordert wiederum das Einfügen zusätzlicher Items in den Fragebogen.

Einige der genannten Ansätze zur Erfassung und Kontrolle von Antworttendenzen beziehen außerdem den Einfluss des zu messenden Merkmals nicht mit ein. So kann das Antwortmuster einer Testperson auf das Vorhandensein von Antworttendenzen hindeuten; allerdings könnte dieses Muster auch durch die tatsächliche Merkmalsausprägung entstanden sein. Wie bereits erwähnt, beeinflusst dies die Eindimensionalität der Testung, da neben der Antworttendenz eventuell noch ein zweites Konstrukt, nämlich das eigentlich mit dem Fragebogen zu erfassende Merkmal, gemessen wird. Um dieses Problem zu umgehen, kann die Item Response Theorie (IRT) eingesetzt werden.

2.3.2.1 Item Response Theorie Ansätze

Der große Vorteil der IRT ist, dass hier für eine Gruppe von Items deren Eindimensionalität überprüft werden kann und in diesem Fall sichergestellt ist, dass tatsächlich nur eine Dimension, also das interessierende Merkmal oder eine bestimmte Antworttendenz, gemessen wird. Antworttendenzen können mit den Charakteristika von Items und den Merkmalsausprägungen einer Testperson interagieren (Bolt & Johnson, 2009; De Jong, Steenkamp, Fox & Baumgartner, 2008). Hier bietet die IRT einen zusätzlichen Vorteil, da in ihren Modellen die Item- und Personenparameter getrennt erfasst werden.

Johnson und Bolt (2010) gehen in ihrem mehrdimensionalen IRT Modell (*Factor-Analytic Multinomial Logit Item Response Model*) von einem gleichzeitigen Einfluss der zugrundeliegenden Merkmalsausprägungen einer Testperson und der Antworttendenzen bei der Wahl der Antwortkategorie aus. Bolt und Johnson (2009) sowie Bolt und Newton (2011) zeigten in ihren Studien einen erweiterten, mehrdimensionalen Ansatz von Bocks *Nominal Response Model* (1972) auf, der ebenfalls diesen Einfluss berücksichtigt. Dieses Modell geht dabei von einem kontinuierlichen Faktor in der Erklärung von Antworttendenzen aus, was es von anderen IRT-Modellen unterscheidet. Die Verwendung eines kontinuierlichen Faktors ermöglicht dabei nicht nur die Untersuchung und Kontrolle der Effekte von Antworttendenzen bezüglich des Gesamttestergebnisses einer Testperson, sondern auch bezüglich DIF- (*differential item functioning*) Analysen. Mit der Hilfe von DIF-Analysen kann untersucht werden, ob sich ein Item im Hinblick auf unterschiedliche Subgruppen einer Population oder den Kontext unterschiedlich auswirkt (Böckenholt, 2012).

Wiederum andere IRT Modelle (*mixed Rasch models*) ermöglichen eine Identifizierung latenter Subgruppen von Personen bzw. Klassen, die unterschiedliche Präferenzen und Nutzung von Antworttendenzen zeigen. Zusätzlich wurden hierfür noch der Einfluss dieser Subgruppen auf die Eindimensionalität der Messung (Rost, Carstensen & von Davier, 1999) und der Beitrag von Antworttendenzen zu Korrelationen auf dem Skalenlevel (Austin, Deary & Egan, 2006) untersucht.

Ein weiterer IRT Ansatz zur Erfassung und Kontrolle von Antworttendenzen (in diesem Fall die Tendenz zu extremen Antworten) stammt von De Jong, Steenkamp, Fox & Baumgartner (2008). Die Autoren gehen in ihrer Studie davon aus, dass Items sich in Bezug auf das Provozieren von Antworttendenzen ebenso unterscheiden wie Personen in ihrer Tendenz extreme Antworten zu wählen; ihr Ansatz berücksichtigt diese Effekte und erlaubt es somit verschiedenen Items unterschiedlich nützlich in der Erfassung von Antworttendenzen zu sein und bezieht auch die Unterschiede dieser Nützlichkeit für verschiedene Gruppen mit ein. Gleichzeitig mit der Tendenz zu extremen Urteilen werden auch individuelle und Gruppeneinflussfaktoren auf diese Antworttendenz gemessen.

2.3.2.2 Neuere Item Response Theorie Ansätze

Böckenholt (2012) präsentiert in seinem Artikel zur Messung multipler Antwortprozesse in Beurteilungs- und Auswahlsituationen einen der neuesten IRT Ansätze zur Erfassung und Kontrolle von Antworttendenzen. Darin wird aufgezeigt, wie mit Hilfe von ein- und mehrdimensionalen IRT (MIRT) Modellen Antworttendenzen und merkmalsbezogene, aufeinander folgende Antwortprozesse unterschieden werden können.

Böckenholt (2012) geht davon aus, dass der Antwortfindung multiple, ineinander verschachtelte Subprozesse zugrunde liegen. Um diese messen zu können, werden die nominalen oder ordinalen Antworten in binäre Pseudoitems zerlegt, mit Hilfe derer jeder mögliche Antwortsubprozess dargestellt wird. Dies geschieht auf der Basis eines multinominalen Entscheidungsbaums (siehe Abbildung 3). In einer fünfstufigen Antwortskala entscheidet sich die Testperson zum Beispiel zuerst für oder gegen die neutrale Mitte; ist Letzteres der Fall, erfolgt danach eine Entscheidung bezüglich der Richtung der Antwort (zustimmend oder ablehnend) und in einem dritten Schritt in Bezug auf die Intensität (moderat oder stark).

Durch die Zerlegung in binäre Pseudoitems können danach *Simple Structure* MIRT Modelle angewandt werden. In diesen wird davon ausgegangen, dass jedes Item auf nur einem Faktor lädt, und dass Itemgruppen, die auf demselben Faktor laden, diesen eindimensional erfassen. Dieser Prozess wird durch die Differenz zwischen der

Schwierigkeit der Items und der Fähigkeit der Testperson dargestellt. Somit werden für jede Phase des Antwortprozesses individuelle Differenzen im Antwortverhalten durch ein Set von personenspezifischen Parametern beschrieben (Khorramdel & von Davier, 2014).

Böckenholt (2012) verwendet für die Modellierung der Entscheidungen einer Testperson (Mitte vs. Nicht-Mitte, Zustimmung vs. Ablehnung, moderat vs. stark) das *Graded Response* Modell von Samejima (1969, 1997). In diesem Modell werden die gegebenen Antworten als ein Ergebnis einer zugrunde liegenden, normalverteilten latenten Variable gesehen, die auf einer diskreten Antwortskala abgebildet wird. Die Antwortkategorien sind durch Schwellenwerte voneinander abgegrenzt; liegt der latente Wert zwischen dem oberen und unteren Schwellenwert einer Antwortkategorie, so wird diese ausgewählt. Die Wahrscheinlichkeit, dass eine zufällig bestimmte Person *i* die Antwortkategorie *k* ($k = 1, \ldots, K$) des Items *j* mit dem Schwellenwert τ_k wählt, kann dann durch die Differenz zwischen zwei normal kumulativen Verteilungsfunktionen wie folgt dargestellt werden:

$$\Pr(y_{ij} = k) = \Phi(\tau_{k+1} + \theta_i - \gamma_j) - \Phi(\tau_k + \theta_i - \gamma_j)$$

y_j beschreibt dabei die Lage des Items *j* auf dem latenten Kontinuum, definiert durch den Personenparameter θ (Böckenholt, 2012).

Ein *Two-Process Graded Response* Modell dient Böckenholt (2012) zur Untersuchung der Tendenz zur Mitte. Ausgehend von einer fünfstufigen Antwortskala (A, B, C, D, E) wird hier der Antwortprozess in ein binäres Pseudoitem (Entscheidung für oder gegen die Mitte) und ein ordinales vierkategorielles Pseudoitem mit den restlichen Antwortkategorien aufgeteilt. Letzteres Pseudoitem beinhaltet somit sowohl die Richtung (Zustimmung oder Ablehnung) als auch die Intensität (moderat oder stark) der Einstellung der Testperson. Die Wahrscheinlichkeit der Auswahl der neutralen Mitte (mit *C* betitelt) kann wie folgt beschrieben werden:

$$\Pr(y_{ij} = C) = \Phi(\theta_i^{(I)} - \gamma_j^{(I)})$$

Die Wahrscheinlichkeit, eine der verbleibenden Antwortkategorien auszuwählen, stellt sich daraufhin folgendermaßen dar:

$$\Pr(y_{ij} = A) = [1 - \Phi(\theta_i^{(I)} - \gamma_j^{(I)})]\Phi(\tau_B + \theta_i^{(II)} - \gamma_j^{(II)})$$

$$\Pr(y_{ij} = B) = [1 - \Phi(\theta_i^{(I)} - \gamma_j^{(I)})][\Phi(\tau_D + \theta_i^{(II)} - \gamma_j^{(II)}) - \Phi(\tau_B + \theta_i^{(II)} - \gamma_j^{(II)})]$$

$$\Pr(y_{ij} = D) = [1 - \Phi(\theta_i^{(I)} - \gamma_j^{(I)})][\Phi(\tau_E + \theta_i^{(II)} - \gamma_j^{(II)}) - \Phi(\tau_D + \theta_i^{(II)} - \gamma_j^{(II)})]$$

$$\Pr(y_{ij} = E) = [1 - \Phi(\theta_i^{(I)} - \gamma_j^{(I)})][1 - \Phi(\tau_E + \theta_i^{(II)} - \gamma_j^{(II)})]$$

$\gamma_j^{(I)}$ und $\gamma_j^{(II)}$ beschreiben dabei die Itemparameter der beiden Subprozesse; $\theta_i^{(I)}$ und $\theta_i^{(II)}$ sind die jeweiligen Personenparameter.

Dieser IRT-Ansatz hat mehrere Vorteile: durch das Aufteilen in Pseudoitems wird nicht nur ein besseres Verständnis, wie die Antworten auf Items in einer Ratingskala entstehen, gefördert; durch die Bildung von Pseudoitems ergibt sich auch eine Datenstruktur, die leicht zu handhaben ist (vor allem im Vergleich mit anderen, komplexen Ansätzen, die die Originaldaten zur Erfassung von Antworttendenzen verwenden). Weiters sind die Schätzungen der latenten Variablen eindeutig zu interpretieren. Durch den Einsatz von *Simple Structure* MIRT Modellen kann außerdem noch die Eindimensionalität der Pseudoitems und ihrer zugrunde liegenden Subprozesse überprüft werden. Somit wird zuerst getestet, ob eindimensionale Antworttendenzen in den Ratingskala-Daten vorhanden sind und danach können diese korrigiert werden. Auch das Einbeziehen von Kovariaten wie Bildung oder Alter in Bezug auf die Antworttendenzen, als zusätzliche Erklärung für individuelle Unterschiede in multiplen Antwortprozessen, ist mit diesem Ansatz möglich (Khorramdel & von Davier, 2014).

Meiser und Böckenholt (2011, zitiert nach Khorramdel & von Davier, 2014) wählen einen ähnlichen Zugang wie Böckenholt (2012), indem sie die Antworten zu einer sechsstufigen Antwortskala wie folgt aufteilen: Pseudoitems für klare Entscheidungen (werden die Kategorien 0, 1, 4 oder 5 gewählt, wird mit 1 kodiert; sonst mit 0), solche für extreme Entscheidungen (bei den Kategorien 0 oder 5 wird mit 1 kodiert, die

restlichen Kategorien mit 0) und Pseudoitems, bei denen die Antwort in Richtung des zu messenden Merkmals geht (beim Wählen der Kategorien 3, 4 oder 5 wird mit 1 kodiert, ansonsten mit 0). Im Vergleich mit einem eindimensionalen *Partial Credit*-Modell konnten die Autoren danach mit dem von ihnen angewandten vierdimensionalen IRT Modell (das sich auf die vier Antwortsubprozesse bezieht) zeigen, dass Letzteres eine bessere Modellpassung auf die Daten hatte. Weiters konnte auch aufgezeigt werden, wie Eigenschaften, die auf Pseudoitems oder Antworttendenzen basieren, mit Kovariaten wie Persönlichkeitseigenschaften oder Positionseffekten zusammenhängen können.

Khorramdel und von Davier (2014) beschäftigten sich ebenfalls mit der Problematik der Antworttendenzen in Ratingskalen und berechneten hierfür ein- und mehrdimensionale IRT Modelle in Bezug auf die Tendenz zur Mitte und die Tendenz zu extremen Urteilen. Wie Böckenholt (2012) verglichen sie einen eindimensionalen Merkmalsfaktor mit mehrdimensionalen Antworttendenzfaktoren. Im Gegensatz zu Böckenholt (2012) verwendeten sie jedoch nicht nur eine einzelne Fragebogenskala, sondern einen mehrdimensionalen Fragebogen (mit mehreren Skalen), und erweiterten den Ansatz noch um den zusätzlichen Vergleich von mehrdimensionalen Merkmalsfaktoren mit eindimensionalen Antworttendenzfaktoren. Der Vorteil der eindimensionalen Messung von Antworttendenzen liegt darin, dass dies zu einer eindeutigeren Interpretation und damit zu einer höheren Validität führt; außerdem können die Antworttendenzen so leichter und eindeutiger von Eigenschafts- oder Persönlichkeitsmerkmalen unterschieden werden.

Angelehnt an den Entscheidungsbaum von Böckenholt (2012) wurden die beobachteten Antworten zu einer fünfstufigen Antwortskala (1 = *very inaccurate* bis 5 = *very accurate*) in drei binäre Pseudoitems zerlegt. Einen Überblick bietet hier Abbildung 3:

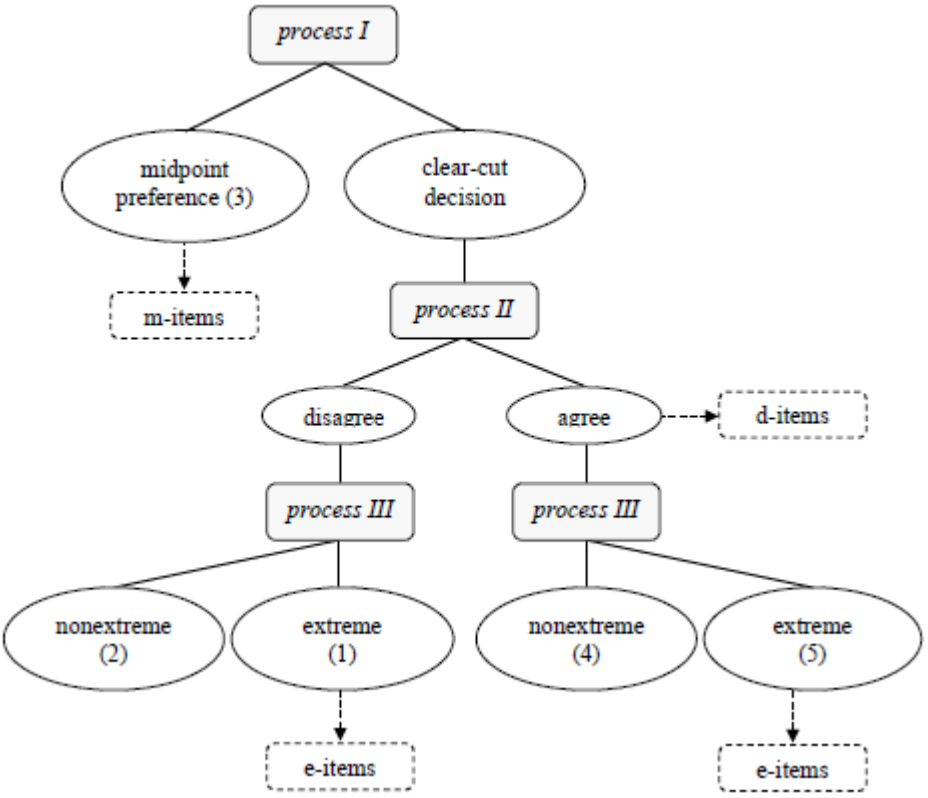


Abbildung 3: multinominaler Prozessbaum (Khorramdel & von Davier, 2014)

Wie aus der obigen Abbildung ersichtlich, betrifft der erste Antwortsubprozess die Entscheidung für oder gegen die neutrale Mitte. Daraus wird das Pseudoitem *m (= middle,* für die Tendenz zur Mitte) gebildet – entscheidet sich die Testperson für die mittlere Antwortkategorie (2), wird mit 1 kodiert, bei allen anderen Antworten (0, 1, 3, 4) erfolgt eine Kodierung mit 0. Wählt die Testperson im ersten Schritt eine eindeutige Antwort von der Mitte weg, folgt der zweite Antwortsubprozess, nämlich ob eine Zustimmung oder Ablehnung in Richtung der Skala erfolgt. Dies betrifft das Pseudoitem *d (= direction,* in Richtung der Skala, also des intendierten Merkmals) – bei der Entscheidung in Richtung Skala erhält das Item eine Kodierung von 1 (3 und 4), ansonsten wird mit 0 kodiert (0 und 1). Der neutralen Mitte wird eine 9 (= fehlender Wert) zugeteilt, der Grund hierfür wird im folgenden Absatz erklärt. Schließlich erfolgt noch der letzte Antwortsubprozess, welcher die Intensität der Zustimmung oder Ablehnung betrifft, und das Pseudoitem *e (= extreme,* Tendenz zu extremen Urteilen) bildet. Werden die extremen Enden der Antwortskala gewählt (0 und 4), wird mit 1

kodiert, bei moderaten Antworten (1 und 3) mit 0. Auch hier erhält die neutrale Mitte wieder eine Kodierung von 9. Eine Übersicht hierzu bietet Tabelle 1:

Tabelle 1: Beispiel für die Kodierung von Pseudoitems (Khorramdel & von Davier, 2014)

Originalitem (fünfstufige Ratingskala)	*Pseudoitem e* (extreme Antworten)	*Pseudoitem d* (Antwort in Richtung der Skala)	*Pseudoitem m* (mittlere Antwort-kategorie)	*Wahrscheinlich-keiten für jede Antwortkategorie*
0	1	0	0	P(0)
1	0	0	0	P(1)
2	–	–	1	P(2)
3	0	1	0	P(3)
4	1	1	0	P(4)

Die ursprüngliche mittlere Antwortkategorie in den Pseudoitems *d* und *e* wurde mit einem fehlenden Wert (9) kodiert, um das Problem der Abhängigkeiten zwischen den Pseudoitems auszuschalten und die drei Pseudoitems *d*, *e* und *m* als Indikatoren für drei latente Variablen nutzen zu können. Aufgrund der speziellen Kodierung wird eine Quasi-Unabhängigkeit (Goodman, 1994; Gail, 1972; Fienberg, 1970) der Pseudoitems erreicht (somit gibt es keine implizierte Abhängigkeit zwischen den drei genannten Pseudoitems). Details und eine diesbezügliche Übersicht bieten die Studien von Khorramdel und von Davier (2014) und von Davier und Khorramdel (2013).

Khorramdel und von Davier (2014) konnten durch die Umkodierung in Pseudoitems und die Berechnung mit ein- und mehrdimensionalen IRT Modellen Antworttendenzen in den Daten nachweisen. Während bei der Tendenz zu extremen Urteilen das Ausschließen eines Items dazu führte, dass Eindimensionalität bei der Messung erreicht werden konnte, war die Tendenz zur Mitte nicht eindeutig von der zu messenden Eigenschaft zu trennen. Beide Messungen der Antworttendenzen zeigen jedoch hohe IRT-bezogene (marginale) Reliabilitäten. Bei der Tendenz zu extremen Urteilen konnten weiters kulturelle Unterschiede aufgezeigt werden. Durch das Nachweisen von Antworttendenzen in den Daten wird auch eine Korrektur in Bezug auf diese möglich – verrechnet man nur jene Antwortkategorien, die nicht durch Antworttendenzen verfälscht werden, erhält man eine faire Messung des intendierten Merkmals.

In der oben angeführten Studie stellte sich heraus, dass manche Pseudoitems nicht nur die Antworttendenzen messen, sondern die Messung (teilweise) mit jener von Antworten in Bezug auf die interessierenden Merkmale vermischt ist. Um diese Problematik zu korrigieren, führten von Davier und Khorramdel (2013) eine weitere Untersuchung durch, in der sie wie oben angeführt vorgingen (Zerlegung der Daten in Pseudoitems, Berechnung mit ein- und mehrdimensionalen IRT Modellen) und zusätzlich *Bifactor* und *Second-Order* Modelle berechneten. Diese Modelle beziehen die Möglichkeit mit ein, dass Items von mehr als einem Faktor abhängig sein können. Eine genauere Beschreibung des *Bifactor* Modells (das für die vorliegende Diplomarbeit wichtig ist) findet sich in Kapitel 2.3.3 (verwendete IRT Modelle).

Von Davier und Khorramdel (2013) konnten in ihrer Untersuchung wieder Antworttendenzen (Tendenz zur Mitte, Tendenz zu extremen Urteilen) nachweisen; auch hier zeigte sich, dass sich eine eindimensionale Messung derselben mit *Simple Structure* IRT-Modellen als problematisch darstellt, da auch die intendierten Merkmale mitgemessen werden. Durch weitere Berechnungen mit den *Bifactor* und *Second Order* Modellen konnte für Ersteres eine bessere Modellpassung erzielt werden, da das *Bifactor* Modell sowohl Ladungen auf dem Antworttendenzfaktor als auch zusätzlich auf den Merkmalsfaktoren für jedes Item erlaubt. Somit konnte gezeigt werden, dass die Tendenz zu extremen Urteilen als Hauptdimension mehr Varianz erklärt als die Merkmalsmessungen als spezifische Dimensionen (im vorliegenden Fall die Persönlichkeitsmerkmale aus den verwendeten Fragebögen). Das heißt, das Pseudoitem *e* misst tatsächlich die Tendenz zu extremen Urteilen und nicht die intendierten Merkmale. Für die Tendenz der Mitte konnte kein eindeutiges Ergebnis in diese Richtung gefunden werden.

2.3.3 Verwendete IRT Modelle

Die vorliegende Diplomarbeit basiert auf den bereits beschriebenen Studien von Böckenholt (2012), Khorramdel und von Davier (2014) und von Davier und Khorramdel (2013), in welchen ein- und mehrdimensionale IRT Modelle miteinander verglichen wurden. Zum besseren Verständnis werden in diesem Kapitel jene IRT

Modelle genauer beschrieben, die auch in der Auswertung im empirischen Teil dieser Arbeit Anwendung fanden.

2.3.3.1 Das Rasch Modell für dichotome Daten

In eindimensionalen IRT Modellen wird davon ausgegangen, dass eine latente Variable zur Beschreibung der Daten ausreichend ist. Das vermutlich bekannteste dieser eindimensionalen Modelle ist das sogenannte Rasch Modell für dichotome Daten (Rasch, 1960) oder auch Ein Parameter Logistisches Modell (1PL Modell). In diesem wird angenommen, dass nur zwei Parameter die Wahrscheinlichkeit der Lösung eines Items bestimmen: die Fähigkeits- oder Eigenschaftsausprägung der Testperson und die Schwierigkeit des Items.

Die allgemeine Modellgleichung für das Rasch-Modell für dichotome Daten stellt sich wie folgt dar:

$$\mathrm{P}(x \mid \theta_v, \beta_i) = \frac{e^{\theta_v - \beta_i}}{1 + e^{\theta_v - \beta_i}}$$

Hierbei bezeichnet θ_v den Personenfähigkeitsparameter und β_i den Itemschwierigkeitsparameter. Khorramdel und von Davier (2104) beschreiben die Wahrscheinlichkeit ein Item zu lösen, als strikt monoton ansteigend in θ_v und absteigend in β_i. Stellt θ_v nun eine Antworttendenz dar (also die Neigung einer Testperson eine gewisse Antworttendenz zu zeigen), kann β_i als die Tendenz des Items gesehen werden, eine gewisse Antworttendenz zu provozieren. Als wichtige Eigenschaften des Rasch-Modells für dichotome Daten gelten weiters noch die Suffizienz (der Personengesamtscore beinhaltet alle Informationen über die Ausprägung des intendierten Merkmals der Testperson und der Itemgesamtscore alle über das Item) und die spezifische Objektivität (der Vergleich zweier Testpersonen ist unabhängig davon, welche Items dafür verwendet werden und umgekehrt).

2.3.3.2 Das Zwei Parameter Logistische Modell

Die Verallgemeinerung des Rasch Modells für dichotome Daten, das Zwei Parameter Logistische Modell (2PL Modell) von Birnbaum (1968), enthält einen weiteren Parameter, den sogenannten Itemdiskriminationsparameter. Zusätzlich zur

Schwierigkeit des Items wird durch diesen noch die Trennschärfe desselben (also der Anstieg der Itemfunktion) miteinbezogen. Verschiedene Items können unterschiedlich gut zwischen stärkeren und schwächeren Merkmalsausprägungen trennen – dies wird im 2PL -Modell durch die Gewichtung jedes Items mit dem Itemdiskriminationsparameter α_i berücksichtigt. Dadurch sind allerdings – anders als im Rasch-Modell für dichotome Daten – keine objektiven spezifischen Vergleiche mehr möglich (Rost, 2004).

Die Modellgleichung des Zwei-Parameter-Logistischen-Modells lässt sich demnach wie folgt darstellen:

$$P(x \mid \theta_v, \beta_i, \alpha_i) = \frac{\exp(\alpha_i(\theta_v - \beta_i))}{1 + \exp(\alpha_i(\theta_v - \beta_i))}$$

2.3.3.3 Mehrdimensionale IRT Modelle

Anders als in den eindimensionalen geht man in den mehrdimensionalen IRT Modellen von mehreren zugrunde liegenden latenten Variablen zur Beschreibung der Daten aus. Durch die Erweiterung in mehrdimensionale Modelle können das Rasch-Modell und das 2PL Modell auch für solche Items berechnet werden, die mehrere Skalen umfassen. Dies ist unter der Bedingung möglich, dass jedem Satz von voneinander unterscheidbaren Subsets von Items (in diesem Fall Skalen) ein unterschiedlicher Personenparameter zugeteilt wird. Reicht zur Lösung eines Items eine Fähigkeits- oder Fertigkeitskomponente aus, so spricht man von einer sogenannten *between-item multidimensionality*; *within-item multidimensionality* bezeichnet umgekehrt den Fall, dass mehrere Fähigkeits- oder Fertigkeitskomponenten zur Lösung notwendig sind (von Davier, Rost & Carstensen, 2007).

Geht man nun davon aus, dass jedes Item nur auf einer Skala lädt (*between-item multidimensionality*) und ein mehrdimensionales Rasch-Modell zur Berechnung genutzt wird, sieht die Wahrscheinlichkeit für Antwort x auf Item i (mit $x = 1,...,m_i$) in Skala k von Testperson v wie folgt aus:

$$P(x \mid \theta_v, \beta_i) = \frac{\exp(x\theta_{vk} - \beta_{ix})}{1 + \sum_{y=1}^{m_i} \exp(y\theta_{vk} - \beta_{iy})}$$

2.3.3.4 Das Bifactor Modell

Im *Bifactor* Modell für binäre Daten (Gibbons & Hedeker, 1992), einem hierarchischen IRT-Modell, misst jedes Item eine Hauptdimension und eine von *K* spezifischen Dimensionen. Während die Hauptdimension jene latente Variable darstellt, die im betreffenden Fall am meisten interessiert, und die Kovarianz zwischen allen Items erfasst, bilden die spezifischen Dimensionen zusätzliche Abhängigkeiten zwischen bestimmten Itemgruppen ab. [Anmerkung: Von Davier und Khorramdel (2013) modellieren bei der Verwendung des *Bifactor* Modells die spezifischen Dimensionen (die Fragebogenskalen) als statistisch unabhängig von der Hauptdimension (die Antworttendenz), um Antworttendenzen getrennt von den eigentlich zu erfassenden Merkmalen modellieren zu können.] Weiters wird eine Normalverteilung der latenten Variablen angenommen.

Das *Bifactor* Modell kann als Gleichung wie folgt dargestellt werden:

$$P(y \mid \theta) = \prod_{i=1}^{I} P(y_{i(k)} \mid \theta_g + \theta_k)$$

y.....Vektor aller binär verrechneten Antworten

$y_{i(k)}$....Antwort auf Item i ($i = 1,...,I$) in Dimension k ($k = 1,...,K$)

θ_k.....dimensionsspezifische Variable

θ_g.....latente Hauptvariable, die allen Items gleich ist

3 Empirischer Teil

3.1 Ziele der Untersuchung

Die vorliegende Diplomarbeit beschäftigt sich mit der Erfassung von Antworttendenzen mit Hilfe eines MIRT Ansatzes von Khorramdel und von Davier (2014), welcher wiederum auf einem Ansatz von Böckenholt (2012) basiert. Dafür werden die erfassten Daten in binäre Pseudoitems aufgespalten, welche dann durch die Berechnung von ein- und mehrdimensionalen IRT Modellen verglichen werden. Dadurch soll festgestellt werden, ob Antworttendenzen (im vorliegenden Fall die Tendenz zur Mitte und die Tendenz zu extremen Urteilen) in den Daten vorhanden sind; ist dies der Fall, wird zusätzlich nach der Studie von von Davier und Khorramdel (2013) ein *Bifactor* Modell berechnet, um noch eindeutiger zwischen Antworttendenzen und konstruktrelevanten Antworten unterscheiden zu können. Das Ziel ist eine faire Erfassung der intendierten Konstrukte (in diesem Fall Persönlichkeitseigenschaften), ohne eine Verzerrung durch vorhandene Antworttendenzen.

Wie in den Studien von Khorramdel und von Davier (2014) und von Davier und Khorramdel (2013) wird in der vorliegenden Diplomarbeit ein Persönlichkeitsfragebogen mit einer fünfstufigen Antwortskala (Ratingskala) zur Erhebung der Daten verwendet. Der Unterschied liegt im Antwortformat – während in den genannten Studien ein verbales Antwortformat verwendet wurde, beschäftigt sich diese Diplomarbeit mit der Erfassung von Antworttendenzen in einem Fragebogen mit einem optischen Antwortformat. Das Ziel dieses Vorgehens ist die Untersuchung, inwieweit Antworttendenzen bei Verwendung eines solchen Antwortformats auftreten.

3.2 Methode

3.2.1 Datenerhebung

Die Datenerhebung wurde gemeinsam mit einer anderen Diplomandin im Rahmen ihrer und der vorliegenden Diplomarbeiten durchgeführt. Die Annahme war, dass mehr Personen gewillt sind, gleich zwei Fragebögen hintereinander zu bearbeiten als zweimal einen Fragebogen. Die Erhebung fand im Oktober und November 2012 über die kostenpflichtige Internet-Plattform www.onlineumfragen.com unter dem Thema „Stress

und Persönlichkeit“ als *Online-Self-Assessment (OSA)* statt. Die Testpersonen wurden in einem selbst verfassten Begrüßungstext (siehe Anhang) auf die Beantwortung der Fragebögen vorbereitet und auf die anonyme Durchführung hingewiesen. Bei vollständiger Bearbeitung beider Fragebögen erhielten sie im Anschluss nicht nur eine Übersicht über ihre Gesamtergebnisse (siehe Anhang), sondern auch eine individuelle Rückmeldung in Bezug auf die Themen Stress, Stressreduktion und Umgang mit Stress, sowie deren Zusammenhang mit dem Thema Persönlichkeit (siehe Anhang).

Die Rekrutierung von Testpersonen erfolgte hauptsächlich über die Internetseite *Facebook*. Der Link zu den beiden Fragebögen wurde im Freundes- und Bekanntenkreis mit der Bitte um Weiterleitung – falls die Untersuchung und anschließende Rückmeldung als hilfreich angesehen wurde – verbreitet und auch auf diversen Hochschulseiten, die auf *Facebook* vertreten sind, gepostet. Außerdem wurde in Lehrveranstaltungen zur psychologischen Diagnostik des Studiengangs Psychologie an der Universität Wien direkt Werbung gemacht – interessierte Studierende konnten sich hier in eine Liste eintragen und erhielten dann den Link zur Untersuchung per E-Mail.

3.2.2 Erhebungsinstrument

Der erste auszufüllende Fragebogen für die Erhebung der Daten war das Differentielle Stress Inventar HR (DSIHR), das sich mit der Erfassung von Stressauslösern, Stressmanifestation, verfügbaren Copingstrategien und Risiken der Stressstabilisierung beschäftigt (Lefevre & Kubinger, 2004). Zur Beantwortung wurde eine fünfstufiges, verbales Antwortformat (trifft nicht zu – trifft eher nicht zu – weder noch – trifft eher zu – trifft zu) verwendet. Weiters wurden zu einer Subskala des DSIHR (Coping in Stresssituationen) nach der Theorie des geplanten Handelns (Ajzen, 1985) jeweils drei Zusatzfragen pro Item vorgegeben, die mit „nein“ oder „ja“ zu beantworten waren. Als zweiter Fragebogen in der Untersuchung wurde das NEO-Fünf-Faktoren-Inventar (NEO-FFI, Borkenau & Ostendorf, 1993) – adaptiert mit einer optischen Ratingskala – vorgegeben. Da nur dieses in der vorliegenden Diplomarbeit verwendet wurde, wird im Folgenden näher darauf eingegangen.

3.2.2.1 NEO-Fünf-Faktoren-Inventar (NEO-FFI)

Das NEO-Fünf-Faktoren-Inventar (NEO-FFI) ist ein multidimensionaler Persönlichkeitsfragebogen, der in der deutschen Version von Borkenau und Ostendorf (1993) vorliegt. Damit werden fünf Persönlichkeitsdimensionen erfasst, die allgemein als „Big Five" bezeichnet werden. Der Fragebogen enthält pro Dimension 12 Items, insgesamt also 60 Items. Diese sind zum Teil negativ, zum Teil positiv gepolt.

Die Skalen umfassen folgende fünf Persönlichkeitskonstrukte (Borkenau & Ostendorf, 1993):

- Neurotizismus

Personen mit hohen Werten in dieser Skala neigen dazu, verlegen, traurig, nervös, unsicher und ängstlich zu sein und sorgen sich um ihre Gesundheit. Sie können ihre Bedürfnisse weniger gut kontrollieren und reagieren in Stresssituationen oft unangemessen. Weiters kann eine Neigung zu unrealistischen Ideen bestehen. (Beispielitem: „Ich fühle mich anderen oft unterlegen.")

- Extraversion

Personen mit hohen Werten in dieser Skala sind oft heiter, gesellig, personenorientiert, aktiv, optimistisch, herzlich und gesprächig; Aufregungen und Anregungen werden als positiv angesehen. (Beispielitem: „Ich habe gerne viele Leute um mich herum.")

- Offenheit für Erfahrung

Personen mit hohen Werten in dieser Skala schätzen neue Erfahrungen und Abwechslung und haben großes Interesse an öffentlichen Ereignissen und Kultur. Sie sind meist phantasievoll, wissbegierig, kreativ und unabhängig in ihrem Urteil. (Beispielitem: „Ich probiere oft neue und fremde Speisen aus.")

- Verträglichkeit

Personen mit hohen Werten in dieser Skala sind mitfühlend, wohlwollend, altruistisch und verständnisvoll. Sie vertrauen anderen Menschen, neigen zur Nachgiebigkeit, haben ein starkes Harmoniebedürfnis und arbeiten gerne mit anderen zusammen. (Beispielitem: „Ich versuche zu jedem, dem ich begegne, freundlich zu sein.")

- Gewissenhaftigkeit

Personen mit hohen Werten in dieser Skala neigen zu Zuverlässigkeit und Ordentlichkeit, arbeiten hart und systematisch und gelten als pünktlich, ehrgeizig, diszipliniert und penibel. (Beispielitem: „Ich halte meine Sachen ordentlich und sauber.“)

In Bezug auf die Reliabilität des NEO-FFI liegen die internen Konsistenzen der fünf Skalen zwischen r = .71 und r = .85.

Im Original erfolgt die Erfassung der Antworten der Testpersonen über eine fünfstufige, verbale Antwortskala (starke Ablehnung – Ablehnung – neutral – Zustimmung – starke Zustimmung). Dies wurde im Rahmen der vorliegenden Diplomarbeit insofern abgeändert, als dass statt der verbalen eine selbst erstellte optische Antwortskala benutzt wurde. Abbildung 4 zeigt dieses geänderte Antwortformat.

Abbildung 4: als fünfstufiges, optisches Antwortformat verwendete Smileys

Die Smileys wurden in einem Zeichenprogramm erstellt und zehn Personen in verschiedenen Farben im Sinne einer Vorauswahl vorgelegt – dabei wurde schnell deutlich, dass blau als eher negativ, gelb als eher positiv wahrgenommen wird; die restlichen Farben wurden als eher ambivalent angesehen (so empfanden manche Personen die Farbe Rot als positiv, andere hingegen sahen darin eher ein Warnsignal und daher eher etwas Negatives). Auch in Hinblick auf die Deutlichkeit der Gesichtsausdrücke der Smileys wurden die Anmerkungen und Eindrücke der Personen, die zur Vorauswahl befragt wurden, berücksichtigt. Die Endauswahl mit den Farben Blau für Ablehnung und Gelb für Zustimmung sowie den in Abbildung 4 gezeigten Gesichtsausdrücken wurde im Rahmen des *Online-Self-Assessments* im Internet vorgegeben.

3.2.3 Stichprobe

Von 900 Personen, die die Bearbeitung des *Online-Self-Assessments* zumindest begonnen hatten, beantworteten insgesamt 603 Personen beide Fragebögen vollständig. Da eine Person erst 14 Jahre alt war und somit das festgelegte Mindestalter von 16 Jahren unterschritten hatte, musste sie aus dem Datensatz ausgeschlossen werden. Zwei Personen machten keine Angaben bezüglich ihrer demographischen Daten – diese machen im folgenden Absatz immer die fehlenden 0,6% in den Angaben aus.

Insgesamt nahmen 430 Frauen (71,3%) und 171 Männer (28,4%) an der Untersuchung teil. Davon hatten 501 Personen (83,1%) die österreichische Staatsbürgerschaft, 82 (13,6%) stammten aus Deutschland und eine Person (0,2%) gab an, aus der Schweiz zu kommen. Die restlichen 17 Personen (2,8%) bestimmten ihre Staatsbürgerschaft mit „andere“. Die Muttersprache wurde von 572 Personen als Deutsch angegeben (94,9%), 29 Personen (4,8%) wählten eine andere Muttersprache. Während 356 Personen (59%) ihren Familienstand als ledig bezeichneten, leben 224 Personen (37,1%) in einer Beziehung bzw. sind verheiratet, 20 Personen (3,3%) sind geschieden und eine Person (0,2%) gab an, verwitwet zu sein. In Bezug auf die Ausbildung gab es zwei Personen ohne Schulabschluss (0,3%), 18 Personen mit einem Abschluss der Pflichtschule (3%) und 55 Personen (9,1%) haben eine Lehre beendet. 336 Personen (55,7%) gaben als höchste Ausbildungsstufe Matura an, 190 Personen (31,5%) einen Universitätsabschluss. Insgesamt wählten 351 Personen (58,2%) ein Studium als derzeitige Hauptbeschäftigung aus. 148 Personen (24,5%) sind laut Eigenaussage nicht beschäftigt, 137 Personen (22,7%) geringfügig, 116 Personen (19,2%) geben Teilzeitarbeit als Beschäftigung an und 200 Personen (33,2%) arbeiten Vollzeit. Die jüngste Testperson ist 16 Jahre, die älteste Testperson 65 Jahre – das Durchschnittsalter lag bei 29,5 Jahren. Eine grafische Aufbereitung der demografischen Daten findet sich im Anhang.

3.2.4 Datenaufbereitung

Da die vorliegende Diplomarbeit – wie bereits erwähnt – auf den respektiven Studien von Khorramdel und von Davier (2014) bzw. von Davier und Khorramdel (2013) basiert, wurde die Datenaufbereitung nach deren Vorbild durchgeführt. Eine genaue

Beschreibung der Vorgehensweise findet sich im Theorieteil dieser Diplomarbeit; folgend sei nur nochmalig ein kurzer Überblick gegeben.

Nach Umpolung der negativ formulierten Items, erfolgte die Umkodierung in drei binäre Pseudoitems – Pseudoitems *e* für die Tendenz zu extremen Antworten, Pseudoitems *m* für die Tendenz zur Mitte und Pseudoitems *d* für die betreffenden Persönlichkeitsmerkmale im NEO-FFI. Wie auch in den genannten Studien wurde die mittlere Antwortwortkategorie in den Pseudoitems *e* und *d* mit 9 (fehlender Wert) kodiert, um Quasi-Unabhängigkeit zu erreichen. Danach wurden die so umkodierten Daten für Berechnungen mit ein- und mehrdimensionalen IRT-Modellen verwendet. Die dazu benötigte spezielle Software wird im Folgenden näher erläutert.

3.2.5 Auswertungssoftware

Die Schätzung der ein- und mehrdimensionalen IRT Modelle erfolgte über das *mixture general diagnostic modeling framework* (MGDM; von Davier, 2010), welches eine Spezifikation eines diskreten *mixture* Modells mit einer hierarchischen Komponente erlaubt. Dafür wurde die Software *mdltm* (*multidimensional discrete latent trait models*) verwendet. Diese Software kann für Analysen in Bezug auf eine Vielzahl von Modellen genutzt werden, wie z.B. *latent class*-Modelle, diagnostische Modelle oder eben auch ein- und mehrdimensionale IRT Modelle. Die Implementierung eines EM-Algorithmus ermöglicht *marginal maximum likelihood*-Schätzungen der Parameter. Durch verschiedene Einschränkungen, die in Bezug auf die Itemparameterschätzungen getroffen werden können, wird der gemeinsame Einsatz von *mdltm* und IRT Modellen vereinfacht (Hee Seo, Xu & von Davier, 2011).

Mit dieser Software können auch mehrdimensionale IRT Modelle geschätzt werden, die sowohl auf dem Rasch oder dem 2PL Modell, als auch auf dem generalisierten *Partial Credit* Modell basieren können. Die Modellgleichung in Bezug auf die inkludierten Modelle sieht wie folgt aus:

$$P(X = x \mid \beta_i, a, q_i, \gamma_i, c) = \frac{\exp(\beta_{ixc} + \sum_{k=1}^{K} x\gamma_{ikc} h(q_{ik} a_k))}{1 + \sum_{y=1}^{m_i} \exp(\beta_{iyc} + \sum_{k=1}^{K} y\gamma_{ikc} h(q_{ik} a_k))}$$

Dabei bezeichnet $c = 1,...,C$ verschiedene Klassen. Die interessierenden Eigenschaften oder Fertigkeiten (Skills), deren Anzahl K ausmacht, werden als latente Eigenschaftsmuster oder Fertigkeitsprofile $a = (a_1,...,a_K)$ dargestellt, während eine binäre Q-Matrix (q_{ik}) die Fertigkeitsfaktoren mit den Items verbindet. Die Funktion $h(q_{ik},a_k)$ bildet die Q-Matrix und die Fertigkeiten für das Item ab. Die klassenspezifischen Itemschwierigkeitsparameter sind in der obigen Gleichung mit β_{ixc} bezeichnet, während der Steigungsparameter, der die Fertigkeiten und Items verbindet, mit $\gamma_{ikcx} = x\gamma_{ikc}$ dargestellt wird.

3.2.6 Hypothesen

Sollten Antworttendenzen in den Daten vorhanden sein, wird – wie bei Khorramdel und von Davier (2014) – davon ausgegangen, dass die Pseudoitems *e* und *m* diese (respektive die Tendenz zu extremen Urteilen und die Tendenz zur Mitte) über alle Skalen hinweg messen, während die Pseudoitems *d* die interessierenden Persönlichkeitsmerkmale – weitgehend unverzerrt von der Antworttendenz zur Mitte oder zu Extremurteilen – abbilden. Dementsprechend wird die Hypothese aufgestellt, dass im Falle, dass die Pseudoitems *e* und *m* tatsächlich nur die Antworttendenzen messen, eindimensionale IRT Modelle eine bessere Passung an die Daten aufweisen als mehrdimensionale IRT Modelle (hier fünfdimensionale Modelle aufgrund der fünf zu messenden Eigenschaften). Umgekehrt sollen die Pseudoitems *d* (fast) unbeeinflusst von den Antworttendenzen sein und somit nur die fünf Persönlichkeitseigenschaften des NEO-FFI messen, weshalb ein fünfdimensionales IRT Modell eine bessere Passung an die Daten aufweisen sollte als ein eindimensionales IRT Modell. Dies spiegelt sich auch in den zu erwartenden niedrigen Korrelationen der Pseudoitems *d* zwischen diesen fünf Skalen wieder, während die Korrelationen der Pseudoitems *e* und *m* zwischen den Skalen hoch sein sollten.

Wie bei von Davier und Khorramdel (2013) kann bei nicht eindeutigen Ergebnissen (sollten also die Items sowohl auf dem Antworttendenzfaktor als auch auf den fünf Persönlichkeitsdimensionen laden) ein *Bifactor* Modell zur weiteren Klärung des Sachverhalts herangezogen werden. Dieses erlaubt, dass jedes Item zwei Ladungen auf je zwei Faktoren (einem Generalfaktor und einem spezifischen Faktor) aufweist. Damit wird berücksichtigt, dass womöglich einige Testpersonen in ihrem Antwortverhalten Antworttendenzen zeigen, während andere wiederum die Fragen tatsächlich in Bezug auf den Fragebogeninhalt beantwortet haben. Insofern stellen im *Bifactor* Modell die Antworttendenzen den Generalfaktor dar, die Persönlichkeitsdimensionen im Gegenzug die spezifischen Faktoren. Das Modell kann Aufschluss darüber geben, wie viel Varianz von den einzelnen Faktoren erklärt wird und ob die meiste Varianz durch den Generalfaktor (Tendenz zur Mitte oder Tendenz zu extremen Urteilen) oder durch die spezifischen Faktoren erklärt werden kann. Die Hypothese dahingehend lautet, dass der Generalfaktor einen höheren Anteil an Varianz erklärt, wenn entsprechenden Antworttendenzen vorliegen, und diese mit den Pseudoitems *e* und *m* erfasst werden.

Schlussendlich kann noch die Hypothese formuliert werden, dass die Pseudoitems *d* ein faireres Maß zur Erfassung der Persönlichkeitsdimensionen darstellen als die Originaldaten, wenn die entsprechenden Antworttendenzen in den Daten vorliegen. Wenn Pseudoitems *e* und *m* Indikatoren für die Antworttendenzen darstellen, kann davon ausgegangen werden, dass die Daten, welche auf den Pseudoitems *d* basieren, (fast) nicht von eben diesen Antworttendenzen beeinflusst sind (da Pseudoitems *d* entsprechend kodiert wurden). Dies könnte sich unter anderem darin zeigen, dass die Skaleninterkorrelationen basierend auf den Pseudoitems *d* geringer ausfallen als auf der Basis der Originaldaten, da die Big Five Skalen so konstruiert sind, dass sie weitgehend unterschiedliche Persönlichkeitskonstrukte erfassen sollen (vgl. Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013).

3.2.7 Analysen

Zusammenfassend soll also festgestellt werden, ob die Pseudoitems *d* die fünf Persönlichkeitsdimensionen des NEO-FFI erfassen und die Pseudoitems *e* und *m* jeweils die Tendenz zu extremen Urteilen und die Tendenz zur Mitte. Dafür wurden die

folgenden Analysen – wie bei Khorramdel und von Davier (2014) und von Davier und Khorramdel (2013) – durchgeführt (jeweils auf der Basis des 2PL Modells und mit der Skalenzugehörigkeit der Items als fixem Faktor):

- Analyse der Gesamtdaten

In einem ersten Schritt kamen mehrdimensionale Antworttendenz- und Merkmalsfaktoren zum Einsatz, um festzustellen, ob tatsächlich Antworttendenzen in den Daten vorhanden sind. Hierzu wurden ein 3-dimensionales, ein 5-dimensionales und ein 7-dimensionales IRT-Modell miteinander verglichen. Im 3-dimensionalen Modell wurden die drei verschiedenen Arten von Antwortprozessen untersucht – das heißt, die Pseudoitems *e* (Tendenz zu extremen Urteilen), *m* (Tendenz zur Mitte) und *d* (Persönlichkeitsdimensionen) wurden je einem Faktor zugeteilt. Eine entsprechende Tabelle zur Übersicht findet sich bei Khorramdel und von Davier (2014).

Im Falle des 5-dimensionalen IRT-Modells wurden alle Pseudoitemtypen den fünf Merkmalsfaktoren zugeteilt, um die Big Five-Persönlichkeitsdimensionen abbilden zu können. Tabelle 4 zeigt wieder die jeweilige Zuordnung. Auch hier findet sich eine entsprechende Tabelle zur Übersicht bei Khorramdel und von Davier (2014).

Im 7-dimensionalen IRT-Modell wurden sowohl die Persönlichkeitsdimensionen als auch die beiden Antworttendenzen eingearbeitet. Gleich bleibt hier die Zuordnung der Pseudoitems *e* und *m* zu den jeweiligen Antworttendenzfaktoren; der Unterschied liegt in der Zuteilung der Pseudoitems *d*, die hier nicht einem generellen, sondern allen fünf Persönlichkeitsdimensionen zugeordnet werden. Khorramdel und von Davier (2014) bieten hier wieder eine entsprechende Tabelle zur Übersicht.

- Analysen in Bezug auf die Pseudoitems

Zusätzlich zu den oben angeführten mehrdimensionalen IRT-Modellen in Bezug auf die Antworttendenzen wurden diese auch noch eindimensional für jede Art von Pseudoitems (die Pseudoitems werden also jeweils nur einem Faktor zugeordnet) berechnet und mit einem mehr- (5-) dimensionalen IRT Modell für die Persönlichkeitsdimensionen des NEO-FFI (die Pseudoitems werden den Faktoren der Persönlichkeitsdimensionen zugeordnet) verglichen. Zusätzlich wurden im Sinne der

Modellpassung noch Rasch Modelle für die Pseudoitems *e* und *m* berechnet. Das Rasch Modell enthält – wie weiter oben bereits beschrieben – keinen zusätzlichen Itemdiskriminationsparameter und misst somit restriktiver als das 2PL Modell. Damit sind eindeutigere Ergebnisse in Bezug auf die Dimensionalität der Messungen zu erwarten.

- Analysen in Bezug auf das *Bifactor* Modell

Zur Klärung, ob die Items eventuell sowohl auf einen Antworttendenzfaktor als auch auf den Persönlichkeitsdimensionen laden, wurden zusätzlich noch *Bifactor* Modelle berechnet, wobei wie bei von Davier und Khorramdel (2013) die spezifischen Faktoren als statistisch unabhängig vom Generalfaktor modelliert wurden. Diese umfassten wieder eindimensionale Antworttendenzfaktoren für die Pseudoitemtypen *e* und *m*. Während die Antworttendenzen als Hauptfaktor (Tendenz zur Mitte entsprechend den Pseudoitems *m*, Tendenz zu extremen Urteilen entsprechend den Pseudoitems *e*) gewählt wurden, dienten die Persönlichkeitsdimensionen des NEO-FFI als spezifische Faktoren. Die *Bifactor* Modelle wurden mit den jeweiligen 1- bzw. 5-dimensionalen IRT Modellen für die Pseudoitems *e* und *m* verglichen.

- Analysen der Pseudoitems *d* und der Originaldaten

Zuletzt wurde noch ein 5-dimensionales IRT Modell für die Originaldaten in Bezug auf die Persönlichkeitsdimensionen berechnet. Hierfür wurden diese ihrem respektiven Faktor zugeteilt. Damit wird ein Vergleich der Skaleninterkorrelationen der Originaldaten mit jenen des 5-dimensionalen IRT Modells in Bezug auf die Pseudoitems *d* möglich und somit kann überprüft werden, ob die Skaleninterkorrelationen der Pseudoitems *d* wie erwartet niedriger sind als jene der Originaldaten.

Zur Evaluation der Modelle bzw. den Vergleich der Modellpassungen wurden das Akaike Informationskriterium (AIC; Akaike, 1974) und das Bayesian Informationskriterium (BIC; Schwarz, 1978) verwendet. Je geringer der Wert dieser Informationskriterien ist, desto besser ist im Normalfall die Modellpassung. Beide nutzen dabei den Wert der *maximum likelihood* eines Modells (*L*) und die geschätzte

Parameteranzahl (n_p). Der AIC gewichtet die Parameteranzahl mit 2, womit sich folgende Formel ergibt:

$$AIC = -2 \log L + 2 n_p$$

Der BIC gewichtet die Parameteranzahl mit einem variablen Koeffizienten, nämlich mit dem Logarithmus der Stichprobengröße *N*. Der Grund hierfür ist, dass es bei großen Stichprobenumfängen zu einer Überparameterisierung kommen kann (Rost, 2004). Der BIC gebietet diesem Umstand schneller Einhalt als der AIC und zwar sobald Log (N) > 2 ist. Die Formel sieht daher wie folgt aus:

$$BIC = -2 \log L + (\log N) n_p$$

3.3 Ergebnisse

- Analyse der Gesamtdaten

Um festzustellen, ob Antworttendenzen in den Daten vorhanden sind, wurden zuerst Analysen gleichzeitig über alle Skalen und alle Gruppen von Pseudoitems hinweg durchgeführt. Hierzu wurden IRT Modelle mit mehrdimensionalen Antworttendenzfaktoren eingesetzt; konkret handelte es sich hierbei um 3-, 5- und 7-dimensionale IRT-Modelle (basierend auf dem 2PL Modell, als fixer Faktor wird die Skalenzugehörigkeit der Items herangezogen), die geschätzt und anschließend miteinander verglichen wurden. Damit sollte ermöglicht werden, in den Daten vorhandene Antworttendenzen von Persönlichkeitsfaktoren unterscheiden zu können. Während im 3-dimensionalen Modell die Pseudoitems den drei Dimensionen *e*, *d* und *m* zugeordnet wurden, erfolgte im 5-dimensionalen Modell eine Zuteilung aller Typen von Pseudoitems zu den Big Five Persönlichkeitsfaktoren. Die Zuordnung im 7-dimensionalen Modell umfasste die Pseudoitems des Typ *d* zu den eben erwähnten Persönlichkeitsfaktoren, während die Pseudoitems der Typen *e* und *m* jeweils einem eigenen Faktor (also einem 6. und 7. Faktor) zugeteilt wurden. Anders als die Pseudoitems der Typen *e* und *m* sollen jene des Typs *d* im 7-dimensionalen Modell also nicht nur eine Dimension messen, sondern alle fünf Persönlichkeitsdimensionen abbilden.

Die Ergebnisse zeigen entsprechend der Hypothese, dass das 7-dimensionale Modell (AIC = 98134,41; BIC =99776,32) eine bessere Modellpassung aufweist als das 5-dimensionale Modell (AIC = 99152,52; BIC = 100754,81) und das 3-dimensionale

Modell (AIC = 99209,64; BIC = 100789,93). Die genauen Ergebnisse sind Tabelle 2 zu entnehmen. [Anmerkung: je niedriger der AIC oder BIC Wert, desto besser ist die Modellpassung.]

Tabelle 2: Ergebnisse der 3-, 5- und 7-dimensionalen 2PL Modelle mit mehrdimensionalen Antworttendenzfaktoren, inklusive aller Typen von Pseudoitems (180 Items insgesamt)

Alle Skalen, Pseudoitems e, d und m	*7-dimensionales 2PL Modell*	*5-dimensionales 2PL Modell*	*3-dimensionales 2PL Modell*
AIC Index BIC Index log-penalty (modell-basierend, pro Item)	98134,41 99776,32 0,5255	99152,52 100754,81 0,5311	99209,64 100789,93 0,5314

- Analyse in Bezug auf die Pseudoitems

Nachdem in der obigen Analyse Antworttendenzen in den Daten festgestellt werden konnten, wurden weitere Analysen durchgeführt, diesmal allerdings mit eindimensionalen Antworttendenzfaktoren und mehrdimensionalen Persönlichkeitsfaktoren. Diese sollten aufzeigen, ob eine eindimensionale Messung der Antworttendenzen möglich ist oder gleichzeitig auch Persönlichkeitsfaktoren miterhoben wurden. Dafür wurden 1- und 5-dimensionale IRT Modelle (wieder basierend auf dem 2PL Modell und mit der Skalenzugehörigkeit der Items als fixem Faktor) berechnet, allerdings separat für alle Typen von Pseudoitems. Im 1-dimensionalen Modell wurden daher die Pseudoitems der Typen *e* und *m* nacheinander einer Dimension zugeteilt, von der angenommen wurde, dass sie tatsächlich nur eine bestimmte Antworttendenz (also die Tendenz zu extremen Urteilen bzw. die Tendenz zur Mitte) misst. Im 5-dimensionalen Modell hingegen wurden die Pseudoitems der Typen *e* und *m* jeweils den fünf Persönlichkeitsdimensionen zugeordnet.

Die Ergebnisse zeigen, dass das 1-dimensionale 2PL Modell (AIC = 36868,85; BIC = 37405,89) betreffend die Pseudoitems *m* – entsprechend der Hypothese – besser auf die Daten passt als das 5-dimensionale 2PL Modell (AIC = 37279,68; BIC = 37891,54). Ein anderes Bild (widersprechend der aufgestellten Hypothese) zeigt sich für die Pseudoitems *e* – hier zeigt das 5-dimensionale 2PL Modell (AIC = 34633,36; BIC = 35201,18) eine bessere Passung an die Daten als das 1-dimensionale 2PL Modell (AIC = 34880,70; BIC = 35417,74). Auch nach Ausschluss von insgesamt 11 Items mit geringer Passung laut Graphical Item Check (GIC; siehe unten) im 1-dimensionalen

2PL Modell (respektive Items 06, 09, 16, 22, 33, 44, 45, 46, 52, 54 und 56) konnte keine bessere Passung an die Daten des 1-dimensionalen 2PL Modells im Vergleich zum 5-dimensionalen 2PL Modell erreicht werden. Die genauen Ergebnisse sind Tabelle 7 zu entnehmen. Auf die Pseudoitems *d* wird später genauer eingegangen, diese sind der Vollständigkeit halber allerdings bereits jetzt in Tabelle 3 eingefügt.

Tabelle 3: Ergebnisse der 1- und 5-dimensionalen 2PL Modelle und 1-dimensionalen Rasch Modelle inklusive aller Skalen einzeln für jeden Typ von Pseudoitem

	5-dimensionales 2PL Modell	*1-dimensionales 2PL Modell*	*1-dimensionales Rasch Modell*
e-items, alle 5 Skalen:			
AIC Index	34633,33	34880,70	34902,69
BIC Index	35201,18	35417,74	35184,41
log-penalty (modell-basierend, pro Item)	0,6086	0,6132	0,6156
m-items, alle 5 Skalen:			
AIC Index	37279,68	36868,85	36821,16
BIC Index	37891,54	37405,89	37102,88
log-penalty (modell-basierend, pro Item)	0,5114	0,5061	0,5071
d-items, alle 5 Skalen:			
AIC Index	24733,60	26888,27	
BIC Index	25301,45	27435,30	
log-penalty (modell-basierend, pro Item)	0,4333	0,4717	

Um die bisherigen Ergebnisse absichern zu können, wurden noch zusätzlich je ein 1-dimensionales Rasch-Modell für die Pseudoitems der Typen *e* und *m* berechnet. Diese wurden mit den bereits oben angeführten 1-dimensionalen 2PL-Modellen verglichen.

Hierbei zeigen die Ergebnisse für den Pseudoitemtyp *m*, dass das Rasch Modell (AIC = 36821,16; BIC = 37102,88) sogar eine noch bessere Modellpassung aufweisen als das 1-dimensionalen 2PL Modell (AIC = 36868,85; BIC = 37102,88). Für den Pseudoitemtyp *e* ist das Ergebnis nicht ganz so eindeutig; zwar weist laut AIC (AIC = 34880,70) das 1-dimensionale 2PL Modell eine bessere Passung an die Daten auf als das 1-dimensionale Rasch Modell (AIC = 34902,69) – zur Erinnerung: das 5-dimensionale 2PL Modell wies hier eine noch bessere Modellpassung auf –, bezogen auf den BIC ist jedoch das 1-dimensionale Rasch Modell (BIC = 35184,41) besser auf die Daten passend als das 1-dimensionale 2PL Modell (BIC = 35417,74) und das 5-dimensionale 2PL Modell (BIC = 35201,18). Auch hier wurden 10 Items mit geringer Passung laut Grafischem Modelltest (GMT; siehe unten) im 1-dimensionalen Rasch Modell (respektive Items 01, 02, 07, 27, 28, 30, 38, 44, 48 und 54) ausgeschieden, es

konnte jedoch keine bessere Passung im AIC an die Daten des 1-dimensionalen Rasch Modells im Vergleich zum 5-dimensionalen 2PL Modell erreicht werden. Eine genaue Übersicht über die Ergebnisse ist Tabelle 3 zu entnehmen.

Zur grafischen Darstellung der Itemschwierigkeiten in den berechneten 1-dimensionalen 2PL Modellen für die Pseudoitems der Typen *e* und *m* wurden Grafische Modelltests (GMTs) erstellt. Diese Bezeichnung ist jedoch irreführend – der GMT ist ein Modelltest für das Rasch Modell, nicht aber für das hier verwendete 2PL-Modell. Dementsprechend ist hier besser von einem „Graphical Item Check" (GIC) die Rede. Für die beiden 1-dimensionalen Rasch Modelle wurden in weiterer Folge ebenfalls GMTs erstellt. Hierfür wurde die Stichprobe der an der Studie teilnehmenden Personen in zwei Gruppen geteilt; das Teilungskriterium war der Median des Testscores (Gruppe 1 = die Testscores sind kleiner/gleich dem Median, Gruppe 2 = die Testscores sind größer als der Median). Die Ergebnisse dieser beiden Gruppen wurden danach in Punktform in einer Grafik mit einer 45°-Geraden durch den Ursprung aufgetragen. Diese Punkte zeigen an, ob die Itemschwierigkeiten sich für alle Gruppen ähneln (die Punkte liegen auf oder nahe der Geraden) oder ob sie sich systematisch voneinander unterscheiden (die Punkte liegen weiter von der Geraden entfernt). In letzterem Fall könnte dies ein Hinweis darauf sein, dass die Items des verwendeten Tests in Bezug auf die Itemschwierigkeiten für die Gruppen unterschiedlich sind und somit problematisch für die Berechnung verschiedener Modelle sein können (Khorramdel & von Davier, 2014). Sollte somit eine Berechnung ergeben, dass die Messung nicht eindimensional ist, kann ein GMT bzw. GIC herangezogen werden, um optisch zu überprüfen, welche Items eventuell ausgeschieden werden sollten, um eine bessere Modellpassung an die Daten erreichen zu können.

In Abbildung 5 findet sich der GIC für das 2PL Modell, berechnet für die Pseudoitems des Typ *e*. Die meisten der Punkte liegen relativ weit von der 45°-Geraden entfernt, was darauf hindeutet, dass viele der im Test verwendeten Items problematisch sein könnten. Dies kann einen Einfluss auf die Eindimensionalität der Testung haben.

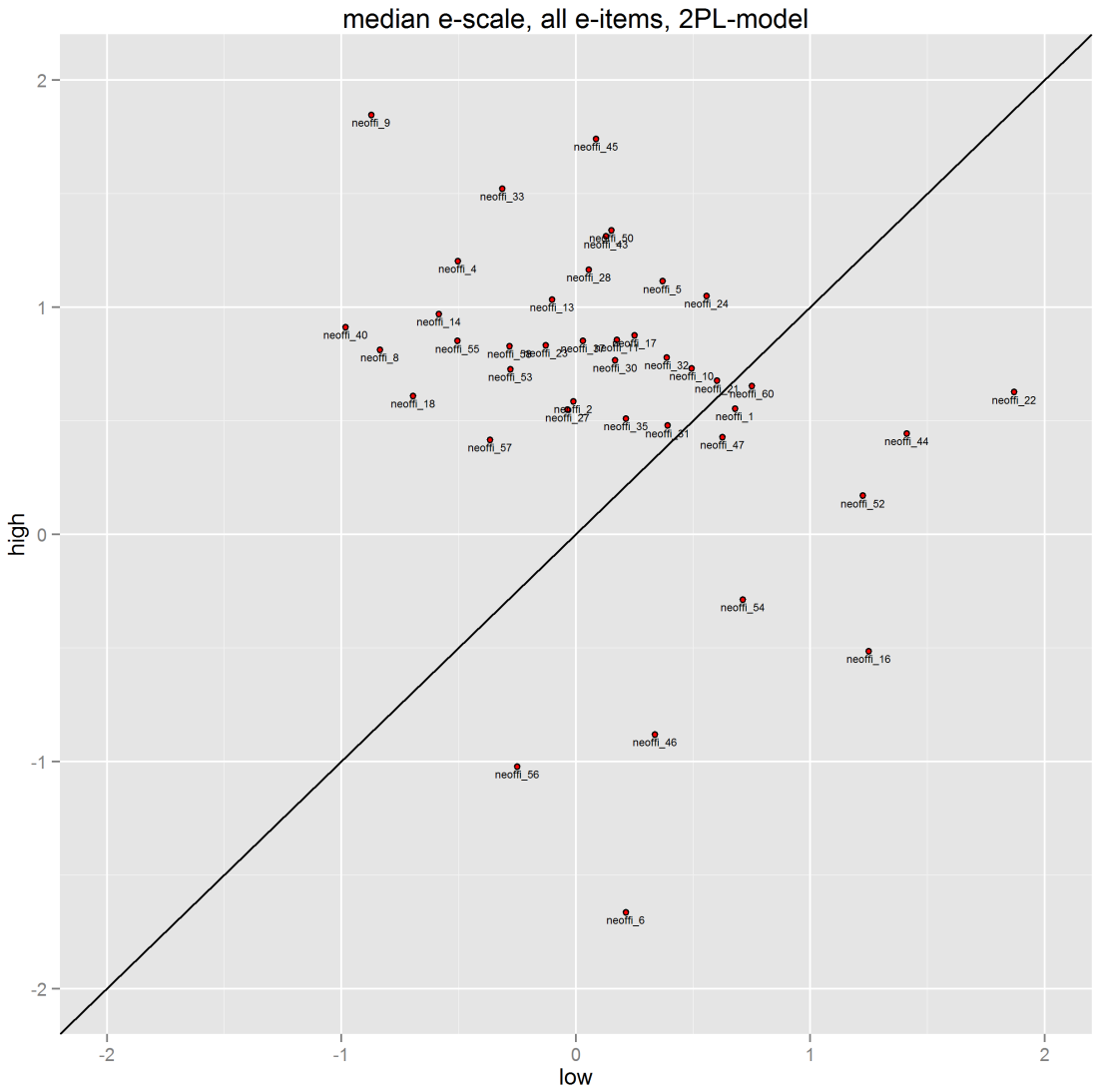


Abbildung 5: Graphical Item Check der Itemschwierigkeitsparameter für das 1-dimensionale 2PL Modell, berechnet für den Pseudoitemtyp e. Teilungskriterium ist der Median der Testscores.

Tabelle 4 gibt einen Überblick über die Diskriminierungsparameter („slope“ Parameter) aus dem 1-dimensionalen 2PL Modell für die Pseudoitems *e*, auf dem der GIC in Abbildung 6 basiert.

Tabelle 4: ***Diskriminierungsparameter aus dem 1-dimensionalen 2PL Modell für die Pseudoitems e***

Pseudoitems e	*Diskriminierungs-Parameter*	*Pseudoitems e*	*Diskriminierungs-Parameter*
Item 1	0.71171	*Item 31*	1.06498
Item 2	0.88778	*Item 32*	1.09957
Item 3	0.84478	*Item 33*	1.03250
Item 4	0.90365	*Item 34*	0.86706
Item 5	0.79554	*Item 35*	1.30156
Item 6	1.23956	*Item 36*	1.20167
Item 7	0.66680	*Item 37*	1.52153
Item 8	0.51998	*Item 38*	0.57629
Item 9	0.82432	*Item 39*	0.85138
Item 10	0.84500	*Item 40*	1.10700
Item 11	0.95168	*Item 41*	1.08110
Item 12	1.26585	*Item 42*	1.25861
Item 13	0.82449	*Item 43*	0.87160
Item 14	0.73267	*Item 44*	1.04259
Item 15	1.03263	*Item 45*	1.38502
Item 16	1.25553	*Item 46*	1.21336
Item 17	0.97719	*Item 47*	1.38413
Item 18	0.90383	*Item 48*	0.71186
Item 19	0.75273	*Item 49*	1.22395
Item 20	1.03545	*Item 50*	1.35796
Item 21	1.05505	*Item 51*	1.02803
Item 22	0.86059	*Item 52*	1.43990
Item 23	1.05095	*Item 53*	0.81184
Item 24	1.08878	*Item 54*	0.84537
Item 25	1.07164	*Item 55*	1.28419
Item 26	0.85923	*Item 56*	1.01606
Item 27	0.93627	*Item 57*	1.55329
Item 28	0.58409	*Item 58*	0.67875
Item 29	1.04649	*Item 59*	0.81912
Item 30	0.76049	*Item 60*	1.10069

Abbildung 6 zeigt den GIC für das 2PL Modell, berechnet für die Pseudoitems des Typ *m*. Auch hier können wieder viele der im Test verwendeten Items als problematisch gesehen werden (viele Punkte liegen relativ weit von der 45°-Geraden entfernt). Wie bereits erwähnt, kann dies kann einen Einfluss auf die Eindimensionalität der Testung haben.

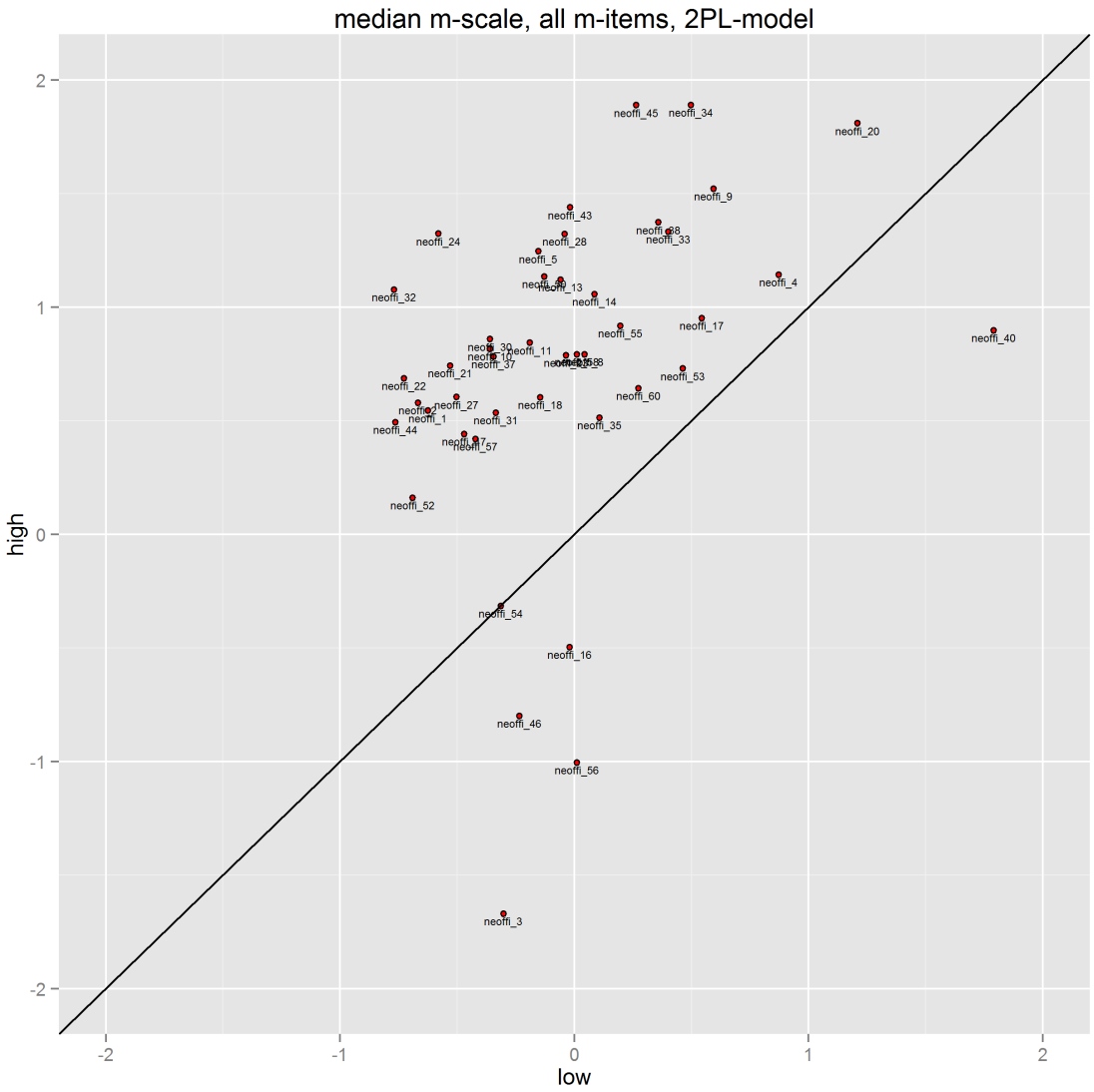


Abbildung 6: Graphical Item Check der Itemschwierigkeitsparameter für das 1-dimensionale 2PL Modell, berechnet für den Pseudoitemtyp m. Teilungskriterium ist der Median der Testscores.

Tabelle 5 gibt einen Überblick über die Diskriminierungsparameter („slope“ Parameter) aus dem 1-dimensionalen 2PL Modell für die Pseudoitems *m*, auf dem der GIC in Abbildung 6 basiert.

Tabelle 5: Diskriminierungsparameter aus dem 1-dimensionalen 2PL Modell für die Pseudoitems m

Pseudoitems m	*Diskriminierungs-Parameter*	*Pseudoitems m*	*Diskriminierungs-Parameter*
Item 1	0.74991	*Item 31*	1.18815
Item 2	0.82875	*Item 32*	0.87755
Item 3	0.77971	*Item 33*	0.93541
Item 4	1.22755	*Item 34*	0.96873
Item 5	0.74352	*Item 35*	1.50603
Item 6	0.93256	*Item 36*	1.01327
Item 7	0.94138	*Item 37*	1.03184
Item 8	0.99416	*Item 38*	0.99702
Item 9	1.03347	*Item 39*	0.96586
Item 10	0.94230	*Item 40*	1.59485
Item 11	0.92121	*Item 41*	1.19254
Item 12	1.26753	*Item 42*	1.16418
Item 13	0.84542	*Item 43*	1.08815
Item 14	0.76463	*Item 44*	0.88689
Item 15	0.71070	*Item 45*	1.33911
Item 16	1.04269	*Item 46*	1.02141
Item 17	1.16315	*Item 47*	1.00971
Item 18	1.02433	*Item 48*	1.08458
Item 19	0.89754	*Item 49*	1.14140
Item 20	1.18881	*Item 50*	1.01346
Item 21	0.73069	*Item 51*	0.87277
Item 22	0.65393	*Item 52*	1.29636
Item 23	1.00885	*Item 53*	1.20355
Item 24	0.89863	*Item 54*	0.89665
Item 25	0.91649	*Item 55*	0.90681
Item 26	0.97437	*Item 56*	0.69585
Item 27	0.84721	*Item 57*	1.04364
Item 28	0.93549	*Item 58*	1.04201
Item 29	1.00870	*Item 59*	0.83347
Item 30	0.92954	*Item 60*	1.31340

Abbildung 7 zeigt den Grafischen Modelltest (GMT) für das Rasch Modell, berechnet für die Pseudoitems *e*. Darin ist zu erkennen, dass die meisten Punkte nahe an der 45°-Geraden liegen und nur wenige Items problematisch gesehen werden müssten.

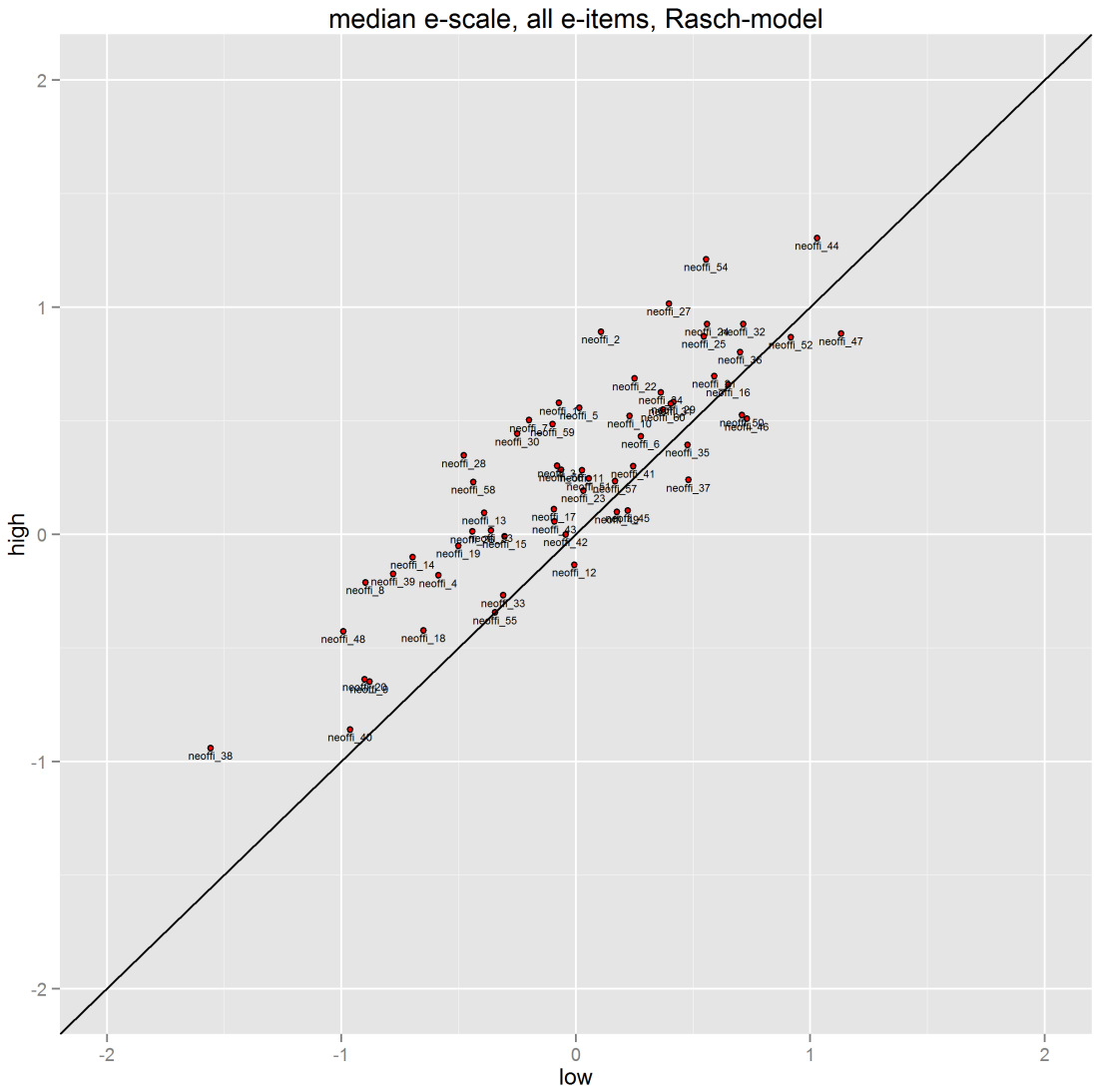


Abbildung 7: Grafischer Modelltest der Itemschwierigkeitsparameter für das 1-dimensionale Rasch Modell, berechnet für den Pseudoitemtyp e. Teilungskriterium ist der Median der Testscores.

Abbildung 8 zeigt den Grafischen Modelltest (GMT) für das Rasch Modell, berechnet für die Pseudoitems *m*. Zwar liegen hier viele Punkte nicht nahe an der 45°-Geraden, bilden aber eine eindeutig zusammengehörige Gruppe.

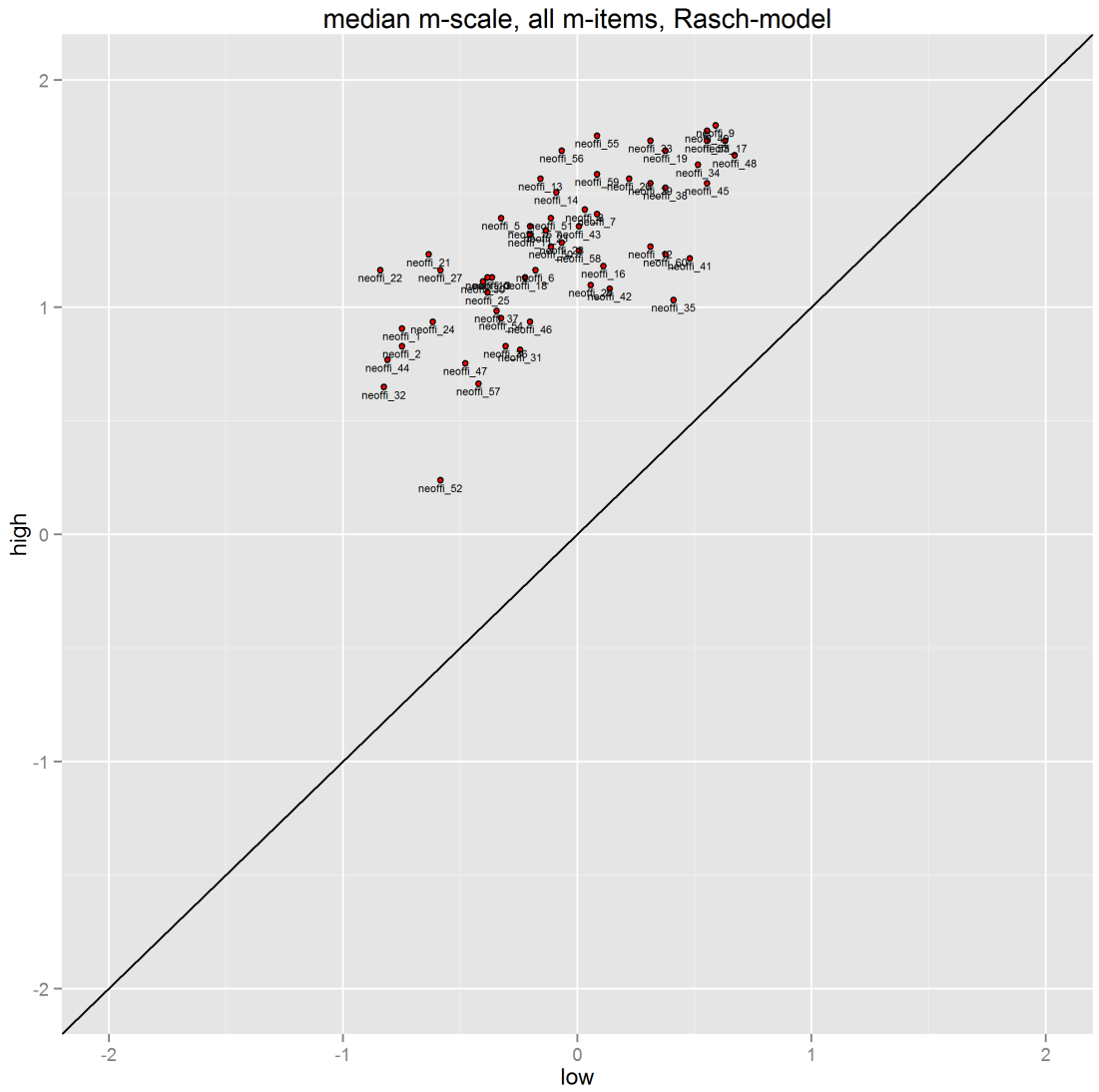


Abbildung 8: Grafischer Modelltest der Itemschwierigkeitsparameter für das 1-dimensionale Rasch Modell, berechnet für den Pseudoitemtyp m. Teilungskriterium ist der Median der Testscores.

Abbildung 9 zeigt den GIC für das 2PL Modell, berechnet für die Pseudoitems des Typ *e*, nach dem Ausscheiden jener 11 Items, die sich im Original GIC (Abbildung 5) als auffällig dargestellt haben. Rein grafisch gesehen zeigt sich hier durch das Ausscheiden der auffälligen Items eine eindeutige Verbesserung der Passung des 1-dimensionalen 2PL Modells.

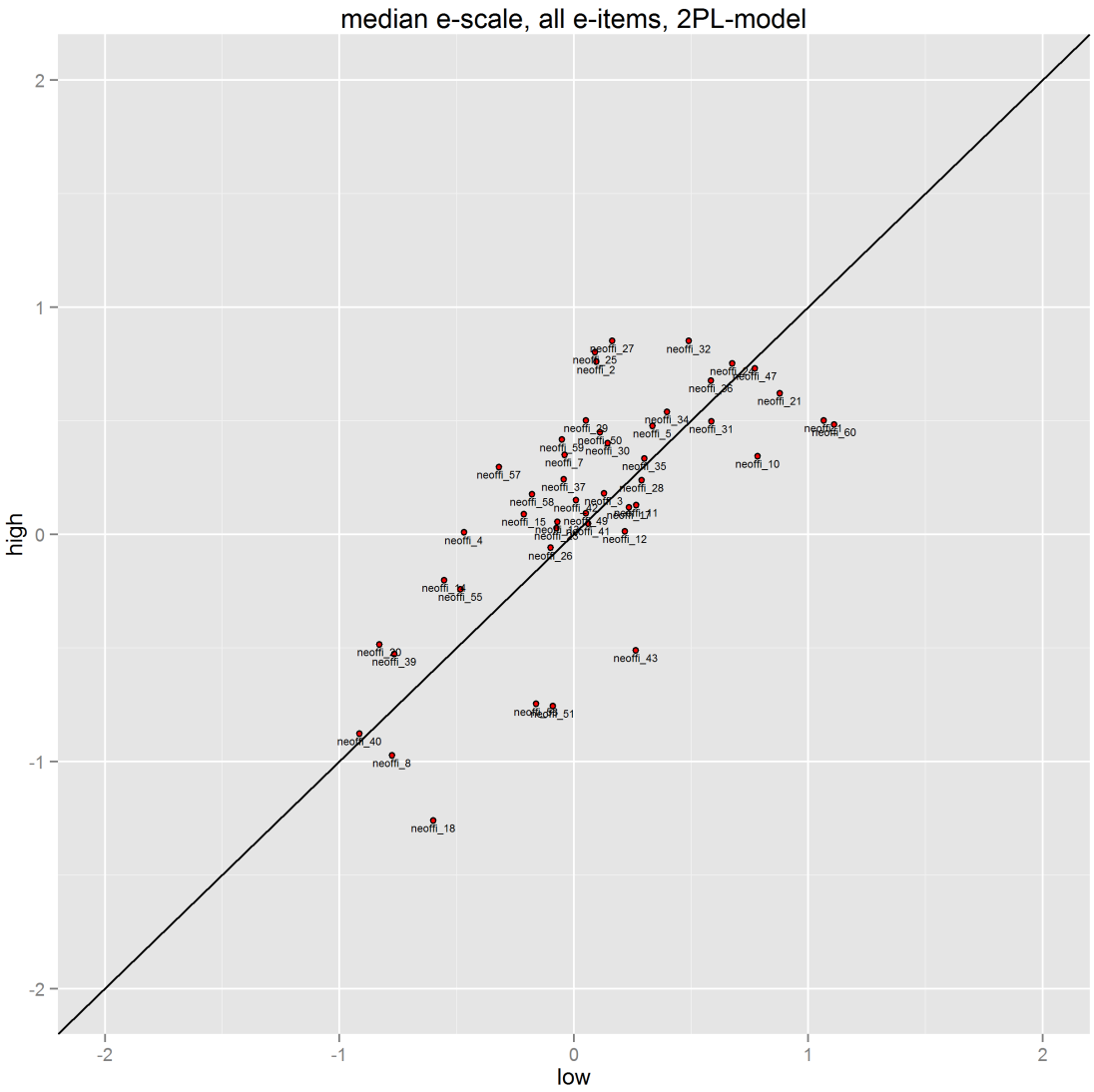


Abbildung 9: Graphical Item Check der Itemschwierigkeitsparameter für das 1-dimensionale 2PL Modell, berechnet für den Pseudoitemtyp e, nach Ausscheiden von 11 auffälligen Items.

Abbildung 10 zeigt den GMT für das Rasch Modell, berechnet für die Pseudoitems des Typ *e*, nach dem Ausscheiden jener 10 Items, die sich im Original GMT (Abbildung 7) als auffällig dargestellt haben. Auch hier zeigt sich nach Ausscheiden der auffälligen Items eine leichte Verbesserung der Passung des 1-dimensionalen Rasch Modells.

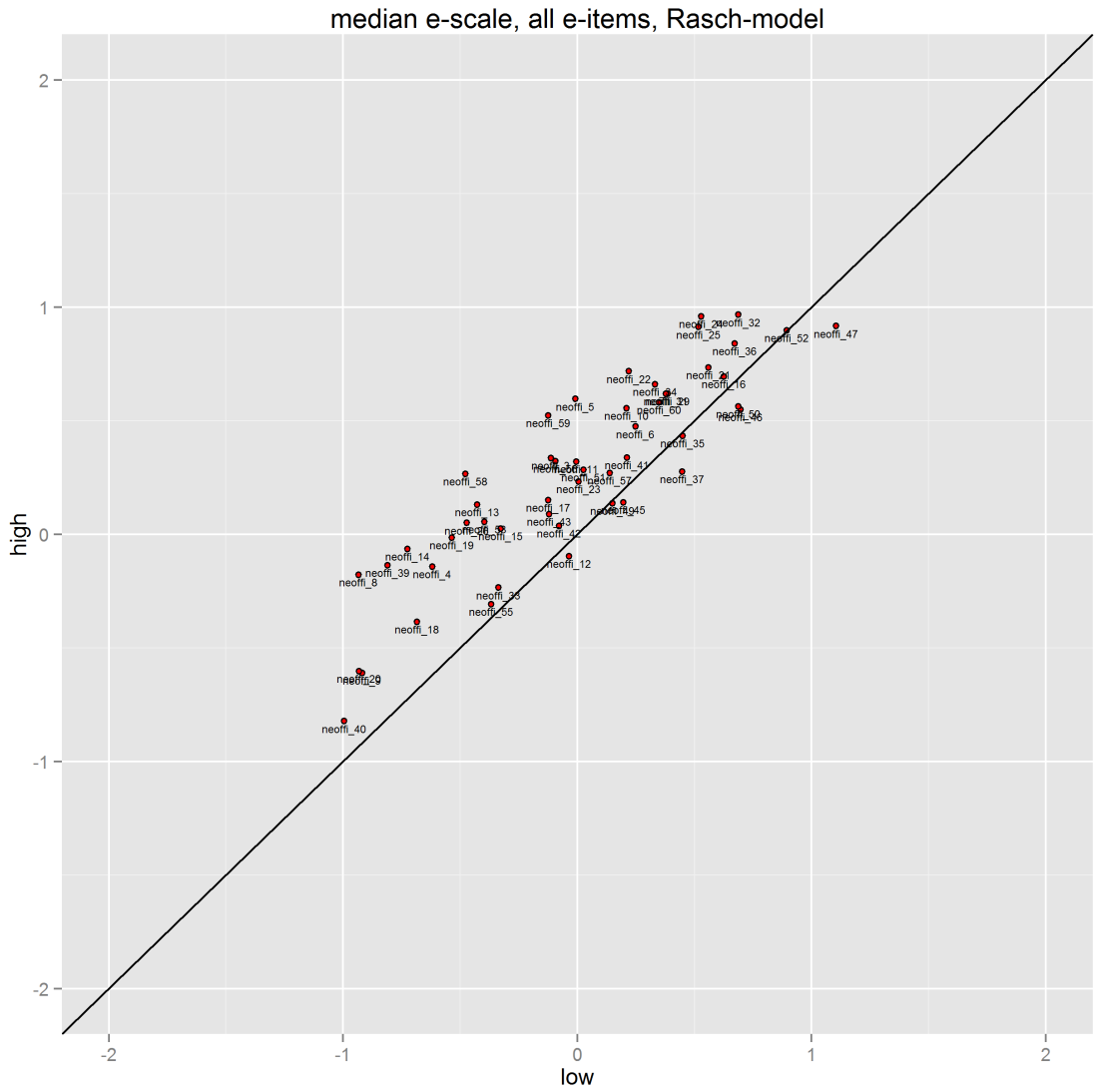


Abbildung 10: Grafischer Modelltest der Itemschwierigkeitsparameter für das 1-dimensionale Rasch Modell, berechnet für den Pseudoitemtyp e, nach Ausscheiden von 11 auffälligen Items.

Da aufgrund der obigen Analysen davon ausgegangen werden kann, dass Antworttendenzen (Tendenz zu extremen Urteilen und Tendenz zur Mitte) in den Daten vorhanden sind, kann es natürlich der Fall sein, dass die Originalitems des NEO-FFI von diesen beeinflusst werden. Dementsprechend wurde in zusätzlichen Analysen überprüft, ob die Pseudoitems des Typs *d*, die nicht von den Antworttendenzen beeinflusst sein sollten, eine bessere Messung der Persönlichkeitsdimensionen des NEO-FFI darstellen. Zu diesem Zweck wurden ein 1-dimensionales (die Pseudoitems wurden nur einem Faktor zugeordnet) und ein 5-dimensionales (die Pseudoitems wurden allen fünf Persönlichkeitsfaktoren zugeordnet) IRT Modell (jeweils basierend

auf dem 2PL Modell, als fixer Faktor gilt die Skalenzugehörigkeit der Items) berechnet und diese wurden im Anschluss miteinander verglichen.

Die Ergebnisse – welche die obige diesbezügliche Hypothese bestätigen – zeigen, dass das 5-dimensionale 2PL Modell (AIC = 24733,60; BIC = 25301,45) eindeutig besser auf die Daten passt als das 1-dimensionale 2PL Modell (AIC = 26888,27; BIC = 27425,30). Eine Übersicht über die genauen Ergebnisse zeigt nochmals die bereits weiter oben angeführte Tabelle 7.

Aus der Berechnung der 5-dimensionalen 2PL Modelle mit jeweils einem eindimensionalen Antworttendenzfaktor können weiters die Skaleninterkorrelationen entnommen werden. Zusätzlich wurde noch ein 5-dimensionales 2PL Modell für die Originaldaten (vor der Aufteilung in die Pseudoitemtypen) vorgenommen, um auch hier die für einen Vergleich benötigten Skaleninterkorrelationen zu erlangen.

Tabelle 6 zeigt die auf Basis des Modells geschätzten Skaleninterkorrelationen der Pseudoitemtypen *e*, *d* und *m* sowie der Items der Originaldaten aus dem NEO-FFI (die empirischen Skaleninterkorrelationen finden sich im Anhang). Es wird deutlich, dass sowohl die Skaleninterkorrelationen der Pseudoitems des Typs *d* (Wertebereich -0,32 bis 0,26) als auch jene der Items der Originaldaten (Wertebereich -0,42 bis 0,37) generell niedriger sind als jene der Pseudoitems des Typs *e* (Wertebereich 0,38 bis 0,56) und auch häufiger niedriger sind als jene des Pseudoitemtyps *m* (Wertebereich -0,01 bis 0,17).

Tabelle 6: Geschätzte Interkorrelationen der Score-Verteilung der Big Five Dimensionen aus dem 5-dimensionalen IRT Modell für die Pseudoitemtypen e, d und m sowie für die Originalitems des NEO-FFI

	Neurotizismus	*Extraversion*	*Offenheit*	*Verträglichkeit*	*Gewissenhaftig-keit*
e-Items:					
Neurotizismus	1,00000				
Extraversion	0,56293	1,00000			
Offenheit	0,49547	0,47202	1,00000		
Verträglichkeit	0,51394	0,56178	0,45202	1,00000	
Gewissenhaftig.	0,52824	0,52483	0,38230	0,55558	1,00000
m-Items:					
Neurotizismus	1,00000				
Extraversion	0,08752	1,00000			
Offenheit	0,03677	-0,01490	1,00000		

Verträglichkeit	0,02765	0,13175	0,05269	1,00000	
Gewissenhaftig.	0,05134	0,14879	0,00643	0,17199	1,00000
d-Items:					
Neurotizismus	1,00000				
Extraversion	-0,42410	1,00000			
Offenheit	-0,14351	0,07563	1,00000		
Verträglichkeit	0,26329	-0,37472	0,01508	1,00000	
Gewissenhaftig.	-0,26300	0,23022	-0,03284	-0,19509	1,00000
Originalitems des NEO-FFI:					
Neurotizismus	1,00000				
Extraversion	-0,42113	1,00000			
Offenheit	-0,10802	0,03227	1,00000		
Verträglichkeit	-0,18925	0,37590	-0,02408	1,00000	
Gewissenhaftig.	-0,26679	0,22616	-0,04003	0,22575	1,00000

Die marginalen Reliabilitäten für die beiden eindimensionalen Messungen der Antworttendenzen waren in Bezug auf die *m*-Items eher im mittleren Bereich (0,67), jene der *e*-Items jedoch – wie bei Khorramdel und von Davier (2014) – eher hoch (0,88).

- Analyse mit dem *Bifactor* Modell

Zwar haben die Analysen zumindest in Bezug auf die Pseudoitems *m* ergeben, dass die in den Daten vorhandene Tendenz zur Mitte tatsächlich eindimensional erfasst werden kann, für die Pseudoitems *e* konnte dies allerdings nicht klar bestätigt werden. Deshalb wurde zusätzlich ein *Bifactor* Modell für Pseudoitems *e* berechnet (wiederum auf Basis des 2PL Modells), da dieses erlaubt, dass Items gleichzeitig auf zwei Dimensionen laden können. Als Generalfaktor (jener Faktor, auf dem alle Items laden) wurde die jeweilige Antworttendenz gewählt, die spezifischen Faktoren waren die Persönlichkeitsdimensionen des NEO-FFI. Das *Bifactor* Modell ermöglicht weiters eine Abklärung darüber, wie viel Varianz durch den Generalfaktor bzw. die spezifischen Faktoren erklärt wird. Für Pseudoitems *m* wurde ebenfalls ein *Bifactor* Modell berechnet, um die Eindimensionalität der Messung nochmals zu bestätigen. Nach der Berechnung wurden die *Bifactor* Modelle mit den jeweils am besten passenden Modellen aus den vorhergehenden Analysen in Bezug auf die Pseudoitems verglichen.

Im Falle der Pseudoitems *e* zeigen die Ergebnisse ein der Hypothese deutlich widersprechendes Bild – hier weisen sowohl das 1-dimensionale 2PL Modell (AIC = 34880,70; BIC = 35417,74) als auch das 5-dimensionale 2PL Modell (AIC = 34633,36;

BIC = 35201,18) eine bessere Modellpassung auf als das *Bifactor* Modell (AIC = 36264,77; BIC = 37004,29). In Bezug auf Pseudoitems *m* zeigt sich erwartungsgemäß, dass sowohl das 1-dimensionale 2PL Modell (AIC 36868,85; BIC = 37405,89) als auch das 1-dimensionale Rasch Modell (AIC = 36821,16; BIC = 37102,88) besser auf die Daten passen als das *Bifactor* Modell (AIC = 36960,20; BIC = 37699,72). Tabelle 7 zeigt nochmals eine Übersicht über die genauen Ergebnisse.

Tabelle 7: Ergebnisse der 1-dimensionalen 2PL und Rasch Modelle inklusive aller Skalen sowie der Bifactor Modelle berechnet für die Pseudoitemtypen e und m

	Bifactor Modell	*5-dimensionales 2PL Modell*	*1-dimensionales 2PL Modell*	*1-dimensionales Rasch Modell*
e-Items: *AIC* *BIC* *log-penalty (modell-basierend, pro Item)*	 36264,77 37004,29 0,6361	 34633,36 35201,18 0,6086	 34880,70 35417,74 0.6132	 34902,69 35184,41 0.6157
m-Items: *AIC* *BIC* *log-penalty (modell-basierend, pro Item)*	 36960,20 37699,72 0,5061	 37279,68 37891,54 0,5114	 36868,85 37405,89 0,5061	 36821,16 37102,88 0,5071

Auf die Darstellung der durch die Faktoren im *Bifactor* Modell erklärten Varianz wird aufgrund der Nichtpassung des Modells verzichtet.

- Analysen der Pseudoitems *d* und der Gesamtdaten

Zuletzt wurden noch die Skaleninterkorrelationen der Pseudoitems *d* und der Originalitems aus den jeweiligen 5-dimensionalen 2PL Modellen miteinander verglichen. Niedrigere Skaleninterkorrelationen deuten hier auf eine bessere Messung der Big Five Persönlichkeitsdimensionen hin. Betrachtet man den gesamten Wertebereich über alle Skalen hinweg, zeigen sich die Skaleninterkorrelationen der Pseudoitems *d* (Wertebereich von -0,43 bis 0,26) etwas niedriger als jene der Originalitems (Wertebereich von -0,42 bis 0,37). Umgekehrt weisen die Originalitems jedoch in den einzelnen Werten öfter niedrigere Skaleninterkorrelationen auf als die Pseudoitems *d*. Deshalb ist keine eindeutige Aussage in Bezug auf die Hypothese möglich, dass die Pseudoitems *d* eine bessere Messung i. S. von niedrigeren Skaleninterkorrelationen der Persönlichkeitsdimensionen darstellen als die

Originalitems. Pseudoitems *d* bieten jedenfalls eine unverfälschtere Messung der Persönlichkeitskonstrukte und damit eine fairere Messung. Tabelle 8 (wie auch Tabelle 6) gibt hier eine Übersicht über die geschätzten Skaleninterkorrelationen (die empirischen Skaleninterkorrelationen finden sich im Anhang).

Tabelle 8: Geschätzte Interkorrelationen der Score-Verteilung der Big Five Dimensionen aus dem 5-dimensionalen IRT Modell für die Pseudoitems d sowie for die Originalitems des NEO-FFI

	Neurotizismus	*Extraversion*	*Offenheit*	*Verträglichkeit*	*Gewissenhaftigkeit*
d-items:					
Neurotizismus	1,00000				
Extraversion	-0,42410	1,00000			
Offenheit	-0,14351	0,07563	1,00000		
Verträglichkeit	0,26329	-0,37472	0,01508	1,00000	
Gewissenhaftigkeit	-0,26300	0,23022	-0,03284	-0,19509	1,00000
Originalitems des NEO-FFI:					
Neurotizismus	1,00000				
Extraversion	-0,42113	1,00000			
Offenheit	-0,10802	0,03227	1,00000		
Verträglichkeit	-0,18925	0,37590	-0,02408	1,00000	
Gewissenhaftigkeit	-0,26679	0,22616	-0,04003	0,22575	1,00000

3.3.1 Interpretation der Ergebnisse

Eine erste Analyse der Gesamtdaten mit mehrdimensionalen Antworttendenz- und Merkmalsfaktoren (also eine Analyse mit allen Typen von Pseudoitems und über alle Skalen hinweg) ergab, – wie auch bei Khorramdel und von Davier (2014) und von Davier und Khorramdel (2013) – dass ein 7-dimensionales 2PL Modell besser auf die Daten passte als ein 3- oder 5-dimensionales 2PL Modell. Im 7-dimensionalen 2PL Modell wurden sowohl die fünf Persönlichkeitsdimensionen des NEO-FFI als auch die beiden Antworttendenzen zur Mitte und zu extremen Urteilen als jeweils eigene Faktoren miteinbezogen; somit kann der aufgestellten Hypothese entsprechend also davon ausgegangen werden, dass in den erhobenen Daten diese Antworttendenzen vorhanden sind und die Pseudoitems der Typen *e* und *m* diese tatsächlich erfassen, während die Pseudoitems des Typ *d* die Persönlichkeitsdimensionen des NEO-FFI abbilden.

Nachdem in den Daten durch die oben erwähnte erste Analyse Antworttendenzen festgestellt werden konnten, wurden zur weiteren Überprüfung Analysen mit

eindimensionalen Antworttendenzfaktoren und mehrdimensionalen Persönlichkeitsfaktoren, durchgeführt. Für die Pseudoitems *m* zeigte sich, dass ein 1-dimensionales 2PL Modell besser auf die Daten passt als ein 5-dimensionales 2PL Modell. Da der GIC in Bezug auf den Pseudoitemtyp *m* auf eine nicht optimale Passung des 1-dimensionalen 2PL Modells hinweist, wurde zur Absicherung der Ergebnisse noch ein 1-dimensionales Rasch-Modell gerechnet. Dieses konnte die Daten sogar noch besser abbilden als das 1-dimensionale 2PL-Modell, was sich auch im Vergleich des GIC für das 1-dimensionale 2PL Modell und des GMT für das 1-dimensionale Rasch Modell zeigt. Da das Rasch Modell restriktiver misst als das 2PL Modell, welches einen Itemparameter mehr schätzt, kann also für die Pseudoitems des Typ *m* davon ausgegangen werden, dass die Messung eindimensional in Bezug auf die Tendenz zur Mitte erfolgt ist.

Für die Pseudoitems des Typs *e* ergab sich ein der oben aufgestellten Hypothese widersprechendes Bild – hier passte das 5-dimensionale 2PL Modell besser auf die Daten als das 1-dimensionale 2PL Modell. Auch nach Ausschluss von insgesamt 11 in Bezug auf die Modellpassung des 1-dimensionalen Modells auffälligen Items auf Basis des GIC zeigten sich keine Änderungen in diesen Ergebnissen. Somit konnte nicht bestätigt werden, dass die Pseudoitems *e* die Antworttendenz eindimensional erfassen – zwar zeigte in weiterer Folge das 1-dimensionale Rasch Modell im BIC eine bessere Modellpassung als die 1- und 5-dimensionale 2PL Modelle und es gibt nur relativ geringe Unterschiede im AIC und BIC der 1- und 5-dimensionalen 2PL Modelle, auf eine Eindimensionalität der Messung der Tendenz zu extremen Urteilen kann daraus allerdings nicht geschlossen werden. Auch der Ausschluss von insgesamt 10 in Bezug auf die Modellpassung des 1-dimensionalen Rasch Modells auffälligen Items auf Basis des GMT änderte nichts an den Ergebnissen. Betrachtet man GIC und GMT auf Basis der Itemschwierigkeitsparameter zeigt sich jedenfalls, dass das Rasch Modell besser zu passen scheint als das 2PL Modell, was für die Ergebnisse des BIC und damit dafür sprechen würde, dass eine Tendenz zu extremen Urteilen in den Daten vorliegt.

Um die Dimensionalität der Pseudoitems genauer zu untersuchen, z.B. ob eine Mischung aus inhaltsrelevanten Antworten und Antworttendenzen vorliegt, wurden weiters noch *Bifactor* Modelle für die Pseudoitems *e* und *m* berechnet. Diese erlauben

es den Items auf mehreren Faktoren zu laden. Die *Bifactor* Modelle zeigen im Vergleich zu den bereits berechneten Modellen keine bessere Passung an die Daten. Die Hypothese, dass die Pseudoitems *e* am besten durch den Einfluss des Generalfaktors und der spezifischen Faktoren (Antworttendenz und Persönlichkeitsmerkmale), also auf mehreren verschiedenen Faktoren ladend, erklärt werden können, kann somit nicht bestätigt werden. Daraus könnte man schließen, dass die Pseudoitems *e* wahrscheinlich eher die Persönlichkeitskonstrukte des NEO-FFI als die Tendenz zu extremen Urteilen erfassen. Dagegen spricht die bessere Passung des BIC im 1-dimensionalen Rasch Modell im Vergleich zum 5-dimensionalen 2PL Modell; dies zeigt an, dass die Pseudoitems nicht eindeutig nur die Fragebogenskalen messen. Betrachtet man außerdem GIC und GMT, scheint das 1-dimensionale Rasch Modell besser zu passen. Das würde für die Ergebnisse des BIC und damit dafür sprechen, dass eine Tendenz zu extremen Urteilen in den Daten vorliegt. Um hier Klarheit zu schaffen, müssten noch andere IRT Modelle zum Einsatz kommen und die Dimensionalität der Pseudoitems *e* weiter untersucht werden. In Punkt 3.4 (Diskussion) wird nochmals genauer darauf eingegangen.

Die Ergebnisse in Bezug auf die Pseudoitems *d* zeigen, dass das 5-dimensionale 2PL Modell eine bessere Passung an die Daten als das 1-dimensionale 2PL Modell aufweist. Damit kann die Aussage getroffen werden, dass die Pseudoitems *d* die fünf Persönlichkeitsmerkmale des NEO-FFI messen.

In Bezug auf die erhaltenen Skaleninterkorrelationen ist anzumerken, dass niedrige Skaleninterkorrelationen auf eine passende Messung verschiedener Dimensionen bzw. Skalen in Bezug auf die Big Five Konstrukte hindeuten (da diese theoretisch so konstruiert sind, dass sie unterschiedliche, wenig überlappende Konstrukte erfassen); das bedeutet, je höher die Skaleninterkorrelationen sind, desto schlechter erfassen die jeweiligen Items bzw. Pseudoitems diese Dimensionen bzw. Skalen wie theoretisch intendiert. Wie in der aufgestellten Hypothese erwartet, zeigten sich die Skaleninterkorrelationen der Pseudoitems *d* und der Items der Originaldaten niedriger als Skaleninterkorrelationen der Pseudoitems *e* und *m*. Dies deutet – die obige Hypothese bestätigend – darauf hin, dass sowohl die Pseudoitems des Typs *d* als auch

die Items der Originaldaten bessere Messungen der Persönlichkeitsskalen darstellen als die Pseudoitems der Typen *e* und *m*.

Zuletzt wurde noch die Hypothese überprüft, dass die Pseudoitems *d*, welche von den Antworttendenzen (fast) unbeeinflusst sein sollten, eine bessere Messung der Persönlichkeitsdimensionen des NEO-FFI darstellen (und somit auch eine faire Messung dieser Dimensionen) als die Items der Originaldaten. Einzeln betrachtet zeigen sich die Skaleninterkorrelationen der Originalitems gehäuft niedriger als jene der Pseudoitems *d* – letztere weisen allerdings insgesamt einen niedrigeren Wertebereich auf. Eine eindeutige Aussage in Bezug auf die Hypothese kann somit nicht getroffen werden, allerdings kann zumindest festgestellt werden, dass die Pseudoitems *d* jedenfalls eine unverfälschtere Messung der Persönlichkeitskonstrukte und damit eine fairere Messung bieten.

3.4 Diskussion

Wie im Theorieteil bereits ausführlich beschrieben, haben sich eine Vielzahl von Autorinnen und Autoren mit der Erfassung von Antworttendenzen befasst und sich dafür verschiedenster stilistischer und statistischer Methoden bedient. Besonders interessant zeigt sich hier die Studie von von Davier und Khorramdel (2013), da die vorliegende Diplomarbeit unter anderem darauf basiert. Durch den Einsatz der gleichen statistischen Methoden und desselben Fragebogens, des NEO-FFI, wird somit ein vorsichtiger Vergleich mehr oder weniger möglich, solange man in Betracht zieht, dass sich die Stichproben in Größe, Zusammensetzung und Gewinnung unterscheiden.

Der große Unterschied in der erwähnten Studie und der vorliegenden Diplomarbeit ist hierbei das verwendete Antwortformat für die Items des NEO-FFI. Während die Beantwortung der Items in der Studie von von Davier und Khorramdel (2013) über eine fünfstufige verbale Ratingskala (Starke Ablehnung – Ablehnung – Neutral – Zustimmung – Starke Zustimmung) erfolgte, kam im Rahmen der vorliegenden Diplomarbeit eine fünfstufige optische Ratingskala zum Einsatz (siehe auch Punkt 3.2.2.1.).

Die Ergebnisse beider Untersuchungen zeigten, dass Antworttendenzen (Tendenz zur Mitte und Tendenz zu extremen Urteilen) in den Daten vorhanden waren; während in der Studie von von Davier und Khorramdel (2013) jedoch keine Eindimensionalität für die Pseudoitems *e* und *m* festgestellt werden konnte, gelang dies in der vorliegenden Diplomarbeit wenigstens für die Pseudoitems des Typs *m*. In ihrer Studie wiesen von Davier und Khorramdel (2013) nach, dass die Pseudoitems *e* und *m* sowohl auf den Fragebogendimensionen als auch auf einem Antworttendenzfaktor laden, was in der vorliegenden Diplomarbeit nicht der Fall ist. Die Feststellung von von Davier und Khorramdel (2013), dass die Pseudoitems *d* eine bessere Messung der Persönlichkeitsmerkmale i. S. von niedrigeren Skaleninterkorrelationen darstellen als die Originalitems, kann in der vorliegenden Diplomarbeit nicht getätigt werden. In der aktuellen Untersuchung gibt es auch keinen Hinweis darauf, dass die Pseudoitems *e* auf mehreren Faktoren laden, während dies in der Studie von von Davier und Khorramdel (2013) mittels *Bifactor* Modellen nachgewiesen werden konnte.

Wie bereits oben erwähnt, ist ein direkter Vergleich aufgrund der Unterschiedlichkeit der Stichproben natürlich schwierig – ein eindeutiger Vorteil des verbalen Antwortformats gegenüber des optischen oder umgekehrt ist allerdings nicht zu erkennen.

In der vorliegenden Diplomarbeit ergaben sich insofern Probleme in Bezug auf die Pseudoitems *e*, als dass keine konkrete Aussage darüber getroffen werden konnte, was diese genau messen. Zusammengefasst kann Folgendes festgehalten werden: eine eindimensionale Messung der Tendenz zu extremen Urteilen ist nicht möglich, die Pseudoitems laden aber auch nicht auf zwei Faktoren, sind also keine Mischung aus Antworttendenz und inhaltsrelevanten Antworten. Daraus könnte geschlossen werden, dass die Pseudoitems *e* wahrscheinlich eher die Persönlichkeitskonstrukte des NEO-FFI als die Tendenz zu extremen Urteilen erfassen. Damit ist nun die Kodierung der Pseudoitems *d* kritisch zu sehen, die ja darauf aufbaut, dass sowohl die Tendenz zur Mitte als auch die Tendenz zu extremen Urteilen in den Daten vorliegen. Messen nun die Pseudoitems *e* mehr die Fragebogenskalen, gehen wichtige Informationen bei der Kodierung der Pseudoitems *d* verloren. Hier passt somit eher eine Verrechnung, welche die extremen Antworten miteinbezieht und nur auf die mittlere Antwortkategorie

verzichtet. Betrachtet man allerdings die bessere Passung des BIC im 1-dimensionalen Rasch Modell im Vergleich zum 5-dimensionalen 2PL Modell, spricht dies dafür, dass die Pseudoitems nicht eindeutig nur die Fragebogenskalen erfassen. Auch der GIC und GMT auf Basis der Itemschwierigkeitsparameter zeigen, dass das 1-dimensionale Rasch Modell besser zu passen scheint, was für die Ergebnisse des BIC und damit dafür sprechen würde, dass eine Tendenz zu extremen Urteilen in den Daten vorliegt. Insgesamt müssten noch andere IRT Modelle zum Einsatz kommen und die Dimensionalität der Pseudoitems *e* weiter untersucht werden (eventuell auch durch das Ausscheiden von auffälligen Items aufgrund anderer Itemfitmaße als GIC und GMT), um hier Klarheit zu schaffen.

Kritisch anzumerken in Bezug auf die vorliegende Untersuchung ist außerdem, dass – wie in Punkt 3.2.1. beschrieben – zwei Fragebögen hintereinander in der Form eines *Online-Self-Assessments* vorgegeben wurden. Der in der vorliegenden Diplomarbeit verwendete Fragebogen wurde im Anschluss an einen anderen bearbeitet. Deshalb kann nicht ausgeschlossen werden, dass es eventuell zu Reihenfolgeeffekten bei der Beantwortung gekommen ist. Bei der Durchführung einer ähnlichen Studie sollte hier nur einer der Fragebogen vorgegeben werden, um solche eindeutig ausschließen zu können.

Ebenfalls kritisch zu betrachten ist die Tatsache, dass die Vorgabe der beiden Fragebögen ausschließlich als *Online-Self-Assessment* stattfand – somit wurden eventuell Personen aus der Untersuchung ausgeschlossen, die keinen Zugang zu einem PC hatten. Dies könnte sich eventuell auf die Repräsentativität der Stichprobe auswirken. Weiters birgt dieser Vorgabemodus das Problem des fehlenden Testleiters / der fehlenden Testleiterin – die Durchführung des *Online-Self-Assessments* fand fachpsychologisch unkontrolliert (Kubinger, 2009) statt, womit z.B. keine Kontrolle über die Rahmenbedingungen der Untersuchung (wie eigenständige Durchführung, Identität der Testperson, Lärm etc.) gegeben war.

In Bezug auf die Berechnungen der mehrdimensionalen IRT Modelle muss außerdem noch auf die niedrige Rechenleistung des verwendeten PCs aufmerksam gemacht werden. Hier konnten die Berechnungen nicht immer so genau durchgeführt werden

(z.B. mit mehr maximal levels in *mdltm*), wie es wünschenswert gewesen wäre. Die Ergebnisse haben somit zwar Aussagekraft, manche Modellunterschiede wären bei einer noch genaueren Schätzung aber womöglich stärker zur Geltung gekommen. Bei weiteren Studien sollte daher an einen PC mit hoher Rechenleistung gedacht werden.

Da bisher vor allem der Einfluss von Antworttendenzen bei fünfstufigen Ratingskalen untersucht wurde, sollten in zukünftigen Studien Ratingskalen mit einer anderen Anzahl von Antwortkategorien untersucht werden. Ein interessanter Ansatz für die Zukunft z.B. für die Tendenz zu extremen Urteilen wäre hier unter anderem die Verwendung einer Ratingskala ohne eine neutrale Mitte (z.B. vierstufig) oder Antwortskalen mit mehr Antwortstufen.

Auch in Bezug auf numerische Ratingskalen stehen noch Ergebnisse aus bzw. weitere Ergebnisse in Bezug auf visuelle Ratingskalen. Letzteres könnte auch in anderer Form vorgegeben werden (z.B. statt Smileys andere Symbole wie + oder -). Auch Mischformen von Antwortformaten (z.B. verbal/optisch) stellen interessante Untersuchungsansätze dar.

Schlussendlich ist noch auf die Möglichkeit hinzuweisen, andere Antworttendenzen, wie z.B. die Akquieszenz – soweit möglich – mit dem neuen IRT-Ansatz nach Böckenholt (2012) und den Erweiterungen nach Khorramdel und von Davier (2014) sowie von Davier und Khorramdel (2013) zu untersuchen.

3.5 Zusammenfassung

Antworttendenzen und deren verfälschende Wirkung auf Fragebogenergebnisse in Untersuchungen konnten in einer Vielzahl von Studien belegt werden, die sich verschiedenster Methoden der Statistik bedienten (vgl. Baumgartner & Steenkamp, 2001; De Jong, Steenkamp, Fox & Baumgartner, 2008; Dolnicar & Grun, 2009; Weijters, Schillewaert & Geuens, 2008). Ein neuerer Ansatz zur Erfassung und Kontrolle von Antworttendenzen mit Hilfe der Item Response Theorie (IRT) stammt von Böckenholt (2012) und wurde von Khorramdel und von Davier (2014) und von Davier und Khorramdel (2013) erweitert. Dieser erweiterte Ansatz soll nun in der vorliegenden Diplomarbeit ausgebaut werden. Dafür wurden die in der Untersuchung

gewonnenen polytomen Daten in binäre Pseudoitems zerlegt und mit IRT Modellen über ein- und mehrdimensionale Faktoren für Antworttendenzen und Merkmale modelliert. Das Ziel war die Feststellung, ob die Antworttendenz zur Mitte und zu extremen Urteilen in den Daten vorhanden bzw. eindimensional messbar sind (erste bzw. zweite aufgestellte Hypothese), sowie ob die Daten von diesen bereinigt werden können. Sollte eine eindimensionale Messung der untersuchten Antworttendenzen nicht möglich sein, könnte dies daran liegen, dass die Items eventuell auf zwei Faktoren (einem Generalfaktor und einem spezifischen Faktor) laden (dritte Hypothese). Hierfür wurde die Berechnung von *Bifactor* Modellen eingeplant. Schlussendlich wurde noch untersucht, ob die von den Antworttendenzen bereinigten Daten tatsächlich eine faire Messung der Persönlichkeit darstellen (vierte Hypothese).

Der Unterschied im Untersuchungsdesign zu den oben erwähnten Studien liegt im verwendeten Antwortformat – während in den zitierten Studien Fragebögen mit einer verbalen Ratingskala vorgegeben wurden, kam in der vorliegenden Diplomarbeit eine optisches Ratingskala zur Anwendung. Dafür wurden die Items des NEO-Fünf-Faktoren-Inventar (NEO-FFI; Borkenau & Ostendorf, 1993) in Form eines *Online-Self-Assessments* im Internet bereitgestellt (und nach der Datenerhebung auch wieder offline genommen). Verwertbare Daten lagen schlussendlich von 603 Testpersonen vor, die über die Internetseite *Facebook* und per E-Mail sowie in einer Lehrveranstaltung an der Universität Wien angeworben wurden.

Die durch die Testung gewonnenen Daten wurden in binäre Pseudoitems (*e* = *extreme*, Tendenz zu extremen Urteilen; *d* = *direction*, in Richtung des Merkmals gehend, *m* = *middle*, Tendenz zur Mitte) zerlegt, die nacheinander ablaufende Antwortsubprozesse der Testpersonen darstellen; im Anschluss daran wurden diese Pseudoitems im Rahmen von IRT Analysen mit ein- und mehrdimensionalen Antworttendenz- und Merkmalsfaktoren modelliert. In einem ersten Schritt (mehrdimensionale Antworttendenz- und Merkmalsfaktoren) wurde festgestellt, dass die untersuchten Antworttendenzen tatsächlich in den erfassten Daten vorhanden sind, womit die erste Hypothese bestätigt werden konnte. Wie sich im Zuge der weiteren Berechnungen jedoch herausstellte, ist das Vorhandensein der Tendenz zu extremen

Urteilen insofern fraglich, als dass in Bezug auf Pseudoitems *e* nicht eindeutig klar ist, was diese genau messen.

Weitere Analysen (eindimensionale Antworttendenz- und mehrdimensionale Merkmalsfaktoren) bestätigten die zweite Hypothese nur zum Teil – für die Pseudoitems *m* konnte die eindimensionale Messung der Tendenz zur Mitte eindeutig nachgewiesen werden, für die Pseudoitems *e* gelang dies jedoch nicht. Zusätzlich zu 2PL Modellen wurden auch 1-dimensionale Rasch Modelle berechnet. Die Eindimensionalität der Tendenz zur Mitte konnte hiermit noch weiter belegt werden, da ein 1-dimensionales Rasch Modell die beste Modellpassung zeigte; für die Pseudoitems *e* zeigten sich hier Unterschiede in den AIC und BIC bezüglich der Passung der verschiedenen Modelle (in Bezug auf den AIC passte das 5-dimensionale 2PL Modell am besten, in Bezug auf den BIC passt das 1-dimensionale Rasch Modell am besten), wodurch keine genaue Aussage getroffen werden konnte. Auch durch den Ausschluss von insgesamt 11 (in Bezug auf die Modellpassung) auffälligen Items laut GIC konnte keine eindimensionale Messung im 2PL Modell in Bezug auf die Pseudoitems *e* erreicht werden; dasselbe galt für den Ausschluss von insgesamt 10 (in Bezug auf die Modellpassung) auffälligen Items im Rasch Modell laut GMT. Es ist somit eher nicht davon auszugehen, dass die Pseudoitems *e* die Tendenz zu extremen Urteilen eindimensional messen.

Um die Dimensionalität der Pseudoitems *e* weiter zu untersuchen, wurde die Hypothese verfolgt, dass diese vielleicht eine Mischung aus Antworttendenzen und inhaltsrelevanten Antworten darstellen, da womöglich ein Teil der Testpersonen Antworttendenzen zeigt und andere nicht (vgl. von Davier & Khorramdel, 2013). Daher wurde für Pseudoitems *e* – und als zusätzliche Bestätigung der Eindimensionalität auch für Pseudoitems *m* – noch ein *Bifactor* Modell gerechnet. Dieses erlaubt es den Items gleichzeitig auf zwei Faktoren zu laden (im vorliegenden Fall auf der Antworttendenz und jeweils einer der Persönlichkeitsdimensionen). In diesem Modell wurde die jeweilige Antworttendenz als Generalfaktor bestimmt, die fünf Persönlichkeitsdimensionen des NEO-FFI stellten die spezifischen Faktoren dar. Die Hypothese, dass die Pseudoitems *e* sowohl auf dem Faktor der Antworttendenz als auch auf je einem der Persönlichkeitsdimensionen laden, konnte nicht bestätigt werden (das

Bifactor Modell zeigte keine entsprechende Modellpassung). Für Pseudoitems *m* konnte erwartungsgemäß gezeigt werden, dass das *Bifactor* Modell die Daten nicht besser beschreiben kann als das 1-dimensionale 2PL und Rasch Modell.

Insgesamt kann keine klare Aussage in Bezug darauf getroffen werden, was die Pseudoitems *e* genau messen: einiges (bessere Passung des AIC im 5-dimensionalen 2PL Modell, Nichtpassung des *Bifactor* Modells) deutet darauf hin, dass sie eher die Persönlichkeitskonstrukte des NEO-FFI erfassen. Ist dies der Fall, so muss die Kodierung der Pseudoitems *d* kritisch gesehen werden – wenn die Pseudoitems *e* mehr die Fragebogenskalen als die Tendenz zu extremen Urteilen messen, gehen bei der Kodierung der Pseudoitems *d* wichtige Informationen verloren. Andererseits gibt es auch Hinweise darauf, dass eine Tendenz zu extremen Urteilen in den Daten vorliegt (bessere Passung des BIC im 1-dimensionalen Rasch Modell, GIC und GMT auf Basis der Itemschwierigkeitsparameter). Es müssten noch andere IRT Modelle zum Einsatz kommen und die Dimensionalität der Pseudoitems *e* weiter untersucht werden, um hier Klarheit zu schaffen.

Die Analyse in Bezug auf die Pseudoitems *d* – von denen angenommen wird, dass sie weitgehend unbeeinflusst von den untersuchten Antworttendenzen sind – ergab, dass diese eine passende Messung für die fünf untersuchten Persönlichkeitsmerkmale darstellten (ein 5-dimensionales 2PL Modell zeigte eine bessere Passung als ein 1-dimensionales 2PL Modell; außerdem zeigten sich niedrige Skaleninterkorrelationen auf Basis dieser Kodierung).

In Bezug auf die letzte Hypothese, dass Pseudoitems *d* eine bessere Messung der Persönlichkeitsmerkmale darstellen als die Originalitems (i. S. von geringeren Skaleninterkorrelationen), konnte keine eindeutige Aussage getroffen werden; die Skaleninterkorrelationen der Originalitems zeigten häufiger niedrigere Werte, die Skaleninterkorrelationen der Pseudoitems *d* waren insgesamt (über den gesamten Wertebereich betrachtet) niedriger. Pseudoitems *d* bieten jedenfalls eine unverfälschtere Messung der Persönlichkeitskonstrukte und damit eine fairere Messung.

Ein Vergleich mit einer Studie von von Davier und Khorramdel (2013), die mit derselben Berechnungsmethode und demselben Fragebogen (allerdings mit einer verbalen Ratingskala) arbeiteten, zeigte keinen eindeutigen Vorteil einer visuellen Ratingskala, da es scheinbar bei beiden Antwortformaten zu Antworttendenzen kommen kann.

Weitere Studien zum Thema sind notwendig und sollten nicht nur verschiedenstufige Ratingskalen in Betracht ziehen, sondern auch andere Antwortformate und Antworttendenzen.

4 Literaturverzeichnis

Ajzen, I. (1985). *From intentions to actions: A theory of planned behavior*. In J. Kuhl, & J. Beckmann (Hrsg.), Springer series in social psychology (S. 11-39). Berlin: Springer.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Arnold, W., Eysenck, H. J. & Meili, R. (Hrsg.). (1996). *Lexikon der Psychologie*. Freiburg: Herder Verlag GmbH.

Asendorpf, J. B. & Neyer F. J. (2012). *Psychologie der Persönlichkeit.* Berlin, Heidelberg: Springer-Verlag

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,* 1235-1245.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, Nay-saying, and Going to Extremes: Black-White Differences in Response Styles. *Public Opinion Quarterly, 48*, 491-509.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research*, *38*, 143-156.

Becker, P. (2003). Persönlichkeitsfragebogen. In K. Kubinger (Hrsg.), *Schlüsselbegriffe der psychologischen Diagnostik* (S. 332-337). Weinheim, Basel, Berlin: Beltz Verlag.

Billiet, J. B. & McClendon, M. J. (2000). Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items. *Structural Equation Modeling, 7*, 608-628.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M., & Novick, M. R. (Hrsg.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Böckenholt, U. (2012). Modeling Multiple Response Processes in Judgment and Choice. *Psychological Methods*. Advance online publication.

Bolt, D. M., & Johnson, T. R. (2009). Addressing Score Bias and Differential Item Functioning Due to Individual Differences in Response Style. *Applied Psychological Measurement, 33*, 335-352.

Bolt, D. M., & Newton, J. R. (2011). Multiscale Measurement of Extreme Response Style. *Educational and Psychological Measurement, 71*, 814-833.

Borkenau, P. & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI) nach Costa und McCrae. Handanweisung*. Göttingen: Hogrefe.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler.* Heidelberg: Springer Medizin Verlag

Buckley, J. (2009, June). *Cross-National Response Styles in International Educational Assessments: Evidence from PISA 2006*. NCES Conference on the Program for International Student Assessment: What We Can Learn from PISA, Washington DC. Heruntergeladen von: http://edsurveys.rti.org/PISA/

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion.* München: Pearson Education Deutschland GmbH.

Chen, C., Lee, S.-Y., & Stevenson, H. W. (1995). Response Style and Cross-Cultural Comparisons of Rating Scales Among East Asian and North American Students. *Psychological Science, 6*, 170-75.

Chun, K.-T., Campbell, J. B., & Yoo, J. H. (1974). Extreme response style in cross-cultural research: A reminder. *Journal of Cross-Cultural Psychology, 5*, 465-480.

De Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P. & Baumgartner, H. (2008). Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation. *Journal of Marketing Research, 45*, 104-115.

Dolnicar, S. & Grun, B. (2009). Response Style Contamination of Student Evaluation Data. *Journal of Marketing Education, 31*, 160-172.

Ekman, P. & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17 (2),* 124-129.

Elfering, A. & Grebner, S. (2011). On the Intra- and Interindividual Differences in the Meaning of Smileys. *Swiss Journal of Psychology, 70 (1)*, 13-23.

Faulbaum, F., Prüfer, P. & Rexroth, M. (2009). *Was ist eine gute Frage? Die systematische Evaluation der Fragenqualität.* Wiesbaden: VS Verlag.

Fienberg, S. E. (1970). Quasi-Independence and Maximum Likelihood Estimation in Incomplete Contingency Tables. *Journal of the American Statistical Association, 65*, 1610-1616.

Gail, M. H. (1972). Mixed Quasi-Independent Models for Categorical Data. *Biometrics, 28(3)*, 703-712.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423-436.

Goodman, L. A. (1994). On Quasi-Independence and Quasi-Dependence in Contingency Tables, with Special Reference to Ordinal Triangular Contingency Tables. *Journal of the American Statistical Association, 89*, 1059-1063.

Hamamura, T., Heine, S. J., & Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences, 44*, 932-942.

Hambros, K. (2003). Zumutbarkeit. In K. Kubinger (Hrsg.), *Schlüsselbegriffe der psychologischen Diagnostik* (S. 445-448). Weinheim, Basel, Berlin: Beltz Verlag.

Harzing, A.-W. (2006). Response Styles in Cross-national Survey Research. A 26-country Study. *International Journal of Cross Cultural Management, 6(2)*, 243-266.

Hee Seo, M., Xu, X. & von Davier M. (2007, 2009). *Software for Multidimensional Discrete Latent Trait Models. Manual*.

Hofstede, G. (2001). *Culture's consequences* (2nd ed.). Thousand Oaks, CA: Sage.

Hox, J. J., Leeuw, E. D. de, & Kreft, G. G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Hrsg.), *Measurement Error in Surveys* (pp. 445-448). Hoboken, New Jersey: John Wiley & Sons, Inc.

Hui, C. H., & Triandis, H. C. (1989). Effects of Culture and Response Format on Extreme Response Style. *Journal of Cross-Cultural Psychology, 20*, 296-309.

Jäger, R. (2004). Konstruktion einer Ratingskala mit Smilies als symbolische Marken. *Diagnostica, 50 (1)*, 31-38.

Javaras, K. N., & Ripley, B. D. (2007). An "Unfolding" Latent Variable Model for Likert Attitude Data. *Journal of the American Statistical Association, 102*, 453-463.

Johnson, T. R., & Bolt, D. M. (2010). On the Use of Factor-Analytic Multinomial Logit Item Response Models to Account for Individual Differences in Response Style. *Journal of Educational and Behavioral Statistics, 35*, 92-114.

Johnson, T., Kulesa, P., Cho, Y. L. & Shavitt, S. (2005). The Relation Between Culture and Response Styles: Evidence From 19 Countries. *Journal of Cross-Cultural Psychology, 36*, 264-277.

Kersting, M. (2003). Augenscheinvalidität. In K. Kubinger (Hrsg.), *Schlüsselbegriffe der psychologischen Diagnostik* (S. 54-55). Weinheim, Basel, Berlin: Beltz Verlag.

Khorramdel, L. & Kubinger, K. D. (2006). The Effect of Speediness on Personality Questionnaires: An Experiment on Applicants within a Job Recruiting Procedure. *Psychological Test and Assessment Modeling* [formerly: *Psychology Science Quarterly*]*, 48*, 378-397.

Khorramdel, L. & von Davier, M. (2014). Measuring Response Styles across the Big Five: A Multiscale Extension of an Approach using Multinomial Processing Trees. *Multivariate Behavioral Research, 49*, 161-177.

Kieruj, N. D. & Moors, G. (2010). Variations in response style behavior by response scale format in attitude research. *International Journal of Public Opinion Research, 22(3)*, 320-342.

Kieruj, N. D. & Moors, G. (2012). Response style behavior: question format dependent or personal style?. *Quality & Quantity, 47(1)*, 193-211.

Kubinger, K. D. (2003). (Un-) Verfälschbarkeit. In K. Kubinger (Hrsg.), *Schlüsselbegriffe der psychologischen Diagnostik* (S. 29-32). Weinheim, Basel, Berlin: Beltz Verlag.

Kubinger, K. D. (2009). *Psychologische Diagnostik*. Göttingen: Hogrefe Verlag GmbH & Co.KG.

Kunin, T. (1955). The construction of a new type of attitude measure. *Personnel Psychology, 8*, 65-78.

Kunin, T. (1998). The construction of a new type of attitude measure. *Personnel Psychology, 51*, 823-824.

Lefevre, S. & Kubinger, K. D. (2004). *Differentielles Stressinventar HR. Manual.* Mödling: Schuhfried GmbH.

McCrae, R. R. & Costa, P. T. Jr. (1983). Joint factors in self-reports and ratings. Neuroticism, extraversion, and openness to experience. *Journal of Personality and Social Psychology, 52*, 1258-1265.

Moors, G. (2003). Diagnosing Response Style Behavior by Means of a Latent-Class Factor Approach. Socio-Demographic Correlates of Gender Role Attitudes and Perceptions of Ethnic Discrimination Reexamined. *Quality & Quantity, 37*, 277-302.

Moors, G. (2009). Ranking the ratings: a latent-class regression model to control for overall agreement in opinion research. *International Journal of Public Opinion Research, 22(1)*, 93-119.

Moors, G. (2012). The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. *European Journal of Work and Organizational Psychology, 21(2)*, 271-298.

Moosbrugger, H. & Kelava, A. (2012). *Testtheorie und Testkonstruktion*. Berlin, Heidelberg: Springer-Verlag.

Morren, M., Gelissen, J., & Vermunt, J. (2012). The Impact of Controlling for Extreme Responding on Measurement Equivalence in Cross-Cultural Research. *Methodology, 8(4)*, 159-170.

Paulhus, D. L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Hrsg.), *Measures of personality and social psychological attitudes* (S. 17-59). San Diego, CA: Academic Press.

Perleth, C. (2003). Psychologisch-diagnostische Verfahren. In K. Kubinger (Hrsg.), *Schlüsselbegriffe der psychologischen Diagnostik* (S. 433-439). Weinheim, Basel, Berlin: Beltz Verlag.

Pothmann, R. (1996). Klinische Schmerzdiagnostik bei Kindern. In H.-D. Basler, C. Franz, B. Kröner-Herwig, H. P. Rehfisch & H. Seemann (Hrsg.), *Psychologische Schmerztherapie* (S. 307-315). Berlin, Heidelberg: Springer Verlag

Rasch, G. (1960). *Studies in mathematical psychology: Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche (Expanded Edition, Chicago, University of Chicago Press, 1980).

Reynolds, N. & Smith, A. (2010). Assessing the Impact of Response Styles on Cross-Cultural Service Quality Evaluation: A Simplified Approach to Eliminating the Problem. *Journal of Service Research, 13(2)*, 230-243.

Rost, J. (2004). *Lehrbuch Testtheorie-Testkonstruktion*. Bern: Verlag Hans-Huber.

Rost, J., Carstensen, C. H. & von Davier, M. (1999). Sind die Big Five Rasch-skalierbar? Eine Reanalyse der NEO-FFI-Normierungsdaten. *Diagnostica, 45(3)*, 119-127.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34(4, Pt. 2)*.

Samejima, F. (1997). Graded Response Model. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of Modern Item Response Theory* (S. 85-100). New York, NY: Springer.

Saris, W. E., & Aalberts, C. (2003). Different Explanations for Correlated Disturbance Terms in MTMM Studies. *Structural Equation Modeling, 10 (2)*, 193-213.

Saris, W. E., Satorra, A., & Coenders, G. (2004). A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design. *Sociological Methodology, 34(1)*, 311-437.

Schaarschmidt, U. & Fischer, A. W. (1999). *IPS – Inventar zur Persönlichkeitsdiagnostik in Situationen. Handanweisung*. Frankfurt: Swets & Zeitlinger.

Schwarz, G. (1978). Estimating the dimension of a model. The *Annuals of Statistics, 6*, 461-464.

Seiwald, B. (2003). Antwortformat. In K. Kubinger (Hrsg.), *Schlüsselbegriffe der psychologischen Diagnostik* (S. 23-28). Weinheim, Basel, Berlin: Beltz Verlag.

Seiwald, B. (2003). Antworttendenzen. In K. Kubinger (Hrsg.), *Schlüsselbegriffe der psychologischen Diagnostik* (S. 29-32). Weinheim, Basel, Berlin: Beltz Verlag.

Seiwald, B. (2003). Item. In K. Kubinger (Hrsg.), *Schlüsselbegriffe der psychologischen Diagnostik* (S. 235-237). Weinheim, Basel, Berlin: Beltz Verlag.

van Herk, H., Poortinga, Y. H. & Verhallen, T. M. M. (2004). Response Styles in Rating Scales: Evidence of Method Bias in Data From Six EU Countries. *Journal of Cross-Cultural Psychology*, *35*, 346-360.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8-28.

von Davier, M. & Khorramdel, L. (2013). The problem of differentiating between response styles and construct related responses: A new IRT approach using bifactor and second-order IRT models. In: *New Developments in Quantitative Psychology: Presentations from the 77th Annual Psychometric Society Meeting* (S.463-488). New York: Springer.

von Davier, M., Rost, J., & Carstensen, C. H. (2007). Introduction: Extending the Rasch Model. In M. von Davier & C. H. Carstensen (Hrsg.), *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications* (S. 1-12). New York: Springer.

Weijters, B., Geuens, M., & Schillewaert, N. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*, 409-422.

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The Stability of Individual Response Styles. *Psychological Methods, 15*, 96-110.

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The Individual Consistency of Acquiescence and Extreme Response Style in Self-Report Questionnaires. *Applied Psychological Measurement, 34*, 105-121.

Welkenhuysen-Gybels, J., Billiet, J., & Cambré, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-Cultural Psychology, 34(6)*, 702-722.

5 Tabellenverzeichnis

6 Abbildungsverzeichnis

7 Anhang

Anhang I: Begrüßungstext für den Online-Fragebogen

Hallo!

Hast du Stress in Studium oder Arbeit? Nutzt du deine Ressourcen optimal? Weißt du, wo deine Stärken und Schwächen liegen? Im Rahmen unserer Diplomarbeiten* kannst du ein Self-Assessment zum Thema Stress und Persönlichkeit ausfüllen, bei dem du eine individuelle Rückmeldung erhältst.

Mit dem folgenden Fragebogen wird dein Umgang mit Stress im Alltag erhoben sowie Persönlichkeitseigenschaften, die für ein erfolgreiches Studieren oder im Arbeitsleben relevant sein können. Mit Stress gut umzugehen, ist beispielsweise bei Prüfungen oder Projektabgaben wichtig. Auch Persönlichkeitseigenschaften, wie gute soziale Kompetenzen, ein positiver Umgang mit den eigenen Emotionen oder ein gewissenhafter Arbeitsstil können hilfreich sein. Probleme in diesen Bereichen können unter anderem ein Vorankommen im Studium bzw. die Zusammenarbeit mit Kollegen und Kolleginnen erschweren.

Die Bearbeitungsdauer liegt ungefähr zwischen 20 und 30 Minuten. Wir würden dich ersuchen, den Fragebogen vollständig zu bearbeiten. Dies ist auch in deinem Sinne, da wir dir nur in diesem Fall eine individuelle Rückmeldung deiner Ergebnisse mit hilfreichen Links und Tipps anbieten können.

Selbstverständlich werden deine Daten vertraulich behandelt.

Falls Fragen oder Unklarheiten auftauchen, kannst du dich gerne an uns wenden.

Wir bedanken uns für deine Mitarbeit!

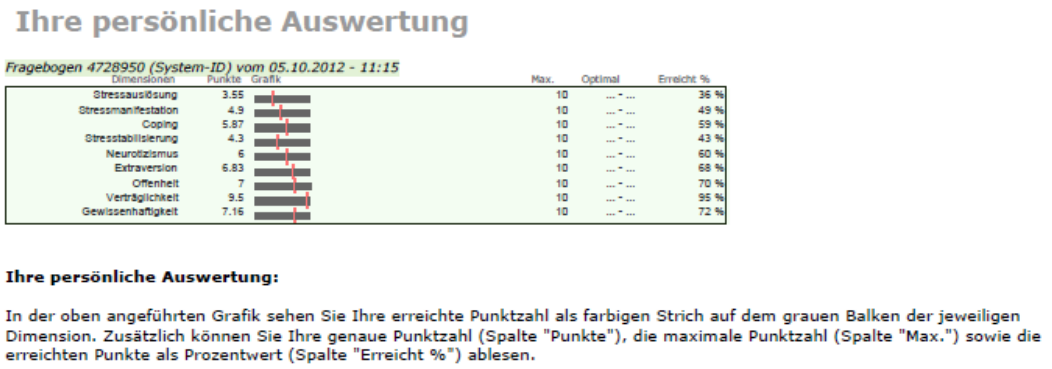
Unsere Kontaktdaten:

Alina Bugelnig
(alina.bugelnig@univie.ac.at)

Claudia Krutis
(claudia.krutis@univie.ac.at)

* Es handelt sich um Diplomarbeiten bezüglich der Qualitätssicherung von Verfahren am Fachbereich psychologische Diagnostik der Fakultät für Psychologie der Universität Wien.

Anhang II: Ergebnispräsentation und Rückmeldetexte des Online-Fragebogens



Ihre persönliche Auswertung

Fragebogen 4728950 (System-ID) vom 05.10.2012 - 11:15

Dimensionen	Punkte	Grafik	Max.	Optimal	Erreicht %
Stressauslösung	3.55		10	... - ...	36 %
Stressmanifestation	4.9		10	... - ...	49 %
Coping	5.87		10	... - ...	59 %
Stresstabilisierung	4.3		10	... - ...	43 %
Neurotizismus	6		10	... - ...	60 %
Extraversion	6.83		10	... - ...	68 %
Offenheit	7		10	... - ...	70 %
Verträglichkeit	9.5		10	... - ...	95 %
Gewissenhaftigkeit	7.16		10	... - ...	72 %

Ihre persönliche Auswertung:

In der oben angeführten Grafik sehen Sie Ihre erreichte Punktzahl als farbigen Strich auf dem grauen Balken der jeweiligen Dimension. Zusätzlich können Sie Ihre genaue Punktzahl (Spalte "Punkte"), die maximale Punktzahl (Spalte "Max.") sowie die erreichten Punkte als Prozentwert (Spalte "Erreicht %") ablesen.

Darstellung der Scores für jede Skala aus dem Online-Fragebogen.

--- Stressauslösung

Sie haben **2 - 4 von 10 Punkten.**

Die Dimension Stress*Auslösung* beschäftigt sich damit, durch welche Bereiche Stress ausgelöst wird. Diese Stressauslöser können sich auf drei unterschiedliche Bereiche beziehen. Diese sind das Alltagsgeschehen (unvermeidliche Angelegenheiten des täglichen Lebens), Interaktion (berufliche und private Kontakte mit anderen Menschen) sowie Existenzängste und Zukunftssorgen. Ein **hoher Punktwert** gibt an, dass die Gesamtanzahl an vorhandenen Auslösern von Stress hoch ist. Personen mit einem **niedrigen Punktwert** geben an, durch wenige Dinge gestresst zu sein.

Tipps
Bei einem hohen Punktwert gibt es unterschiedliche Möglichkeiten den Stressauslösern entgegenzuwirken. Das Ziel besteht darin, äußere Belastungen zu verändern. Der Alltag sollte möglichst stressfrei sein. Dies wiederum kann auf unterschiedlichen Wegen passieren: Zum einen ist es wichtig, egal ob beruflich, im Studium oder im privaten Bereich, die (Arbeits-)Organisation zu optimieren. Dazu gehört, dass Prioritäten gesetzt werden, eine realistische Zeiteinteilung bereits im Vorfeld geplant wird und auch das Delegieren von Aufgaben, die nicht unbedingt selbst erledigt werden müssen. Des Weiteren können Kommunikations-, Sozial und Problemlösekompetenzen verbessert bzw. kritisch reflektiert werden, um eine Atmosphäre mit Mitmenschen zu schaffen, in der man sowohl um Hilfe bitten, als auch Grenzen setzen kann (z.B. „nein“ sagen oder „ohne mich“), um zusätzlichen Stress zu vermeiden.

--- Stress*Manifestation*

Sie haben **4 - 6 von 10 Punkten.**

Die Dimension Stress*Manifestation* beschreibt die Anzahl der auftretenden Ausprägung von Stress. Diese Ausprägungen können sich sowohl auf der körperlichen (z.B. Schmerzen, Infektionsanfälligkeit, Magenbeschwerden) als auch auf der emotional-kognitiven Ebene (z.B. Unlustgefühl, Konzentrationsschwierigkeiten, Stimmungsschwankungen) zeigen. Ein **hoher Punktwert** gibt an, dass die

Gesamtanzahl an auftretenden Ausprägungen von Stress hoch ist und viele unterschiedliche Formen von Stress verspürt werden. Personen mit einem **niedrigen Punktwert** hingegen beschreiben, wenige unterschiedliche Formen von Stress zu erleben.

Tipps
Bei einem hohen Punktwert ist das Ziel, die körperlichen sowie die emotional-kognitiven *Manifestation*en von Stress zu verringern. Wichtig ist dabei, regelmäßig für Entspannungs- und Erholungsmöglichkeiten zu sorgen, um langfristig die eigene Belastbarkeit zu stärken. Eine gesunde Ernährung, regelmäßiger Ausgleich durch Hobbys, Freizeitaktivitäten und Sport sowie außerberufliche soziale Kontakte können die Stress*Manifestation*en verringern. Des Weiteren helfen kleine Pausen, die im Tagesablauf eingeplant und auch ausgeführt werden sowie genügend Schlaf. Ein weiterer Tipp sind Entspannungstechniken, die bei regelmäßiger Anwendung eine Verbesserung des Stresserlebens bewirken können. (z.B. Progressive Relaxation nach Jacobson http://de.wikipedia.org/wiki/Progressive_Muskelentspannung)

--- Bewältigungsmöglichkeiten

Sie haben **4 - 6 von 10 Punkten.**

Die Dimension Bewältigungsmöglichkeiten beschreibt, wie viele unterschiedliche Arten und Strategien mit Stress umzugehen und ihn zu verringern man tatsächlich verwendet. Diese Möglichkeiten teilt man in jene auf, die durch positive Emotionen und Kognitionen eine Bewältigung ermöglichen (z.B. sich selber Mut zusprechen, an etwas Schönes denken) und jene, die dies durch ein aktives Vorgehen gegen die Stressursache tun (z.B. Prioritätensetzung, um Hilfe bitten, Deligieren von Aufgaben). Personen mit einem **hohen Punktwert** geben an, dass sie viele verschiedene Bewältigungsmöglichkeiten anwenden, um ihren Stress zu verringern. Ein **niedriger Punktwert** hingegen sagt aus, dass wenige unterschiedliche Stressbewältigungsstrategien verfolgt werden.

Tipps
Bei einem niedrigen Punktwert ist es hilfreich, sich mit sich selbst und dem erlebten Stress auseinanderzusetzen. Unter dem unten angeführten Link findet man kurz zusammengefasst die hier bereits unter den Punkten Stress*Manifestation*, Stress*Auslösung* und Stress*Stabilisierung* angeführten Möglichkeiten zur Stressbewältigung.
http://www.studentenberatung.at/studentenberatung/de/stressbewaeltigung-i-iv.htm#stressbewaeltigung Für eine detaillierte Beschreibung empfehlen wir das Buch: Gert Kaluza, Stressbewältigung , ISBN 978-3-642-13719-8

--- Stress*Stabilisierung*

Sie haben **4 - 6 von 10 Punkten.**

Die Dimension Stress*Stabilisierung* beschreibt, welche Faktoren dazu führen, dass es zu einer chronischen Belastung durch Stress kommen kann. Diese Faktoren können von

außen (z.B. Zeitdruck, Gefühl von Machtlosigkeit) oder von innen (z.B. gedankliche Beschäftigung mit Stress, die Annahme, Stress sei wichtig um Erfolg zu haben) eine Stressbelastung verstärken. Personen mit einem **hohen Punktwert** geben an, dass sie durch viele persönliche Einstellungen ein stärkeres Beanspruchungsgefühl erleben als andere und sich dadurch eine chronische Belastung ergeben könnte. Ein **niedriger Punktwert** hingegen sagt aus, dass die persönlichen Einstellungen das Belastungsgefühl nicht erhöhen.

Tipps
Bei einem hohen Punktwert besteht das Ziel darin, sich kritisch mit den eigenen stresserzeugenden und verstärkenden Einstellungen auseinanderzusetzen und diese im Lauf der Zeit zu verändern, um förderliche Denkweisen zu entwickeln. Einige Beispiele für Einstellungsveränderung: Leistungsansprüche sollen überprüft werden, um sie den eigenen Leistungsgrenzen anpassen zu können. Schwierigkeiten sollen positiv, als Herausforderung gesehen, anstatt als Bedrohung verstanden zu werden. Persönliche Distanzierung von alltäglichen Aufgaben solle gefördert werden, anstatt sich mit den Problemen zu identifizieren.

--- **Neurotizismus**

Sie haben **4 - 6 von 10 Punkten.**

Die Dimension Neurotizismus erfasst die emotionale Stabilität eines Menschen. Personen mit einem **hohen Punktwert** geben an, dass sie öfters negative Gefühle haben und diese zum Teil als überwältigend empfinden. Sie geben an beispielsweise traurig, ängstlich, nervös, erschüttert, betroffen, beschämt, unsicher und verlegen zu reagieren. Sie beschreiben, eher leicht aus dem seelischen Gleichgewicht zu geraten. Sie geben an, sich viele Sorgen zu machen und eine Neigung zu unrealistischen Ideen zu haben. Personen mit einem **niedrigen Punktwert** beschreiben sich als eher emotional stabil. Sie schreiben sich selbst Attribute wie sorgenfrei, ruhig und ausgeglichen zu und geben an, dass sie nur schwer aus der Ruhe zu bringen sind und in Stresssituationen nicht so schnell aus der Fassung geraten.
In Bezug auf einen positiven Umgang mit Stress (bzw. dessen Reduktion oder Vermeidung) ist hier ein niedriger Punktwert vorteilhaft. Dieser spricht für eine hohe Frustrationstoleranz bzw. Belastbarkeit, die für die positive Bewältigung von unvorhergesehenen Hindernissen und Rückschlägen in Studium oder Beruf wichtig ist. Geringer Neurotizismus deutet weiters auf eine gute Distanzierungsfähigkeit hin – dies bedeutet, dass man nicht ständig über die Arbeit nachgrübelt, sondern sich gedanklich auch mit anderen Dingen, die nicht mit Stress verbunden sind, beschäftigen kann.

Tipps
Bei einem hohen Punktwert kann es hilfreich sein, den „Blick auf das Positive“ in Stresssituationen zu verstärken und die positiven Aspekte bewusst hervorzuheben („Finde ich etwas Gutes in dieser Situation?“). Ebenso solle man sich seine eigenen Stärken und Ressourcen („Hat es in meinem Leben schon schwierige Situationen gegeben? Und wie habe ich diese gemeistert?“) bewusst machen, um dadurch ein „Nachgrübeln“ über einen negativen Ausgang der Situation zu verringern. Zusätzlich kann es hilfreich sein, den Blickwinkel zu verändern und zu versuchen, das stressauslösende Problem in einem anderen Licht zu betrachten („Wie sehen andere

Menschen die Situation? Warum sind sie dadurch weniger belastet?“). Zu guter Letzt ist es förderlich, sich mit den positiven Konsequenzen eines Problems auseinanderzusetzen und sich darüber Gedanken zu machen, wie es sein wird, wenn man die Situation erfolgreich bewältigt hat.

--- Extraversion

Sie haben **6 - 8 von 10 Punkten.**

Die Dimension Extraversion erfasst die Geselligkeit und Aktivität eines Menschen. Personen mit einem **hohen Punktwert** in dieser Dimension erleben sich selbst als optimistisch, heiter, energisch, gesprächig, aktiv und selbstsicher. Sie geben an gerne unter Menschen, in Gruppen und auf gesellschaftlichen Versammlungen zu sein. Sie beschreiben außerdem eine Neigung zu einem heiteren Gemüt. Personen mit einem **niedrigen Punktwert** in dieser Dimension beschreiben sich als eher zurückhaltend, unabhängig und weniger aktiv bzw. ausgeglichen. Sie geben an gerne allein Zeit zu verbringen, jedoch ohne dabei pessimistisch oder unglücklich zu sein. In Bezug auf einen positiven Umgang mit Stress (bzw. dessen Reduktion oder Vermeidung) kann Extraversion wichtig sein, weil es eher extravertierten Personen oft leichter fällt, Kontakte zu knüpfen und sich ein soziales Netzwerk aufzubauen. Ein soziales Netzwerk kann wiederum in bestimmten Lebenssituationen und vor allem bei Stress wichtige und hilfreiche Unterstützung bieten. Hohe Extraversion wird oft von hoher sozialer Kompetenz begleitet. Mit einer eher hohen sozialen Kompetenz fällt es auch leicht, gut mit Kollegen und Kolleginnen in Studium oder Beruf auszukommen und zusammenzuarbeiten, sowie diesen die eigenen Meinungen und Ideen mitzuteilen. Ein gutes Auskommen mit anderen Personen kann wiederum dazu beitragen, sozialen Stress zu reduzieren.

Tipps
Bei niedrigen Punktwerten kann ein aktiveres Zugehen auf anderen Menschen in Stresssituationen insofern förderlich sein, als dass diese Aufgaben übernehmen oder emotionale Unterstützung anbieten können. Auch wenn Personen mit niedrigen Werten oft
lieber alleine arbeiten, sollte man sich darüber bewusst sein, dass in Stresssituationen soziale Kontakte eine Entlastung bringen können. Diesbezüglich kann eventuell auch ein soziales Kompetenztraining helfen
(http://de.wikipedia.org/wiki/Training_sozialer_Kompetenzen).

--- Offenheit für Erfahrungen

Sie haben **6 - 8 von 10 Punkten.**

Die Dimension Offenheit für Erfahrungen erfasst den Umgang mit neuen Erfahrungen, Eindrücken und Erlebnissen. Personen mit einem **hohen Punktwert** beschreiben sich als fantasievoll, wissbegierig, künstlerisch interessiert, intellektuell und experimentierfreudig. Sie geben an, häufig die eigenen positiven und negativen Gefühle sehr deutlich wahrzunehmen und sich für viele öffentliche und persönliche Begebenheiten zu interessieren. Sie beschreiben, dass sie eine höhere Bereitschaft zur kritischen Hinterfragung bestehender Normen und zum Eingehen auf neuartige

Wertvorstellungen im ethischen, sozialen und politischen Bereich haben. Sie geben an, dass ihre Urteile eher unabhängig sind, sie Abwechslung bevorzugen und gerne neue Handlungsweisen ausprobieren. Ihr Verhalten beschreiben sie als unkonventionell. Personen mit einem **niedrigen Punktwert** beschreiben dagegen eher eine Neigung zu konservativen Einstellungen und konventionellem Verhalten. Sie geben an, häufig Bewährtes und Bekanntes lieber zu mögen als Neues und tragen ihre Emotionen weniger stark nach außen. In Bezug auf den Umgang mit Stress kann eine niedrige Offenheit zu Frustration und Stress führen, da man oft in alten Lösungsansätzen und Ideen verhaftet bleibt, selbst wenn diese sich nicht umsetzen lassen.

Tipps
Bei einem niedrigen Punktwert sollte man versuchen, andere Meinungen, Erfahrungen und Ansichten zu berücksichtigen, um zum Beispiel neue Lösungsmöglichkeiten für Probleme zu finden. So können auch eigene neue Ideen angeregt werden. Außerdem erweitert das Interesse an neuen Inhalten das eigene Wissen und somit kann die Leistung in Studium oder Beruf verbessert werden. Dies kann zu Reduktion oder Vermeidung von Stress führen. Dabei ist jedoch zu beachten, dass nicht zu viele Veränderungen auf einmal vorgenommen werden sollten, um für eine gewisse Stabilität in Lern- bzw. Berufsumfeld zu sorgen.

--- Verträglichkeit

Sie haben **8 - 10 von 10 Punkten.**

Die Dimension Verträglichkeit erfasst den eigenen Umgang mit anderen Menschen. Personen mit einem **hohen Punktwert** beschreiben sich als verständnisvoll, wohlwollend und mitfühlend gegenüber anderen Menschen und versuchen, diesen zu helfen. Dabei gehen sie davon aus, dass diese ebenso hilfsbereit sind. Ein hoher Punktwert steht für ein starkes Harmoniebedürfnis und die Neigung zu Nachgiebigkeit, Kooperation und zwischenmenschlichen Vertrauen. Personen mit einem **niedrigen Punktwert** beschreiben sich als eher misstrauisch gegenüber den Absichten anderer Menschen und als egozentrisch und antagonistisch. Sie geben an, dass sie lieber mit anderen wetteifern, als mit diesen zusammenzuarbeiten. Sie beschreiben, dass es ihnen oft leichter fällt für ihre eigenen Interessen zu kämpfen und neigen meist zu Skepsis und Misstrauen. In Bezug auf einen positiven Umgang mit Stress (bzw. dessen Reduktion oder Vermeidung) ist Verträglichkeit – ebenso wie Extraversion – für den sozialen Umgang wichtig. Wie oben bereits erwähnt (siehe Dimension Extraversion), geht es hier um ein gute Zusammenarbeit mit Kollegen und Kolleginnen, das Knüpfen von neuen Kontakten und eine gute Kommunikation. Bei zu großem Misstrauen oder Skepsis gegenüber anderen kann es zu Konflikten kommen, die in weitere Folge wieder den Stress erhöhen.

Tipps
Bei einem niedrigen Punktwert ist (ähnlich wie in der Dimension Extraversion) das Zugehen auf andere Personen wichtig. So kann man zum Beispiel Kollegen und Kolleginnen um Feedback bitten, wie man auf Andere wirkt beziehungsweise wie das Gesagte von Anderen aufgefasst/aufgenommen wird. So kann man Missverständnissen vorbeugen oder diese klären und zu einer Konfliktlösung beitragen. Auch kann man üben, anderen zuzuhören oder sie ausreden zu lassen und Konfliktgespräche auf eine

eher sachliche Ebene zu verlagern. Hier kann ebenfalls ein soziales Kompetenztraining helfen (http://de.wikipedia.org/wiki/Training_sozialer_Kompetenzen).

--- Gewissenhaftigkeit

Sie haben **6 - 8 von 10 Punkten.**

Die Dimension Gewissenhaftigkeit erfasst eine Art der Selbstkontrolle, nämlich die Planung, Organisation und Durchführung von Aufgaben. Personen mit einem **hohen Punktwert** beschreiben sich als ausdauernd, systematisch, zielstrebig, ordentlich, ehrgeizig, fleißig, penibel, genau, willensstark, pünktlich, zuverlässig und diszipliniert. Extrem hohe Ausprägungen können sich eventuell in zwanghafter Ordentlichkeit, einem übertrieben hohen Anspruchsniveau oder in Formen von Arbeitssucht äußern. Personen mit einem **niedrigen Punktwert** beschreiben eher eine Neigung zur Nachlässigkeit, Gleichgültigkeit und Unbeständigkeit. Sie geben an, meist ein geringeres Engagement im Verfolgen der eigenen Ziele zu zeigen. In Bezug auf einen positiven Umgang mit Stress (bzw. dessen Reduktion oder Vermeidung) erweist sich eine hohe Gewissenhaftigkeit oft als vorteilhaft, da Personen mit einem genauen Arbeitsstil Fehler vermeiden, die zu Stress führen können. Auch ein gutes Zeitmanagement sowie eine gute Planungs- und Organisationsfähigkeit können Stress im Studium oder im Arbeitsalltag vermeiden oder verringern.

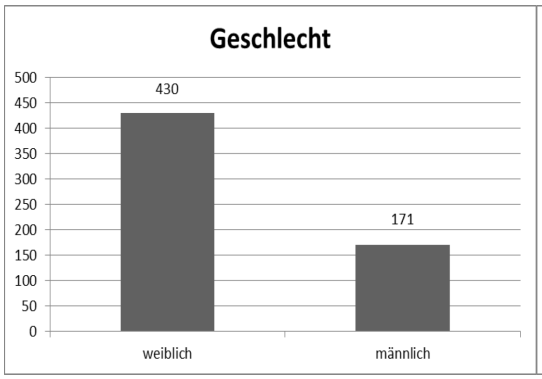
Tipps
Bei einem niedrigen Punktwert kann es sehr hilfreich sein, ein gezieltes Zeitmanagement und eine gute Prioritätensetzung in den Alltag zu integrieren. Hier sind hilfreiche Links zu den angesprochenen Themen für Studium und Beruf:
http://www.metacom.com/zeitmanagement-tipps.php
http://www.studentenberatung.at/studentenberatung/de/zeitmanagement.htm#Bedeutung
www.studentenberatung.at/studentenberatung/de/lernen-mit-erfolg.htm#Lern-und-Arbeitsplanung

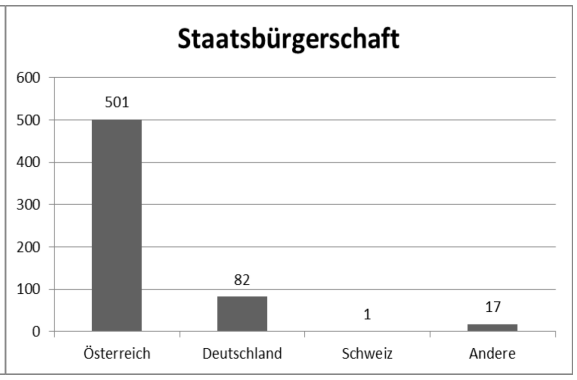
---
Im Internet gibt es viele hilfreiche Tipps und Tricks zur Stressbewältigung sowie Erklärungen, was Stress ist bzw. wie er zustande kommt. Um Ihnen eine mühsame Suche zu ersparen, haben wir unten eine Sammlung von passenden Links zusammengestellt und hoffen, dass diese in Kombination mit den obigen Ergebnissen der Fragebögen für Sie hilfreich sind:

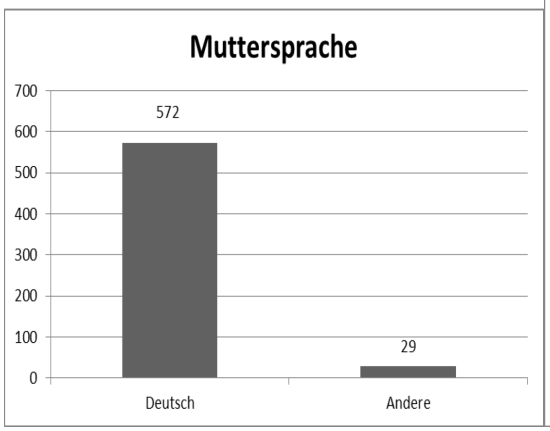
Weiterführende Links:
http://de.wikipedia.org/wiki/Stress
http://arbeitsblaetter.stangl-taller.at/TEST/STRESS/Test.shtml
http://arbeitsblaetter.stangl-taller.at/STRESS/Stressbewaeltigung-Uebungen.shtml
https://www.gesundheit.gv.at/Portal.Node/ghp/public/content/gesund-leben-stress-arbeit.html
http://www.arbeiterkammer.com/online/stress-19509.html
http://www.pfiffprojekt.de/workshop/fundus/stress_stressbewaeltigung/uebungen%20Stress%20und%20*Coping*%20komplett.pdf
http://www.zeitblueten.com/news/2010/entspannungsuebungen/

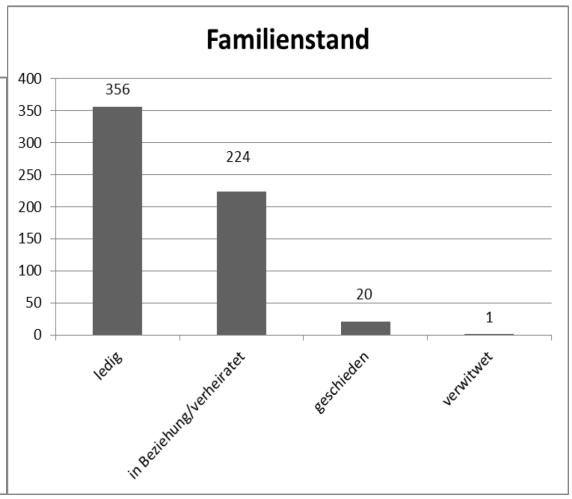
http://www.hauptsache-bildung.de/2012/burnout-praevention-10-uebungen-zur-stressbewaeltigung-am-arbeitsplatz
http://www.stresscoach.at/tipps/index.html
http://studium.lerntipp.at/stress/stressuebungen.shtml

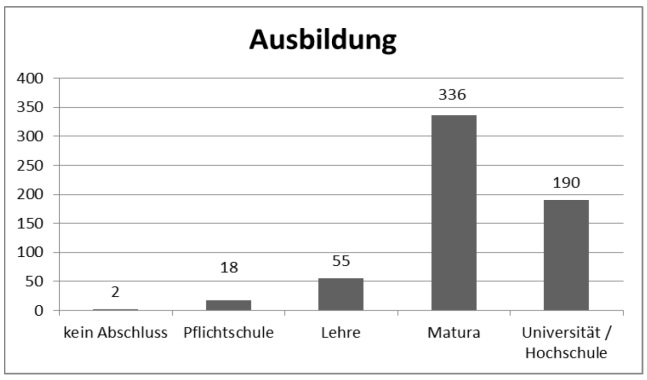
Anhang III: Übersicht über die demografischen Daten der an der Untersuchung mit dem NEO-FFI teilnehmenden Testpersonen

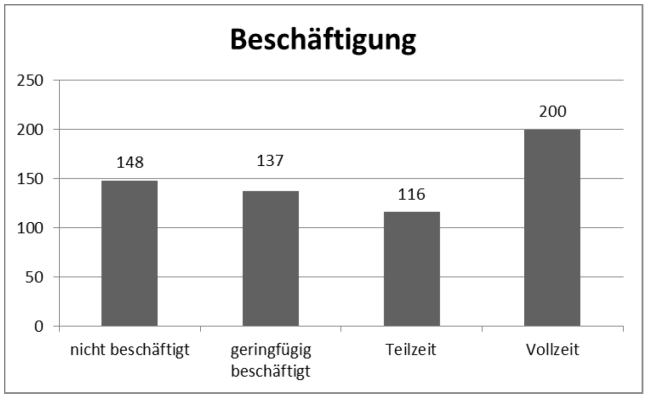












Anhang IV: empirische Skaleninterkorrelationen

Empirische Interkorrelationen der Score-Verteilung der Big Five-Dimensionen aus dem 5-dimensionalen IRT-Modell für die Pseudoitemtypen e, d und m sowie für die Originalitems des NEO-FFI

	Neurotizismus	*Extraversion*	*Offenheit*	*Verträglichkeit*	*Gewissenhaftig-keit*
e-Items:					
Neurotizismus	1,00000				
Extraversion	0,59452	1,00000			
Offenheit	0,52625	0,49671	1,00000		
Verträglichkeit	0,54003	0,59752	0,47749	1,00000	
Gewissenhaftig.	0,55387	0,54858	0,39119	0,58928	1,00000
m-Items:					
Neurotizismus	1,00000				
Extraversion	0,09080	1,00000			
Offenheit	0,03600	-0,01666	1,00000		
Verträglichkeit	0,02697	0,13751	0,05623	1,00000	
Gewissenhaftig.	0,04962	0,15452	0,00774	0,18780	1,00000
d-Items:					
Neurotizismus	1,00000				
Extraversion	-0,45086	1,00000			
Offenheit	-0,15166	0,07969	1,00000		
Verträglichkeit	0,27079	-0,39891	0,01991	1,00000	
Gewissenhaftig.	-0,27117	0,23855	-0,03840	-0,20292	1,00000
Originalitems des NEO-FFI:					
Neurotizismus	1,00000				
Extraversion	-0,42322	1,00000			
Offenheit	-0,10872	0,03306	1,00000		
Verträglichkeit	-0,19001	0,37962	-0,02261	1,00000	
Gewissenhaftig.	-0,26799	0,22726	-0,03954	0,22814	1,00000

Claudia Krutis
Curriculum Vitae

Ausbildung

1985-1989	Volksschule Tulln
1989-1993	Hauptschule Tulln
1993-1998	Höhere Lehranstalt für wirtschaftliche Berufe Tulln
Seit Oktober 2004	Studium der Psychologie / Universität Wien; *Spezialisierung: Psychologische Diagnostik*
Oktober 2007 bis Jänner 2008	Studium der Japanologie / Universität Wien

Berufliche Erfahrung

Juli – August 1998	Volontärin bei der APA (Austria Presse Agentur) in Wien
Oktober 1998 – Mai 1999	Front Office-Tätigkeit bei der RTR GmbH (Rundfunk- und Telekom Regulierungs-GmbH) in Wien
Mai 1999 – August 2004	Assistentin der Streitschlichtung bei der RTR GmbH (Rundfunk- und Telekom Regulierungs-GmbH) in Wien
Februar – März 2006 und 2007	AGRANA Tulln (Dateneingabe)
August 2006 – April 2012	Assistenz und Buchhaltung bei DIE GARTEN TULLN
August 2011 – Februar 2012	Pflichtpraktikum bei Frau Mag. Riedler (Praxisgemeinschaft Riedler & Schubert) in Tulln
März 2012 – Jänner 2014	Organisation der Computerdiagnostik am Arbeitsbereich der Psychologischen Diagnostik an der Universität Wien