universität wien

Diplomarbeit

Titel der Arbeit

Effekte des Itempool-Präsentationsformats: Exemplarische Analyse eines Big Five-Selbstbeschreibungsfragebogens

Verfasser:

Maximilian Armin Hetzel

Angestrebter akademischer Grad

Magister der Naturwissenschaften (Mag. rer. nat.)

Wien, im Mai 2013

Studienkennzahl:	298
Studienrichtung:	Psychologie
Betreuerin:	Mag. Dr. Christine Hohensinn

Danksagung

Ich möchte meiner Familie, allen voran meinen Eltern, tiefsten Dank ausdrücken. Die Möglichkeit des Studiums – mehr noch, die Möglichkeit eines Studiums in Wien – wäre ohne die Unterstützung meiner Mutter Elisabeth und meines Vaters Armin undenkbar gewesen. Dies umfasst bei weitem nicht nur finanzielle Unterstützung, sondern vor allem die engagierte Förderung der Findung meines Weges in der Welt.

Auch anderen Mitglieder meiner Familie möchte ich Dank sagen: meiner Großmutter Luzia, meinem Großvater Armin und meinem Patenonkel Hubert, welche mich mit ihrem aufrichtigem Interesse den ganzen Weg lang unterstützt haben.

Außerdem möchte ich mich bei meiner Diplomarbeitsbetreuerin Doktor Christine Hohensinn bedanken. Ihre Geduld und stetige Bereitschaft, sich Zeit für nicht nur knappe Hilfestellungen, sondern auch längere Gespräche und Diskussionen zu den Themen meiner Arbeit zu nehmen waren eine unschätzbare Hilfe dabei, jene Arbeit zu schreiben, die ich mir vorgestellt hatte. Mehr noch, sie und Professor Doktor Klaus Kubinger waren die beiden Personen, welche mein Interesse für die psychologische Diagnostik weckten.

„Last, but not least“, möchte ich ganz besonderen Dank an Alexander Bachinger aussprechen, der mit seiner Expertise im Programmieren und tatkräftigen Unterstützung die Realisierung meines Online-Assessments erst überhaupt möglich machte.

Zum Zwecke des Vergleiches verschiedener Möglichkeiten des Präsentationsformats eines gegeben Itempools wurden 392 Personen mit einem Big Five-Selbstbeschreibungsfragebogen in je einem von drei Präsentationsformaten in einer Online-Erhebung befragt. Untersucht wurden die Präsentationsformate eines Itempools als Liste, als disjunkte Abfolge von Seiten (One-Item-a-Time) sowie als leicht adaptierte Version des Q-Sorts. Mittelwerte der Skalensummen des Fragebogens, Ausnutzung des verfügbaren Antwortspektrums sowie separat erhobene Akzeptanz des Verfahrens durch die Teilnehmer/Innen waren zwischen den Gruppen äquivalent. Unterschiede zeigten sich in den Bearbeitungszeiten sowie den inneren Konsistenzen der Itempools der Faktoren innerhalb der jeweiligen Präsentationsformate, wobei letztere als Anwendung des Rasch-Modells implementiert wurden. Überlegenheit gegenüber den anderen Formaten war alleinig für das One-Item-a-Time-Format feststellbar, welches bessere Rasch-Modell-Passung als die anderen Formate zeigte. Deutlich erhöhte Bearbeitungszeiten als jene der anderen Formate (δ = 1.17 zum Listen-, δ = .59 zum One-Item-a-Time-Format) ohne Anwesenheit von expliziten Vorteilen lassen von der Verwendung des Q-Sorts abraten.

In order to compare different means of itempool-presentation, 392 people participated in an online study which employed a Big-Five-questionnaire, each one having received the questionnaire in a different presentation format. The compared formats were the presentation of items as a list, as a disjunctive sequence of single pages (“One-Item-a-Time”), and as a variant on the Q-Sort format. Means of the questionnaire’s scales, usage of the spectrum of answers as well as the separately measured acceptance of the instrument were equivalent between groups. Differences were found concerning the time necessary to complete the questionnaire. Groups also varied in regards to inner consistency, the analysis for which was implemented by usage of the Rasch-Model. Only the One-Item-a-Time-format proved to be slightly superior to the others, by virtue of achieving superior fit to the Rasch-Model. Since the Q-Sort-format takes considerably longer to finish in comparison to the other formats ($\delta = 1.17$ compared to listwise presentation, $\delta = 0.59$ compared to One-Item-a-Time-presentation), and since it did not show any explicit advantages over the other formats, its usage is advised against.

Inhaltsverzeichnis

I. Einleitung

Selbstbeschreibungsfragebögen gehören seit jeher zum festen methodischen Repertoire der Psychologie und all ihrer Disziplinen; sei es zur Erfassung von Facetten der Persönlichkeit, relevanter Einstellungen, Details zur Familiensituation, zur Erfragung von Meinungen über diverseste Sachverhalte (Bortz & Döring, 2006), oder zur Registrierung möglicher klinischer Symptome und deren Intensitäten (Derogatis, Lipman, Rickels, Uhlenhuth, & Covi, 1974); eine solche Auflistung möglicher Inhalte könnte lang fortgesetzt werden. Neben dem Inhalt der Fragen, Aussagen oder simplen Adjektiven der psychologisch-diagnostischen Instrumente, welche den zu diagnostizierenden Personen vorgelegt werden, interessiert die psychologische Diagnostik jedoch immer auch die Bedeutung peripherer Merkmale des zur Debatte stehenden Instruments. So ging beispielsweise vor mehr als einer Dekade Schwarz (1999) in seinem vielzitierten Artikel auf die Problematik von Item-Kontexteffekten in der psychologischen Diagnostik ein: „Self-reports of behaviours and attitudes are strongly influenced by features of the research instrument, including question wording, format and context." (S.93). In Folge dieser Aussage geht Schwarz auf die Effekte der unterschiedlichen Designs von Antwortmöglichkeiten, Ratingskalen, und vielen weiteren peripheren Merkmalen von Selbstbeschreibungsinstrumenten ein, unter anderem auch den möglichen Implikationen des Kontextes der jeweiligen Items. Obwohl die Forschung zu diesen Themen bis zum heutigen Tage anhält, gibt es doch Bereiche die wenig bis gar keine Aufmerksamkeit erhalten haben. Zu diesen vernachlässigten Randbereichen der diagnostischen Situation gehören auch die Unterschiede der Formate der Präsentation des gegebenen Itempools eines Instruments.

Zwar wurden in der Vergangenheit oftmals Fragen nach den Effekten der Reihenfolge der jeweiligen Items (zum Beispiel in Sparfeldt et al., 2006) gestellt – diese beschränken sich aber zumeist auf die Analyse der linear-sukzessiven Abfolge der Items. Dabei wird übersehen, dass derart befragte Personen in keiner Weise gezwungen sind, die gegebenen Fragen in einer derartigen Reihenfolge zu beantworten. Vielmehr wird ein Zurückkehren zu vorher gegebenen Antworten und deren Fragen, beziehungsweise ein vorläufiges Überspringen bestimmter Teile, von den meisten gängigen Instrumenten ermöglicht, in manchen Formaten sogar dazu ermutigt.

Im Zuge der Betrachtung der Forschung über Item-Kontexteffekte fällt auf, dass die Übersicht, die den befragten Personen über den gegebenen Itempool gewährt wird, oft stark variiert. Dies beginnt mit Formaten, die den Personen außer dem aktuellen Item keinerlei Überblick über die Items und die bereits gegebenen Antworten erlauben, bis hin zu Formaten, welche nicht nur einen Überblick über den gesamten Itempool gewähren, sondern die gegebenen Antworten auf eine Art und Weise darstellen, welche eine optische Repräsentation der Relationen aller

gegebenen Antworten zueinander ermöglicht. Auch in den Manualen der meisten Verfahren werden selten die Gründe für die Wahl eines bestimmten Formates angesprochen.

Da die Existenz von Verzerrungen des Antwortverhaltens durch Item-Kontexteffekte eine Realität des psychologisch-diagnostischen Alltags ist, stellt sich die Frage, ob verschiedene Formate der Itempoolpräsentation nicht zu messbaren Unterschieden im Auftreten solcher Effekte führen.

Die folgende Diplomarbeit beschäftigt sich nun mit eben dieser Fragestellung: übt das gewählte Format der Präsentation des gegebenen Itempools eines psychologisch-diagnostischen Instrumentes Einflüsse auf das Antwortverhalten der befragten Personen aus?

Im Zuge dessen sollen zunächst im theoretischen Teil der Arbeit einige unterschiedliche Formate der Itempoolpräsentation besprochen werden. Anschließend soll aus der bestehenden einschlägigen Literatur ein Überblick über verwandte Untersuchungen von Item-Kontexteffekte und deren Ergebnisse gegeben werden. Außerdem wird auf die Rolle der Akzeptanz eines Fragebogens durch die Teilnehmer von Befragungen eingegangen werden. Auch die mögliche Rolle der Item Response Theory zur Analyse solcher möglicherweise auftretenden Effekte soll angesprochen werden.

Im empirischen Teil soll daraufhin die der Fragestellung zugehörige Untersuchung dargestellt werden, in der 392 Personen einen Big-Five-Selbstbeschreibungsfragebogen in je einer von drei Itempool-Präsentationsarten sowie einen Fragebogen zu ihrer Akzeptanz des bearbeiteten Verfahrens ausfüllten. Unterschiede in innerer Konsistenz, Skalenmittelwerten, Akzeptanz und weiterer Merkmale zwischen den drei Bedingungen werden untersucht und besprochen.

II. Theoretischer Teil

1. Die verschiedenen Itempool-Präsentationsarten

Wie bereits in *I. Einleitung* angesprochen, sind die Beschreibungen des Präsentationsformats beziehungsweise Layouts diverser Fragebogen nicht selbstverständlich Teil der Dokumentation ihrer Konstruktion. Daher sollen in den folgenden Kapiteln die existierenden Präsentationsformate in Hinsicht auf ihren Aufbau und ihrer Implikationen auf die Situation der Bearbeitung durch die befragten Personen kurz dargestellt werden.

1.1. Das Listen-Format

Seit der Einführung von Selbstbeschreibungsfragebögen in die wissenschaftliche Psychologie zu Beginn des 20. Jahrhunderts (Schönpflug, 2004) waren und sind diese immer ein fester Bestandteil der normorientierten psychologischen Diagnostik von Persönlichkeitseigenschaften gewesen. Aufgrund der naheliegenden Gegebenheiten von gedruckten Fragebögen war das „Listen"-Format – sprich die Aneinanderreihung von zu verschiedenen Skalen zugehörigen Items auf einem oder mehreren bedruckten Blättern – der Status Quo. Praktisch alle heute populären Persönlichkeitsinventare verwenden in ihrer Paper-Pencil-Fassung das „Listen-Format", wie zum Beispiel der 16 PF-R (16 Persönlichkeits-Faktoren-Test, revidierte Fassung; Schneewind & Graf, 1998), das NEO-PI-R (NEO Persönlichkeitsinventar, revidierte Fassung; Ostendorf & Angleitner, 2004) oder das TIPI (Trierer Integriertes Persönlichkeitsinventar; Becker, 2003).

Auch im Bereich der computergestützten psychologischen Diagnostik ist das Format weitestgehend erhalten geblieben. So zählen eine immense Zahl von Online-Erhebungen sowie die jeweiligen Computerversionen der eben genannten Persönlichkeitsinventare zu ihren aktuelleren Vertretern.

Dabei hat das Listen-Format zwei spezifische Attribute, welche es im Kontext der hier gestellten Forschungsfrage herauszustellen gilt. Um spezifischere Aussagen treffen zu können, wird von dem häufig anzutreffenden Fall ausgegangen, dass entweder alle Items des Fragebogens auf einer Seite präsentiert sind, oder es den befragten Personen erlaubt ist, zwischen den verschiedenen Seiten vor- und zurückzublättern.

- Übersicht über den Itempool

  Es ist den befragten Personen jederzeit vor Abgabe (oder Absenden) des Fragebogens möglich, jedes einzelne Item des Itempools einzusehen. Dazu ist jedoch Vor- oder Zurückblättern, beziehungsweise Hoch- oder Hinunterscrollen, notwendig. Somit ist die Übersicht des Listen-Formats über den Itempool im Vergleich zum „One-Item-a-Time"-Format höher und im Vergleich zum Q-Sort-Format geringer (für nähere Ausführungen zu beiden siehe unten).
- Übersicht über bereits gegebene Antworten

  Die befragten Personen können sich jederzeit, nicht nur über die Items, sondern auch über die bereits von ihnen gegebenen Antworten, einen Überblick verschaffen. Dieser Überblick ist höher, wenn die Items in einer Item-mal-Ratingskala-Matrix abgebildet sind und niedriger, wenn die Items disjunkt voneinander mit einer Ratingskala pro Item hintereinander vorgegeben werden; letztere Alternative ist bei Fragebögen, welche ohnehin meist dieselbe Ratingskala für ihre Items verwenden, schon aus Platzgründen eher selten. Zudem ist es den Personen jederzeit möglich, bereits gegebene Antworten im Nachhinein nochmals zu verändern.

  Damit ist die Übersicht für die befragten Personen über ihre bereits gegebenen Antworten im Vergleich zum „One-Item-a-Time"-Format höher, im Vergleich zum Q-Sort-Format aber geringer.

1.2. Das One-Item-a-Time-Format

Durch das Aufkommen der Möglichkeit der computerisierten Vorgabe von Fragebögen ergeben sich neue Möglichkeiten des Präsentationsformats von Fragebögen, welche für Paper-Pencil-Verfahren höchst unpraktisch gewesen wären. So ist es bei einem digital vorgegebenen Fragebogen ein leichtes, jedes Item auf einer eigenen Seite zu präsentieren sowie Vor- und Zurückblättern komplett zu unterbinden, was im Folgenden als „One-Item-a-Time"-Format bezeichnet werden soll (wie z.B. im Big Five Plus One, Holocher-Ertl, Kubinger, & Menghin, 2003; und im Big Five Plus One Revised, Siptroth, 2011). Diese Möglichkeit ist zum Beispiel dann interessant, wenn es dem Diagnostiker darauf ankommt, dass die befragte(n) Person(en) auf jedes Item einzeln antworten und diese so wenig wie möglich miteinander in Verbindung bringen (siehe dazu auch unter *3. Die Rolle der Item Response Theory im Vergleich der unterschiedlichen Präsentationsarten)*.

In Analogie zum Listen-Format sollen die relevanten Attribute des One-Item-a-Time-Formats kurz dargestellt werden:

- Übersicht über den Itempool

  Im One-Item-a-Time-Format ist es den befragten Personen nicht gestattet, ein anderes als das aktuell angezeigte Item einzusehen. Weder Vor- noch Zurückblättern sind möglich; die einzige Möglichkeit, das nächste Item einzusehen, ist die Beantwortung des aktuell angezeigten. Hinzu kommt, dass die Zahl von Items in den meisten gegebenen Itempools zu hoch ist, um sie sich inzidentell oder sogar durch bewusste Anstrengung, durchgehend einzuprägen.

  Somit muss die Übersicht über den gegebenen Itempool im One-Item-a-Time-Format im Vergleich zum Listen-, wie auch zum Q-Sort-Format, als niedriger eingestuft werden.
- Übersicht über bereits gegebene Antworten

  Hier trifft der eben bereits angesprochene Sachverhalt weiterhin zu: die befragten Personen können ihre bereits gegebenen Antworten nach dem Fortschreiten zum nächsten Item nicht mehr einsehen, und aufgrund der hohen Zahl an Items ist auch ein aktives Einprägen, beziehungsweise inzidentelles Merken der vorher gegebenen Antworten, eher unwahrscheinlich. Auch das nachträgliche Abändern der Antwort auf ein Item ist nach dem Fortschreiten zum nächsten Item nicht mehr möglich. Die Übersicht über die bereits gegebenen Antworten ist also beim One-Item-a-Time-Format niedriger als bei den anderen beiden Formaten.

1.3. Das Q-Sort-Format

1935 vom britischen Physiker und Psychologen William Stephenson entwickelt (Brown, 1996), ist der Q-Sort heute im Bereich der psychologischen Persönlichkeitsdiagnostik verhältnismäßig selten anzutreffen. Nichtsdestotrotz existieren zahlreiche interessante Anwendungen der Technik, sei es in der Diagnostik der Persönlichkeit, der Evaluation von Psychotherapiefortschritten, oder in der Beschreibung psychopathologischer Wesenszüge (Bambery & Porcerelli, 2010). Trotzdem ist der Q-Sort in gängigen Lehrbüchern der psychologischen Diagnostik – wenn überhaupt – nur noch mit einer kurzen Erwähnung vertreten (zum Beispiel in Kubinger, 2009, oder auch schon früher in Lösel, 1999). Der Grund hierfür könnte sein, dass der Q-Sort ursprünglich in das zugehörige methodische Repertoire der Q-Methodologie eingebettet ist. Q-Methodologie unterscheidet sich von dem normalerweise in der Fragebogenkonstruktion üblichen Vorgehen durch den Einsatz der „Q-Faktorenanalyse“

anstatt der üblichen „R-Faktorenanalyse“. Die R-Faktorenanalyse beschreibt die Möglichkeit, aus den gefundenen Korrelationen zwischen den Items des Verfahrens ihnen übergeordnete Supra-Variablen, sogenannte Faktoren, zu extrahieren (Kubinger, 2009). Die Q-Faktorenanalyse hingegen extrahiert die übergeordneten Variablen nicht aus den Korrelationen zwischen den Items, sondern aus jenen zwischen den Antwortmustern der Personen (Block, 2008). Das Ergebnis sind also keine Supra-Variablen auf denen jede Person durch eine bestimmte Ausprägung beschrieben werden kann, sondern eine Art „Gruppen-Variablen“, die die Ähnlichkeit der Antworten der befragten Personen abbilden. Auch der in der psychologischen Diagnostik übliche nomothetische Ansatz der Beschreibung von Personen tritt hier gegenüber der detaillierten, ideographischen Beschreibung in den Hintergrund (Fluckinger, 2010).

Im Kontext der Persönlichkeitsdiagnostik bezeichnet nun das Q-Sort-Format die Vorgehensweise, der Person die zu bewertenden Adjektive oder Aussagen – dort auch Deskriptoren genannt – nicht auf einem oder mehreren Blättern, sondern auf jeweils einer Karte pro Deskriptor vorzulegen. Der Einfachheit halber wird auch im digitalen Kontext weiterhin von Karten gesprochen, obwohl es sich de facto um aktive Felder handelt. Jene Karten werden anschließend in einem Vorbereitungsschritt von der zu befragenden Person durchgesehen und in drei Stapel (je einen für generelle Zustimmung, Ambivalenz und generelles Widersprechen) sortiert. Im nächsten, eigentlich hauptsächlichen Bearbeitungsschritt, wird die befragte Person gebeten, die Karten auf einem Feld anzuordnen, welches in mehrere Spalten unterteilt ist (Abbildung 1.1). Dabei repräsentieren die verschiedenen Spalten verschiedene Stufen der Zustimmung zu den jeweiligen Deskriptoren der Karten. Die vertikale Anordnung der Karten hat dabei normalerweise keine Bedeutung (Prasad, 2001).

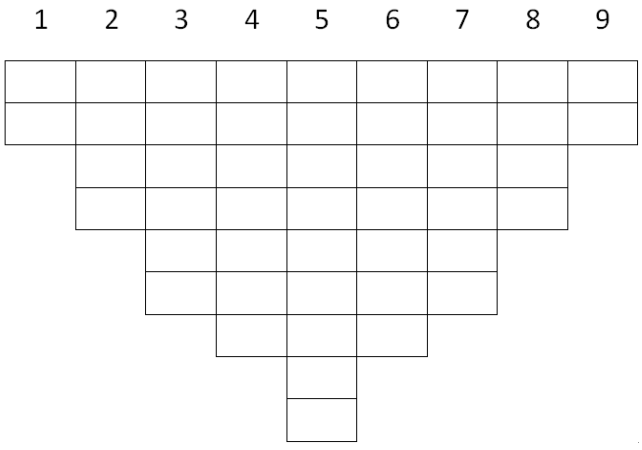


Abbildung 1.1. Exemplarisches Kartenfeld eines Q-Sorts mit Quasi-Normalverteilung

Die Verteilung, mit welcher die Personen die Karten einordnen sollen, ist dabei jedoch, wie aus obiger Graphik erkennbar, nicht frei wählbar. Die vorgegebenen Verteilungen sind je nach Q-Sort-Verfahren unterschiedlich und variieren meistens zwischen einer Gleich- und einer Quasi-Normalverteilung (Block, 2008). Auch gängige Online-Tools, welche eine einfache digitale Implementierung der Q-Sort-Methode erlauben (vgl. Q-Assessor von The Epimetrics Group, 2010; Hackert & Brählers flashQ, 2007), nötigen den Forscher zu einer solchen a-Priori-Festlegung der zu erstellenden Verteilung der Karten. Dies ist natürlich mit der Q-Methodologie kongruent, welche schließlich auf die Erhebung ipsativer Daten abzielt.

Allerdings macht diese Generierung ipsativer Daten den Q-Sort für aktuelle Modelle von normorientierter Persönlichkeitsdiagnostik unattraktiv; vorausgesetzt, man möchte die zu befragende Person bezüglich mehrerer Facetten ihrer Persönlichkeit befragen, muss eine extreme Einschätzung eines Deskriptors immer auf Kosten der Möglichkeit erfolgen, einem anderen Deskriptor eine ebenso extreme Einschätzung zukommen zu lassen. Somit würden sämtliche Antworten auf die Items nicht mehr unabhängig voneinander erfolgen können, sondern stünden notwendigerweise in einem komplexen Zusammenhang miteinander. Das einfache Aufsummieren der Ratingantworten pro Skala sowie der folgende Vergleich mit einer Normstichprobe würde so ad Absurdum geführt werden. Auch die Verwendung von – zumindest so angestrebter – voneinander unabhängiger, übergeordneter Persönlichkeitsfacetten in Form von Skalen wäre damit nicht mehr möglich. Aus diesen Gründen wurde im empirischen Teil dieser Arbeit darauf verzichtet, den befragten Personen eine fixe Verteilung der Karten vorzuschreiben. Von hier ab folgende Bezüge auf den Q-Sort beziehen sich auf die für die vorliegende Untersuchung verwendete adaptierte Version des Verfahrens.

Wie oben für die anderen beiden Formate bereits geschehen, sollen die relevanten Attribute des Q-Sort-Formats herausgestellt werden:

- Übersicht über den Itempool

  Im Q-Sort-Format ist die Übersicht über den gegebenen Itempool im Vergleich zum Listen- und zum One-Item-a-Time-Format maximal ausgeprägt. Die zur Verfügung stehenden Karten sind jederzeit komplett einsehbar, und nach der ersten Aufteilung der Karten auf die Spalten sind alle Items des Itempools auf kleinstmöglichem Raum anschaulich aufgereiht.
- Übersicht über bereits gegebene Antworten

  Auch die Übersicht über die bereits gegebenen Antworten ist im Q-Sort höher als in den anderen Formaten; nach dem Einordnen der Karten hat die befragte Person nicht nur

einen genauen Überblick über die gegebenen Antworten in Form einer räumlichen Repräsentation, sondern auch einen Überblick über die Relation ihrer Antworten untereinander, sowie einen räumlichen Überblick über bis zu diesem Punkt wenig oder gar nicht verwendete Antwortkategorien.

Relevant ist natürlich auch die Frage, ob das Q-Sort-Format prinzipiell mit den anderen Formaten vergleichbar ist, oder ob die vorhergegangene Forschung Hinweise auf dissonante testtheoretische Eigenschaften hinweist. Wenige Studien beschäftigten sich bis dato mit dieser Frage, jedoch konnte Fluckinger (2010) in seiner ausführlichen Dissertation zeigen, dass dasselbe persönlichkeitsdiagnostische Verfahren, einmal mittels Likert-Ratingskala und einmal mittels Q-Sort vorgelegt, akzeptabel konvergente Korrelationen sowie eine sehr ähnliche Faktorenstruktur aufweist. Zusätzlich zeigten Fluckingers Analysen, dass das Q-Sort-Format bei in seiner Erhebung auf höhere Akzeptanz stieß als das übliche Likert-Ratingverfahren. Dieses Thema soll im Vergleich der verschiedenen Präsentationsarten unter *4. Auswirkungen des Präsentationsformats auf die Akzeptanz von Fragebögen* nochmals im Detail aufgegriffen werden.

1.4. Zusammenfassung der Eigenschaften der drei Formate

Zum Abschluss der Beschreibung der drei unterschiedlichen Präsentationsarten von Itempools sollen die für diese Arbeit relevanten Merkmale noch einmal vergleichend in tabellarischer Form zusammengestellt werden (Tabelle 1.1):

Tabelle 1.1

Überblick über die relativen Eigenschaften der Itempool-Präsentationsarten

	Listen-Format	One-Item-a-Time-Format	Q-Sort-Format
Übersicht über den Itempool	Mittel	Niedrig	Hoch
Übersicht über bereits gegebene Antworten	Mittel	Niedrig	Hoch

2. Verwandte Item-Kontext-Effekte

Sparfeldt, Schilling, Rost, und Thiel (2006) stellten zum Auftakt ihres Artikels folgende Prämisse in den Raum: „The notion of item context effects implies that psychometric properties of an item or scale are altered by the presentation format […].“ Dabei beziehen sie sich mit dem Term “presentation format” zwar nicht direkt auf die Manipulation der möglichen Übersicht über den Itempool, sondern auf die sogenannte Blockbildung von Items. Item-Blockbildung, im Englischen auch als Blocked Format oder Item Grouping bezeichnet, beschreibt die Vorgehensweise, sämtliche Items einer Skala nacheinander vorzugeben – im Gegensatz zur normalerweise üblichen disjunkten Vorgabe durch (quasi-) randomisierte Vorgabereihenfolge der Items der verschiedenen Skalen (Franke, 1997). Dies ist von den in der vorliegenden Arbeit untersuchten verschiedenen Präsentationsformaten (Listen-, One-Item-a-Time- und Q-Sort-Format) zwar auf den ersten Blick verschieden, jedoch existieren suffiziente Gemeinsamkeiten der beiden Ansätze, um einen Vergleich nahezulegen.

Sparfeldt et al. knüpfen mit den Prämissen zu ihrer Studie zur Item-Blockbildung argumentativ unter anderem an einer Studie von Tourangeau und Rasinski (1988) an: letztere diskutierten in ihrem überblickshaften Artikel Befunde zu der Frage, ob der Inhalt vorhergehender beantworteter Items einen Einfluss auf die Beantwortung späterer Items haben könnte. Im Zuge ihrer Argumentation kommen die Autoren zur Feststellung, dass die Bearbeitung früherer Items den Interpretationskontext des aktuellen Items sowie den Abruf von für die Beantwortung eines Items relevanten Informationen verändert. Franke (1997) geht in der Interpretation der Ergebnisse ihrer Studie zur Item-Blockbildung noch weiter: laut ihrer Darstellung führt die Item-Blockbildung dazu, dass die Messintention der jeweiligen Skala für die befragten Personen klar sichtbar wird, was wiederum in ungewöhnlichem Antwortverhalten resultieren soll.

Bei diesen Darstellungen der Item-Kontexteffekte bei Item-Blockbildung lässt sich die Parallele zur Manipulation der ermöglichten Übersicht über einen Itempool ziehen: wenn die gewährte Übersicht über einen Itempool denkbar hoch ist, kann nicht mehr sicher davon ausgegangen werden, in welcher Reihenfolge die Personen die Items bearbeiten, oder ob sich die Kontexteffekte im Falle hoher Übersicht auf die lineare Abfolge der Items beschränken. Dem folgend könnten sich die Personen im Listen-Format zunächst einen groben Überblick über die Items verschaffen, wodurch der Kontext der jeweiligen Items ein anderer wäre als bei linear-sukzessiver Bearbeitung ohne eines vorhergehenden Überblicks über den Itempool. Zum Beispiel im Falle der Vorgabe eines Itempools mittels Q-Sort-Format ist dieser Überblick, wie bereits in *1.3. Das Q-Sort-Format* dargelegt, sogar unumgänglich. Im One-Item-a-Time-Format hingegen ist

es ausgeschlossen, sich einen solchen initialen Überblick über die enthaltenen Items zu verschaffen.

Auf jeden Fall wäre bei gleichzeitigem Vorlegen aller Items die Messintention und, bei weitläufig bekannten psychologischen Konstrukten, wie zum Beispiel Extraversion, auch die Skalenzugehörigkeit vieler Items wesentlich offensichtlicher als bei disjunkt-randomisierter Vorgabe.

Aus diesen Gründen sollen im Folgenden einige ausgewählte Ergebnisse zur Item-Blockbildung dargestellt werden, mit dem Ziel, daraus möglicherweise zu erwartende Effekte der Manipulation der Übersicht über einen gegebenen Itempool zu modellieren.

Einleitend sollen die Ergebnisse der eben bereits erwähnten Studie von Sparfeldt et al. (2006) betrachtet werden: Die Forscher untersuchten Effekte auf die Bearbeitung eines Inventars zum akademischen Selbstkonzept, welches sie einer Kontrollgruppe in klassisch-randomisierter Form und einer Experimentalgruppe in skalenbezogen geblockter Form vorgaben. Relevant ist im Zuge der hiesigen Fragestellungen unter anderem, dass die Autoren bei der Beschreibung ihrer Instrumente angeben dass die Art wie die Blockbildung optisch repräsentiert wird keineswegs trivial ist. Verwendet wurde sozusagen eine Fragebogen-Matrix in der Form „Itemstamm x Skala“, was dem unter *1.1. Das Listen-Format* besprochenen Präsentationsformat entspricht. Die Autoren geben an, diese optische Komponente - welche ebenfalls die Übersicht über den Itempool erhöht – würde potentielle Blockbildungs-Effekte maximieren.

Trotz der gewählten optischen Repräsentation der Blockbildung, einer suffizient großen Stichprobe und vielseitiger Auswertungsmethoden, fanden die Autoren keine relevanten Unterschiede zwischen der Kontroll- und Experimentalgruppe; weder in Bezug auf die Faktorenstruktur oder psychometrischen Eigenschaften des Fragebogens, noch in Bezug auf die Skalenmittelwerte der beiden Gruppen. In dieser Studie konnten keine konsistenten Unterschiede zwischen geblockter und randomisierter Vorgabe eines Fragebogens gefunden werden.

Anders bei der Studie von Franke (1997): ihr Artikel beschäftigt sich mit der geblockten versus randomisierten Vorgabe der Items der Skalen einer klinisch-psychologischen Symptomcheckliste. Die Autorin fand signifikante Beeinflussungen des Antwortverhaltens durch den Präsentationsmodus des Instruments. Namentlich waren ein Großteil der Skalenmittelwerte, die Streuungen der jeweiligen Summenscores der Skalen sowie die Reliabilität der Skalen (operationalisiert durch Cronbachs α) in der Versuchsbedingung mit geblockten Items signifikant

niedriger als in der Standardanordnung der Items. Auch die Interkorrelationsmatrizen der Skalen, welche die Autorin als Hinweis auf die Validität des Instruments anführt, waren in der Versuchsbedingung mit geblockten Items untypisch und verzerrt – es zeigten sich deutliche Unterschiede zur Interkorrelationsmatrix der Kontrollbedingung und zu jener der Normstichprobe.

Im Gegensatz zu Frankes Studie merkten Lam, Green und Bordignon (2002) in ihrer Studie an, dass die Blockbildung thematisch zusammengehöriger Items zu erhöhter Reliabilität (im Sinne der Interitem-Korrelationen und Cronbachs α) der jeweiligen Skalen führen sollte. Allerdings spekulierten die Autoren, dass dieser Anstieg von Reliabilität zu Lasten der Validität stattfindet: da die befragten Personen sich des Themas der Items einer Skala bewusst sind, reagieren sie mehr auf dieses übergeordnete Thema als auf den eigentlichen Iteminhalt und geben somit generalisierte, homogenere Antworten, was die Interitem-Korrelationen und Cronbachs α freilich erhöhen, die Validität der Skala jedoch absinken lässt. Über dies hinaus muss die hier gefundene Erhöhung des Reliabilitätsmaßes im Lichte der hiesigen Fragestellung kritisch betrachtet werden, da das von den Autoren verwendete Instrument die Möglichkeit einer „Weiß ich nicht“-Antwort zuließ. Diese Antwort wurde in der Gruppe mit geblockten Items signifikant seltener gewählt, was zwangsläufig eine positive Auswirkung auf das errechnete Reliabilitätsmaß des Cronbachs α hat und die interpretierbare Wirkung der Item-Blockbildung selbst auf die Reliabilität relativiert.

Schriesheim, Kopelman und Solomon (1989) gingen ebenfalls davon aus, dass die Skalenreliabilität positiv durch Item-Blockbildung beeinflusst werden könnte, und dass diese Reliabilitätssteigerung auf Kosten der diskriminanten Validität der jeweiligen Items geschehen würde. Trotz einer umfassenden Analyse über drei separate Untersuchungen hinweg konnten sie diesen Effekt jedoch nicht eindeutig bestätigen; zwar gab es stellenweise Anzeichen für die vermuteten Effekte der Änderung des Formats der Itempoolpräsentation, jedoch mit zahlreichen Einschränkungen: die gefundenen Effekte waren durchwegs eher klein, selten statistisch signifikant und keineswegs durch alle drei Studien hindurch konsistent.

Bezüglich der Reliabilität prognostizierten Krampen, Hense und Schneider (1992) den Annahmen von Schriesheim et al. ähnliche Effekte der Itemblockbildung, jedoch waren die Ergebnisse unerwartet inkonsistent; unter der Bedingung der geblockten Items stieg die Reliabilität dreier Skalen signifikant an, fiel jedoch für zwei weitere Skalen. Die Validität (hier operationalisiert durch Interskalenkorrelationen) war hingegen, wie anfangs angenommen, merklich von der Itemblockbildung beeinflusst. Weiterhin bemerkenswert ist in dieser Studie,

dass die Autoren eine zusätzliche unabhängige Variable inkludierten, nämlich eine Instruktion zur Generierung von Antwortverhalten im Sinne sozialer Erwünschtheit.

Hier zeigte sich, dass die Effekte welche durch die „Instruktion zu sozial erwünschtem Antworten“ entstanden, in der Gruppe mit Itemblockbildung stärker auftraten als in jener mit herkömmlicher Reihenfolge der Items. Dieses Ergebnis ist konsistent mit der Interpretation von Franke (1997), nach der die Item-Blockbildung die Erfassung der Messintention erleichtert. Zur interpretativen Einordnung ihrer Ergebnisse muss jedoch auch angemerkt werden, dass die Stichprobe der beschriebenen Studie ausschließlich aus Studierenden der Psychologie bestand.

Vor dem Fazit sollen zum besseren Überblick die oben beschriebenen Studien in ihren für die gegenständliche Fragestellung relevanten Merkmalen in tabellarischer Form nochmals kurz, und der Einfachheit halber verallgemeinert, zusammengefasst werden (Tabelle 2.1).

Tabelle 2.1

Generalisierter Überblick über Studienergebnisse der Itemblockbildung. Zellbesetzungen beschreiben jeweils die Eigenschaften der Gruppe unter der Bedingung mit Itemblockbildung.

Autoren	Thema des verwendeten Instruments	Skalenmittel-werte	Streuung der Scores	Reliabilitäten der Skalen	Validität des Instruments
Sparfeldt et al. (2006)	Akademisches Selbstkonzept	gleich	gleich	gleich	gleich
Franke (1997)	Klinische Symptom-checkliste	niedriger	niedriger	niedriger	niedriger
Lam et al. (2002)	Einstellung zu einem Schultest	N/A	N/A	höher (kritisch zu betrachten; s.o.)	niedriger
Schriesheim et al. (1989)	Job-Charakteristika & Lebens-zufriedenheit	N/A	N/A	keine signifikanten/ konsistenten Ergebnisse	keine signifikanten/ konsistenten Ergebnisse
Krampen et al. (1992)	Persönlichkeits-facetten	in Richtung sozialer Erwünschtheit verschoben	N/A	Inkonsistente Verschiebungen	niedriger

N/A: betreffende Werte waren nicht Teil der jeweiligen berichteten Ergebnisse

Nach diesem kurzen Überblick über die bisherige Forschung im Bereich möglicher Effekte der Manipulation des Formats der Itempoolpräsentation im Sinne der Itemblockbildung ist zusammenfassend zu sagen, dass kaum von einheitlichen Effekten einer solchen Manipulation gesprochen werden kann. Mittelwerte, Streuungen, Reliabilitäten und Validitäten der jeweiligen Skalen, beziehungsweise Instrumente, zeigten – wenn überhaupt – inkonsistente Veränderungen in den Bedingungen mit Itemblockbildung. Dies könnte freilich daher rühren, dass die wenigsten Studien aus diesem Bereich diagnostische Instrumente mit derselben Thematik verwenden. Es wäre die Annahme naheliegend, dass die Itemblockbildung bei einer klinischen Symptomcheckliste zu anderen Effekten führt als bei einem Persönlichkeitsfragebogen im Kontext der Personalauswahl. So könnte zum Beispiel die Itemblockbildung bei ersterer Thematik dazu führen, dass die Implikation einer psychischen Erkrankung stärker präsent ist, was die niedrigeren Mittelwerte in der Studie von Franke (1997) als Reaktanz-Phänomen erklären würde. Dahingegen würde dieselbe Anordnung bei den Fragen nach dem akademischen Selbstkonzept in der Studie von Sparfeldt et al. (2006) vermutlich keine derartige Reaktanz erzeugen, da die Skalenzugehörigkeit der Items hier auch in randomisierter Form bereits offensichtlich ist.

Jener Kennwert, welcher über die besprochenen Studien hinweg am deutlichsten von der Item-Blockbildung betroffen war, war die Reliabilität des jeweiligen Verfahrens. Dieser Umstand soll im folgenden Kapitel *3. Die Rolle der Item Response Theory im Vergleich der unterschiedlichen Itempool-Präsentationsarten* weiter aufgegriffen werden.

3. Die Rolle der Item Response Theory im Vergleich der unterschiedlichen Itempool-Präsentationsarten

Obwohl sich die Ergebnisse der Studien, die unter *2. Verwandte Item-Kontext-Effekte* besprochen wurden, nur in wenigen Punkten decken, gibt es einen Aspekt, den die Studien in den gefundenen, beziehungsweise erwarteten, Ergebnissen teilen: die Veränderung der Reliabilität durch die Manipulation der Übersicht über den Itempool, wobei die Richtungen dieses Effekts unterschiedlich sind. Dabei verlassen sich die Autoren zur Operationalisierung der Reliabilität zumeist auf den Kennwert des Cronbachs α, konzentrieren sich also auf die innere Konsistenz der Skalen. Ebenso auffallend ist, dass keine/r der Autor/innen zur Untersuchung

seiner, beziehungsweise ihrer, Fragestellung die Methoden der Item Response Theory (IRT) heranzog, obwohl sie für die Fragestellung besser geeignet wären. Das Maß des Cronbachs α gehört zum Repertoire der klassischen Testtheorie, welche von der Annahme bereits vorliegender, fehlerbehafteter Messwerte ausgeht (Rost, 2004). Die IRT hingegen setzt bei den Items selbst, sprich dem Zustandekommen dieser Messwerte an. Dabei erlaubt die Verwendung von IRT-Methoden im Vergleich zu jenen der klassischen Testtheorie in manchen Fällen ein vorteilhafteres Vorgehen, was im Folgenden kurz dargestellt werden sollen.

Statt von IRT wird auch oft von probabilistischen Latent-Trait-Modellen gesprochen, was vielleicht der anschaulichere Begriff ist. „Latent-Trait"-Modelle heißen deshalb so, weil den Antworten einer Person in einem psychologisch-diagnostischen Messinstrument als Ursache eine latente, das heißt nicht direkt beobachtbare, Eigenschaft als zugrundeliegend angenommen wird. Diese latente Eigenschaft steht im direkten Interesse des diagnostischen Prozesses, da es genau jene zu messen gilt. Der Titel „probabilistisch" rührt daher, dass die Antwort einer Person auf ein Item des Messinstruments als Ausgang einer Wahrscheinlichkeitsfunktion modelliert wird; je nach Ausprägung der latenten Eigenschaft der Person und der Charakteristik des Items resultiert eine unterschiedliche Wahrscheinlichkeit für die jeweils möglichen Antworten der Person auf das Item (Roskam, 1996). Im Kontext dieser Argumentation ist vor allem das dichotome, logistische Testmodell von Rasch von Interesse, welches der Einfachheit halber im Folgenden als Rasch-Modell bezeichnet werden wird. Das Rasch-Modell ist ein Spezialfall der IRT, welches von drei Grundannahmen ausgeht (vgl. Rost, 2004):

- Das Vorliegen von dichotomen oder post-hoc dichotomisierten Itemantworten
- Die alleinige Berücksichtigung einer eindimensionalen Eigenschaft auf der Personenseite, das heißt im Bereich von Leistungstests die Fähigkeit und im Bereich von Persönlichkeitsfragebögen die betreffende Eigenschaft
- Die alleinige Berücksichtigung der Charakteristik des Items, das heißt seine „Schwierigkeit" bei Leistungstests beziehungsweise sein „Aufforderungscharakter" bei Persönlichkeitsfragebögen

Einer der großen Vorteile des Modells ist nun die Tatsache, dass die Gültigkeit des Modells innerhalb eines Itempools tatsächlich mittels einem statistischen Signifikanztest möglich ist (Roskamp, 1996); demgegenüber können die meisten Testmodelle der IRT nur die relative Güte ihrer Passung zu den gegebenen Daten mittels deskriptiver Kennwerte argumentieren. Die in diesem Kontext wichtigste Eigenschaft des Rasch-Modells betrifft die notwendige Eindimensionalität des betreffenden Itempools. Um Homogenität im Sinne des Rasch-Modells zu

erlangen, dürfen die Antworten auf die Items des Itempools ausschließlich durch eine einzige latente Eigenschaft bedingt sein; erfassen die Items mehrere verschiedene Eigenschaften, so ist die Voraussetzung der Eindimensionalität verletzt und der statistische Test zur Passung des Rasch-Modells auf die Daten resultiert in einem signifikanten Ergebnis. Somit besitzt ein Itempool, welcher sich nach einem gegebenen Modelltest als Rasch-Modell-homogen erwiesen hat, notwendigerweise innere Konsistenz (Kubinger, 2009), sprich genau jenen Aspekt der Reliabilität, den die Autor/innen obiger Studien mittels Cronbachs α darzustellen suchten. Würde beispielsweise die Vorgabereihenfolge oder die Art der Darstellung eines Itempools neben der zu messenden, latenten Eigenschaft der Person ebenfalls einen Einfluss auf die Wahrscheinlichkeit einer zustimmenden oder ablehnenden Antwort auf die Items ausüben, so würde dies in testbaren Abweichungen vom Rasch-Modell zu Buche schlagen.

Dass dieser Ansatz der Verwendung von Cronbachs α überlegen wäre, lässt sich bereits an der Problematik erkennen, welche die Interpretation des Kennwertes in der Studie von Lam et al. (2002) erschwert (siehe oben unter *2. Verwandte Item-Kontext-Effekte*).

Somit wäre es naheliegend, einen Selbstbeschreibungsfragebogen, dessen Itempool bereits als Rasch-Modell-konform bestätigt ist, in verschiedenen Itempoolpräsentationsarten vorzulegen und zu prüfen, ob die Geltung des gegebenen IRT-Modells über die verschiedenen Gruppen hinweg noch gegeben ist. Einer der Hauptgründe, warum eine solche Untersuchung noch nicht vorgenommen wurde, mag darin liegen, dass derartige Verfahren im Vergleich zu jenen aus dem methodischen Repertoire der klassischen Testtheorie relativ rar sind. Nichtsdestotrotz existieren solche Verfahren, zum Beispiel der Big 5 Plus One von Holocher-Ertl et al. (B5PO, 2003) beziehungsweise dessen Revision B5PO-R von Siptroth (2011). Hierzu sollen in *6.2.2. B5PO-R: Big Five Plus One Revised* weitere Ausführungen folgen. Auch zeigten Rost, Carstensen und von Davier (1999), dass Big-Five-Selbstbeschreibungsfragebögen prinzipiell Rasch-Modell-skalierbar sind, wenn auch mit Einschränkungen. Daher fiel die Wahl des Verfahrens, mittels welchem die Unterschiede bezüglich der inneren Konsistenz der Skalen zwischen den verschiedenen Itempoolpräsentationsarten untersucht werden sollen, auf den B5PO-R. Denn dieser erlaubt es möglicherweise, die Frage nach der inneren Konsistenz eines Verfahrens aus einem anderen methodischen Blickwinkel zu betrachten, als dem üblicherweise verwendeten Cronbachs α und anderen korrelativen Reliabilitätskriterien.

4. Das Big Five-Faktorenmodell der Persönlichkeit

Wie unter *3. Die Rolle der Item-Response Theory im Vergleich der unterschiedlichen Präsentationsarten* dargestellt, wird im empirischen Teil der Diplomarbeit das Verfahren B5PO-R (Siptroth, 2011) zur Anwendung kommen. Daher soll an dieser Stelle der theoretische Kern des Verfahrens kurz erörtert werden: das Big-Five-Faktorenmodell der Persönlichkeit.

Das Big Five-Faktorenmodell beschreibt fünf Supra-Variablen, welche zusammen die mittlerweile vielleicht am besten bekannte und am weitesten verbreitete psychologische Strukturtheorie der Persönlichkeit bilden. Ursprünglich in seinen Grundzügen auf den deutschen Psychologen Hans Eysenck zurückzuführen (Eysenck, 1944, nach Maltby, Day & Macaskill, 2011), konnte das Modell seitdem in zahlreichen Untersuchung bestätigt werden (Maltby et al., 2011). Dabei entstammt das Big Five-Faktorenmodell nicht wie in der Psychologie sonst üblich einer zunächst theoretisch konzeptualisierten Hypothese, welche anschließend empirisch bestätigt wurde. Tatsächlich ist das Modell das Produkt von zunächst exploratorischer und später konfirmatorischer Faktorenanalysen, welche wiederholt aus für die Persönlichkeitsbeschreibung relevant angesehenen Verhaltensweisen fünf Faktoren extrahierten (Maltby et al., 2011). Diese fünf Faktoren werden als grundlegende Pfeiler der Struktur der Persönlichkeit angesehen, nicht zuletzt aufgrund der starken Beeinflussung ihrer Ausprägungsgrade durch genetische Grundlagen (Riemann, Angleitner & Strelau, 1997). Ebenso ist die Unabhängigkeit und Eindimensionalität der Faktoren gut belegt, so dass es möglich war, einen Rasch-Modell-homogenen Fragebogen zur Kurzerfassung der Big Five zu entwickeln (Holocher-Ertl et al., 2003). Sogar die Skalierung eines bereits bestehenden Fragebogens nach dem Rasch-Modell war – wenn auch unter Einschränkungen – möglich, obwohl dieser in seiner ursprünglichen Konstruktion nicht nach dem Rasch-Modell skaliert worden war (Rost et al., 1999).

Da die Benennung der resultierenden Faktoren der wohl subjektivste Teil einer Faktorenanalyse ist, existieren weit voneinander abweichende Bezeichnungen für die gut belegten fünf Faktoren des Strukturmodells der Big Five, wenngleich die Faktoren inhaltlich selten variieren. Im Folgenden sollen die Fünf Faktoren kurz nach Maltby et al. (2011) paraphrasierend beschrieben werden.

- Offenheit für Erfahrungen

  Dieser Faktor umfasst Eigenschaften der intellektuellen Neugier, des divergenten Denkens und der Vorstellungskraft. Individuen hoher Offenheit präferieren neue Lösungsansätze und suchen auch aktiv nach ihnen. Personen mit niedriger Offenheit sind dagegen eher konservativ und auf die Aufrecht- und Beibehaltung bewährter Ideen ausgerichtet.

- Gewissenhaftigkeit

  Gewissenhaftigkeit beschäftigt sich in seinem Kern mit Selbstdisziplin und Kontrolle. Hohe Ausprägungen stehen für Personen, die gerne und viel planen sowie hochgradig organisiert sind. Geringe Ausprägungen hingegen beschreiben Personen, welche eher sorglos und leichter ablenkbar sind und die wenige Zeit mit Planen verbringen. Gewissenhaftigkeit ist aufgrund seines starken Bezugs zu Verhalten im Berufskontext regelmäßig eines der Augenmerke von diagnostischen Fragestellungen zur Bewerberauswahl.

- Extraversion

  Dieser Faktor beschreibt die Geselligkeit einer Person. Personen, welche auf diesem Faktor hoch scoren, werden als tatkräftig, optimistisch und gesellig beschrieben. Sie werden auch Extravertierte genannt. Im Gegenzug werden Personen, welche auf diesem Faktor niedrig scoren, als Introvertierte bezeichnet und als reserviert sowie gesellschaftlich unabhängig beschrieben. Der Vollständigkeit halber soll angemerkt werden, dass Rost et al. (1999) bei diesem Faktor die deutlichsten Abweichungen von Eindimensionalität im Sinne der IRT vorfanden.

- Verträglichkeit

  Dieser Faktor dreht sich um die Art der sozialen Interaktion einer Person; Personen hoher Ausprägungen werden als vertrauensvoll, hilfsbereit, sympathisch und weichherzig beschrieben, wohingegen Personen geringer Ausprägung eher als misstrauisch, feindselig, skeptisch und unkooperativ gelten.

- Neurotizismus

  Die Messintention dieses Faktors beinhaltet die emotionale Stabilität und Anpassung einer Person. Personen mit hohen Neurotizismuswerten erleben starke Stimmungsschwünge und sind sprunghaft in ihren Emotionen; Personen mit niedrigeren Ausprägungen sind im Vergleich ruhig, gut angepasst und neigen eher weniger zu maladaptiven emotionalen Zuständen. In vielen fünf-Faktoren-Lösungen wird dieser

Faktor als emotionale Stabilität bezeichnet, um die negative Konnotation der Skala abzumildern.

5. Auswirkungen des Präsentationsformat auf die Akzeptanz von Fragebögen

In seiner ausführlichen Dissertation untersuchte Fluckinger (2010) die Unterschiede in den Resultaten eines Big-Five-basierten Selbstbeschreibungsfragebogens, wenn dieser einmal mittels klassischer Likert-Ratingskala und einmal mittels Q-Sort vorgegeben wurde. Neben extensiver Untersuchung der resultierenden Faktorenstruktur und weiterer testtheoretischer Merkmale lag sein hauptsächliches Augenmerk auf der Akzeptanz des Fragebogens durch die Teilnehmer der Studie.

Fluckinger hypothetisierte, dass das Verfahren in der Q-Sort-Bedingung von den Teilnehmern als angenehmer und unterhaltsamer erlebt werden würde als jenes in der Likert-Bedingung. Zudem traf er die Annahme, dass ersteres Verfahren von den Teilnehmern als messgenauer und interaktiver erlebt werden würde, und dass die Teilnehmer die Leichtigkeit, mit der die Antworten in dem Verfahren gefälscht werden könnten, in der Q-Sort-Bedingung als geringer einstufen würden.

Bei Auswertung seiner Ergebnisse fand Fluckinger, dass das Itempool-Präsentationsformat einen starken Effekt auf die wahrgenommene Interaktivität des Verfahrens ausübte; generell empfanden die Teilnehmer die Q-Sort-Bedingung als deutlich interaktiver als die Likert-Bedingung. Auch die Annahme, dass der Q-Sort als schwieriger zu fälschen wahrgenommen werden würde, wurde mit einem moderaten bis starken Effekt bestätigt. Die Annahme, dass die Q-Sort-Bedingung als angenehmer und unterhaltsamer empfunden würde, konnte ebenfalls bestätigt werden, allerdings nur mit einem kleinen Effekt. Lediglich die Vorannahme, dass der Q-Sort über all dies hinaus als messgenauer wahrgenommen werden würde, konnte anhand von Fluckingers Daten nicht bestätigt werden.

Dementsprechend lieferte Fluckinger deutliche Hinweise darauf, dass die Akzeptanz des „Big Five“-Verfahrens durch die Teilnehmer – mit Ausnahme der erlebten Messgenauigkeit – deutlich vom Itempool-Präsentationsformat beeinflusst wurde. Zusätzlich fand er Ergebnisse, die darauf hindeuteten, dass der Q-Sort – vielleicht aufgrund seiner geringeren Verbreitung als die üblichen

Likert-Ratingskalen – positivere Reaktionen in der Verwendung als Selektionsinstrument im betrieblichen Kontext erzeugte.

Auch in der bereits unter *2. Verwandte Item-Kontext-Effekte* erwähnten Studie von Krampen et al. (1992) zum Thema der Itemblockbildung interessierten sich die Forscher für das unterschiedliche subjektive Erleben der Teilnehmer während des Ausfüllens des Fragebogens. Die aus der Studie resultierenden Daten zeigten eine leichte Tendenz der Teilnehmer, das Ausfüllen der geblockten Items im Vergleich zur Standardreihenfolge als etwas langweiliger, entscheidungsschwieriger, situationsabhängiger und fragwürdiger einzustufen. Auch insgesamt zeigte sich ein leichter Trend in Richtung einer negativeren Beurteilung des Verfahrens unter der Bedingung der Itemblockbildung.

Obwohl der Ansatz der Untersuchung der Wahrnehmungen der befragten Personen im Bereich des Vergleichs unterschiedlicher Itempool-Präsentationsarten relativ selten verfolgt wird, zeigen jene Studien die einer solchen Frage Platz einräumen, ein definitives Vorhandensein solcher Effekte. Dementsprechend wird auch im empirischen Teil dieser Arbeit auf die Akzeptanz eines Verfahrens unter verschiedenen Itempool-Präsentationsarten durch die Teilnehmer eingegangen werden (siehe auch unter *6.5. Hypothese 5: Unterschiede in der Akzeptanz des Verfahrens durch die befragten Personen*).

III. Empirischer Teil

6. Ziel der Untersuchung

Ziel der Untersuchung ist die Klärung der Forschungsfrage, inwieweit das Präsentationsformat des gegebenen Itempools eines Selbstbeschreibungsfragebogens Effekte auf das Antwortverhalten der befragten Personen sowie auf ihre Akzeptanz des bearbeiteten Fragebogens ausübt (siehe auch unter *I. Einleitung*). Zu diesem Zwecke werden den befragten Personen der hier besprochenen Untersuchung zwei aufeinander folgende Online-Fragebögen vorgelegt werden: zunächst ein Big Five-Selbstbeschreibungsinstrument (siehe 7*.2.2. B5PO-R: Big Five Plus One Revised*) in je einer von drei Bedingungen mit variierender Übersicht über den Itempool, gefolgt von einem Fragebogen zur differenzierten Erfassung der Akzeptanz des eben ausgefüllten Selbstbeschreibungsfragebogens durch die befragten Personen (siehe 7*.2.3. Akzept!-P*).

Im Folgenden sollen jedoch zuerst die aus der Forschungsfrage und der besprochenen Literatur abgeleiteten Hypothesen detailliert dargestellt werden. Nach der Darstellung der Hypothesen und der jeweils angemessenen statistischen Verfahren wird separat kurz auf die A-Priori-Powerberechnungen eingegangen werden. Aufgrund der inkonsistenten Ergebnisse der Studien zur Item-Blockbildung (siehe *2. Verwandte Item-Kontext-Effekte*) kann keine eindeutige Erwartung bezüglich der hier untersuchten Fragestellung getroffen werden; die untersuchten Hypothesen werden daher weitestgehend ungerichtet gestellt werden.

6.1. Hypothese 1: Veränderungen der inneren Konsistenz der Faktoren des Fragebogens

Wie unter *2. Verwandte Item-Kontext-Effekte* und *3. Die Rolle der Item Response Theory im Vergleich der unterschiedlichen Itempool-Präsentationsarten* bereits besprochen betreffen jene zu erwartenden Effekte, welche sich aus bisherigen Untersuchungen ableiten lassen, in erster Linie Veränderungen in der Reliabilität im Sinne der inneren Konsistenz der Itempools der Faktoren des Fragebogens. Wo sich jedoch vorhergehende Untersuchungen der Manipulation der Übersicht über einen gegebenen Itempool und die bereits gegebenen Antworten auf korrelative Maße wie Cronbachs α beschränkten, soll hier auf die Methoden der Item Response Theory zurückgegriffen werden. Dies ist bei der Beantwortung der Frage nach Verschiebungen der inneren Konsistenz der Itempools naheliegend, da beim hier verwendeten Rasch-Modell jene innere Konsistenz im Sinne der lokalen stochastischen Unabhängigkeit der Items eines Faktors eine notwendige Voraussetzung für die Modellgültigkeit ist (Kubinger, 2009).

Dementsprechend soll für das verwendete Verfahren, welches bereits als Rasch-Modell-konsistent bestätigt wurde (Siptroth, 2011), untersucht werden ob die enthaltenen Items über die drei Bedingungen mit verändertem Präsentationsformat des Itempools hinweg noch Modellgeltung besteht. Sollte für die Items der Faktoren über alle Bedingungen hinweg keine Rasch-Modell-Gültigkeit bestehen, jedoch eine oder mehrere Bedingungen in sich Modellgültigkeit erreichen, muss dies als Hinweis auf die Beeinflussung der inneren Konsistenz durch die Präsentationsart des Itempools gelten.

Je nachdem ob alle Gruppen gemeinsam, nur eine, oder zwei der Gruppen Modellgültigkeit beweisen, werden weitere den Ergebnissen entsprechende statistische Analysen nachfolgen; im Falle fehlender Modellgültigkeit in allen Bedingungen muss auf Methoden der klassischen Testtheorie zurückgegriffen werden.

Aufgrund der Überlegenheit des Skalenniveaus der Personenparameter gegenüber der normalerweise üblichen Summe der Ratingantworten der befragten Personen soll die Rasch-Modell-Analyse vor allen anderen Auswertungen stattfinden, um im Falle von Modellgültigkeit über alle Präsentationsformate hinweg die nach dem Rasch-Modell geschätzten Personenparameter für weitere Analysen verwenden zu können.

Die Modelltests zur Überprüfung der Geltung des Rasch-Modells in den jeweiligen Gruppen sollen mittels Andersens Likelihood-Ratio-Tests erfolgen.

6.2. Hypothese 2: Unterschiedliche Nutzung des verfügbaren Antwortspektrums

Die zweite Hypothese bezieht sich auf die Nutzung des Spektrums der gegebenen Antwortmöglichkeiten. Allgemein wird bei der Verwendung von Ratingskalen darauf abgezielt, der befragten Person ein angemessenes Antwortspektrum zu bieten, so dass sie ihre Selbsteinschätzung oder Meinung differenziert darstellen kann (Bühner, 2001). Bedenkt man nun, dass manche Personen dazu neigen, bestimmte Response Sets zu verwenden (wie zum Beispiel die Tendenz zum extremen oder zum moderaten Urteil, siehe zum Beispiel in Rost et al., 1999) liegt der Gedanke nahe, dass bei steigender Übersicht über die bereits gegebenen Antworten auch den Personen selbst eine solche Nutzung des zur Verfügung stehenden Antwortspektrums offensichtlich werden muss.

Dementsprechend stellt sich nun die Frage, ob ein größerer Überblick über das verfügbare Antwortspektrum und die bereits gegebenen Antworten die befragten Personen dazu verleiten

könnte, noch nicht beziehungsweise selten verwendete Antwortmöglichkeiten „aufzufüllen“. Ein Indikator für ein solches Vorgehen seitens der befragten Personen wäre die Verletzung der Homogenitätsannahme der resultierenden Verteilungen von gewählten Antwortmöglichkeiten zwischen den drei Gruppen der verschiedenen Itempool-Präsentationsmodi.

Der korrespondierende statistische Test der zweiten Hypothese ist der χ^2-Test zur Überprüfung der Homogenitätsannahme zur Verteilung der Zellenbesetzungen von den Präsentationsarten. Dieser wird über eine drei-mal-sechs-Kontingenztafel mit den Spalten „Antwortspektrum, eins bis sechs“ und den Zeilen „Präsentationsart: Listen-, One-Item-a-Time- und Q-Sort-Format“ gerechnet werden. Als praktisch relevant wird a-priori eine mittlere Effektstärke von $w \geq .3$ angesetzt. Für die Verwendung exakter Tests beziehungsweise Korrekturen der Teststatistik gibt es keine Notwendigkeit, da die Testgröße unter der Annahme von erwarteten Zellbesetzungen von $n \geq 5$ ausreichend asymptotisch χ^2-verteilt ist (Bortz & Schuster, 2010). Dies sollte selbst im Falle stark linksschiefer Verteilungen der Ratingantworten gegeben sein.

6.3. Hypothese 3: Unterschiedliche Mittelwerte & Interkorrelationsmatrizen der Skalen des Selbstbeschreibungsfragebogens

Obwohl die Effekte der Manipulation des Präsentationsformats auf die Mittelwerte der jeweiligen Skalen keineswegs eindeutig oder konsistent sind, fanden manche Forscher (unter anderem Franke, 1997) Veränderungen in den Skalenmittelwerten unter der Bedingung mit Itemblockbildung. Zusätzlich fanden mehrere Autoren im Zuge der Untersuchung der Validität der jeweiligen Instrumente Veränderungen in den Beziehungen der Skalen untereinander (siehe *2. Verwandte Item-Kontext-Effekte*). Auch wenn in manchen der die Itemblockbildung betreffenden Studien auf eine genauere Untersuchung der resultierenden Mittelwerte, ihrer Beziehungen und deren Interpretation verzichtet wurde, sollen hier die Mittelwerte der verschiedenen Bedingungen einander gegenübergestellt werden. Mangels konsistenter oder direkt vergleichbarer Vorergebnisse lässt sich dabei keine Richtung der Effekte im Vornherein annehmen.

Somit erwartet die dritte Hypothese Unterschiede zwischen den Gruppen der Präsentationsformate in den Skalenmittelwerten des verwendeten Big Five-Selbstbeschreibungsfragebogens B5PO-R (siehe 7*.2.2. B5PO-R: Big Five Plus One Revised*). Als Auswertungsmethode wird eine einfaktorielle, multivariate Varianzanalyse (One-Way-MANOVA) vorgesehen, um nicht nur etwaigen Mittelwertsunterschieden sondern auch den Beziehungen

der abhängigen Variablen untereinander in den drei verschiedenen Bedingungen Rechnung zu tragen. Die One-Way-MANOVA wird mit den drei Präsentationsformaten als unabhängiger Gruppenvariable, und den drei verwendeten Skalen des B5PO-R als abhängigen Variablen durchgeführt werden. Diese überspannende Unterschiedshypothese soll im Falle signifikanter Unterschiede von entsprechenden post-hoc Tests nachgefolgt werden. Im Falle relevanter Verletzungen der Voraussetzungen des Tests soll auf eine robuste, also nonparametrische Version des Verfahrens zurückgegriffen werden.

6.4. Hypothese 4: Unterschiede in den Bearbeitungszeiten

Dass die Q-Sort-Bedingung eine höhere Bearbeitungszeit mit sich bringt als die anderen Settings ist naheliegend. Schließlich enthält jenes Setting eine weitaus längere und kompliziertere Instruktion als die anderen beiden. Außerdem ist davon auszugehen dass die befragten Personen vermutlich mit dem klassischen Fragebogenformat vertraut sind; diese Annahme kann man beim Q-Sort ob seiner geringeren Verbreitung nicht treffen. Es stellt sich jedoch die Frage, wie stark dieser Unterschied in den Bearbeitungszeiten ist. Sollte sich herausstellen, dass das Q-Sort-Format explizite testtheoretische oder akzeptanzbezogene Vorteile gegenüber den anderen Formaten aufweist, wird die erhöhte Bearbeitungszeit ein Faktor sein der auf jeden Fall im Sinne der Ökonomie gegenüber anderen möglichen Vorteilen abgewogen werden muss.

Da aufgrund der komplexeren Instruktion und geringeren Verbreitung des Q-Sorts-Formats kleine bis mittlere Unterschiede in den Mittelwerten der Bearbeitungszeiten trivial wären, soll für den Omnibus-Test eine Mindesteffektstärke von $f \geq .40$ als praktisch relevant angesehen werden. Ein solcher Effekt soll mittels einfaktorieller, univariater Varianzanalyse (One-Way-ANOVA) mit drei Gruppen untersucht werden. Sollten die Voraussetzungen für das Verfahren verletzt werden, wird die Analyse der betreffenden Daten mittels eines nonparametrischen Verfahrens sowie allfälligen nonparametrischen Einzelverglichen erfolgen.

Aufgrund der oben geschilderten Annahmen kann detaillierter gesprochen angenommen werden dass die Bedingung des Q-Sort höhere Bearbeitungszeiten als die anderen Bedingungen produzieren wird; zusätzlich könnte es auch sein dass die Bedingung des One-Item-a-Time-Formats höhere Bearbeitungszeiten als das Listenformat produziert, da die befragten Personen hier zwischen ihren Antworten auf die Items einen zusätzlichen „Schritt“ einlegen müssen um zur nächsten Seite zu gelangen. Um diesen Annahmen gerecht zu werden wird neben der

herkömmlichen ANOVA eine Analyse mit dementsprechend vordefinierten, orthogonalen Kontrasten durchgeführt werden.

6.5. Hypothese 5: Unterschiede in der Akzeptanz des Verfahrens durch die befragten Personen

Wie bereits unter *5. Auswirkungen des Präsentationsformat auf die Akzeptanz von Fragebögen* besprochen deuten mehrere Untersuchungen zu verwandten Themen auf Unterschiede in der Akzeptanz des ausgefüllten Fragebogens durch die befragten Personen hin (Fluckinger, 2010, sowie Krampen et al., 1992). Um auch in dieser Studie derartige Effekte identifizieren zu können wurde an den zentralen Big-Five Selbstbeschreibungsfragebogen ein Instrument zur Erhebung der Einschätzung des ersteren Fragebogens durch die befragten Personen angehängt, namentlich der Akzept!-P (Kersting, 2010, sowie Kersting, 2012, siehe auch unter *7.2.3. Akzept!-P*). Jener Fragebogen enthält mehrere Skalen, welche inhaltlich mit den Ergebnissen Fluckingers korrespondieren – trotzdem wird von einer Anpassung der erwarteten Ergebnisse in die jeweiligen Richtungen abgesehen, da zum einen Fluckingers Versuchsanordnung den Vergleich beider Formate (Liste und Q-Sort) durch dieselben Personen enthielt und die Anordnung der vorliegenden Untersuchung jeder Person nur je ein Format vorlegt, sowie eine weitere, von Fluckinger nicht untersuchte Präsentationsart enthält (One-Item-a-Time). Außerdem interessiert hier zunächst – mangels genauerer Vorergebnisse zu allen drei Formaten – die prinzipielle Existenz eines Unterschiedes in den Skalensummen des Akzept!-P. Die fünf beziehungsweise sechs Faktoren des Instrumentes sind unter *7.2.3. Akzept!-P* weiterhin detaillierter beschrieben.

Um den möglichen Zusammenhängen zwischen den Skalensummen des Akzept!-P Geltung zukommen zu lassen und die Kumulierung des Risikos eines Fehlers erster Art zu kontrollieren, sollen zum Vergleich der Mittelwerte der Skalensummen des Fragebogens über die drei Gruppen eine multivariate, einfaktorielle Varianzanalyse gerechnet werden. Dabei werden die drei verschieden Gruppen des Präsentationsformats (Listen-, One-Item-a-Time- und Q-Sort-Format) als unabhängige, die Skalensummen des Akzept!-P als abhängige Variablen einbezogen werden. Im Falle substantieller Abweichungen der Daten von den Anforderungen der MANOVA wird auf eine robuste, also nonparametrische Version des Verfahrens zurückgegriffen werden.

6.6. A-Priori-Powerberechnungen

Vor der Durchführung der Untersuchung wurden zunächst A-Priori-Powerberechnungen durchgeführt. Dieses Vorgehen dient dem Ziel, einen angemessenen Rahmen für jenen Stichprobenumfang zu eruieren, der für die Feststellung der interessierenden Effekten notwendig wäre (Rasch, Kubinger, und Yanagida, 2012). Jene derartigen Berechnungen, die direkt möglich waren, wurden mittels dem Statistikprogramm R (R Development Core Team, 2011), genauer mit dem Programmpaket „pwr" (Champely, 2009) durchgeführt.

Zu Beginn soll in Bezug auf Hypothese 1 erwähnt werden, dass zum aktuellen Zeitpunkt exakte a-Priori-Powerberechnungen von Rasch-Modelltests noch ein rechnerisches Problem darstellen; lediglich eine Abschätzung der zu erwartenden Risikos eines Fehlers erster und zweiter Art ließe sich unter der Voraussetzung theoretischer Vorannahmen zur Passung des Itempools sowie der Verteilung der Itemparameter vornehmen (Kubinger, Rasch, und Yanagida, 2011). Mangels Verfügbarkeit solcher Vorannahmen wurde hier beschlossen, für diese Tests eine Mindestgruppengröße von je 100 Personen anzusetzen und das Signifikanzniveau der Tests auf $\alpha \leq .01$ festzulegen, um der üblichen mehrfachen Durchführung des Modelltests für verschiedene Trennungskriterien Rechnung zu tragen.

Ebenso ist die a-Priori-Powerberechnung von multivariaten Designs wie in Hypothese 3 und 5 aktuell noch problematisch; eine genaue Abschätzung der resultierenden Power ist nur unter der Vorkenntnis von Populationsmittelwerten und –streuungen möglich, sowie gewissen Vorannahmen zur Struktur der Interkorrelationsmatrizen der abhängigen Variablen (D'Amico, Neilands, & Zambarano, 2001). Da die hier vorgegebenen Fragebögen in adaptierten Versionen verwendet werden sollen (siehe *7.2.2. B5PO-R: Big Five Plus One Revised* und *7.2.3. Akzept!-P*), können solche Vorannahmen nicht ohne Weiteres getroffen werden. Nichtsdestotrotz sollte die Verwendung der MANOVAs in Hinsicht auf die resultierende Power unproblematisch sein, da aufgrund der Rasch-Modell-Analysen ohnehin ein relativ großer Stichprobenumfang benötigt wird. Zudem kann für die relevanten Daten beider betreffender Hypothesen davon ausgegangen werden, dass die abhängigen Variablen zumindest mittlere Korrelationen miteinander aufweisen – im Falle des B5PO-Rs aufgrund des hohen Aufforderungscharakters der Items, und im Falle des Akzept!-Ps aufgrund der gemeinsamen inhaltlichen Richtung der Skalen. Dies ist ebenfalls der Power der zu verwendenden MANOVAs zuträglich, da das Verfahren höhere Power bei höher korrelierten abhängigen Variablen erreicht (Field, Miles, & Field, 2012).

Die Kennwerte der Powerberechnungen der verbleibenden statistischen Tests – des χ^2-Tests von Hypothese 2 und der ANOVA von Hypothese 4 – konnten ohne Komplikationen durchgeführt werden und sind in Tabelle 6.1 aufgelistet.

Tabelle 6.1

Kennwerte der a-priori Powerberechnungen unter Annahme der Mindestgruppengröße von 100 Personen (N = 300)

(α = .05; 1 – β = .80)	Als relevant festgesetzte Effektstärke	Kleinste noch feststellbare Effektstärke
Hypothese 2 – χ^2-Test (*df* = 10)	$w \geq .30$	$w = .04$
Hypothese 4 – ANOVA	$f \geq .40$	$f = .18$

7. Methode

7.1. Untersuchungsplan

Um nun die oben aufgestellten Hypothesen untersuchen zu können, wurde eine dreiteilige Fragebogenbatterie zusammengestellt, dessen Beantwortung die dazu notwendigen Daten liefern sollte. Aufgrund der unter *1.3. Das Q-Sort-Format* besprochenen Problematiken der Online-Implementierung war es notwendig, ein speziell für die Zwecke dieser Untersuchung angefertigtes Online-Tool entwerfen zu lassen, welches es erlaubt einen Q-Sort einerseits ohne einen Vor-Sortierungsschritt und andererseits ohne eine fixierte Verteilung für dessen Karten vorzugeben. Der Vor-Sortierungsschritt wurde hier unterlassen, da für die Items des B5PO-R lediglich 30 Deskriptor-Karten notwendig waren (siehe *7.2.2. B5PO-R: Big Five Plus One Revised*), und somit weitaus weniger Karten als in anderen Implementierungen des Q-Sorts verwendet werden (zum Beispiel beim California Q-Sort von Block, 2008). Das Aufheben der fixen Verteilung eines Q-Sorts aufgrund des Wunsches die durch das Verfahren generierten Daten wie gewohnte, ratingbasierte Verfahren zu behandeln wurde bereits in der Literatur thematisiert; so argumentiert beispielsweise Fluckinger (2010) – sehr kurz gefasst – dass die durch eine fixierte Verteilung entstehenden Interdependenzen zwischen den einzelnen Items rechnerisch keinen gravierenden Unterschied machen würden. Da jedoch unbekannt ist wie genau sich jene Interdependenzen im Falle der Verwendung von Auswertungsmethoden der Item-Response-

Theory niederschlagen könnten, wurde trotz Fluckingers Argumentation beschlossen, die fixierte Verteilung zu verwerfen und den befragten Personen freie Wahl über die Verteilung der Items, oder auch „Karten“, zu überlassen. Auch zur Untersuchung der interessierenden Hypothese 2 – welche nach der Ausnutzung des verfügbaren Spektrums von Antwortmöglichkeiten fragt – war diese Aufhebung notwendig.

Das fertige Instrument soll hier nun kurz in seinem Ablauf beschrieben werden, ist jedoch auch im Anhang unter *11. Screenshots des Online-Assessments* in auszugsweisen Screenshots einzusehen.

Nach einer kurzen Begrüßung und Instruktion der Teilnehmer wurden sie zu einigen kurzen Fragen bezüglich ihrer demographischen Daten weitergeleitet. Daraufhin folgend verzweigte sich der Pfad für die Teilnehmer: jeweils abwechselnd teilte der Server die Personen entweder zu einer Version des B5PO-Rs im Listen-Format, einer Version des B5PO-Rs im One-Item-a-Time-Format oder einer Version des B5PO-Rs im Q-Sort-Format zu. Damit sollte sichergestellt werden dass jede der drei B5PO-R-Versionen von gleich vielen Personen ausgefüllt werden, wobei jedoch am Ende der Erhebung aufgrund von auszuschließenden Personen nicht exakt gleiche Gruppengrößen resultierten. Die verwendeten Items des B5PO-Rs wurden selbstverständlich in den drei Gruppen in jeweils derselben, quasi-randomisieren Reihenfolge vorgegeben, um etwaige Item-Sequenz-Effekte zu vermeiden. Schlussendlich gelangten sämtliche Teilnehmer zum letzten Schritt der Befragung, nämlich dem Akzept!-P, welcher ihre Meinung zur soeben ausgefüllten B5PO-R-Version erfragte. Die Existenz weiterer Formate neben dem, das sie selbst ausgefüllt hatten, war den Teilnehmern dabei nicht bewusst. Somit folgte die vorliegende Untersuchung einem einfaktoriellen, balancierten Mehrgruppenplan mit Randomisierung (Bortz & Döring, 2006).

Am Ende der Erhebung hatten die Teilnehmer noch die Möglichkeit, freie Anmerkungen zur Untersuchung zu hinterlassen. Außerdem wurde den Teilnehmern angeboten, als Dank für ihre Teilnahme eine Email zu erhalten in der sie bezüglich ihrer Antworten im B5PO-R nach Abschluss der Auswertungen Feedback bekommen würden.

Die Teilnehmer wurden sowohl vor als auch nach dem Ausfüllen der Online-Fragebögen darüber informiert, dass ihre Daten komplett vertraulich behandelt werden würden und dass vor allem ihre Emailadresse nicht an Dritte weitergegeben würde.

7.2. Erhebungsinstrumente

Im Folgenden sind die drei verschiedenen Abschnitte des Erhebungsinstrumentes detailliert dargestellt. Im Anhang finden sich unter *11. Screenshots des Online-Assessments* Screenshots der jeweiligen Schritte der Befragung.

7.2.1. Demographie

Nach einer Seite mit initialen Instruktionen wurden die Teilnehmer/Innen der Untersuchung gebeten einige Angaben zu ihren demographischen Daten zu machen. Dies geschah einerseits um ein besseres Bild der resultierenden Stichprobe zeichnen zu können, und andererseits um für spätere Rasch-Modell-Analysen Teilungskriterien zur Verfügung zu haben. Erhoben wurde (in dieser Reihenfolge) das Geschlecht, das Alter in Jahren, die höchste abgeschlossene Ausbildung, ob Deutsch die eigene Muttersprache sei und die Nationalität der jeweiligen befragten Person.

7.2.2. B5PO-R: Big Five Plus One Revised

Die Wahl eines Selbstbeschreibungsfragebogens fiel für die vorliegende Untersuchung auf den B5PO-R (Big Five Plus One Revised, Siptroth, 2011), der Neufassung des B5PO (Big Five Plus One, Holocher-Ertl et al., 2003), ein computerbasiertes Verfahren zur Big-Five-Persönlichkeitsdiagnostik durch Selbstbeschreibung anhand von Adjektiven. Für dessen Wahl gab es zahlreiche Gründe: zum einen, wie bereits unter *6.1. Hypothese 1: Veränderungen der inneren Konsistenzen der Skalen* angesprochen, wurde der Itempool unter anderem mit den Mitteln der Item-Response-Theory entwickelt und ist nachweislich Rasch-Modell-konsistent (Siptroth, 2011). Dies erlaubt weiterführendere Analysen zur inneren Konsistenz des Pools als bei den meisten gängigen Selbstbeschreibungsfragebögen möglich wäre. Zudem handelt es sich bei den erhobenen Persönlichkeitsfacetten um wohl bekannte und untersuchte Konstrukte der psychologischen Persönlichkeitsmessung, welche auch bereits im Vorfeld auf ihre prinzipielle Rasch-Modell-Skalierbarkeit untersucht wurden (Rost et al., 1999). Ein weiterer Vorteil des Instruments ist seine Äquivalenz für den deutschen sowie für den österreichischen Raum: da prinzipiell bei Online-Verteilung eines Links zu erwarten ist, dass die resultierende Stichprobe Personen verschiedener Nationalitäten mit der gewählten Sprache beinhalten wird, ist die Tatsache, dass bei der Konstruktion des B5PO-R die Äquivalenz für beide Länder bedacht wurde, ein großer Vorteil des Instruments. Auch die Ökonomie des Verfahrens war für das Akquirieren

der Stichprobe ein willkommener Aspekt. Zusätzlich zu diesen Vorteilen des Verfahrens soll noch erwähnt werden, dass der B5PO-R in seiner Standardfassung eines der wenigen Verfahren ist, welche das unter *1.2. Das One-Item-a-Time-Format* besprochene One-Item-a-Time-Format der Itempoolpräsentation realisieren.

Der B5PO-R erhebt in seiner ursprünglichen Fassung sechs voneinander unabhängige Facetten der Persönlichkeit. Namentlich handelt es sich dabei um die Faktoren Gefühlsbetontheit (Empathie), Verträglichkeit, Extraversion, Sorgfältigkeit (Gewissenhaftigkeit), Emotionale Kontrolle und Offenheit. Hierbei ist auffallend, dass es sich nicht um die unter *4. Das Big Five-Faktorenmodell der Persönlichkeit* beschriebenen üblichen fünf Faktoren handelt, sondern um eine adaptierte Faktorenstruktur. Der zusätzliche Faktor der Empathie ergab sich bei der Konstruktion des B5PO-R im Zuge der angestrebten Eindimensionalität sensu Rasch-Modell aus dem Faktor der Verträglichkeit heraus (Siptroth, 2011). Er beschreibt das gefühlsbetonte Einfühlungsvermögen einer Person. Da der zentrale Punkt der Fragestellung dieser Untersuchung jedoch die Übersichtlichkeit über das auszufüllende Instrument ist, wurde entschieden nur drei der sechs Faktoren des Instruments vorzulegen. Jene drei Faktoren sollten für sich in der Lage sein etwaige Veränderungen im Antwortverhalten der befragten Personen durch die Manipulation der Übersicht über den Itempool und bereits gegebene Antworten abzubilden. Auch wurde dadurch die Zumutbarkeit und Ökonomie des vorgelegten Fragebogens erhöht. Die Wahl fiel auf die Faktoren Extraversion, Verträglichkeit und Gewissenhaftigkeit; Extraversion wurde gewählt, da jener Faktor zu den in der Allgemeinbevölkerung bekanntesten zählt, was durch den Niederschlag der Vokabel in der nicht-fachlichen Sprache deutlich wird. Verträglichkeit wurde verwendet, da dieser Faktor unter Umständen Informationen über die Tendenz zu Antworten in Richtung sozialer Erwünschtheit abzubilden vermag (vgl. in Kubinger, 2002). Aufgrund der bei der Fragebogenkonstruktion intendierten Unabhängigkeit der Skalen des B5PO-R ist die Extraktion der Skalen ohne resultierenden Informationsverlust – über den offensichtlichen hinaus – möglich.

Obwohl unter *4. Das Big Five-Faktorenmodell der Persönlichkeit* bereits allgemeine Beschreibungen der Faktoren erfolgten, sollen die hier verwendeten Faktoren mit direkten Bezug zu den Beschreibungen der ursprünglichen Verfahrenskonstruktion nach Siptroth (2011) paraphrasiert erfolgen:

- Extraversion mit 9 Items

  Personen, welche auf diesem Faktor hoch scoren, werden als gesellig, optimistisch und sozial aktiv beschrieben. Niedrige Werte beschreiben eher Personen welche eine Vorliebe für Rückzug an sich haben, wobei dies nicht eine Abwesenheit sozialer Kompetenzen bedeuten muss.
- Sorgfältigkeit (Gewissenhaftigkeit) mit 9 Items

  Dieser Faktor beschreibt gewissermaßen die Fähigkeiten einer Person zum Selbstmanagement und beinhaltet Eigenschaften wie Fleiß, Zuverlässigkeit und Ordnung, diametral entgegengesetzt zu Planlosigkeit, Gleichgültigkeit und Nachlässigkeit.
- Verträglichkeit mit 12 Items

  Verträglichkeit ist ein Faktor mit zwischenmenschlichem Fokus. Er beschreibt Personen die hilfsbereit, mitfühlend und altruistisch veranlagt, sowie an einem warmen zwischenmenschlichen Verhältnis orientiert sind. Demgegenüber sprechen niedrige Ausprägungen in dieser Skala für tendenziell eher egoistische, kühle Personen.

Üblicherweise verwendet der B5PO-R eine Analogskala zur Beantwortung seiner Adjektivpaare durch die befragten Personen. Der betreffenden Person wird pro Seite ein Adjektivpaar über einer Analogskala vorgegeben, woraufhin die Person den Regler der Analogskala in Richtung des Adjektivs verschieben soll, welches sie ihrer Meinung nach besser beschreibt. Hierbei wird der durch die befragte Person markierte Bereich im Nachhinein digital dichotomisiert. Jene Dichotomisierung erfolgt pro Faktorzugehörigkeit eines Items an unterschiedliche Stellen des Antwortkontinuums der Analogskala, abhängig von der Schiefe der resultierenden Verteilung der summierten Antworten. Die Dichotomisierung der drei hier verwendeten Skalen erfolgt in der Standardfassung des B5PO-Rs an den Stellen 90:10 (Gewissenhaftigkeit und Verträglichkeit) beziehungsweise 80:20 (Extraversion) des Antwortspektrums.

Im Fall der vorliegenden Untersuchung wurde jedoch gegen die Verwendung der üblichen Analogskala entschieden, da die Verwendung des Q-Sort-Präsentationsformats dieses Vorgehen ausschließt. Stattdessen wurde, um optimale Vergleichbarkeit zu ermöglichen, in jeder der drei Bedingungen eine sechsstufige Ratingskala implementiert. Abbildung 7.1 enthält als Beispiel für die (durchgehend idente) Benennung der Antwortkategorien die verwendete Ratingskala des Listen-Formats.

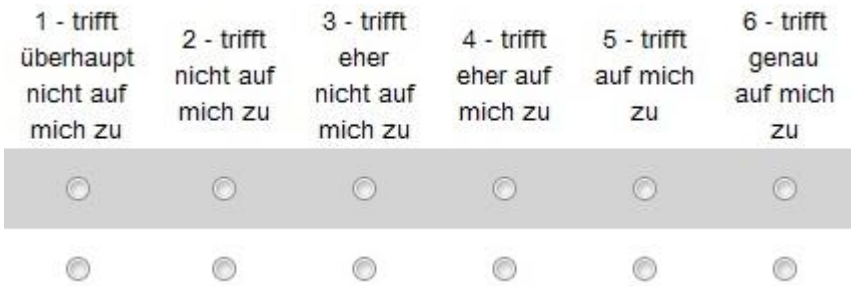


Abbildung 7.1. exemplarische Darstellung der Ratingskala beim Listen-Format mit Benennungen der Antwortkategorien

Durch das Fehlen einer mittleren Antwortkategorie erlaubt jenes Vorgehen trotz der Veränderung des Antwortmodus weiterhin Dichotomisierungen der gegebenen Antworten an der 50/50-Marke sowie an weiteren Stellen des Spektrums der wählbaren Antwortmöglichkeiten. Der Verzicht auf die mittlere Antwortkategorie hatte neben der Erleichterung der Dichotomisierung auch den Grund, den Rost et al. (1999) in ihrer Studie zur IRT-Skalierung eines Big-Five Selbstbeschreibungsfragebogens ansprechen:

> *„[…]das fünfstellige Antwortformat erweist sich als nicht kompatibel mit der Eindimensionalität der betreffenden Skalen. Die mittlere Antwortkategorie wird nicht dann gewählt, wenn die Person einen mittleren Ausprägungsgrad der gemessenen Dimension hat, sondern generell zu selten und wahrscheinlich aus anderen Gründen als einer mittleren Dimensionsausprägung."* (Rost et al., 1999; S.125)

Zusätzlich zu den soeben besprochenen Veränderungen musste eine weitere Adaption der Standardfassung des B5PO-Rs vorgenommen werden: da üblicherweise Adjektivpaare verwendet werden mussten diese auf einzelne Adjektive reduziert werden, da die Implementation des Q-Sort-Präsentationsformats zwar mit einzelnen, nicht jedoch mit Adjektivpaaren funktioniert. Um die besprochene Übersichtlichkeit möglichst hoch zu halten wurde entschieden sich dabei auf die „positiven" Adjektive zu beschränken. Tabelle 7.1 gibt ein Itembeispiel zur vorgenommen Adaptierung der Items wider.

Tabelle 7.1

Beispiel für ursprüngliche und verwendete Adjektive des B5PO-R

Zugehörige Skala	Ursprüngliches Adjektivpaar	Verwendetes Adjektiv
Gewissenhaftigkeit	verantwortungsvoll – verantwortungslos	verantwortungsvoll

Sowohl der B5PO-R in seiner ursprünglichen Fassung als auch die hier verwendete adaptierte Version beschreiben hohe Ausprägungen in den jeweiligen Faktoren beziehungsweise Persönlichkeitsfacetten mit hohen Skalensummen (beziehungsweise Personenparametern des Rasch-Modells).

Natürlich ist wegen der Verwendung ausschließlich „positiver" Items ein gewisser Grad an Akquieszenz in den Antworten der Teilnehmer/Innen der Studie zu erwarten; jedoch war diese Beschränkung für die Untersuchung der Hypothese 2 (siehe *6.2. Hypothese 2: Unterschiede in der Nutzung des verfügbaren Antwortspektrums*) und der Implementierung des Q-Sort-Formats zwingend notwendig.

7.2.3. Akzept!-P

Zur differenzierten Erfassung der Akzeptanz des B5PO-Rs in den verschiedenen Itempool-Präsentationsmodi durch die Studienteilnehmer/Innen wurden sie im Anschluss an dessen Bearbeitung zum Akzept!-P[1] (Kersting, 2012) weitergeleitet. Der Akzept!-P aus der Gruppe der Akzept!-Verfahren (Kersting, 2010) fordert die befragten Personen auf, ihre Meinung zu bestimmten Aspekten eines Instrumentes mittels einer sechsstufigen Ratingskala zu beantworten. Dabei enthält der Fragebogen auch zahlreiche negativ gepolte Items, welche derart verrechnet werden, dass schlussendlich hohe Summenscores für hohe Akzeptanz des Verfahrens stehen. Sämtliche Skalensummen errechnen sich aus je vier Items, einzige Ausnahme bildet die Skala „Antwortformat" mit drei Items. Zusätzlich erfragt das Verfahren am Ende eine Gesamtbewertung des zuvor aufgefüllten Instruments mittels des deutschen Schulnotensystems. Im Zuge der Voruntersuchung wurde allerdings klar, dass zur sinnvollen Verwendung des Akzept!-P mehrere kleinere Adaptionen vorgenommen werden müssen: in

[1] An dieser Stelle soll Dr. Martin Kersting für die Bereitstellung des Akzept!-P für die vorliegende Diplomarbeit herzlich gedankt sein.

erster Linie mussten einige Formulierungen geändert werden, so wie beispielsweise der übliche Bezug auf „Fragen bzw. Aussagen“ auf „Eigenschaftsworte“ geändert wurde, um mit dem B5PO-R zu korrespondieren. Weitere Ausführungen hierzu finden sich unter *7.3. Voruntersuchung*. Tabelle 7.2 führt die fünf verschiedenen Skalen zusammen mit je einem Beispielitem an.

Tabelle 7.2

Beispielitems des Akzept!-P[2]

Skala	(adaptiertes) Beispielitem
Kontrollierbarkeit	"Bei der Bearbeitung der Eigenschaftsworte wusste ich jederzeit, was ich tun muss."
Messqualität	"Das Verfahren ermöglicht es, die zwischen verschiedenen Menschen bestehenden Unterschiede in dem vom Verfahren erfassten Merkmal exakt zu messen."
Augenscheinvalidität	"Dass man mit den Eigenschaftsworten geeignete Personen für einen Job herausfinden kann, ist zu bezweifeln." (negativ gepolt)
Wahrung der Privatsphäre	„Wie ich mich bezüglich solcher Eigenschaften einschätze, geht diejenigen, die die Verfahrensergebnisse erhalten, nichts an.“ (negativ gepolt)
Antwortfreiheit	"Aufgrund der vorgegebenen Antwortmöglichkeiten hatte ich nicht die Freiheit, so zu antworten, wie es für meine Person zutreffend ist." (negativ gepolt)

Zum Zeitpunkt dieser Untersuchung befindet sich der Akzept!-P noch in seiner Evaluierungsphase. Dementsprechend existieren derzeit noch keine veröffentlichten Ergebnisse bezüglich der Faktorenstruktur oder testtheoretischen Eigenschaften. Allerdings korrespondiert der Akzept!-P weitestgehend mit den bereits evaluierten Akzept!-L und Akzept!-AC (Kersting, 2012; Kersting, 2010) und weist sehr deutliche Plausibilität und inhaltliche Gültigkeit auf. Aufgrund dieser Gründe wurde entschieden, im hiesigen Kontext keine separate

[2] Bei den dargestellten Beispielitems handelt es sich um die öffentlich einsehbaren Beispielitems der Website der Akzept!-Verfahrensfamilie (Kersting, 2012)

Faktorenanalyse durchzuführen. Stattdessen wird über die konkrete Verwendbarkeit der jeweiligen Skalen mittels deskriptiver Parameter wie Cronbachs α entschieden werden.

7.3. Vorerhebung

Um Verständnisproblemen bei den Instruktionen vorzubeugen wurde eine Voruntersuchung in zwei Wellen mit kleinem Umfang durchgeführt (pro Welle: n = 4). Außerdem sollte untersucht werden, ob das Ausfüllen für die Teilnehmer sowohl mit Maus als auch mit Touchpad oder Tablet problemlos möglich ist. In der ersten Welle wurden vier Personen gebeten, das Instrument im Setting Q-Sort auszufüllen und anschließend einige Fragen zu beantworten. Außerdem hatten die Teilnehmer/Innen noch die Möglichkeit, freie Anmerkungen zum Instrument zu machen. Da in der ersten Welle der Voruntersuchung mehrere Probleme auftraten, wurden einige Änderungen am Instrument vorgenommen und eine zweite Welle durchgeführt; nachdem in der zweiten Welle besagte Probleme nicht mehr auftraten, wurden keine weiteren Voruntersuchungen durchgeführt. Teilnahmebedingung für die Voruntersuchung war einerseits kein/e Studierende/r der Psychologie zu sein und noch nie an einer Befragung mittels Q-Sort teilgenommen zu haben.

Die Probleme welche sich im ersten Durchgang zeigten bezogen sich wider Erwarten hauptsächlich auf den Akzept!-P. Zum einen wurde kritisiert, dass die Beschriftung der Ratingskala beim Scrollen nicht mehr sichtbar sei, was das Ausfüllen behindere; zum anderen merkten die befragten Personen an, dass nicht vollkommen klar sei ob sich die Aussagen des Akzept!-P auf den B5PO-R oder auf den Akzept!-P selbst beziehen würden. Nach Anpassung der Instruktionen und des technischen Layouts der Seite traten in der zweiten Welle besagte Probleme nicht mehr auf, woraufhin zur Haupterhebung fortgeschritten wurde.

Der Vollständigkeit halber soll angemerkt werden, dass beim B5PO-R-Q-Sort in keiner der Wellen nennenswerte Probleme auftraten, die Teilnehmer jedoch konsistent den Akzept!-P kritisierten; häufig genannte Punkte beinhalten doppeltes Abfragen ähnlicher Thematiken sowie verwirrende und unklare Formulierungen durch negativ gepolte Items, deren Beantwortung eine doppelte Verneinung durch die Personen erforderte.

7.4. Durchführung der Untersuchung

Die Durchführung der Untersuchung begann am 23.10.2012 und wurde am 11.12.2012 abgeschlossen, und fand wie oben unter *7.1. Untersuchungsplan* beschrieben statt. Der Zugang zum Fragebogen wurde mittels Online-Links über diverse soziale Netzwerke ermöglicht; außerdem wurden Links zur Umfrage über mehrere Email-Verteilerlisten versendet.

7.5. Stichprobe

Entsprechend der Kategorisierung von Stichprobenursprüngen nach Bortz und Döring (2006) wurde für die hier durchgeführte Untersuchung eine Ad-Hoc-Stichprobe herangezogen. Ursprünglich wurden insgesamt *N* = 392 Personen mit dem endgültigen Instrument befragt, wobei es keine Beschränkungen bezüglich Alter oder anderer demographischer Variablen gab. Die Beschreibung der Stichprobe bezüglich ihrer demographischen Kennwerte erfolgt in *8.1. Beschreibung der demographischen Variablen*.

8. Ergebnisse

In diesem Abschnitt der Arbeit folgen nun die Ergebnisse der besprochenen Untersuchung. Die statistischen Analysen der Daten erfolgten durchgehend durch das Basispaket des Auswertungsprogramms R (R Core Development Team, 2011); an jenen Stellen, an denen weitere Pakete oder Formeln zur Anwendung kamen werden jene explizit erwähnt werden.

8.1. Beschreibung der demographischen Variablen

Wie bereits erwähnt enthielt die Stichprobe nach Beendigen der Erhebungen insgesamt *N* = 392 Personen; allerdings musste die Person mit der ID = 82 aufgrund unplausibler demographischer Angaben und überdeutlicher Verwendung von stereotypen Antwortmustern aus der Stichprobe entfernt werden. Zusätzlich wurden zwei weitere Personen von der Auswertung ausgeschlossen (ID = 394 und ID = 306), da ihre Bearbeitungszeiten des B5PO-R extreme zeitliche Ausreißer darstellten, und bei derart hohen Abweichungen nicht mehr sichergestellt werden kann, dass die betreffenden Personen das Instrument ohne substantielle Pausen beantwortet haben. Die

beiden ausgeschlossenen Personen lagen in ihren Bearbeitungszeiten je 9,6 und 73,0 Standardabweichungen vom Mittelwert entfernt.

Somit resultierte eine tatsächliche Stichprobengröße von $N_{korrigiert}$ = 389. Von diesen 389 Personen bearbeiteten 128 (32.90%) den B5PO-R im Listen-, 130 (33.42%) im One-Item-a-Time-, und 131 Personen (33.68%) im Q-Sort-Format. Somit ergibt sich aufgrund der oben angesprochenen Entfernungen einiger Personen ein sehr leicht unbalanciertes Design der Untersuchung.

Des Weiteren waren 261 (67.10%) Personen der Stichprobe weiblich und 128 (32.90%) männlich. Die Range des Alters erstreckte sich von 15 bis 74 Jahren, mit einem Mittelwert von M = 36,23 und einer Standardabweichung von SD = 12,73. 96,14% (*n* = 374) der befragten Personen gaben Deutsch als ihre Muttersprache an; 3,86% (*n* = 15) Personen wählten die Kategorie „andere“. Die Nationalitäten der befragten Personen teilen sich wie folgt auf: 77,63% (*n* = 302) gaben Deutsch als ihre Nationalität an, 19,28% (*n* = 75) nannten Österreich und 3,08% (*n* = 12) der befragten Personen wählten die Kategorie „andere“.

Zuletzt sind in Tabelle 8.1 die Angaben zur höchsten abgeschlossenen Ausbildung der befragten Personen aufgeschlüsselt.

Tabelle 8.1

Verteilung der höchsten abgeschlossenen Ausbildungen der Stichprobe

	Kein Schulabschluss	Pflichtschule	Realschule/ berufsbildende mittlere Schule	Lehre	Matura/ Abitur/ Fachabitur	Akademie/ Kolleg	Universität / Fachhochschule
	0,77%	0,51%	13,62%	5,40%	31,62%	4,11%	43,96%
n =	3	2	53	21	123	16	171

8.2. Rasch-Modell-Analysen (Hypothese 1)

Um, wie bereits unter 6.*1. Hypothese 1: Veränderungen in den inneren Konsistenzen der Faktoren* besprochen, die Reliabilitäten der Faktoren des B5PO-R im Rahmen der inneren Konsistenz im Wechsel der Präsentationsformate genauer zu untersuchen, wurden pro Faktor Modellpassungstests zur Überprüfung der Rasch-Modell-Konsistenz durchgeführt – zunächst für die Daten der drei Präsentationsmodi gemeinsam, und anschließend für die jeweiligen Präsentationsmodi in sich. Dazu benötigt das Rasch-Modell Daten in Form von dichotomen beziehungsweise dichotomisierten Itemantworten, welche im Kontext eines Persönlichkeitsfragebogens die Form der Kodierung „0 – trifft nicht auf mich zu“ „1 – trifft auf mich zu“ annehmen sollten.

In der Konstruktionsdokumentation des B5PO und des B5PO-R wurde diesbezüglich folgendermaßen vorgegangen: Nach Betrachtung der resultierenden Verteilungen der Faktoren wurde aufgrund Verzerrungen der Antworten auf der verwendeten Analogskala entschieden, die Dichotomisierung nicht am Mittelpunkt der Analogskala vorzunehmen. Stattdessen wurde der Median der jeweiligen Skalen berechnet und das Verhältnis des Bereiches der Range über und unter dem Median auf die Analogskala umgelegt (Holocher-Ertl et al., 2003; Siptroth, 2011). Damit resultierten neue Mittelpunkte auf den Analogskalen. Die Dichotomisierung nach jenen neuen Mittelpunkten korrigierte die Verteilungsschiefe weitgehend. Mit den so gewonnenen neu dichotomisierten Daten wurde dann das Rasch-Modell berechnet.

Im Falle der vorliegenden Untersuchung wurde nun zunächst die Dichotomisierung an der Mitte der sechsstufigen Ratingskala vollzogen, zwischen „3 – trifft eher nicht auf mich zu“ und „4 – trifft eher auf mich zu“. Also wurde die Ratingantwort aller Personen pro Item bei einer Antwort von „1 – trifft überhaupt nicht auf mich zu“ bis „3 – trifft eher nicht auf mich zu“ mit „0 – trifft nicht auf mich zu“ kodiert, und bei einer Antwort von „4 – trifft eher auf mich zu“ bis „6 – trifft genau auf mich zu“ mit „1 – trifft auf mich zu“. Dies korrespondiert mit dem inhaltlichen Labeling der Ratingstufen, da die Stufen von 1 bis 3 recht eindeutig einer ablehnenden, die Stufen von 4 bis 6 einer zustimmenden Haltung entsprechen. Wie während der Konstruktion des ursprünglichen Fragebogens entstanden dadurch stark linksschiefe Verteilungen der Summen der dichotomisierten Itemantworten. Abbildung 8.1 verdeutlicht dies am Beispiel des Faktors Extraversion.

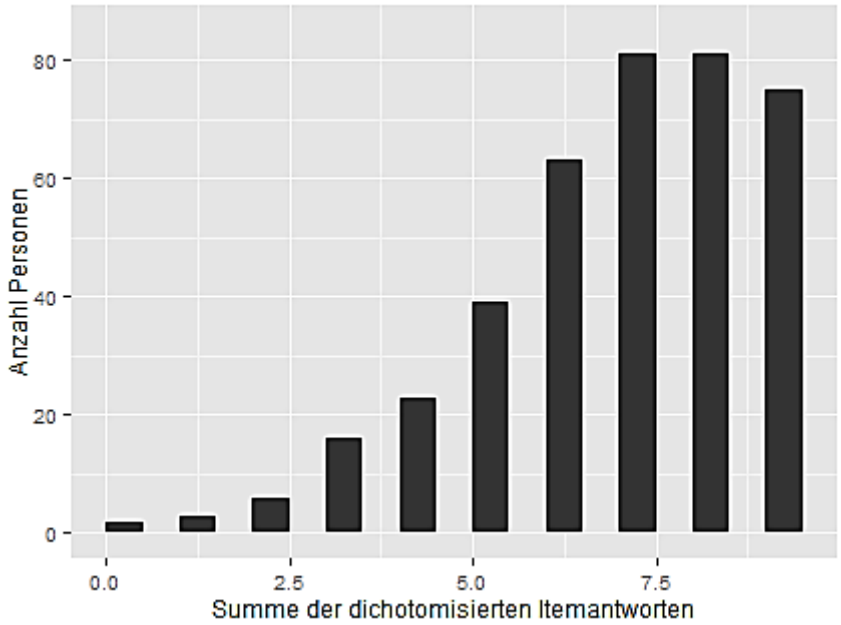


Abbildung 8.1. Verteilung der Rohscores der ursprünglich dichotomisierten Ratingantworten des Faktors Extraversion in der gesamten Stichprobe

Bei der Überprüfung der Antwortmuster der dichotomisierten Itemantworten wurde das Problem der ursprünglich dichotomisierten Daten noch deutlicher: durch die Dichotomisierung am gewählten Mittelpunkt zwischen „3 – trifft eher nicht auf mich zu“ und „4 – trifft eher auf mich zu“ waren nunmehr derart viele Personen pro Item mit „1 – trifft auf mich zu“ kodiert, dass die Schätzung der Parameter des Rasch-Modells in manchen der intendierten Untergruppen nicht mehr auf sinnvolle Art möglich war. Somit würden bei Verrechnung der Antworten nach ursprünglicher Dichotomisierung viele der Items ihre Differenzierungsfähigkeit verlieren. Dies rührt vermutlich daher, dass das Instrument – wie in der Dokumentation seiner Konstruktion bereits angemerkt – leichte Antwortverzerrungen in Richtung der zustimmenden Seite der Skala provoziert. Tabelle 8.2 veranschaulicht diese Verzerrung der Antworten.

Tabelle 8.2

Häufigkeitsverteilung der gewählten Antwortkategorien in der gesamten Stichprobe

Ratingantwort	1	2	3	4	5	6
relative Häufigkeit	.01	.04	.13	.28	.35	.18
absolute Häufigkeit	107	476	1571	3262	4111	2143

Somit wurde beschlossen, die Dichotomisierung in Analogie zur Konstruktion des B5PO-R an einem höheren Punkt der Ratingskala durchzuführen um die Schiefe der Verteilungen der summierten dichotomisierten Itemantworten zu korrigieren. Da die hier verwendete sechsstufige Ratingskala allerdings deutlich weniger Abstufungen erlaubt als die ursprüngliche Analogskala (dort waren 21 Inkremente der Analogskala in Verwendung), entspricht die neue Dichotomisierung der Ratingskala nicht immer ganz dem Verhältnis der Bereiche über und unter des Medians der jeweiligen Summenscores der Faktoren. Kriterium für den neuen Mittelpunkt der Dichotomisierung war das Entstehen einer möglichst wenig schiefen Verteilung der summierten dichotomisierten Itemantworten. Tabelle 8.3 gibt einen Überblick über die Kennwerte der neuen Dichotomisierungen, während Abbildung 8.2 am Beispiel des Faktors Extraversion die korrigierte, weniger schiefe Verteilung exemplarisch wiedergibt. Die Schiefe der Verteilungen wurde mittels des R-Pakets „psych“ (Revelle, 2011) berechnet.

Tabelle 8.3

Kennwerte der ursprünglichen und neuen Dichotomisierung

Faktor	Schiefe vorher	Median vorher	Range	$\frac{\text{Median}}{\text{Range}} \times 6$	Neuer Mittelpunkt der Ratingskala	Neue Schiefe	Neuer Median
Extraversion	-0.87	7	9	4.7	[1:4] ; [5:6]	0.25	3
Gewissenhaftigkeit	-1.05	8	9	5.3	[1:4] ; [5:6]	-0.14	5
Verträglichkeit	-1.25	11	12	5.5	[1:4] ; [5:6]	-0.22	7

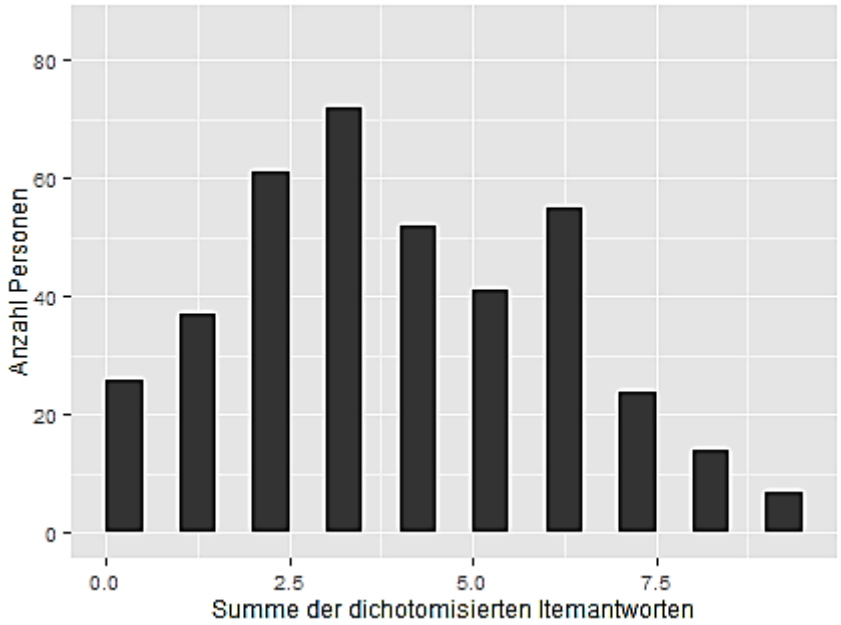


Abbildung 8.2. Verteilung der Rohscores der neu dichotomisierten Ratingantworten des Faktors Extraversion in der gesamten Stichprobe

Nach der beschriebenen Dichotomisierung wurden verschiedene Modelltests durchgeführt, welche im Folgenden beschrieben werden sollen. Da die Skalierung des Itempools eines Faktors nach dem Rasch-Modell eine hohe Anzahl an wiederholten Modelltests mit sich bringt wurde das angestrebte Signifikanzniveau im Gegensatz zu den anderen Signifikanztests dieser Studie auf $\alpha \leq .01$ gesetzt. Zusätzlich wurde darauf verzichtet Waldtests durchzuführen, da diese die Kumulierung des Risikos für einen Fehler erster Art weiterhin fördern. Außerdem liefern sie kaum Informationen, welche nicht auch aus den Goodness-of-Fit-Plots entnommen werden können. Um jene Items zu identifizieren, deren Parameterschätzungen nicht dem Rasch-Modell entsprechen, wurden eben dieses Graphiken zur Überprüfung der Modellpassung (Goodness-of-Fit-Plots, im Folgenden kurz mit GOF-Plots abgekürzt) verwendet (Rasch, Kubinger und Yanagida, 2011). Dabei sind die Skalierungen der GOF-Plots nicht einheitlich gewählt, sondern immer in einem Maßstab gehalten der die Identifizierung von dem Modell abweichender Items möglichst einfach halten soll.

Sämtliche Rasch-Modell-Schätzungen, GOF-Plots und Modelltests wurden mit dem R-Paket „eRm“ (Mair, Hatzinger, & Maier, 2012) durchgeführt beziehungsweise erstellt. Aufgrund urheberrechtlicher Gründe wurde auf die Angabe konkreter Iteminhalte verzichtet.

8.2.1. Faktor 1: Extraversion

Zunächst wurden die Schätzungen der Parameter des Rasch-Modells für den Faktor Extraversion über alle Gruppen hinweg vorgenommen. Daraufhin wurden mittels der erhaltenen Schätzungen anhand mehrerer Trennungskriterien Andersens Likelihood-Ratio-Tests durchgeführt. Das erste Trennungskriterium stellten die drei Gruppen der unterschiedlichen Präsentationsformate dar. Für das zweite Trennungskriterium wurde die Stichprobe – wie so üblich – in die beiden Hälften mit Werten jeweils unter beziehungsweise über dem Median geteilt. Das dritte Trennungskriterium stellte das Geschlecht der befragten Personen dar, das vierte war die Aufteilung der Personen in Gruppen, die jeweils unter beziehungsweise über dem Median des Alters lagen. Die Ergebnisse jener initialen Tests sind in Tabelle 8.4 wiedergegeben.

Tabelle 8.4

Ergebnisse der Andersens Likelihood-Ratio-Tests für den gesamten Itempool des Faktors Extraversion in allen Präsentationsformaten

Trennungskriterium	χ^2	*df*	*p*
Präsentationsformat	8.96	16	.915
Anzahl zutreffender Adjektive (Median)	37.34	8	< .001*
Geschlecht	34.34	8	< .001*
Alter	60.26	8	< .001*

*: $p \leq .01$

Aus Tabelle 8.4 geht hervor, dass alle Trennungskriterien außer „Präsentationsformat“ signifikante Modellabweichungen feststellen. Dementsprechend wurden graphische Überprüfungen der Modellgeltung für jene Trennungskriterien durchgeführt, welche in den Abbildungen 8.3, 8.4 und 8.5 wiedergegeben sind.

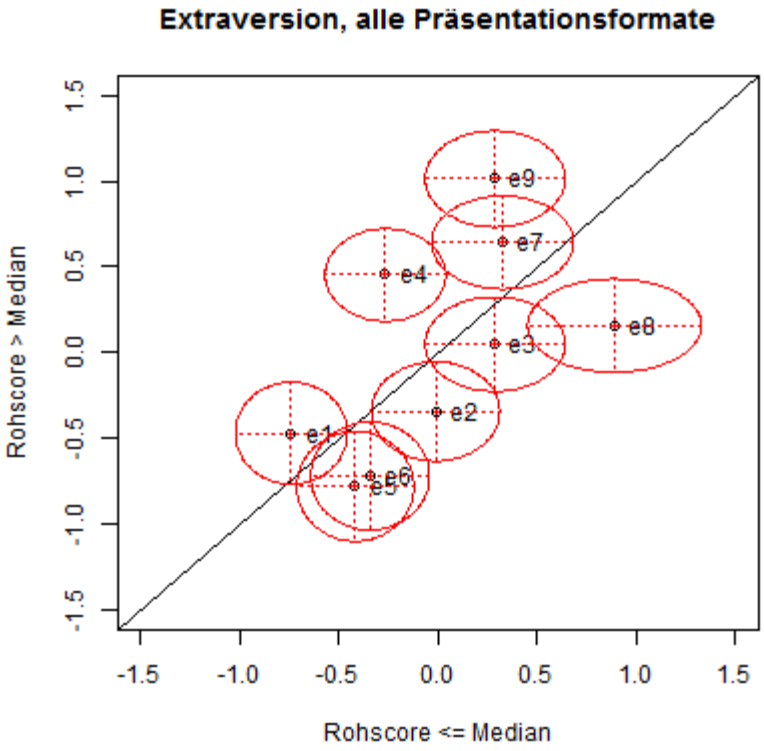


Abbildung 8.3. GOF-Plot für den gesamten Itempool des Faktors Extraversion in allen Präsentationsformaten nach dem Teilungskriterium „Anzahl zutreffender Items (Median)“

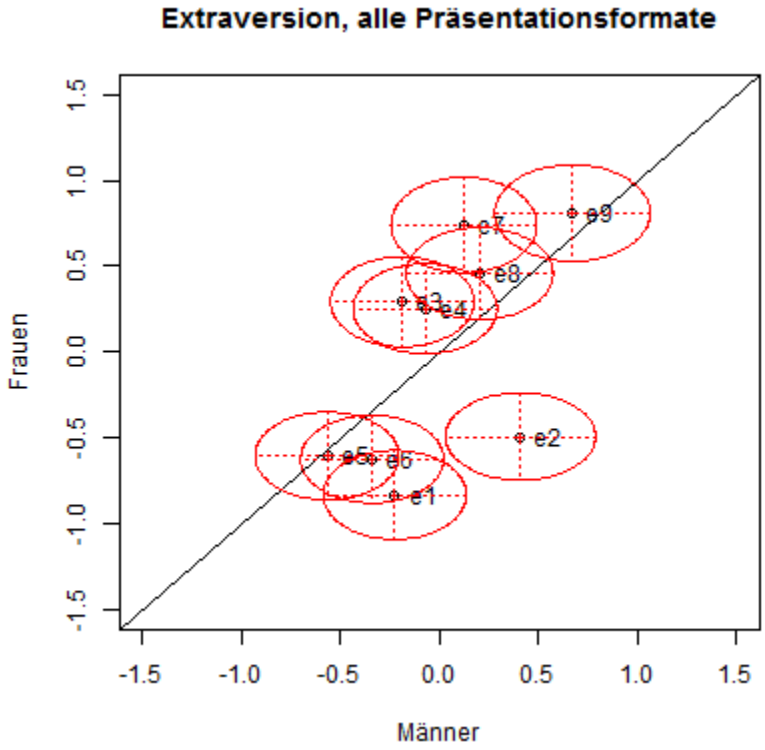


Abbildung 8.4. GOF-Plot für den gesamten Itempool des Faktors Extraversion in allen Präsentationsformaten nach dem Teilungskriterium „Geschlecht“

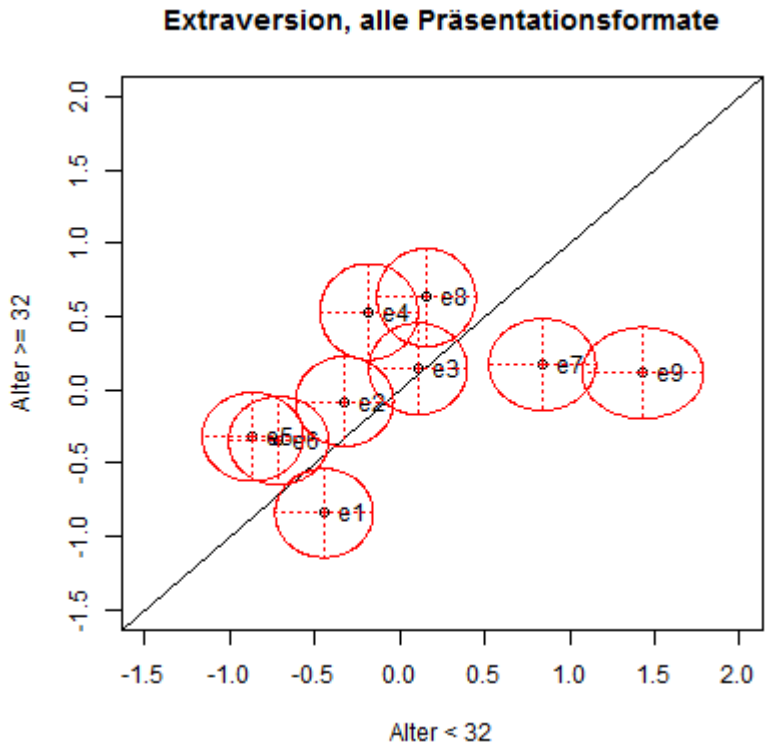


Abbildung 8.5. GOF-Plot für den gesamten Itempool des Faktors Extraversion in allen Präsentationsformaten nach dem Teilungskriterium „Alter“

Wie aus den graphischen Überprüfungen der Modellgeltung hervorgeht, waren pro Teilungskriterium mehrere Items auffällig in ihrer Abweichung. Die Items mit den auffälligsten Modellabweichungen waren dabei jeweils Item „e4“ beim Teilungskriterium „Anzahl zutreffender Items (Median)“, Item „e2“ beim Teilungskriterium „Geschlecht“ und Item „e9“ beim Teilungskriterium „Alter“.

Als nächstes wurden die auffälligen Items zuerst einzeln, danach in Kombination, aus dem Itempool entfernt und erneute Rasch-Modell-Schätzungen und Andersens Likelihood-Ratio-Tests durchgeführt. Keine der Kombinationen verbleibender Items erreichte dabei erheblich bessere Modellpassung. Die genauen Parameter und GOF-Plots der jeweiligen Schätzungen sollen hier aus Platzgründen nicht angegeben werden. Es soll jedoch angemerkt werden, dass unter Verwendung des Trennungskriteriums „Präsentationsformate“ bei keinem der durchgeführten Modelltests signifikante Modellabweichungen festgestellt werden konnten.

Nachdem für den Itempool des Faktors Extraversion über alle Präsentationsformate hinweg, auch unter der Entfernung weiterer Items, keine zufriedenstellende Rasch-Modellgültigkeit

bestätigt werden konnte, wurde mit der Untersuchung der Modellpassung innerhalb der einzelnen Gruppen der unterschiedlichen Präsentationsformate fortgefahren.

a) Rasch-Modell-Passung in der Gruppe Listen-Format, Faktor Extraversion

Analog zur Berechnung des Rasch-Modells über alle Gruppen hinweg wurde nun die Schätzung des Modells innerhalb der Gruppe des Präsentationsformats Liste für alle Items des Faktors Extraversion vorgenommen. Die Ergebnisse der Modelltests sind in Tabelle 8.5 zusammengefasst. Das Trennungskriterium „Präsentationsformat“ fehlt nun selbstverständlich, da hier ja nur die Daten des Präsentationsformats Liste zum Tragen kamen.

Tabelle 8.5

Ergebnisse der Andersens Likelihood-Ratio-Tests für den gesamten Itempool des Faktors Extraversion im Listen-Präsentationsformat

Trennungskriterium	χ^2	*df*	*p*
Anzahl zutreffender Adjektive (Median)	27.43	8	.001*
Geschlecht	14.76	8	.064
Alter	24.96	8	.002*

*: $p \leq .01$

Wie Tabelle 8.5 zu entnehmen ist, erzeugen die Modelltests nach den Teilungskriterien „Anzahl zutreffender Adjektive (Median)“ und „Alter“ signifikante Modelltests, nicht jedoch der Modelltest nach dem Kriterium „Geschlecht“. Abermals wurden Graphiken der Modellpassung (Abbildungen 8.6 und 8.7) angefertigt, um problematische Items zu identifizieren.

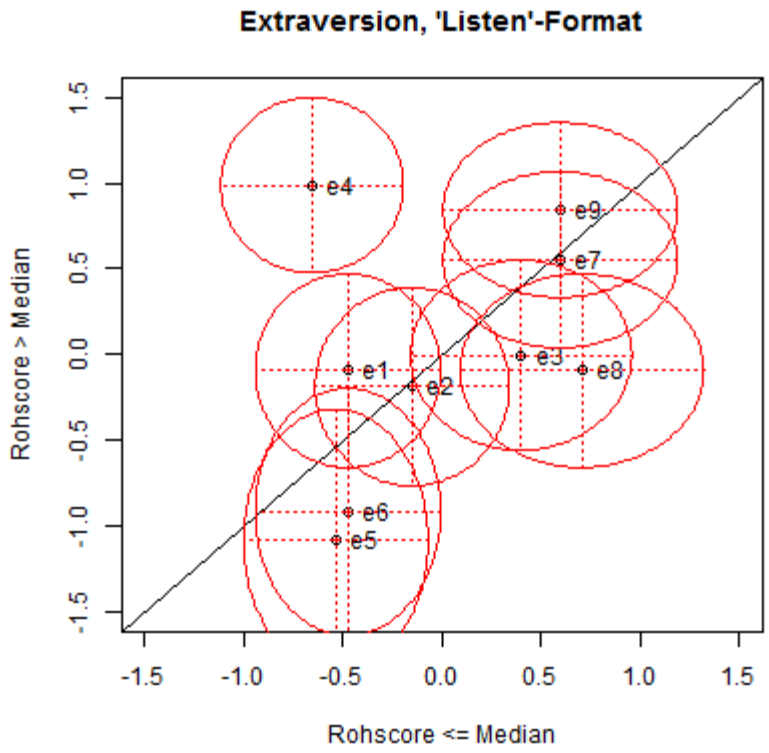


Abbildung 8.6. GOF-Plot für den gesamten Itempool des Faktors Extraversion im Listen-Präsentationsformat nach dem Teilungskriterium „Anzahl zutreffender Adjektive (Median)“

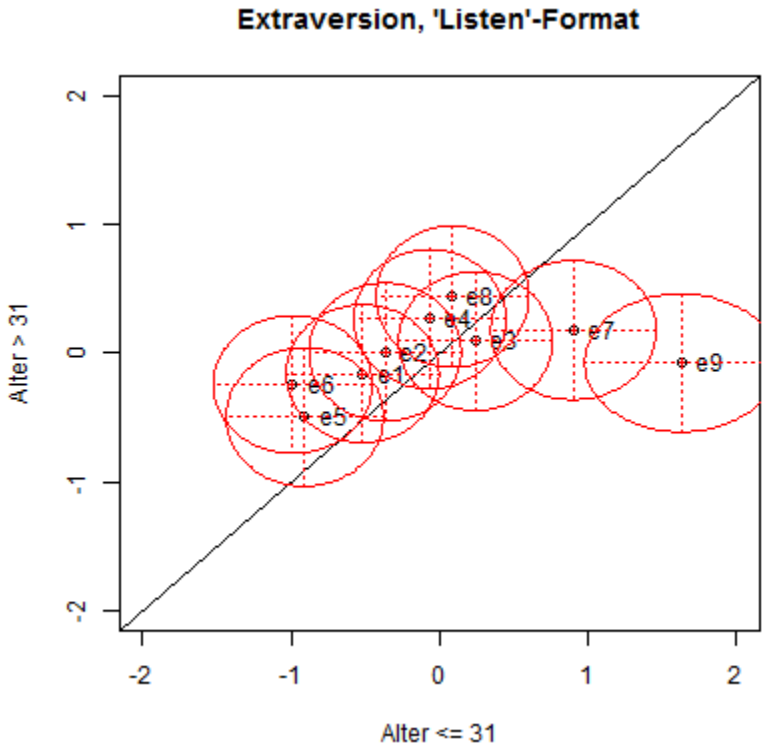


Abbildung 8.7. GOF-Plot für den gesamten Itempool des Faktors Extraversion im Listen-Präsentationsformat nach dem Teilungskriterium „Alter“

Hier waren jene Items mit den stärksten Abweichungen das Item „e4“ für das Teilungskriterium „Anzahl zutreffender Adjektive“, und das Item „e9“ für das Teilungskriterium „Alter“. Folglich wurden die beiden Items zunächst einzeln, dann gemeinsam aus dem Itempool entfernt und nochmals Andersens Likelihood-Ratio-Tests durchgeführt. Die Ergebnisse der Tests nach den jeweiligen Schritten sind in Tabelle 8.6 wiedergegeben.

Tabelle 8.6

Ergebnisse der Andersens Likelihood-Ratio-Tests des Faktors Extraversion im Listen-Präsentationsformat jeweils unter Ausschluss des Items „e4“ (Schritt 1), des Items „e9“ (Schritt 2), und beiden zusammen (Schritt 3)

	Trennungskriterium	χ^2	*df*	*p*
Schritt 1	Anzahl zutreffender Adjektive (Median)	7.95	7	.337
	Geschlecht	13.30	7	.065
	Alter	26.37	7	< .001*
Schritt 2	Anzahl zutreffender Adjektive (Median)	34.42	7	< .001*
	Geschlecht	13.24	7	.067
	Alter	8.19	7	.316
Schritt 3	Anzahl zutreffender Adjektive (Median)	8.74	6	.189
	Geschlecht	11.10	6	.085
	Alter	9.04	6	.171

*: $p \leq .01$

Wie sich aus Tabelle 8.6 entnehmen lässt, führte die Entfernung jeweils einzelner Items nicht zu umfassender Modellgültigkeit. Unter Ausschluss der beiden Items „e4“ und „e9“ jedoch konnte in allen Teilungskriterien Rasch-Modell-Gültigkeit für den Faktor Extraversion in der Gruppe des Listen-Präsentationsformats erreicht werden.

b) Rasch-Modell-Passung in der Gruppe One-Item-a-Time-Format, Faktor Extraversion

Im nächsten Schritt wurde die Rasch-Modell-Passung innerhalb der Gruppe des One-Item-a-Time-Präsentationsformats überprüft. Hierzu wurden zunächst die Parameter des Modells geschätzt und dem folgend Andersens Likelihood-Ratio-Tests mit den Trennungskriterien „Anzahl zutreffender Adjektive“, „Geschlecht“ und „Alter“ berechnet. Tabelle 8.7 gibt die Kennwerte der Berechnungen wieder.

Tabelle 8.7

Ergebnisse der Andersens Likelihood-Ratio-Tests für den gesamten Itempool des Faktors Extraversion im One-Item-a-Time-Präsentationsformat

Trennungskriterium	χ^2	*df*	*p*
Anzahl zutreffender Adjektive (Median)	13.78	8	.088
Geschlecht	17.56	8	.025
Alter	27.98	8	< .001*

*: $p \leq .01$

Wie in Tabelle 8.7 erkennbar, produzierte alleinig der Modelltest nach dem Teilungskriterium “Alter” eine signifikante Teststatistik. Dementsprechend wurde eine Graphik zur Überprüfung der Modellpassung angefertigt (Abbildung 8.8).

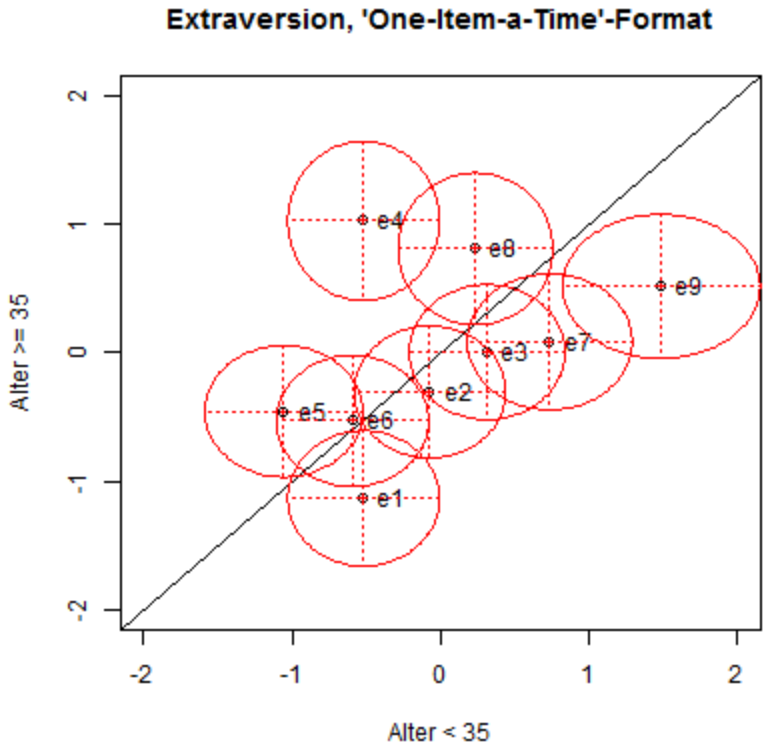


Abbildung 8.8. GOF-Plot für den gesamten Itempool des Faktors Extraversion im One-Item-a-Time-Präsentationsformat nach dem Teilungskriterium „Alter“

Wie sich der Graphik entnehmen lässt war das einzige deutlich auffällige Item das Item „e4“. Nach der Entfernung des Items aus dem Itempool wurden erneute Andersens Likelihood-Ratio-Tests durchgeführt. Die Ergebnisse jener Tests sind in Tabelle 8.8 wiedergegeben.

Tabelle 8.8

Ergebnisse der Andersens Likelihood-Ratio-Tests für den Itempool des Faktors Extraversion ohne das Item „e4“ im One-Item-a-Time-Präsentationsformat

Trennungskriterium	χ^2	*df*	*p*
Anzahl zutreffender Adjektive (Median)	8.38	7	.301
Geschlecht	17.78	7	.013
Alter	13.34	7	.064

*: $p \leq .01$

Wie Tabelle 8.8 entnehmbar, produziert keiner der Modellpassungstests mehr signifikante Ergebnisse. Somit besteht nun nach dem Ausschluss des Items “e4” für die Gruppe des Präsentationsformats One-Item-a-Time Rasch-Modellgeltung bezüglich aller Teilungskriterien.

c) Rasch-Modell-Passung in der Gruppe Q-Sort-Format, Faktor Extraversion

Zuletzt wurde das Rasch-Modell in der Gruppe des Q-Sort-Präsentationsformates geschätzt. Tabelle 8.9 gibt die Ergebnisse der durchgeführten Andersens Likelihood-Ratio-Tests wieder.

Tabelle 8.9

Ergebnisse der Andersens Likelihood-Ratio-Tests für den gesamten Itempool des Faktors Extraversion im Q-Sort-Präsentationsformat

Trennungskriterium	χ^2	*df*	*p*
Anzahl zutreffender Adjektive (Median)	19.75	8	.011
Geschlecht	15.87	8	.044
Alter	34.78	8	< .001*

*: $p \leq .01$

Aus der Tabelle wird klar, dass alleine der Modelltest bezüglich des Trennungskriteriums „Alter” signifikante Modellabweichungen feststellt. Abbildung 8.9 gibt die Graphik zur Überprüfung der Modellpassung wieder.

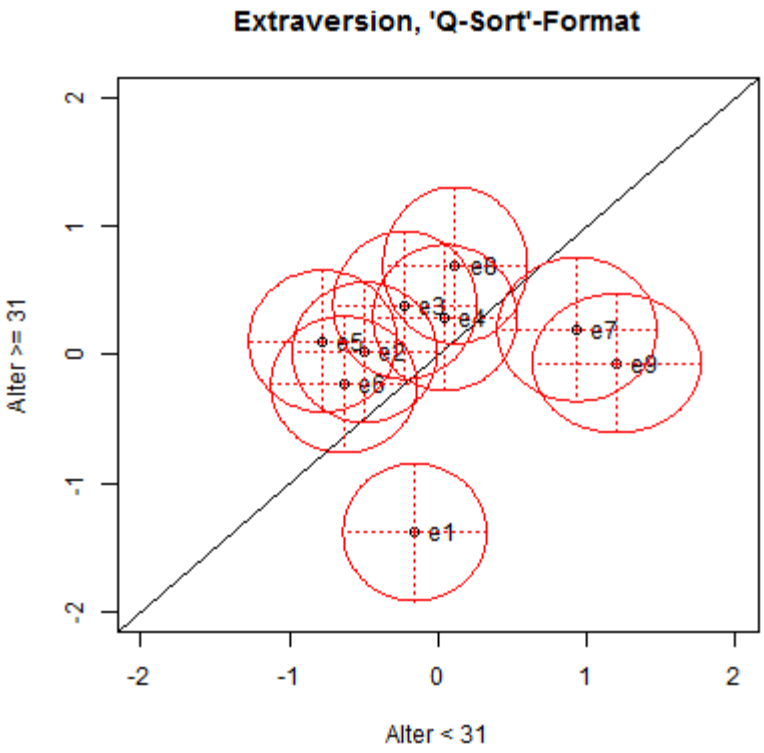


Abbildung 8.9. GOF-Plot für den gesamten Itempool des Faktors Extraversion im Q-Sort-Präsentationsformat nach dem Teilungskriterium „Alter“

Abbildung 8.9 identifiziert die Items „e1“ und „e9“ als die auffälligsten Items. Die Items wurden schrittweise in dieser Reihenfolge aus dem Itempool entfernt; neuerliche Andersens Likelihood-Ratio-Tests wurden berechnet.

Dabei zeigte sich, dass das Entfernen der für das Teilungskriterium „Alter“ auffälligen Items mit einem Abfall der Modellpassung bezüglich der anderen Teilungskriterien einhergeht. Modellgeltung bezüglich aller Teilungskriterien konnte erst nach dem Entfernen zwei weiterer Items (Item „e7“ und „e4“, in dieser Reihenfolge) erzielt werden. Da allerdings die Reduktion des Itempools von neun auf fünf verbleibende Items relativ schwerwiegend ist, muss gefolgert werden dass der Itempool des Faktors Extraversion im Präsentationsformat Q-Sort nicht dem Rasch-Modell entspricht. Aufgrund des ausufernden Umfangs jener weiteren Analysen ist deren ausführliche Dokumentation hier unterlassen worden.

8.2.2. Faktor 2: Gewissenhaftigkeit

Analog zum obigen Vorgehen für den Faktor Extraversion wurde anschließend die Rasch-Modell-Geltung für den Faktor Gewissenhaftigkeit in der gesamten Stichprobe überprüft. Die Ergebnisse von Andersens Likelihood-Ratio-Test sind in Tabelle 8.10 wiedergegeben.

Tabelle 8.10

Ergebnisse der Andersens Likelihood-Ratio-Tests für den gesamten Itempool des Faktors Gewissenhaftigkeit in allen Präsentationsformaten

Trennungskriterium	χ^2	*df*	*p*
Präsentationsformat	16.94	16	.389
Anzahl zutreffender Adjektive (Median)	9.65	8	.291
Geschlecht	15.38	8	.052
Alter	27.36	8	.001*

*: $p \leq .01$

Somit produziert alleinig der Modelltest bezüglich des Trennungskriteriums “Alter” eine signifikante Teststatistik. Zur Identifizierung modellabweichender Items wurde eine Graphik der Modellpassung bezüglich des Teilungskriteriums „Alter“ angefertigt (Abbildung 8.10).

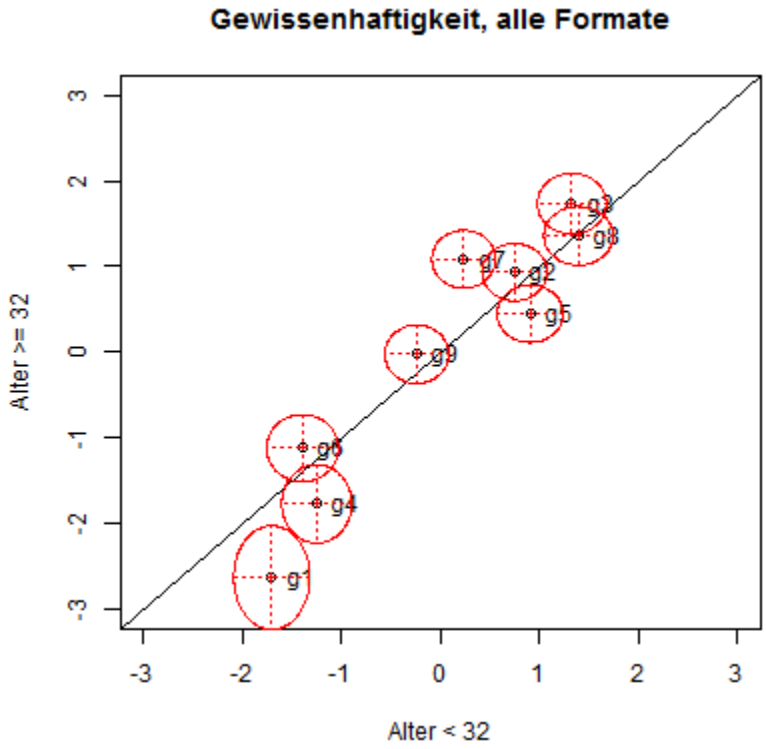


Abbildung 8.10. GOF-Plot für den gesamten Itempool des Faktors Gewissenhaftigkeit in allen Präsentationsformaten nach dem Teilungskriterium „Alter“

Wie aus der Abbildung 8.10 hervorgeht, ist in Bezug auf das Teilungskriterium „Alter“ das auffälligste Item das Item „g7“, dicht gefolgt von Item „g1“. Im nächsten Schritt wurden die beiden Items sukzessive in der genannten Reihenfolge aus dem Itempool entfernt, wobei jeweils Andersens Likelihood-Ratio-Tests für die diversen Teilungskriterien gerechnet wurden. Tabelle 8.11 gibt die Ergebnisse dieser Tests wieder.

Tabelle 8.11

Ergebnisse der Andersens Likelihood-Ratio-Tests des Faktors Gewissenhaftigkeit in allen Präsentationsformaten unter Ausschluss des Items „g7“ (Schritt 1) und Item „g1“ (Schritt 2)

	Trennungskriterium	χ^2	*df*	*p*
Schritt 1	Präsentationsformat	13.12	14	.517
	Anzahl zutreffender Adjektive (Median)	19.55	7	.007*
	Geschlecht	11.94	7	.103
	Alter	18.40	7	.010*
Schritt 2	Präsentationsformat	12.87	12	.378
	Anzahl zutreffender Adjektive (Median)	18.58	6	.005*
	Geschlecht	10.94	6	.090
	Alter	11.76	6	.068

*: $p \leq .01$

Wie Tabelle 8.1. wiedergibt, führte die Entfernung der Items „g7“ und „g1“ zwar zu einem nicht-signifikanten Modelltest bezüglich des Trennungskriteriums „Alter“, allerdings produziert nunmehr der Modelltest bezüglich des Teilungskriteriums „Anzahl zutreffender Adjektive (Median)“ ein signifikantes Ergebnis. Abermals wurde eine Graphik zur Überprüfung der Modellpassung angefertigt um auffällige Items zu identifizieren (Abbildung 8.11).

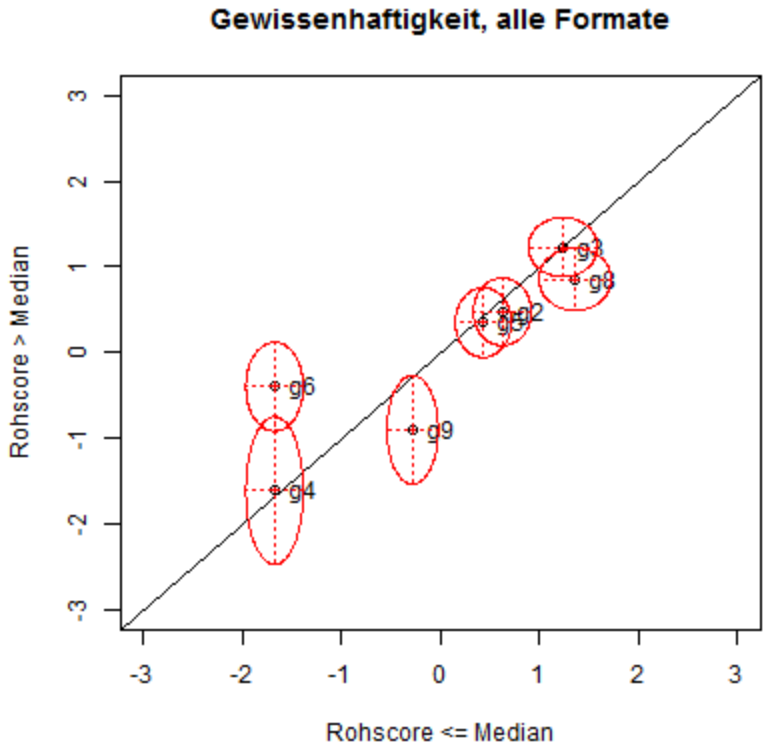


Abbildung 8.11. GOF-Plot des Faktors Gewissenhaftigkeit in allen Präsentationsformaten nach dem Teilungskriterium „Anzahl zutreffender Adjektive (Median)“ unter Ausschluss der Items „g7“ und „g1“

Nunmehr fällt bezüglich des Teilungskriteriums „Anzahl zutreffender Adjektive (Median)“ alleine das Item „g6“ auf. Nach der Entfernung des Items aus dem Pool resultierten für Andersens Likelihood-Ratio-Tests die Ergebnisse in Tabelle 8.12.

Tabelle 8.12

Ergebnisse der Andersens Likelihood-Ratio-Tests des Faktors Gewissenhaftigkeit in allen Präsentationsformaten unter Ausschluss der Items „g7“, „g1“ und „g6“

Trennungskriterium	χ^2	*df*	*p*
Präsentationsformat	11.00	10	.385
Anzahl zutreffender Adjektive (Median)	4.49	5	.482
Geschlecht	11.85	5	.037
Alter	11.76	5	.038

*: $p \leq .01$

Nach der Entfernung der drei Items „g7“, „g1“ und „g6“ produzieren nun die Modelltests bezüglich sämtlicher Trennungskriterien nicht-signifikante Ergebnisse. Somit gilt für die verbleibenden sechs Items des Faktors Gewissenhaftigkeit in der gesamten Stichprobe Rasch-Modell-Gültigkeit. Da jedoch drei von neun Items entfernt werden mussten um Modellgeltung zu erreichen, soll die Geltung des Rasch-Modells für den gesamten Itempool in den jeweiligen Gruppen der Präsentationsformate untersucht werden.

a) Rasch-Modell-Passung in der Gruppe Listen-Format, Faktor Gewissenhaftigkeit

Zunächst wurden für den gesamten Itempool des Faktors Gewissenhaftigkeit in der Gruppe des Listen-Präsentationsformats mehrere Andersens Likelihood-Ratio-Tests bezüglich der Trennungskriterien „Anzahl zutreffender Adjektive (Median)“, „Geschlecht“ und „Alter“ durchgeführt. Die Ergebnisse sind in Tabelle 8.13 zusammengefasst.

Tabelle 8.13

Ergebnisse der Andersens Likelihood-Ratio-Tests für den gesamten Itempool des Faktors Gewissenhaftigkeit im Listen-Präsentationsformat

Trennungskriterium	χ^2	*df*	*p*
Anzahl zutreffender Adjektive (Median)	9.49	7	.219
Geschlecht	10.46	8	.234
Alter	17.45	8	.023

*: $p \leq .01$

Wie aus der Tabelle 8.13 hervorgeht, lassen sich durch keinen der Modelltests signifikante Abweichungen vom Rasch-Modell feststellen. Demnach besteht für den gesamten Itempool des Faktors Gewissenhaftigkeit innerhalb der Gruppe des Listen-Präsentationsformats Rasch-Modell-Gültigkeit. Als einzige Einschränkung muss angeführt werden, dass für das Trennungskriterium „Anzahl zutreffender Adjektive (Median)“ keine Schätzung eines Itemparameters innerhalb der Subgruppen geschätzt werden konnte. Dies rührt daher, dass innerhalb der Subgruppe mit Rohscores oberhalb des Medians sämtliche befragten Personen das Item mit „1 – trifft zu“ beantwortet hatten. Da der Itemparameter allerdings innerhalb aller anderen Subgruppen sowie in der kompletten Gruppe des Listen-Präsentationsformats problemlos geschätzt werden konnte, soll dies der Rasch-Modell-Geltung des Faktors Gewissenhaftigkeit in dieser Gruppe keinen Abbruch tun.

b) Rasch-Modell-Passung in der Gruppe One-Item-a-Time-Format, Faktor Gewissenhaftigkeit

Analog zum obigen Vorgehen in der Gruppe des Listen-Präsentationsformats wurden Modellpassungstests in der Gruppe des One-Item-a-Time-Präsentationsformats durchgeführt. Die Ergebnisse sind in Tabelle 8.14 zusammengefasst.

Tabelle 8.14

Ergebnisse der Andersens Likelihood-Ratio-Tests für den gesamten Itempool des Faktors Gewissenhaftigkeit im One-Item-a-Time-Präsentationsformat

Trennungskriterium	χ^2	*df*	*p*
Anzahl zutreffender Adjektive (Median)	14.26	8	.075
Geschlecht	7.61	8	.472
Alter	14.88	8	.061

*: $p \leq .01$

Wie Tabelle 8.14 zu entnehmen ist produziert keiner der durchgeführten Modelltests signifikante Ergebnisse. Daher wird für den Faktor Gewissenhaftigkeit in der Gruppe des One-Item-a-Time-Präsentationsformats für den vollständigen Itempool Rasch-Modell-Geltung angenommen.

c) Rasch-Modell-Passung in der Gruppe Q-Sort-Format, Faktor Gewissenhaftigkeit

Wie auch schon für die beiden anderen Gruppen wurde nun in der Gruppe des Q-Sort-Präsentationsformats Rasch-Modell-Tests für den Faktor Gewissenhaftigkeit mit all seinen Items durchgeführt. Die Ergebnisse der Signifikanztests sind in Tabelle 8.15 zusammengefasst.

Tabelle 8.15

Ergebnisse der Andersens Likelihood-Ratio-Tests für den gesamten Itempool des Faktors Gewissenhaftigkeit im Q-Sort-Präsentationsformat

Trennungskriterium	χ^2	*df*	*p*
Anzahl zutreffender Adjektive (Median)	5.21	8	.735
Geschlecht	17.43	8	.026
Alter	6.35	8	.608

*: $p \leq .01$

Somit produziert keiner der durchgeführten Andersens Likelihood-Ratio-Tests signifikante Ergebnisse. Damit wird dem gesamten Itempool des Faktors Gewissenhaftigkeit in der Gruppe Q-Sort-Präsentationsformat Rasch-Modell-Geltung attestiert.

8.2.3. Faktor 3: Verträglichkeit

Um die Rasch-Modell-Geltung für den Faktor Verträglichkeit für all seine Items und die gesamte Stichprobe zu bestimmen, wurden nach der Schätzung der Modellparameter vier Andersens Likelihood-Ratio-Tests mit den bereits besprochenen Trennungskriterien berechnet. Die statistischen Kennwerte der Tests sind in Tabelle 8.16 wiedergegeben.

Tabelle 8.16

Ergebnisse der Andersens Likelihood-Ratio-Tests für den gesamten Itempool des Faktors Verträglichkeit in allen Präsentationsformaten

Trennungskriterium	χ^2	*df*	*p*
Präsentationsformat	24.14	22	.340
Anzahl zutreffender Adjektive (Median)	22.45	11	.021
Geschlecht	35.77	11	< .001*
Alter	24.47	11	.011

*: $p \leq .01$

Den Ergebnissen aus Tabelle 8.16 zufolge produziert allein der Modelltest bezüglich des Teilungskriteriums “Geschlecht” signifikante Ergebnisse. Als nächstes wurde eine Graphik zur Überprüfung der Modellpassung erstellt, um durch sie auffällige Items identifizieren zu können (Abbildung 8.12).

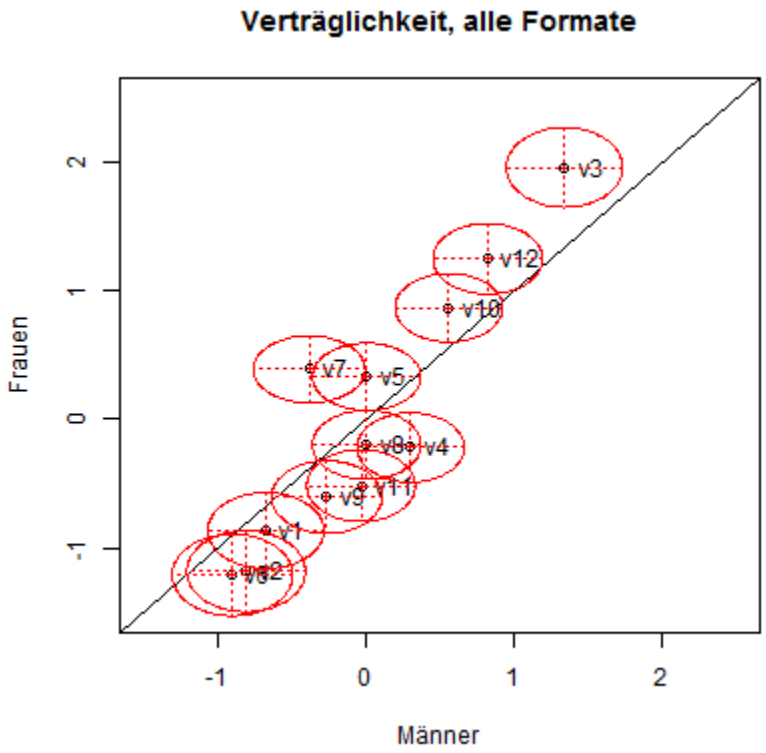


Abbildung 8.12. GOF-Plot für den gesamten Itempool des Faktors Verträglichkeit in allen Präsentationsformaten nach dem Teilungskriterium „Geschlecht“

Abbildung 8.12 identifiziert „v7“, als jenes Item mit den deutlichsten Abweichungen seiner Parameterschätzungen zwischen den Subgruppen. Dem bereits besprochenen Vorgehen folgend wurde das Item aus dem Pool entfernt. Tabelle 8.17 gibt die Kennwerte der wiederholten Andersens Likelihood-Ratio-Tests unter Ausschluss des Items wieder.

Tabelle 8.17

Ergebnisse der Andersens Likelihood-Ratio-Tests des Faktors Verträglichkeit in allen Präsentationsformaten unter Ausschluss des Items „v7“

Trennungskriterium	χ^2	*df*	*p*
Präsentationsformat	22.19	20	.330
Anzahl zutreffender Adjektive (Median)	20.79	10	.023
Geschlecht	23.50	10	.009*
Alter	24.37	10	.007*

*: $p \leq .01$

Nunmehr erreichen die Teststatistiken der Modelltests bezüglich der Trennungskriterien “Geschlecht” und “Alter” signifikante Werte. Abbildungen 8.13 und 8.14 geben die Graphiken zur Überprüfung der Modellpassung wieder.

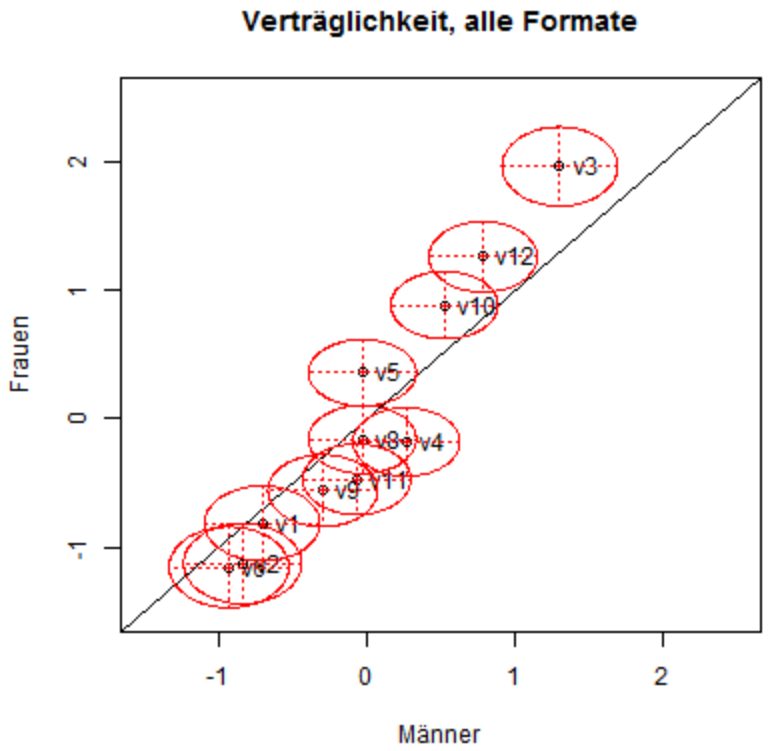


Abbildung 8.13. GOF-Plot des Faktors Verträglichkeit in allen Präsentationsformaten nach dem Teilungskriterium „Geschlecht“ unter Ausschluss des Items „v7“

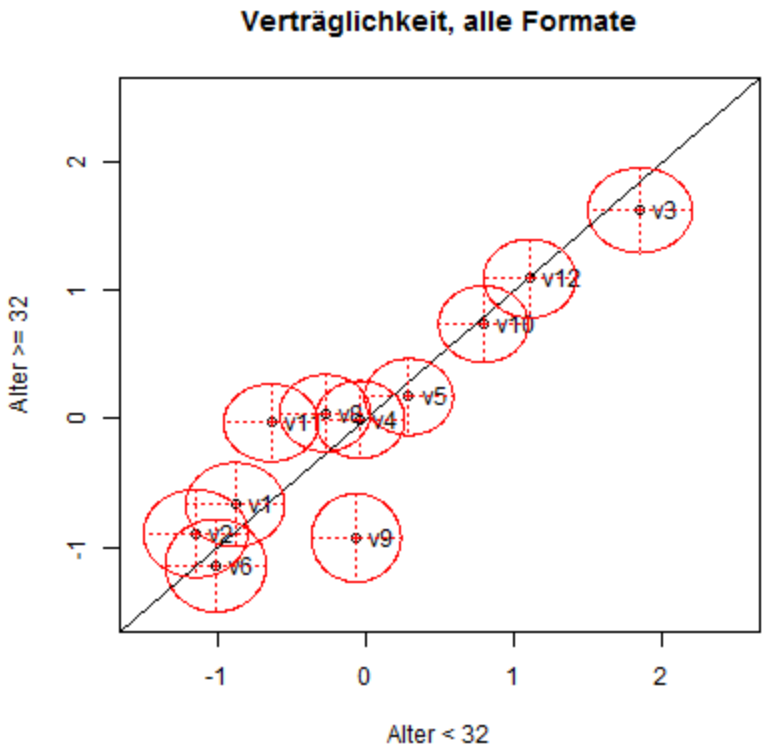


Abbildung 8.14. GOF-Plot des Faktors Verträglichkeit in allen Präsentationsformaten nach dem Teilungskriterium „Alter“ unter Ausschluss des Items „v7“

Abbildungen 8.13 und 8.14 identifizieren das Item „v9“ als jenes mit der größten Abweichung zwischen den Schätzungen seines Parameters zwischen den Subgruppen. Dementsprechend wurde das Item aus dem Pool entfernt und wiederholte Modelltests durchgeführt. Die Ergebnisse sind in Tabelle 8.18 zusammengefasst.

Tabelle 8.18

Ergebnisse der Andersens Likelihood-Ratio-Tests des Faktors Verträglichkeit in allen Präsentationsformaten unter Ausschluss der Items „v7“ und „v9“

Trennungskriterium	χ^2	*df*	*p*
Präsentationsformat	17.79	18	.469
Anzahl zutreffender Adjektive (Median)	23.28	9	.006*
Geschlecht	21.45	9	.011
Alter	10.03	9	.348

*: $p \leq .01$

Nach dem Entfernen der beiden Items „v7" und „v9" produzieren nunmehr die Modelltests bezüglich der Teilungskriterien "Geschlecht" und "Alter" keine signifikanten Ergebnisse, das Teilungskriterium „Anzahl zutreffender Adjektive (Median)" jedoch schon. Abermals wurde eine graphische Überprüfung der Modellpassung vorgenommen (Abbildung 8.15).

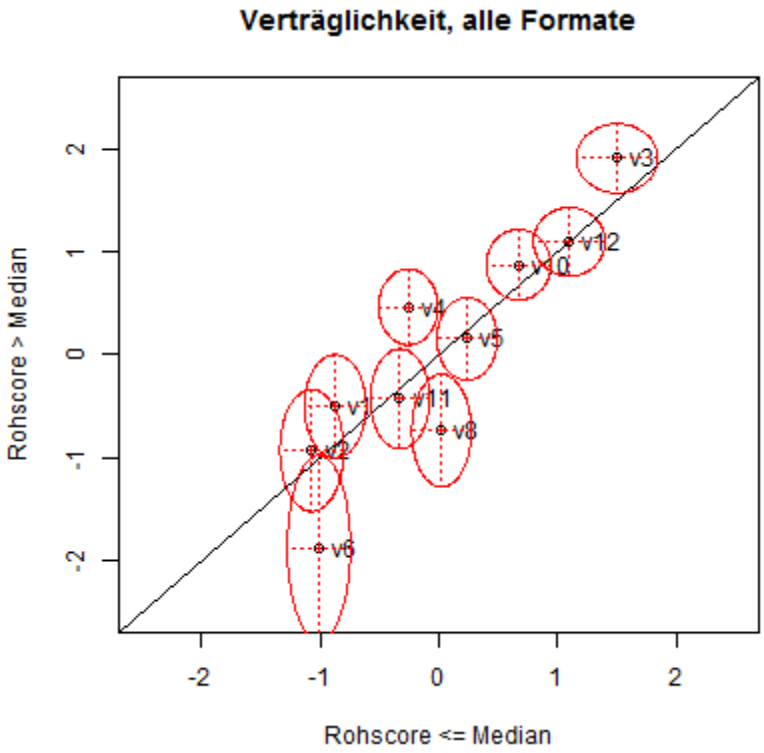


Abbildung 8.15. GOF-Plot des Faktors Verträglichkeit in allen Präsentationsformaten nach dem Teilungskriterium „Anzahl zutreffender Items (Median)" unter Ausschluss der Items „v7" und „v9"

Abbildung 8.15 identifiziert „v4" als jenes Item mit der größten Abweichung der Itemparameterschätzungen zwischen den beiden Subgruppen. Nach der Entfernung des Items aus dem Pool resultierten die in Tabelle 8.19 wiedergegebenen Ergebnisse von Andersens Likelihood-Ratio-Tests.

Tabelle 8.19

Ergebnisse der Andersens Likelihood-Ratio-Tests des Faktors Verträglichkeit in allen Präsentationsformaten unter Ausschluss der Items „v7“, „v9“ und „v4“

Trennungskriterium	χ^2	*df*	*p*
Präsentationsformat	15.37	16	.498
Anzahl zutreffender Adjektive (Median)	8.39	8	.396
Geschlecht	14.82	8	.063
Alter	10.10	8	.258

*: $p \leq .01$

Nach der Entfernung der drei Items „v7”, „v9“ und „v4“ produziert nunmehr keiner der Modelltests eine signifikante Teststatistik. Somit kann angenommen werden, dass für jenen reduzierten Itempool des Faktors Verträglichkeit über alle drei Gruppen der Präsentationsformate hinweg das Rasch-Modell gilt. Da jedoch die Entfernung von drei Items des Itempools aus zwölf Items nicht vollkommen zufriedenstellend ist, soll zusätzlich die Modellgeltung innerhalb der jeweiligen Gruppen der Präsentationsformate untersucht werden.

a) Rasch-Modell-Passung in der Gruppe Listen-Format, Faktor Verträglichkeit

Zunächst wurden, wie bereits für die anderen Faktoren, mehrere Andersens Likelihood-Ratio-Tests für den vollständigen Itempool des Faktors Verträglichkeit innerhalb der Gruppe des Listen-Präsentationsformats berechnet. Die Ergebnisse jener Tests sind in Tabelle 8.20 wiedergegeben.

Tabelle 8.20

Ergebnisse der Andersens Likelihood-Ratio-Tests für den gesamten Itempool des Faktors Verträglichkeit im Listen-Präsentationsformat

Trennungskriterium	χ^2	*df*	*p*
Anzahl zutreffender Adjektive (Median)	17.03	10	.074
Geschlecht	25.53	11	.008*
Alter	19.02	11	.061

*: $p \leq .01$

Wie aus Tabelle 8.20 hervorgeht, erreicht alleine die Teststatistik des Modelltests bezüglich des Teilungskriteriums “Geschlecht” ein signifikantes Ergebnis. Es sei auch angemerkt, dass bezüglich des Teilungskriteriums „Anzahl zutreffender Adjektive (Median)“ für ein Item keine Parameterschätzung vorgenommen werden konnte, da hier sämtliche Personen der Subgruppe „Rohscore > Median“ das Item mit „1 – trifft zu“ beantwortet hatten.

Zur Identifizierung der nicht modellkonformen Items wurde eine Graphik zur Überprüfung der Modellpassung erstellt (Abbildung 8.16).

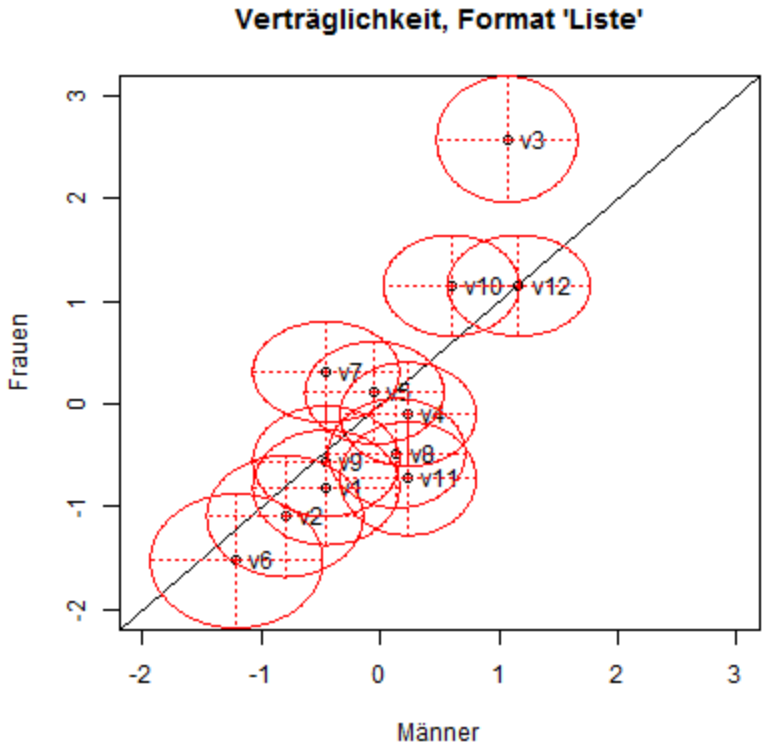


Abbildung 8.16. GOF-Plot für den gesamten Itempool des Faktors Verträglichkeit im Listen-Präsentationsformat nach dem Teilungskriterium „Geschlecht“

Wie In Abbildung 8.16 zu erkennen ist, war das Item mit den stärksten Abweichungen der Schätzungen seiner Itemparameter in den Subgruppen nach dem Teilungskriterium „Geschlecht“ das „v3“. Die Ergebnisse der wiederholten Modelltests nach Entfernen des Items aus dem Itempool sind in Tabelle 8.21 wiedergegeben.

Tabelle 8.21

Ergebnisse der Andersens Likelihood-Ratio-Tests des Faktors Verträglichkeit im Listen-Präsentationsformaten unter Ausschluss des Items „v3“

Trennungskriterium	χ^2	*df*	*p*
Anzahl zutreffender Adjektive (Median)	13.62	10	.191
Geschlecht	13.89	10	.178
Alter	18.48	10	.047

*: $p \leq .01$

Da keiner der in Tabelle 8.21 angeführten Modelltests nunmehr ein signifikantes Ergebnis produziert, kann für den Itempool des Faktors Verträglichkeit in der Gruppe des “Listen”-Präsentationsformats unter Ausschluss des Items „v3“ Rasch-Modell-Homogenität bestätigt werden.

b) Rasch-Modell-Passung in der Gruppe One-Item-a-Time-Format, Faktor Verträglichkeit

Zunächst wurden drei Andersens Likelihood-Ratio-Tests mit verschiedenen Trennungskriterien zur Überprüfung der Rasch-Modell-Homogenität des Itempools des Faktors Verträglichkeit in der Gruppe des One-Item-a-Time-Präsentationsformats durchgeführt. Tabelle 8.22 fasst die Ergebnisse der Modelltests zusammen.

Tabelle 8.22

Ergebnisse der Andersens Likelihood-Ratio-Tests für den gesamten Itempool des Faktors Verträglichkeit im One-Item-a-Item-Präsentationsformat

Trennungskriterium	χ^2	*df*	*p*
Anzahl zutreffender Adjektive (Median)	12.42	11	.333
Geschlecht	17.65	11	.090
Alter	11.06	11	.438

*: $p \leq .01$

Da keiner der in Tabelle 8.22 angeführten Modelltests signifikante Ergebnisse produzierte, kann davon ausgegangen werden, dass für den kompletten Itempool des Faktors Verträglichkeit innerhalb der Gruppe des One-Item-a-Time-Präsentationsformats Rasch-Modell-Homogenität gilt.

c) Rasch-Modell-Passung in der Gruppe Q-Sort-Format, Faktor Verträglichkeit

Zuerst wurden für den gesamten Itempool des Faktors Verträglichkeit innerhalb der Gruppe des Q-Sort-Präsentationsformats drei Andersens Likelihood-Ratio-Tests mit verschiedenen Trennungskriterien zur Bestimmung der Modellpassung berechnet. Tabelle 8.23 gibt die Ergebnisse der Modelltests wieder.

Tabelle 8.23

Ergebnisse der Andersens Likelihood-Ratio-Tests für den gesamten Itempool des Faktors Verträglichkeit im Q-Sort-Präsentationsformat

Trennungskriterium	χ^2	*df*	*p*
Anzahl zutreffender Adjektive (Median)	15.24	11	.172
Geschlecht	21.00	11	.033
Alter	23.16	11	.017

*: $p \leq .01$

Wie den in Tabelle 8.23 zusammengefassten Ergebnissen der Modelltests zu entnehmen ist, produzierte keiner der durchgeführten Andersens Likelihood-Ratio-Tests eine signifikante Teststatistik. Dies lässt annehmen, dass für den vollständigen Itempool des Faktors Verträglichkeit in der Gruppe des Präsentationsformats Q-Sort Rasch-Modell-Homogenität besteht.

8.3. Analyse der Nutzung des Antwortspektrums (Hypothese 2)

Um die Ausnutzung des verfügbaren Antwortspektrums in den Gruppen der Präsentationsformate miteinander zu vergleichen, wurden die Häufigkeiten der Wahl der jeweiligen Antwortkategorien zunächst in einer Kontingenztafel gegeneinander angetragen. Darauf folgend wurde Pearsons χ^2-Test gerechnet, um die Nullhypothese der Unabhängigkeitsannahme der Verteilung der gewählten Antwortmöglichkeiten vom jeweiligen Setting zu testen (siehe auch unter *6.2. Hypothese 2: Unterschiedliche Nutzung des verfügbaren Antwortspektrums*). Tabelle 8.24 gibt jene Häufigkeiten zusammen mit der relevanten Teststatistik wieder. Die Berechnung der Teststatistik erfolgte mittels dem R-Paket „gmodels" (Warnes, 2012).

Tabelle 8.24

Ergebnis und zugehörige Kontingenztafel des χ^2-Test zur Überprüfung der Unabhängigkeitsannahme zwischen den Präsentationsformaten und den Häufigkeiten der gewählten Antwortkategorien (N = 11670)

	gewählte Antwortkategorie									
Setting	1	2	3	4	5	6	χ^2	*p*	*w*	Cramer's *V*
Liste	26	275	468	1033	1421	717				
„One-Item-a-Time"	38	144	537	1085	1406	690	31.38 (*df* = 10)	< .001*	.05	.04
Q-Sort	43	157	566	1144	1284	736				

*: p ≤ .05

Wie aus der Tabelle 8.24 hervorgeht, produziert der χ^2-Test ein signifikantes Ergebnis. Cramer´s *V* als kanonischer Korrelationskoeffizient beträgt *V* = .04, zudem ist im Einklang mit der Untersuchungsplanung (siehe *7.1. Untersuchungsplan*) die Effektstärke *w* (*w* = .05) angegeben, was einem sehr kleinen Effekt entspricht. Die Berechnung der Effektstärke *w* erfolgte nach Volker (2006), die Berechnung von Cramer´s *V* erfolgte mittels dem R-Paket „vcd" (Meyer, Zeileis & Hornik, 2011). Um die relative Verteilung der gewählten Antwortkategorien je Präsentationsformat graphisch zu veranschaulichen, wurde des Weiteren ein Mosaikplot der Tabelle 8.24 erstellt (Abbildung 8.17).

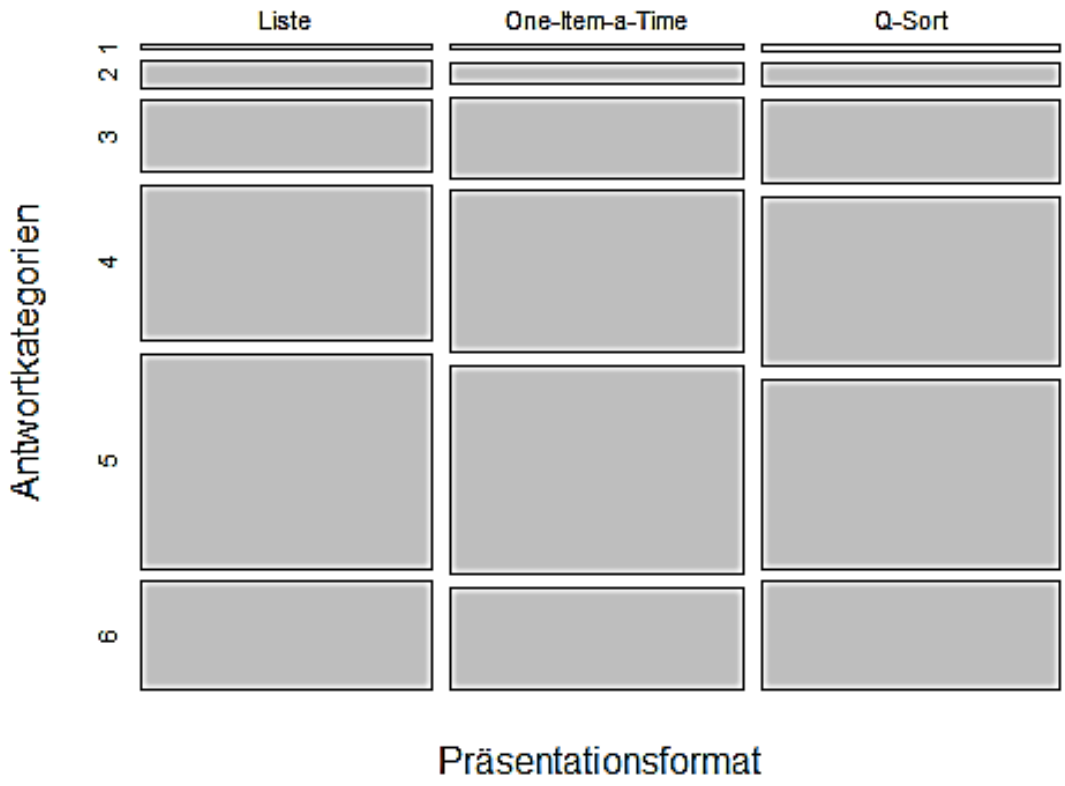


Abbildung 8.17. Mosaikplot der Häufigkeiten der gewählten Antwortkategorien in den jeweiligen Präsentationsformaten

8.4. Analyse der Mittelwertsunterschiede der Skalensummen des Selbstbeschreibungsfragebogens (Hypothese 3)

Um mögliche Mittelwertsunterschiede und Veränderungen in den Interkorrelationsmatrizen der drei Gruppen der verschiedenen Präsentationsformate untersuchen zu können wurde eine einfaktorielle, multivariate Varianzanalyse (One-Way-MANOVA) durchgeführt. Dabei bildeten die drei verschiedenen Präsentationsformate die unabhängige Variable, die drei Summenscores der drei Persönlichkeitsfacetten des B5PO-R bildeten die abhängigen Variablen. Für die Analyse der Summenscores des B5PO-R wurden die summierten Ratingantworten herangezogen. Zwar war im Vorfeld eine ausführliche Rasch-Modell-Analyse durchgeführt worden, jedoch waren die Faktoren des B5PO-R über alle Gruppen hinweg nicht ohne der Entfernung mehrerer Items als Rasch-Modell-homogen zu bestätigen. Aus diesem Grund wird davon abgesehen, die im Zuge der Rasch-Modell-Skalierung geschätzten Personenparameter zu verwenden.

Zunächst folgt die Betrachtung der Voraussetzungen der MANOVA (vgl. Field et al., 2012). Von Unabhängigkeit der Beobachtungen sowie einer randomisierten Stichprobenerhebung und

Intervallskalenniveau der Daten kann approximativ ausgegangen werden. Zur Bestimmung der multivariaten Normalverteilung wurden Shapiro-Wilk-Tests für die Ratingsummen der drei Faktoren zusammen in jeder der Präsentationsformats-Gruppen durchgeführt. Die Ergebnisse sind in Tabelle 8.25 dargestellt, implementiert wurden die Tests mittels dem R-Paket „mvnormtest“ (Slawomir, 2012).

Tabelle 8.25

Ergebnisse der Shapiro-Wilk-Tests zur Überprüfung der Annahme der multivariaten Normallverteilung der Ratingsummen der Faktoren des B5PO-R

Gruppe	*W*	*p*
Liste	0.982	.094
One-Item-a-Time	0.967	.003*
Q-Sort	0.967	.003*

$p \leq .05$

Für die Gruppe des Listen-Präsentationsformats zeigte Shapiro-Wilks Test zur Überprüfung der multivariaten Normalverteilung somit keine Abweichungen, wohl aber für die beiden anderen Gruppen.

Die vierte und letzte Voraussetzung der multivariaten Varianzanalyse ist die Homogenität der Varianz-/Kovarianzmatrizen. Für diese Voraussetzung wurde von der Überprüfung via statistischer Signifikanztests abgesehen; einerseits weil eine derartige Überprüfung das Risiko eines Fehlers erster Art anhebt, und andererseits weil bei lediglich drei Stufen der unabhängigen Variablen und drei abhängigen Variablen diese Überprüfung auch optisch realisierbar ist. Dementsprechend wurden die Varianz-/Kovarianzmatrizen im Verhältnis zueinander betrachtet. Dabei zeigten sich leichte bis starke Abweichungen von der benötigten Homogenität. Besagte Varianz-/Kovarianzmatrizen sind in Tabelle 8.26 wiedergegeben.

Tabelle 8.26

Varianz-/Kovarianzmatrizen der Ratingsummen der Faktoren des B5PO-R

Gruppe Listen-Präsentationsformat			
	Extraversion	Gewissenhaftigkeit	Verträglichkeit
Extraversion	31.24	11.78	11.93
Gewissenhaftigkeit	11.79	47.68	8.87
Verträglichkeit	11.93	8.87	32.69
Gruppe One-Item-a-Time-Präsentationsformat			
	Extraversion	Gewissenhaftigkeit	Verträglichkeit
Extraversion	27.07	5.27	11.7
Gewissenhaftigkeit	5.72	44.79	2.51
Verträglichkeit	11.67	2.51	37.38
Gruppe Q-Sort-Präsentationsformat			
	Extraversion	Gewissenhaftigkeit	Verträglichkeit
Extraversion	30.62	0.75	5.79
Gewissenhaftigkeit	0.75	34.58	9.27
Verträglichkeit	5.79	9.27	37.67

Somit wurden sowohl für die Voraussetzung der multivariaten Normalverteilung als auch der Voraussetzung der Homogenität der Varianz-/Kovarianzmatrizen der abhängigen Variablen Abweichungen festgestellt. Dementsprechend fiel die Wahl der verwendeten statistischen Testgröße der multivariaten Varianzanalyse auf Pillai´s Spur, da diese im Falle annähernd gleicher Gruppengrößen die größte Robustheit besitzt (Field et al., 2012). Die Ergebnisse der MANOVA sind in Tabelle 8.27 zusammengefasst.

Tabelle 8.27

Ergebnis und Kennwerte der One-Way-MANOVA der Ratingsummen der Faktoren des B5PO-R in den jeweiligen Gruppen der Präsentationsformate

	M	*sd*	*Pillai´s Spur V*	*F (3, 385)*	*p*
Extraversion					
Liste	38.12	5.59			
One-Item-a-Time	37.89	5.20			
Q-Sort	37.75	5.53			
Gewissenhaftigkeit					
Liste	41.55	6.91	0.0063	0.82	.485
One-Item-a-Time	40.78	6.69			
Q-Sort	41.10	5.88			
Verträglichkeit					
Liste	55.63	5.72			
One-Item-a-Time	55.53	6.11			
Q-Sort	54.49	6.14			

Für die so beschriebene Durchführung der MANOVA unter der Verwendung von Pillai´s Spur resultierten eine nicht-signifikante Teststatistik von $V = 0.0063$ ($F(3, 385) = 0.82$), $p = .485$. Die Variation des Präsentationsformats übte keine Effekte auf die Ratingsummen in den Faktoren des B5PO-R oder deren Interkorrelationsmatrizen aus.

8.5. Analyse der Bearbeitungszeiten (Hypothese 4)

Zunächst wurden mittels deskriptiver Statistiken und Graphiken die Voraussetzungen der einfaktoriellen, univariaten Varianzanalyse überprüft. Dabei wurden keine Signifikanztest wie zum Beispiel Levenes Test zur Überprüfung der Homogenität der Varianzen durchgeführt, da dieser bei hohen Stichprobenumfängen bereits bei nur leichter Varianz-Inhomogenität sehr leicht signifikante Ergebnisse produziert und seine Durchführung die Kumulierung eines Risikos für einen Fehler erster Art fördert (Field et al., 2012). Tabelle 8.28 gibt die für die Analyse relevanten Varianzen sowie die Mittelwerte der Bearbeitungszeiten in den Gruppen an.

Tabelle 8.28

Varianzen und Mittelwerte der Bearbeitungszeiten des B5PO-R in den drei Präsentationsformaten

	Listen-Format	One-Item-a-Time-Format	"Q-Sort"-Format
sd^2	1.39	2.31	4.08

Wie aus Tabelle 8.28 hervorgeht, zeigten sich substantielle Inhomogenitäten der Varianzen in den drei Gruppen, und auch die graphische Überprüfung der Verteilungen der Bearbeitungszeiten in den Gruppen zeigten jeweils rechtsschiefe Verteilungen (v_{liste} = 1.72, $v_{one\text{-}item}$ = 2.92, $v_{q\text{-}sort}$ = 1.35). Die Kalkulation der Verteilungsschiefen erfolgte abermals durch das R-Paket „psych"(Revelle, 2011).

Da die Varianzen der Bearbeitungszeiten für die verschiedenen Präsentationsformate substantielle Unterschiede aufweisen, wurde beschlossen Welch´s F als Korrektur für die Verzerrung der Teststatistik der ANOVA anzuwenden (Field et al., 2012). Tabelle 8.29 gibt die Kennwerte der Analyse wieder.

Tabelle 8.29

Ergebnis und Kennwerte der Omnibus-ANOVA der Bearbeitungszeiten des B5PO-Rs in den drei Präsentationsformaten

Gruppe	M^a	*sd*	Welch´s *F* (2; 247.8)	*p*	$f_{korrigiert}$
Liste	2.40	1.18			
One-Item-a-Time	2.87	1.52	22.66	< .001*	.33
Q-Sort	3.77	2.02			

* = p ≤ .05

[a]: Mittelwerte in Minuten angegeben

Wie aus Tabelle 8.29 hervorgeht, resultierte der Welch-Test in einem signifikanten Ergebnis ($p < .05$), mit einer mittleren bis großen Effektstärke von $f_{korrigiert} = .33$ (Volker, 2006).

Aufgrund der zu erwartenden Verzerrungen, die durch die Inhomogenität der Varianzen der drei Gruppen entstehen, wurden anstatt der geplanten Kontraste nun als Post-Hoc-Tests multiple zwei-Gruppen-Mittelwertsvergleiche durchgeführt. Aufgrund der Inhomogenität der Varianzen der Gruppen wurden nach den Empfehlungen von Rasch, Kubinger und Moder (2011) zu diesem Zwecke multiple Welch-Tests angewendet. Um die Überhöhung des Risikos für einen Fehler der ersten Art durch multiple Signifikanztests zu unterbinden, wurden die Signifikanzniveaus mittels Bonferroni-Holm-Korrektur angepasst. Ausgehend von den unter *6.4. Hypothese 4: Unterschiede in den Bearbeitungszeiten* dargelegten Erwartungen bezüglich der Richtungen der Effekte konnten einseitige Signifikanztests zur Anwendung verwendet werden. Da aufgrund der inhomogenen Varianzen die Anwendung der üblicherweise verwendeten Effektstärke *d* unangemessen ist, wurde als Korrektur auf Glass´s δ zurückgegriffen. Dabei wurde jeweils die kleinere der beiden Gruppenstandardabweichungen zu Berechnung der Effektstärke herangezogen. Tabelle 8.30 gibt die relevanten Kennwerte der Analysen wieder.

Tabelle 8.30

Kennwerte der Post-Hoc-Tests der Bearbeitungszeiten des B5PO-R in den Gruppen der Präsentationsformate

Verglichene Gruppen	Welch´s *F*	df_1	df_2	*p* [a]	α [b]	Glass´s δ
Liste VS Q-Sort	44.86	1	210.13	< .001	.025	1.17
One-Item-a-Time VS Q-Sort	16.40	1	241.55	< .001	.013	0.59
One-Item-a-Time VS Liste	7.92	1	242.54	.003	.006	0.40

[a]: *p*-Werte einseitig

[b]: nach Bonferroni-Holm-Methode korrigierte Signifikanzniveaus

Wie Tabelle 8.30 zu entnehmen ist, resultierten die Welch-Tests bezüglich aller verglichenen Gruppen in signifikanten Ergebnissen. Der größte Unterscheid zwischen den Mittelwerten der Bearbeitungszeiten bestand zwischen den Gruppen des Q-Sort-Formats und des Listenformats mit einem starken Effekt von δ = 1.17, der zweitgrößte Unterschied zwischen der Gruppe des

One-Item-a-Time-Formats und dem Q-Sort mit einem mittleren Effekt von δ = 0.59, und der kleinste Unterschied zwischen den Gruppen des One-Item-a-Time-Formats und des Listen-Formats mit einem kleinen bis mittleren Effekt von δ = 0.40. Die Berechnung und Interpretation der Effektstärken erfolgte durchgängig nach Volker (2006).

8.6. Analyse der Ergebnisse des Akzept!-P (Hypothese 5)

Zuletzt erfolgte die Analyse der Ergebnisse des Akzept!-P. Da der Fragebogen zur Evaluierung des B5PO-R in der Voruntersuchung (siehe *7.3. Vorerhebung*) sowie auch den freien Anmerkungen am Ende des Instruments in der Hauptuntersuchung auffallend oft Kritik durch die befragten Personen erfuhr, wurde vor weiterführenden Analysen zunächst eine Betrachtung der inneren Konsistenz der Faktoren des Akzept!-P durchgeführt um deren Verwendbarkeit für die Beantwortung der Fragestellung zu prüfen.

Einer der Hauptkritikpunkte der Teilnehmer der Untersuchung waren die zahlreichen doppelten Verneinungen des Fragebogens. Der Akzept!-P enthält von seinen insgesamt 19 Items 11 Items die negativ formuliert sind und auch dementsprechend negativ gepolt verrechnet werden. Da dies bei Missverständnissen die innere Konsistenz der Faktoren gefährden könnte, wurde zunächst für sämtliche Faktoren Cronbachs α berechnet. Tabelle 8.31 gibt jene Maße der Reliabilität im Sinne der inneren Konsistenz wieder.

Tabelle 8.31

Cronbachs α´s der Itempools der Faktoren des Akzept!-Ps

	Cronbachs α	
Skala	vorher[a]	nachher[a]
Wahrung der Privatsphäre	.73	-
Messqualität	.77	-
Augenscheinvalidität	.59	.66
Kontrollierbarkeit	.41	.52
Antwortformat	.67	-

[a]: bezogen auf Cronbachs α vor und nach dem Ausschluss inadäquater Items

Es lässt sich erkennen, dass die meisten der Faktoren – außer „Kontrollierbarkeit“ und „Augenscheinvalidität“ – mit ihren Reliabilitäten zwischen α = .67 und .77 akzeptable Werte aufweisen, zumindest wenn man die Tatsache bedenkt, dass die Faktoren nur aus jeweils drei bis vier Items bestehen. Bei keinem dieser Faktoren hätte das Ausschließen eines Items zu höheren Werten des Cronbachs α geführt, und sämtliche Korrelationen zwischen den einzelnen Items und der Summe der anderen Items ihrer Faktoren waren positiv. Damit wurde beschlossen, jene Faktoren vollständig in die Analyse mit aufzunehmen.

Die eine Ausnahme bildet der Faktor „Kontrollierbarkeit“ mit einem α = .41. Aufgrund dieses eher kleinen Wertes sollen die Kennwerte der einzelnen Items in Tabelle 8.32 detaillierter aufgeschlüsselt werden. In der Tabelle befinden sich auch die Kennwerte des Faktors „Augenscheinvalidität“, welche zwar von vornherein ein höheres α von .59 erreichte, jedoch bei genauerer Betrachtung ebenfalls ein deutlich vom Rest des Itempools abweichendes Item enthält.

Tabelle 8.32

Kennwerte der Items der Skalen Kontrollierbarkeit und Augenscheinvalidität des Akzept!-P

Skala	Item	α_{rest}	r_{rest}
Kontrollierbarkeit	1	.52	.11
	2	.33	.31
	3	.29	.27
	4	.23	.34
Augenscheinvalidität	1	.66	.15
	2	.48	.43
	3	.44	.48
	4	.47	.44

Aus Tabelle 8.32 geht hervor, dass der Ausschluss von Item 1 der Skala „Kontrollierbarkeit“ zu einer Erhöhung von Cronbachs α auf .52 führen würde. Da α = .52 als eine deutliche Besserung im Vergleich zum vorherigen α = .41 angesehen wird, und die Korrelation von Item 1 zur Summe der anderen Items der Skala lediglich *r* = .11 beträgt, soll dieses Item aus der Skala und somit

auch aus den nachfolgenden Analysen ausgeschlossen werden. Das ausgeschlossene Item war eines der Items mit negativer Polung.

Sehr ähnlich verhält sich Item 1 der Skala „Augenscheinvalidität“. Das Entfernen des Items aus dem Pool seines Faktors resultiert in einer Erhöhung des Cronbachs α von .59 auf .66. Während diese Erhöhung zwar nicht unbedingt substantiell hoch ist, beträgt die Korrelation des Items mit dem Rest der Items seiner Skala lediglich *r* = .15, weshalb beschlossen wurde es aus dem Pool zu entfernen. Diesmal war das entfernte Item keines der negativ gepolten.

Bevor nun die Voraussetzungen der multivariaten Normalverteilung und der Homogenität der Varianz-/Kovarianzmatrizen der MANOVA überprüft werden, soll zunächst die letzte erhobene Variable des Akzept!-P betrachtet werden. Diese war ein einzelnes Item, welches die Personen bittet den B5PO-R mit einer Schulnote nach deutschem Notensystem von eins bis sechs zu bewerten. Hierbei handelt es sich theoretisch um eine ordinalskalierte Variable. Multivariate Varianzanalysen sind allerdings nur dann verlässlich, wenn die verwendeten abhängigen Variablen Intervallskalenniveau aufweisen. Laut Field et al. (2012) kann man jedoch davon ausgehen, dass die Testgröße keine substantiellen Verzerrungen erfährt, solange jene Variablen weitestgehend die Normalverteilung einhalten. Daher ist die Verteilung der Schulnoten in Abbildung 8.18 wiedergegeben.

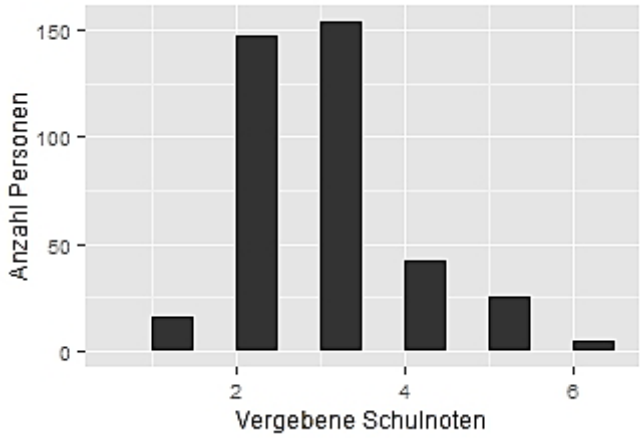


Abbildung 8.18. Häufigkeitsverteilung der vergebenen Schulnoten am Ende des Akzept P

Wie sich in Abbildung 8.18 erkennen lässt, zeigt die Verteilung der vergebenen Schulnoten mit einer Schiefe von *v* = 0.828 zwar Abweichungen von der Normalverteilung; die Ähnlichkeit wäre

jedoch stark genug um für die folgende Auswertung zumindest näherungsweises Intervallskalenniveau der Variable anzunehmen.

Dem folgend wurden drei Shapiro-Wilks Tests zur Überprüfung multivariater Normalverteilung innerhalb der jeweiligen Gruppen durchgeführt. Die Kennwerte der Tests sind in Tabelle 8.33 wiedergegeben.

Tabelle 8.33

Ergebnisse der Shapiro-Wilk-Tests zur Überprüfung der multivariaten Normalverteilung der Ratingsummen der Faktoren des Akzept!-P in den Gruppen der Präsentationsformaten

Gruppe	*W*	*p*
Liste	0.899	< .001*
One-Item-a-Time	0.943	< .001*
Q-Sort	0.898	< .001*

* = $p \leq .05$

Wie sich aus Tabelle 8.33 herauslesen lässt, zeigen sich innerhalb jeder der drei Gruppen starke Abweichungen von multivariater Normalverteilung. Da bei solch starken Abweichungen auch die robusteste Testgröße der MANOVA nicht mehr verwendet werden kann ohne Verzerrungen erwarten zu müssen, wurde nun auf eine robuste Version des Verfahrens zurückgegriffen. Auf die Prüfung der Voraussetzung der Homogenität der Varianz-Kovarianzmatrizen wurde verzichtet, da der hierzu üblicherweise verwendete Box-M-Test sehr sensibel auf Abweichungen der multivariaten Normalverteilung reagiert und in diesem Fall meist signifikante Ergebnisse produziert (Field et al., 2012).

Zunächst seien jedoch in Tabelle 8.34 noch die Kennwerte der Verteilungen der verschiedenen Skalen des Akzept!-P in den drei Gruppen angegeben.

Tabelle 8.34

Kennwerte der Ratingsummen der Faktoren des Akzept!-P in den Gruppen der Präsentationsformate

	M	*Sd*
Wahrung der Privatsphäre		
Liste	18.28	4.06
One-Item-a-Time	19.36	3.94
Q-Sort	19.13	4.00
Messqualität		
Liste	12.41	4.32
One-Item-a-Time	11.92	4.03
Q-Sort	12.50	3.84
Augenscheinvalidität		
Liste	15.35	3.87
One-Item-a-Time	14.69	4.15
Q-Sort	14.92	3.79
Kontrollierbarkeit		
Liste	16.36	2.33
One-Item-a-Time	16.80	1.85
Q-Sort	16.51	2.21
Antwortformat		
Liste	13.36	3.35
One-Item-a-Time	13.62	3.47
Q-Sort	13.83	2.87
Schulnote		
Liste	2.77	1.04
One-Item-a-Time	2.88	1.04
Q-Sort	2.80	0.91

Mittels der Funktion „mulrank()“ aus dem R-Paket „WRS“ (Wilcox & Schönbrodt, 2013) wurde nach Munzel & Brunners Methode (2000, nach Field et al., 2012) eine robuste, nonparametrische Version der MANOVA durchgeführt. Die Ergebnisse sind in Tabelle 8.35 zusammengefasst.

Tabelle 8.35

Ergebnisse der nonparametrischen, robusten MANOVA zum Vergleich der Mittelwerte der Ratingsummen der Faktoren des Akzept!-P in den Gruppen der Präsentationsformate

	Relative mittlere Ränge der Kombinationen Gruppe X Skala							
Gruppe	Wahrung der Privatsphäre	Messqualität	Augenschein-validität	Kontrollier-barkeit	Antwort-format	Schulnote	*F*	*p*
Liste	0.480	0.509	0.529	0.481	0.481	0.486		
One-Item-a-Time	0.517	0.477	0.523	0.523	0.506	0.509	0.684	.693
Q-Sort	0.504	0.517	0.496	0.492	0.507	0.507		

Wie aus Tabelle 8.35 hervorgeht, identifizierte die beschriebene robuste Version der MANOVA keinerlei signifikante Unterschiede in den mittleren Rängen der Skalen des Akzept!-Ps zwischen den drei Gruppen der unterschiedlichen Präsentationsformate.

9. Zusammenfassung und Diskussion

In der vorliegenden Studie wurden die Effekte der Art des Präsentationsformats des gegebenen Itempools eines Selbstbeschreibungsfragebogens auf das Antwortverhalten der befragten Personen, sowie auf die testtheoretischen Eigenschaften des Verfahrens untersucht. Genauer wurden dabei auf Basis einer Online-Erhebung drei verschiedene Präsentationsformate gegenübergestellt: das Listen-Format, welches alle Items auf einer Seite untereinander auflistet, das One-Item-a-Time-Format, welches jedes Item auf einer separaten Seite darstellt, und das Q-Sort-Format, welches in einer Variation des klassischen Q-Sorts die Items als digitale „Karten“ realisiert und deren Beantwortung mittels Einordnung in verschiedene „Fächer“ vorsieht. Dabei stehen diese Fächer, ähnlich einer Ratingskala, für verschiedene Grade der Zustimmung beziehungsweise Ablehnung. Außer der Variation der Präsentationsformate wurde darauf geachtet, maximale Äquivalenz zwischen den drei Bedingungen herzustellen – so waren zum Beispiel Inhalt und Ratingskalen des verwendeten Instruments gleich gehalten. Die zwischen

diesen Bedingungen variierende Eigenschaft der verschiedenen Formate war die Übersicht über den Itempool und über die bereits gegebenen Antworten der befragten Personen.

Dabei wurden, mangels vergleichbarer Vorergebnisse zum Thema, Parallelen zu Studien gezogen, welche sich mit der Itemblockbildung auseinandersetzen (Sparfeldt et al., 2006; Franke, 1997; Lam et al., 2002; Schriesheim et al., 1989; Krampen et al., 1992). Ähnlich der hier durchgeführten Untersuchung variiert bei Vorlage der Itempools mehrerer Skalen in Bedingungen mit und ohne Itemblockbildung die Übersicht, welche den befragten Personen über das Verfahren gewährt wird. Dabei wurde offenbar, dass die Effekte, die eine derartige Variation der gewährten Übersicht über den Itempool provoziert, keineswegs homogene Effekte über die genannten Studien hinweg produzierte. Alleine Veränderungen in der Reliabilität im Sinne der inneren Konsistenz der Itempools der Skalen der verwendeten Instrumente konnte über mehrere Studien hinweg beobachtet werden (vgl. 2. *Verwandte Item-Kontext-Effekte*).

Zum Zwecke der Untersuchung dieses Sachverhalts aus einem neuen methodischen Blickwinkel wurde der in seiner Konstruktion als Rasch-Modell-konsistent bestätigte Selbstbeschreibungsfragebogen B5PO-R (Siptroth, 2011) verwendet. Außerdem wurde, um neben den Ergebnissen des B5PO-Rs auch Aussagen über die Wahrnehmung des Verfahrens durch die Teilnehmer machen zu können, in Folge der Akzept!-P (Kersting, 2012) vorgegeben. Letzterer hat das Ziel, die Akzeptanz eines vorhergehenden Verfahrens durch die Teilnehmer differenziert zu erfassen. Im Zuge der Online-Erhebung wurden die befragten Personen zufällig einer von drei Bedingungen zugeteilt: dabei wurden von sämtlichen Personen zunächst einige demographische Variablen erhoben, woraufhin sie den B5PO-R und anschließend den Akzept!-P ausfüllten. Der Unterschied zwischen den drei Gruppen bestand im Präsentationsformat des B5PO-Rs: je circa einem Drittel der Stichprobe wurde der Fragebogen im Listen-, im One-Item-a-Time-, oder im Q-Sort-Format vorgelegt. Unter anderem im Lichte der angesprochenen Studien zur Itemblockbildung (siehe auch *2. Verwandte Item-Kontext-Effekte*) wurde erwartet, dass – bedingt durch die Variation des Überblicks über den Itempool und die bereits gegebenen Antworten – sich die drei Gruppen in der Verteilung der Antworten über das Spektrum der verwendeten Ratingskala, den Mittelwerten und Interkorrelationsmatrizen der Skalensummen, den Bearbeitungszeiten und der Akzeptanz des adaptierten B5PO-Rs unterscheiden würden. Besonders hervorzuheben ist die zentrale Hypothese der Untersuchung: die Veränderung in der Reliabilitäten im Sinne der inneren Konsistenz der Itempools der Faktoren, welche bis dato in erster Linie unter Verwendung korrelativer Maße wie Cronbachs α untersucht wurde. Im Falle der vorliegenden Studie hingegen wurden die Mittel der Item-Response-Theory implementiert;

da der B5PO-R in seiner Konstruktion explizit unter Gültigkeit des Rasch-Modells skaliert wurde (Siptroth, 2011), war eine eingehende Analyse der inneren Konsistenz der Itempools der Faktoren möglich, da die innere Konsistenz eine notwendige Voraussetzung für die Gültigkeit des Rasch-Modells in einem Itempool ist.

Dem zufolge wurden zwei allgemeine Möglichkeiten der Rasch-Modell-Gültigkeit in Betracht gezogen: zum einen wäre es möglich gewesen, dass die Itempools der Faktoren des B5PO-Rs, welche ja per se als Rasch-Modell-konsistent bestätigt wurden, ihre Modellpassung auch dann behalten wenn man die Daten der kompletten Stichprobe für die Schätzung der Itemparameter nach dem Rasch-Modell zu Rate zieht. Dies würde bedeuten, dass die unterschiedlichen Formate in denen die jeweiligen Itempools präsentiert wurden, keine Effekte auf die innere Konsistenz der Faktoren ausüben. Die andere Möglichkeit wäre mangelhafte Modellpassung für die Daten der gesamten Stichprobe gewesen; dies würde darauf hin deuten, dass die unterschiedlichen Formate sehr wohl einen solchen Effekt ausüben, woraufhin mit Modelltests für die getrennten Itempools der jeweiligen Präsentationsformate fortgefahren werden müsste.

Nun existieren keine standardisierten, objektiven Maße um die unterschiedliche Passung des Rasch-Modells für verschiedene Itempools festzulegen. Um den reinen Effekt der Präsentationsformate auf die innere Konsistenz zu extrahieren kann man allerdings folgendermaßen vorgehen: Bedenkt man die prinzipielle Gültigkeit des Rasch-Modells für die Itempools der drei Faktoren des B5PO-R, so kann man zumindest eine grobe Abschätzung über den Effekt der Präsentationsformate gewinnen, indem man betrachtet wie viele Items in den jeweiligen Gruppen und Faktoren ausgeschlossen werden mussten um Modellgültigkeit herzustellen. Tabelle 9.1 gibt einen solchen Überblick über die Anzahl pro Gruppe und Faktor auszuschließender Items.

Tabelle 9.1

Anzahl der pro Gruppe (Präsentationsformate) und Faktoren des B5PO-R auszuschließenden Items zur Herstellung von Rasch-Modell-Gültigkeit

	Alle Formate gemeinsam	Liste	One-Item-a-Time	Q-Sort	∑ =
Extraversion – 9 Items	> 3	2	1	4	> 10
Gewissenhaftigkeit – 9 Items	3	0	0	0	3
Verträglichkeit – 12 Items	3	1	0	0	4
∑ =	> 9	3	1	4	

Wie Tabelle 9.1 entnommen werden kann, war es für alle drei Faktoren des B5PO-R in allen drei Präsentationsformaten gemeinsam notwendig, mindestens drei Items aus dem Pool auszuschließen, bevor Rasch-Modellgültigkeit hergestellt werden konnte. Dies ist – abermals, unter der Prämisse der prinzipiellen Gültigkeit des Modells für die Itempools – ein deutliches Anzeichen für Veränderungen der inneren Konsistenz durch die unterschiedlichen Arten, die Items zu präsentieren. Für die beiden Faktoren Gewissenhaftigkeit und Verträglichkeit galt das Rasch-Modell in jedem der Präsentationsformate für sich, und das ohne die Notwendigkeit, bedeutsam viele Items auszuschließen. Dies legt den Schluss nahe, dass mit variierender Übersicht über den Itempool durchaus Verschiebungen in der inneren Konsistenz der Faktoren auftreten. Die Messeigenschaften der Items verändern sich je nach der Art ihres Präsentationsformats. Schließt man in diese Betrachtung den Faktor Extraversion mit ein, entsteht ein Bild, das darauf hindeutet, dass die größte innere Konsistenz beim One-Item-a-Time-Präsentationsformat besteht. Die beiden Formate der Liste und des „Q-Sorts“ scheinen hingegen größere Abweichungen der inneren Konsistenz zu erzeugen. Jene Interpretation wäre zum Beispiel mit den Ergebnissen von Franke (1997) kohärent – sie stellte in ihrer Studie zur Itemblockbildung fest, dass mit steigender Übersicht über den Itempool die Reliabilitäten im Sinne der inneren Konsistenzen der Itempools ihrer Faktoren absanken. Dies interpretierte sie dadurch, dass bei hoher Übersicht die Reaktion der erfragten Personen nicht nur auf das jeweilige Item, sondern auch auf die vermutete zugrundeliegende Skala stattfanden, womit die Eindimensionalität der Items kompromittiert werden würde.

Ist also das hauptsächliche Ziel der Fragebogenkonstruktion, ein Instrument mit maximaler innerer Konsistent zu kreieren – wie zum Beispiel bei Rasch-Modell-Skalierungen – so ist scheinbar das One-Item-a-Time-Format die überlegene Wahl.

Trotzdem sind die Ergebnisse aus Tabelle 9.1 mit Vorsicht zu interpretieren: Schließlich waren für die Analysen aller drei Präsentationsformate gemeinsam deutlich größere Gruppenumfänge vorhanden als für die Analysen der einzelnen Formate. Da mit steigender Gruppengrößen auch exaktere Schätzungen der Itemparameter resultieren, kann es sein dass in diesem Falle etwaig vorhandene Modellabweichungen leichter für die Modelltests zu identifizieren waren als im Falle kleinerer Gruppengrößen, welche ja auch größere Standardfehler der Itemparameterschätzungen produzieren. Trotzdem ist diese Annahme eine eher unwahrscheinliche, da ja zum Beispiel für den Itempool des Faktors Extraversion deutliche Unterschiede in den zu entfernenden Items bestehen, welche nicht ausreichend durch die unterschiedliche Gruppengröße erklärt werden kann.

Der letzte auffällige Punkt von Tabelle 9.1 ist der Unterschied zwischen der Anzahl ausgeschlossener Items beim Faktor Extraversion im Vergleich zu den anderen beiden Faktoren. Offensichtlich waren bei Extraversion auffällig größere Reduktionen des Itempools nötig als in den anderen beiden Fällen. Bezieht man sich aber auf frühere Versuche, die Faktoren der Big Five nach dem Rasch-Modell zu skalieren, wird klar, dass jene Auffälligkeiten öfter bei diesem Faktor auftreten, zum Beispiel bei Rost et al. (1999): „[…] lediglich die Skala Extraversion erweist sich – wie so oft – als heterogen“. Unter Betrachtung dieses Fundes zusammen mit den oben angeführten Unterschieden bezüglich der Eindimensionalität der Skala Extraversion und der anderen beiden Skalen könnte man auch den Schluss ziehen, dass zumindest Teile dieser Heterogenität der Skala Extraversion durch das Listen-Präsentationsformat der Studie von Rost et al. zustande gekommen war.

Explizit soll erwähnt werden, dass – wie unter *7.2.2. B5PO-R: Big Five Plus One Revised* dargelegt – der B5PO-R hier nicht in seiner intendierten Form zur Anwendung kam, sondern starke Adaptionen durchlief bevor er im Sinne der vorliegenden Untersuchung verwendet werden konnte. Zum Beispiel die Reduzierung der ursprünglichen diametral gegenübergestellten Adjektivpaare auf einzelne Adjektive war eine der deutlichsten Änderungen des Ursprungsmaterials. Somit stellen die hier gefundenen Ergebnisse ausdrücklich keine Aussage über die Rasch-Modell-Gültigkeit des ursprünglichen Fragebogens dar.

Über die Rasch-Modell-Analysen hinaus wurden die Unterschiede zwischen den drei Gruppen der Präsentationsformate in den Mittelwerten der Faktoren des B5PO-R, den Werten seiner Akzeptanz durch die befragten Personen (erhoben mittels dem Akzept!-P, Kersting, 2012), der Nutzung des verfügbaren Antwortspektrums, und seiner Bearbeitungszeiten untersucht. Die Ergebnisse zeugten von keinerlei Unterschieden zwischen den Mittelwerten und der Akzeptanz der verschiedenen Itempool-Präsentationsformate. Zwar produzierte der statistische Test zur Untersuchung der Unterschiede in der Verwendung des verfügbaren Antwortspektrums zwischen den drei Gruppen ein signifikantes Ergebnis, jedoch war die zugehörige Effektstärke (w = .05) derart klein, dass die Unterschiede kaum als praktisch relevant angesehen werden können. Lediglich die Analyse der Bearbeitungszeiten zeigte deutliche Unterschiede zwischen den drei Präsentationsformaten: die Bearbeitung des Q-Sort-Formats dauerte dabei signifikant länger als die Bearbeitung der anderen beiden Formate, mit einer beachtlichen Effektstärke von Glass´s δ = 1.17 für den Unterschied zum Listen-, und δ = 0.59 zum One-Item-a-Time-Format. Das One-Item-a-Time-Format erforderte längere Bearbeitungszeiten als das Listen-Format, mit einem kleinen bis mittleren Effekt bei Glass´s δ = .40. Interessant war auch die Feststellung, dass mit den höheren Mittelwerten der Bearbeitungszeiten ebenfalls die Streuung der Werte auffällig größer ausfiel, mit $sd_{\text{Q-Sort}} = 1.71\ sd_{\text{Liste}}$ und $sd_{\text{One-Item-a-Time}} = 1.29\ sd_{\text{Liste}}$.

Aus der Zusammenführung dieser Ergebnisse sollen nun die Eigenschaften der verschiedenen Itempool-Präsentationsformate verglichen werden. Zunächst soll festgestellt werden, dass die verschiedenen Formate offenbar keine unterschiedlichen Ergebnisse bezüglich der Skalenmittelwerte produzieren; die Daten deuten auf die Abwesenheit selbst sehr kleiner Effekte hin. Auch die Ausnutzung des verfügbaren Spektrums an Antwortmöglichkeiten unterschied sich nur auf verschwindend geringe Art zwischen den Präsentationsformaten. Offenbar vernachlässigten die befragten Personen zum Beispiel die ablehnendste Antwortkategorie gleich stark in allen Gruppen, unabhängig davon wie offensichtlich ihnen die linksschiefe Verteilung ihrer Antworten beispielsweise während dem Ausfüllen des Q-Sorts gewesen sein mag. Wohl könnte man argumentieren, dass der Aufforderungscharakter der Items so stark war dass er derartige Effekte überdeckt hätte. Jedoch hätte es ob der ungleich höheren Übersicht des Q-Sort-Präsentationsformats im Vergleich zu den anderen beiden Formate zumindest kleine Unterschiede in der Verteilung der gewählten Antwortkategorien geben müssen um die Annahme unterschiedlicher Nutzung der Ratingskala halten zu können. Da dies nicht so war, kann gefolgert werden, dass die höhere Übersicht über die bereits gegebenen Antworten für die befragten Personen keine Auswirkungen auf die weitere Einschätzung der folgenden Adjektive hatte.

Auch bezüglich der Akzeptanz gab es keine identifizierbaren Unterschiede zwischen den Gruppen; zwar konnte, wie in *5. Auswirkungen des Präsentationsformats auf die Akzeptanz von Fragebögen* bereits dargelegt, Fluckinger (2010) deutliche Unterschiede zwischen dem Listen- und dem Q-Sort-Format feststellen, jedoch war der Versuchsaufbau bei ihm ein anderer: eine Hälfte der Stichprobe bearbeitete zunächst ein Verfahren in Listen-Format und einige Zeit später dasselbe Verfahren im Q-Sort-Format, die andere Hälfte bearbeitete das Verfahren in umgekehrter Reihenfolge der Präsentationsformate. Dies könnte zur Folge gehabt haben, dass die Versuchsteilnehmer durch den Kontrast zwischen den beiden Formaten explizit auf das variierte Präsentationsformat aufmerksam gemacht wurden und auch dementsprechend stärker variierende Einstellungen zu ihnen produzierten. Diese Einschränkung seiner Ergebnisse wird dadurch untermauert, dass in mehreren Punkten der Akzeptanz Sequenz-Effekte auftraten:

> *„[…] a number of order effects were found, indicating that the mean of many perceptions and reactions were influenced by the order in which the assessments were experienced. For example, participants viewed the Q-sort as significantly less fakable after having taken the Likert measure first.*" (Fluckinger, 2010, S.62)

In der Untersuchungsanordnung der vorliegenden Studie hingegen fehlte den Personen der Ankerpunkt eines zweiten Formates, da jede befragte Person den B5PO-R nur in einer einzigen Fassung bearbeitete. Für sich alleine genommen bedachten die Teilnehmer und Teilnehmerinnen die Präsentationsformate mit äquivalenten Bewertungen der Akzeptanz.

Trotzdem soll dieses Ergebnis nicht als endgültig angenommen werden, da während der Auswertung der Daten klar wurde, dass teilweise Probleme mit dem verwendeten Verfahren des Akzept!-P bestehen. Zum einen gaben zahlreiche Teilnehmer und Teilnehmerinnen die Rückmeldung, dass die negativ formulierten Items verwirrend und kognitiv anstrengend seien. Dies ist zwar ein bekanntes Problem der Itemgenerierung, fällt aber beim Akzept!-P besonders schwer ins Gewicht, da dieser von seinen neunzehn Items insgesamt elf negativ formulierte Items enthält. Auch besteht eine der Skalen des Instruments ausschließlich aus negativ gepolten Items. Zudem bemängelten die Personen das wiederholte Abfragen desselben Sachverhaltes mit unterschiedlichen Formulierungen. Diese Eigenschaften des Instruments könnten unter Umständen dazu geführt haben, dass die befragten Personen im Zuge einer reaktanten Reaktion keine Anstrengungen leisteten ihre Meinungen differenziert wiederzugeben, sondern verallgemeinerte Antworten auf die Fragen des Verfahrens gaben. Die teils starke Linksschiefe der Verteilungen der Ratingsummen der Skalen des Akzept!-P unterstreicht diese Interpretation.

Zusammenfassend lässt sich über die Präsentationsmodi sagen, dass keines der verwendeten Formate klare und umfassende Überlegenheit gegenüber einem anderen bewies. Der Q-Sort führte, trotz gegenteiliger Erwartungen, zu keiner nennenswert besseren Einschätzung des Verfahrens durch die Teilnehmer oder einer umfassenderen Ausnutzung der verfügbaren Antwortmöglichkeiten. Aufgrund der stark erhöhten Bearbeitungszeiten ist von seiner Verwendung statt einem der anderen Formate sogar eher abzuraten.

Einzig hervorstechend ist das One-Item-a-Time-Format, das bei der Analyse mittels Rasch-Modell-Skalierung deutlich bessere Modellgültigkeit und somit auch höhere innere Konsistenzen der Faktoren des Selbstbeschreibungsfragebogens bewies als die anderen beiden Verfahren, auch wenn die in dieser Studie verwendete Methodik lediglich Hinweise auf jene inneren Konsistenzen liefern kann. Dem muss in der Praxis jedoch gegenübergestellt werden, ob sich die erhöhte Bearbeitungszeit gegenüber dem Listen-Format schlussendlich lohnt. In der vorliegenden Untersuchung waren die zeitlichen Unterschiede zwar nicht gravierend, allerdings muss auch bedacht werden, dass das verwendete Instrument aus lediglich 30 Adjektiven bestand. In der diagnostischen Praxis muss davon ausgegangen werden, dass eines oder mehrere längere Instrumente zur Anwendung kommen, wodurch auch die Unterschiede proportional ansteigen.

Die weiterhin offenen Fragen beziehen sich vor allem auf die Situation der Fragebogenvorgabe und dem Ziel des verwendeten Instruments. Es sind durchaus Situationen denkbar, in denen die Motivation der befragten Personen, sich auf eine gewisse Art und Weise darzustellen, höher ist als in jener der vorliegenden Untersuchung. Eine Person, deren definitives Ziel es ist ein sozial erwünschtes Antwortmuster zu generieren, könnte dies eventuell leichter im Q-Sort-Format als im One-Item-a-Time-Format realisieren. Die unter *2. Verwandte Item-Kontext-Effekte* angesprochene Studie von Krampen et al. (1992) untermauert diese Annahme, da sie starke Hinweise darauf fanden, dass die Effekte einer Instruktion zum sozial erwünschten Antworten durch eine Erhöhung der Übersicht über den Itempool verstärkt wurden. Zur Beantwortung dieser Frage im konkreten Bezug auf die hier verwendeten Präsentationsformate wäre eine Untersuchung unter valideren Umständen notwendig.

Zudem bleibt, wie in *2. Verwandte Item-Kontext-Effekte* bereits angedeutet, offen, ob die hier untersuchten Effekte auch bei Verwendung eines anderen diagnostischen Instruments äquivalente Resultate erzielen. Wie bereits bei der Darstellung der Studien zur Itemblockbildung offenbar wurde, scheinen Effekte welche aus der Übersicht über das Verfahren resultieren, sich

bei Instrumenten mit unterschiedlichen Fragestellungen unterschiedlich auszuwirken. Inwiefern hier bereichsspezifische Effekte bestehen, bleibt zu klären.

10. Literaturverzeichnis

Bambery, M., & Porcerelli, J. (2010). Q-Methodology. In *The Corsini Encyclopedia of Psychology* (Bd. 3, S. 1401-1403). Hoboken, NJ: John Wiley & Sons, Inc..

Becker, P. (2003). TIPI. *Trierer Integriertes Persönlichkeitsinventar*. Göttingen: Hogrefe.

Block, J. (2008). *The Q-Sort in Character Appraisal: Encoding Subjective Impressions on Persons Quantitatively.* Washington, DC: American Psychological Association.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler, 4., überarbeitete Auflage*. Heidelberg: Springer.

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Heidelberg: Springer.

Brown, S. R. (1996). Q Methodology and Qualitative Research. *Qualitative Health Research*, 6 (4), 561-567. Und online unter: http://schmolck.userweb.mwn.de/qmethod/srbqhc.htm (abgerufen am 21.11.2012).

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

Champely, S. (2009). *pwr: Basic functions for power analysis*. R package version 1.1.1. http://CRAN.R-project.org/package=pwr.

D'Amico, E. J., Neilands, T. B., & Zambarano, R. (2001). Power Analysis for multivariate and repeated measures designs: A flexible approach using the SPSS MANOVA procedure. *Behavior Research Methods, Instruments, & Computers*, 33(4), 479-484.

Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., Covi, L. (1974). The Hopkins Symptom Checklist (HSCL): A self-report symptom inventory. *Behavioral Science*, 19(1), 1-15.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. London: SAGE.

Fluckinger, C. D. (2010). *Measurement of Big Five Personality via Q-Sort: Comparison with a Likert Measure and Test-Taker Perceptions and Reactions*. Unveröffentlichte Dissertation, University of Akron.

Franke, G. H. (1997). „The Whole is More than the Sum of its Parts": The Effects of Grouping and Randomizing Items on the Reliability and Validity of Questionnaires. *European Journal of Psychological Assessment*, 13(2), 67-74.

Hackert, C., & Braehler, G. (2007). *FlashQ – Q-Sorting via Internet*. http://www.hackert.biz/flashq/home/ (abgerufen am 28.11.2012).

Holocher-Ertl, S., Kubinger, K. D., & Menghin, S. (2003). *B5PO Big Five Plus One Persönlichkeitsinventar*. Mödling: Schuhfried.

Kersting, M. (2010). Akzeptanz von Assessment-Centern: Was kommt an und worauf kommt es an?. *Wirtschaftspsychologie*, 2, 58-64.

Kersting, M. (2012). *Akzept! Fragebogen zur Messung der Akzeptanz diagnostischer Verfahren*. Online in Internet: URL: http://www.kersting-internet.de/akzept.html (Stand: 10.09.2012, 14:30 Uhr).

Krampen, G., Hense, H., & Schneider, J. F. (1992). Reliabilität und Validität von Fragebogenskalen bei Standardreihenfolge versus inhaltshomogener Blockbildung ihrer Items. *Zeitschrift für experimentelle und angewandte Psychologie*, 39(2), 229-248.

Kubinger, K. D. (2002). On faking personality inventories. *Psychologische Beiträge*, 44, 10-16.

Kubinger, K. D. (2009). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens*. Göttingen: Hogrefe.

Kubinger, K. D., Rasch, D., & Yanagida, T. (2011) A new approach for testing the Rasch Model. *Educational Research and Evaluation: An International Journal on Theory and Practice*, 17(5), 321-333.

Lam, T. C. M., Green, K. E., & Bordignon, C. (2002). Effects of Item Grouping and Position of the "Don´t Know" Option on Questionnaire Response. *Field Methods*, 14(4), 418-432.

Lösel, F. (1999). Persönlichkeitsdaten (Tests). In Jäger, R.S. & Petermann, F. (Hrsg.), *Psychologische Diagnostik* (363-380). Weinheim: Psychologie Verlags Union.

Mair, P., Hatzinger, R., & Maier, M. J. (2012). *eRm: Extended Rasch Modeling*. R package version 0.15-1. http://CRAN.R-project.org/package=eRm.

Maltby, J., Day, L. & Macaskill, A. (2011). *Differentielle Psychologie, Persönlichkeit und Intelligenz*. München: Pearson.

Meyer, D., Zeileis, A., & Hornik, K. (2011). *vcd: Visualizing Categorial Data*. R Package Version 1.2-12. http://cran.r-project.org/web/packages/vcd/index.html.

Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R. NEO Persoenlichkeitsinventar nach Costa und McCrae. Revidierte Fassung*. Göttingen: Hogrefe.

Prasad, R.S. (2001). Development of the HIV/AIDS Q-Sort Instrument to Measure Physician Attitudes. *Family Medicine*, 33(10), 772-8.

R Development Core Team (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rasch, D., Kubinger, K.D. & Moder, K. (2011). The two-sample *t* test: pre-testing its assumptions does not pay off. *Statistical Papers*, 52, 219-231.

Rasch, D., Kubinger, K. D., & Yanagida, T. (2011). *Statistics in Psychology Using R and SPSS*. West Sussex: John Wiley & Sons.

Revelle, W. (2011). *psych: Procedures for Personality and Psychological Research*. R package version 1.01.9. http://cran.r-project.org/web/packages/psych/index.html.

Roskamp, E.E. (1996). Latent-Trait-Modelle. In Erdfelder, E., Mausfeld, R., Meiser, T. & Rudinger, G. (Hrsg.). *Handbuch Quantitative Methoden* (S. 431-458). Weinheim: Psychologie Verlags Union.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber.

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

Schneewind, K. A., & Graf, J. (1998). *Der 16-Persoenlichkeits-Faktoren-Test Revidierte Fassung. 16PF-R.* Bern: Huber.

Schönpflug, W. (2004). *Geschichte und Systematik der Psychologie. Ein Lehrbuch für das Grundstudium*. Weinheim: Beltz.

Schriesheim, C. A., Kopelman, R. E., & Solomon, E. (1989). The Effect of Grouped Versus Randomized Questionnaire Format on Scale Reliability and Validity: A Three-Study Investigation. *Educational and Psychological Measurement*, 49, 487-508.

Schwarz, N. (1999). Self-Reports: How the Questions Shape the Answers. *American Psychologist*, 54(2), 93-105.

Siptroth, V. (2011). *Neubearbeitung der Adjektivliste des Big Five Plus One Persönlichkeitsinventars*. Unveröffentlichte Diplomarbeit, Universität Wien.

Slawomir, J. (2012). *mvnormtest: Normality test for multivariate variables*. R package version 0.1-9. http://CRAN.R-project.org/package=mvnormtest.

Sparfeldt, J. R., Schilling, S. R., Rost, D. H., & Thiel, A. (2006). Blocked Versus Randomized Format of Questionnaires. A Confirmatory Multigroup Analysis. *Educational and Psychological Measurement*, 66(6), 961-974.

The Epimetrics Group (2010). *Q-Assessor: Automated Q-Methodology Studies Online*. http://q-assessor.com (abgerufen am 28.11.2012).

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin*, 103(3), 299-314.

Volker, M. A. (2006). Reporting Effect Size Estimates in School Psychology Research. *Psychology in the Schools*, 43(6), 653-672.

Warnes, G. R. (2012) *gmodels: Various R programming tools for model fitting*. R Package version 2.15.3. http://CRAN.R-project.org/package=gmodels.

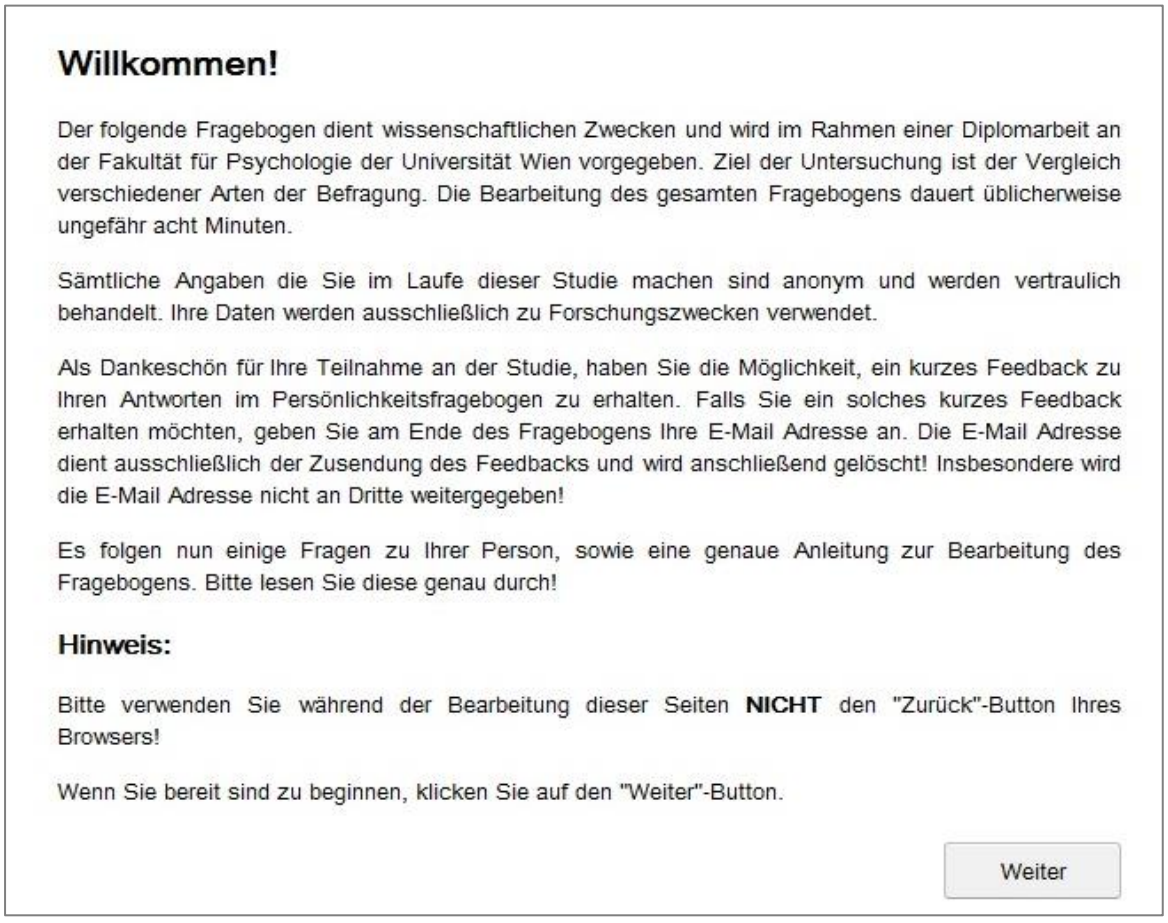
Wickham, H. (2009) *ggplot2: elegant graphics for data analysis*. New York: Springer.

Wilcox, R. R., & Schönbrodt, F.D. (2013). *The WRS package for robust statistics in R.* R Package version 0.20. http://r-forge.r-project.org/projects/wrs/.

IV. Anhang

11. Screenshots des Online-Assessments

Anmerkung: aus Platzgründen sind sämtliche Screenshots des Assessments auf den Bereich der wesentlichen Inhalte zurechtgeschnitten worden. Zudem ist aufgrund des Aufbaus der Studie nicht allen Teilnehmern jede Seite vorgelegt worden – daher ist zu jedem der Screenshots die Gruppenzugehörigkeit angegeben. Aus urheberrechtlichen Gründen sowie um die diagnostische Halbwertszeit der verwendeten Verfahren nicht unnötig zu schmälern, wurden an gegebenen Stellen die konkreten Items mit schwarzen Balken überlegt.



Willkommen!

Der folgende Fragebogen dient wissenschaftlichen Zwecken und wird im Rahmen einer Diplomarbeit an der Fakultät für Psychologie der Universität Wien vorgegeben. Ziel der Untersuchung ist der Vergleich verschiedener Arten der Befragung. Die Bearbeitung des gesamten Fragebogens dauert üblicherweise ungefähr acht Minuten.

Sämtliche Angaben die Sie im Laufe dieser Studie machen sind anonym und werden vertraulich behandelt. Ihre Daten werden ausschließlich zu Forschungszwecken verwendet.

Als Dankeschön für Ihre Teilnahme an der Studie, haben Sie die Möglichkeit, ein kurzes Feedback zu Ihren Antworten im Persönlichkeitsfragebogen zu erhalten. Falls Sie ein solches kurzes Feedback erhalten möchten, geben Sie am Ende des Fragebogens Ihre E-Mail Adresse an. Die E-Mail Adresse dient ausschließlich der Zusendung des Feedbacks und wird anschließend gelöscht! Insbesondere wird die E-Mail Adresse nicht an Dritte weitergegeben!

Es folgen nun einige Fragen zu Ihrer Person, sowie eine genaue Anleitung zur Bearbeitung des Fragebogens. Bitte lesen Sie diese genau durch!

Hinweis:

Bitte verwenden Sie während der Bearbeitung dieser Seiten **NICHT** den "Zurück"-Button Ihres Browsers!

Wenn Sie bereit sind zu beginnen, klicken Sie auf den "Weiter"-Button.

Weiter

Abbildung 11.1. Willkommensbildschirm nach dem Öffnen des Online-Assessment-Links (alle Gruppen)

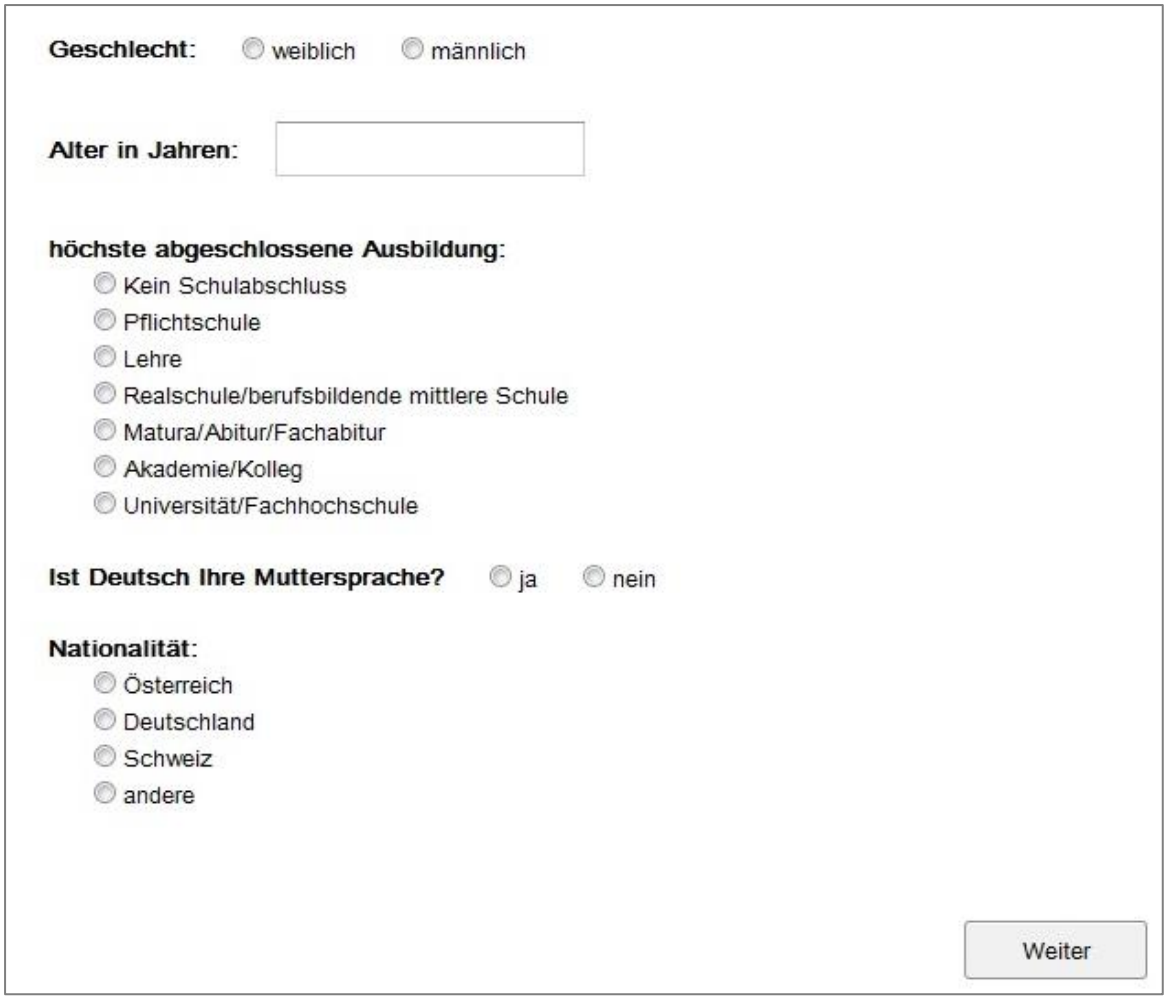


Abbildung 11.2. Demographischer Teil des Fragebogens des Online-Assessments, alle Gruppen

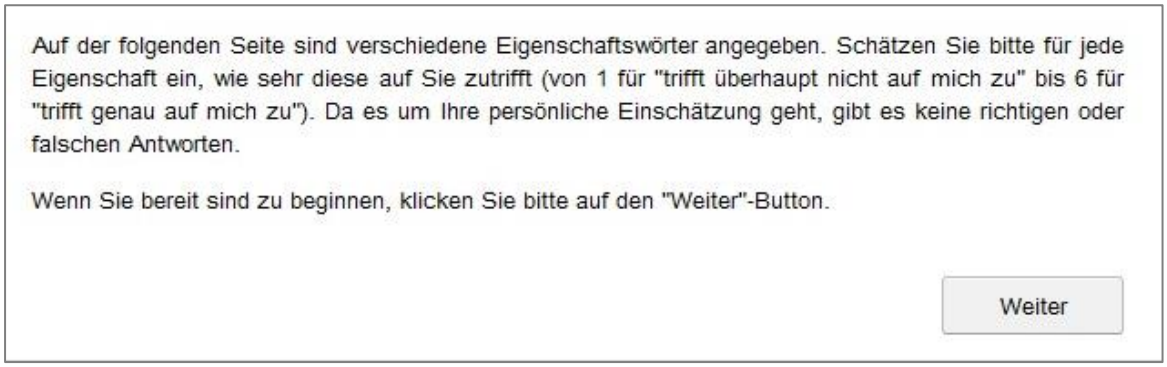


Abbildung 11.3. Instruktionen des Listen-Präsentationsformats (Gruppe Liste)

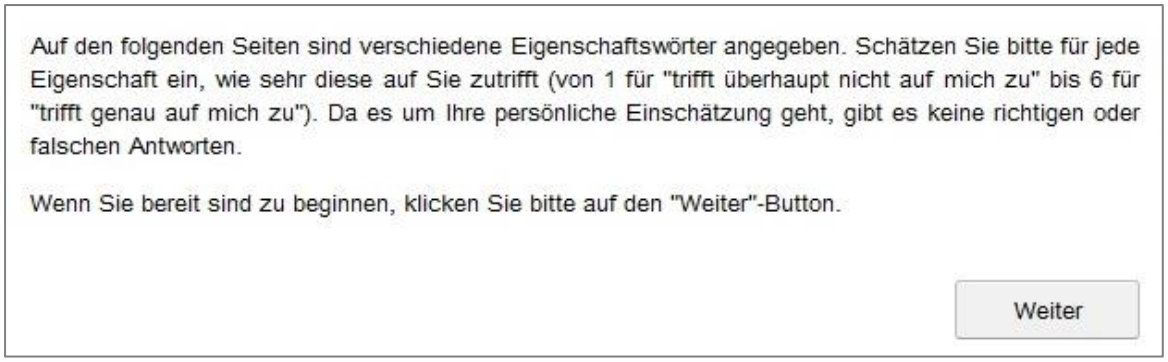


Abbildung 11.4. Instruktionen des One-Item-a-Time-Präsentationsformats (Gruppe One-Item-a-Time)

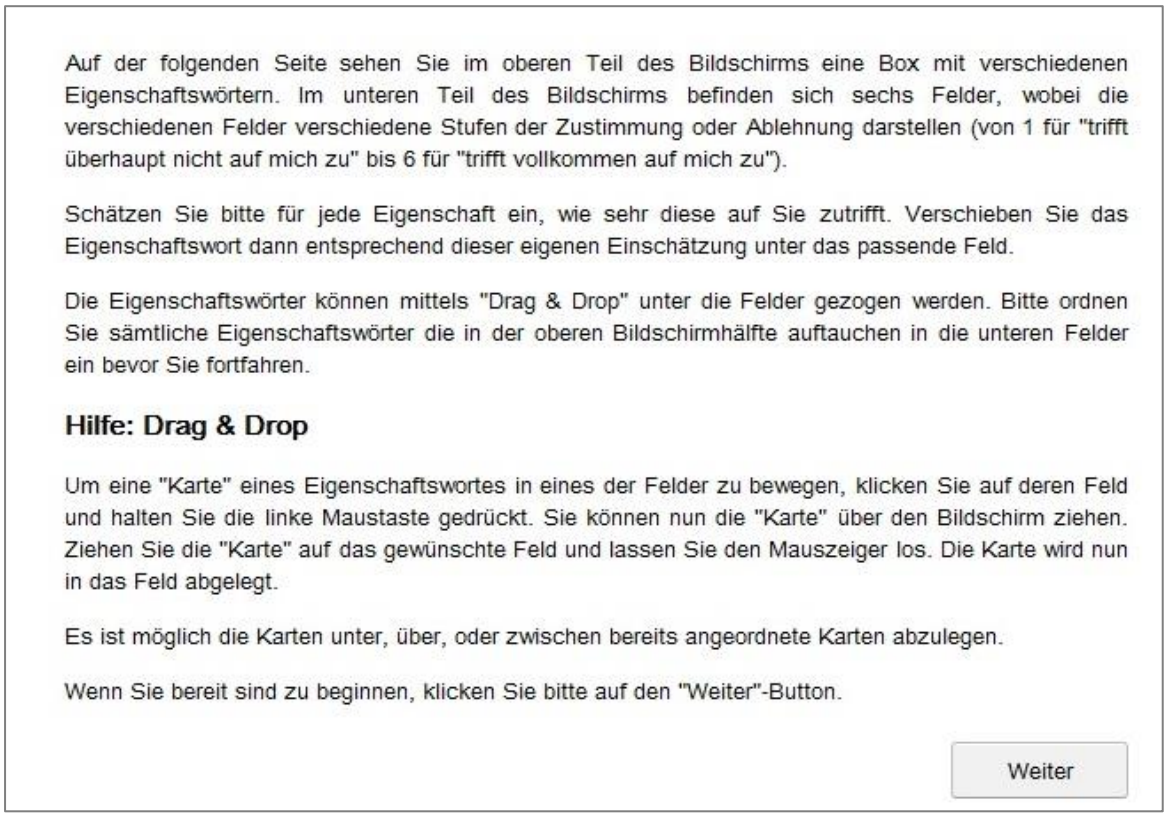


Abbildung 11.5. Instruktionen des Q-Sort-Präsentationsformats (Gruppe „Q-Sort)

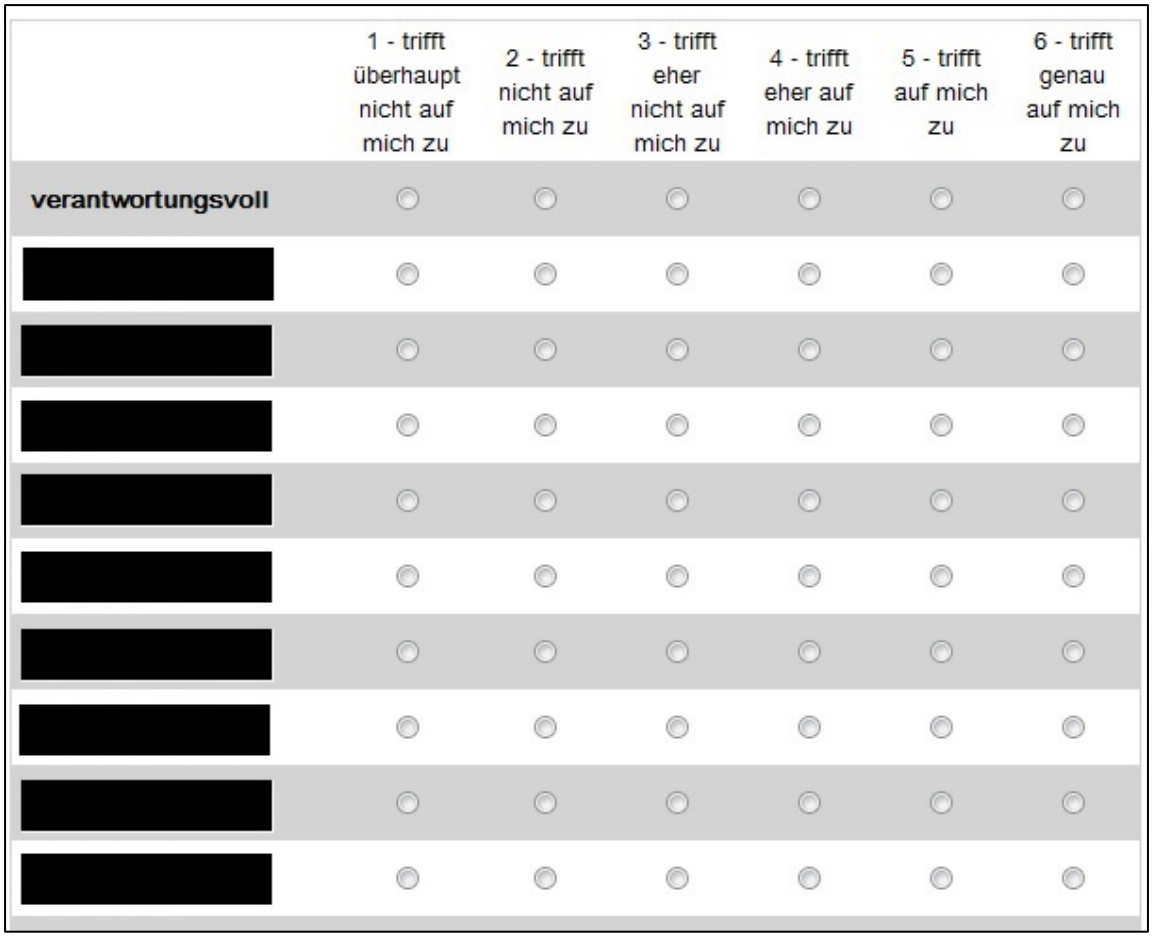


Abbildung 11.6. Auszug der Items des Listen-Präsentationsformats (Gruppe Liste)

Abbildung 11.7. Auszug der Seiten des One-Item-a-Time-Präsentationsformats

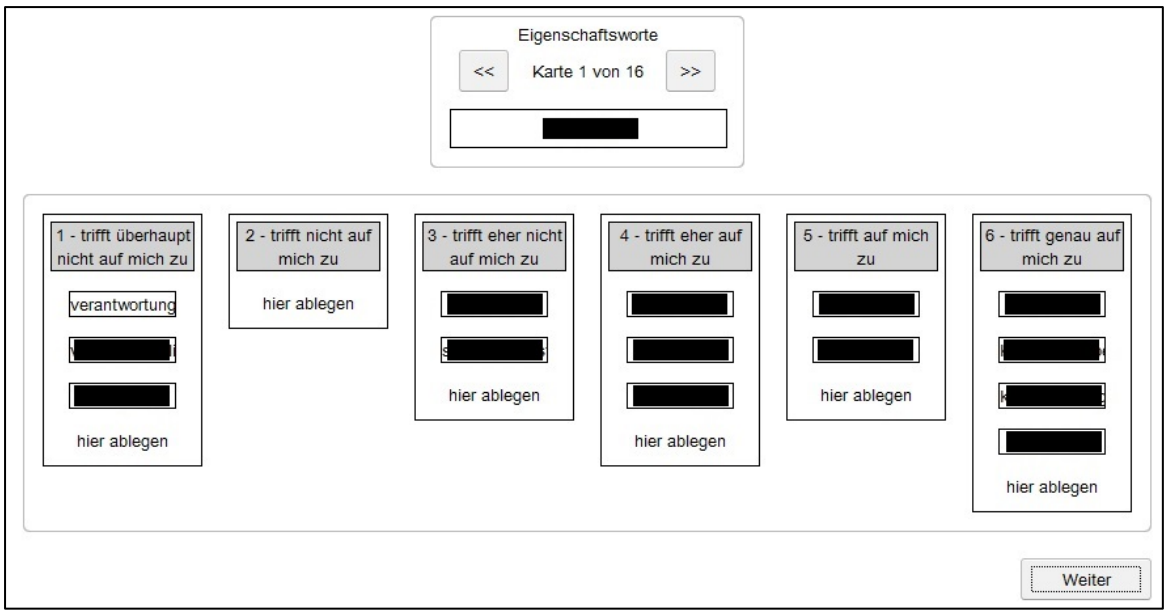
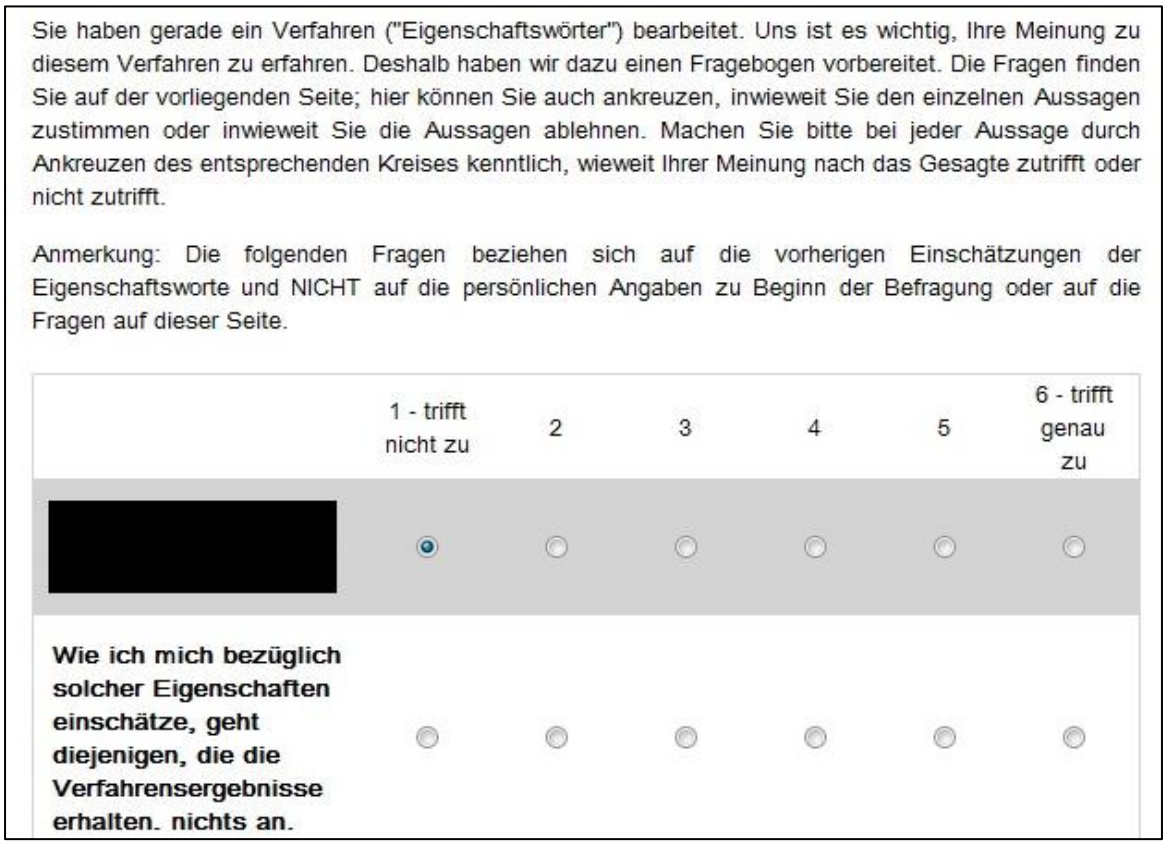


Abbildung 11.8. Beispielhafte Darstellung der Seite des Q-Sort-Präsentationsformats (Gruppe Q-Sort)



Sie haben gerade ein Verfahren ("Eigenschaftswörter") bearbeitet. Uns ist es wichtig, Ihre Meinung zu diesem Verfahren zu erfahren. Deshalb haben wir dazu einen Fragebogen vorbereitet. Die Fragen finden Sie auf der vorliegenden Seite; hier können Sie auch ankreuzen, inwieweit Sie den einzelnen Aussagen zustimmen oder inwieweit Sie die Aussagen ablehnen. Machen Sie bitte bei jeder Aussage durch Ankreuzen des entsprechenden Kreises kenntlich, wieweit Ihrer Meinung nach das Gesagte zutrifft oder nicht zutrifft.

Anmerkung: Die folgenden Fragen beziehen sich auf die vorherigen Einschätzungen der Eigenschaftsworte und NICHT auf die persönlichen Angaben zu Beginn der Befragung oder auf die Fragen auf dieser Seite.

	1 - trifft nicht zu	2	3	4	5	6 - trifft genau zu
[redacted]	◉	○	○	○	○	○
Wie ich mich bezüglich solcher Eigenschaften einschätze, geht diejenigen, die die Verfahrensergebnisse erhalten, nichts an.	○	○	○	○	○	○

Abbildung 11.9. Instruktionen und beispielhafter Auszug der Seite des Akzept!-P (alle Gruppen)

Abbildung 11.10. Dank für die Teilnahme an der Studie (alle Gruppen)

12. Abbildungsverzeichnis

[a]: Graphik wurde erstellt mit dem R-Basispaket (R Core Development Team, 2011)

[b]: Graphik wurde erstellt mit dem R-Paket ggplot2 (Wickham, 2009)

[c]: Graphik wurde erstellt mit dem R-Paket eRm (Mair et al., 2012)

13. Tabellenverzeichnis

Maximilian Armin Hetzel – Lebenslauf

Email: mail@maxhetzel.de

Ausbildung

03/2007 bis voraussichtlich 06/2013	Diplomstudium Psychologie – Universität Wien Schwerpunkte: psychologische Diagnostik, klinische Psychologie Nebenfach: Sozialpsychologie Thema der Diplomarbeit: „Effekte des Itempoolpräsentationsformats: Exemplarische Analyse eines Big-Five-Selbstbeschreibungsfragebogens“
03/2010 bis 01/2011	Zertifizierung als Student Mentor für erstsemestrige Psychologiestudenten der Universität Wien
09/1995 bis 06/2006	Abitur – Ignaz-Taschner-Gymnasium Dachau Fächerschwerpunkt Mathematik & Englisch Nebenfächer: Biologie & Ethik Stufensprecher in 12. & 13. Klasse Klassensprecher in 11. Klasse

---

Berufserfahrung

12/2012 bis 04/2013	Freier Arbeitnehmer für Pabst Science Publishers Eigen-Evaluation des Impact Factors der Journals „Psychological Test and Assessment Modeling“ sowie „Applied Cardiopulmonary Pathophysiology“
11/2012	Leiter eines Einführungsseminars in die Statistiksoftware R sowie das Programmpaket eRm

02/2012 bis 03/2012	Praktikum an der neurologischen Station des Krankenhauses Hietzing, Wien Tätigkeiten: klinisch-psychologische sowie neuropsychologische Diagnostik, Rehabilitation & Training sowie problemzentrierte Patientengespräche im Rahmen des Praktikums besuchte Fortbildungen („Fit für die Praxis“): Phytopharmazie Der neurologische Status Depression im Kindes- und Jugendalter
12/2010 bis 03/2011	Freier Arbeitnehmer für Prof. Dr. Klaus D. Kubinger Rohübersetzer des Buches „Statistics in Psychology“ von Rasch, D., Kubinger, K.D., Yanagida, T. (2012). West Sussex: John Wiley & Sons. (vom Deutschen ins Englische)
2008 bis aktuell	Statistik-Nachhilfe für Studierende

Nebentätigkeiten

08/2011 bis 02/2013	Übersetzer des Buches „Kuen Kuits – die „Essenz“ des Ving Tsun Kung Fu“ von Ulrich Stauner (vom Deutschen ins Englische)
10/2011 bis aktuell	United Chocolates Tätigkeit: Verkauf
10/2010 bis 04/2013	Mitarbeiter des selbstverwalteten Studentenlokals der Bewohner des Studentenheims Gasometer Tätigkeit: Barkeeper
12/2010 bis aktuell	Mitglied der studentischen Heimvertretung des Studentenheims Gasometer Stellvertretender Kassierer Schriftführer Leitung Fitnessraum

Sommersemesterferien 2010 & 2009, sowie 11/2006 bis 02/2007	Autoliv ServiceParts Feldgeding Tätigkeit: Lagerist, Lagerverwaltung, Versand
Sommersemesterferien 2008	MAN Dachau Tätigkeit: Ferienkraft & Archivierungsarbeiten
Sommersemesterferien 2007	Do&Co Cateringservice Tätigkeit: Cateringmitarbeiter bei den Wiener Filmfestwochen

---

Kenntnisse und Fähigkeiten

Fremdsprachen

- Englisch – fließend in Wort und Schrift
- Französisch – Grundkenntnisse

EDV

- Microsoft Office – Basiskenntnisse, fortgeschrittene Kenntnisse in Microsoft Word
- SPSS – fortgeschrittene Kenntnisse
- R – fortgeschrittene Kenntnisse im Anwendungsbereich

Führerschein

- PKW

Interessen & Freizeitbeschäftigungen

Literatur, Film & Musik

Sportliche Aktivitäten

- Kraftsport, Leichtathletik, Laufen, Schwimmen
- Ving Tsun Kung Fu („Ulrich Stauner/Gary Lam Level One“-zertifiziert), Aikido, Aikiken, Tae-Kwon-Do