



universität
wien

Diplomarbeit

Titel der Arbeit

**Zum Einfluss verschiedener Antwortformate bei Tests
zum schlussfolgernden Denken**

Verfasserin

Carmen Pils

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, im April 2013

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

Danksagung

In erster Linie möchte ich mich bei meiner Diplomarbeitsbetreuerin Mag. Dr. Christine Hohensinn für ihre geduldige Unterstützung bedanken, ohne ihre Hilfe und Tipps wäre diese Arbeit nicht zustande gekommen.

Mag. Herbert Poinstingl möchte ich für die rasche und unkomplizierte Zurverfügungstellung der von mir verwendeten Verfahren danken.

Mein besonderer Dank gilt auch den teilnehmenden Schulen, speziell den Direktoren, LehrerInnen und SchülerInnen, die mir die für diese Arbeit notwendigen Testungen erst ermöglichten.

In weiterer Folge gebührt mein Dank in großem Maße natürlich meinen Eltern, Irene und Erwin Pils, da ohne sie mein Studium nicht möglich gewesen wäre. Ich bedanke mich auch für ihren langen Atem, ihre Nachsicht und ihre Ratschläge, diese waren mir eine große Hilfe. Sie haben mir geholfen, dahin zu kommen, wo ich heute bin.

Ein großes Dankeschön auch an meine Großeltern, Franziska und Erwin Pils, dafür, dass sie während meiner Ausbildung immer darauf geachtet haben, dass es mir an nichts fehlt, und für ihr stetiges Interesse an meinen Fortschritten.

Dankbar bin ich auch für die tägliche moralische Unterstützung meines Partners Mag. Wolfgang Böhm, der mich des Öfteren vorm Verzweifeln und Aufgeben bewahrt hat. Danke auch für die höchst kritischen Korrekturlesearbeiten.

Vielen Dank auch an meine StudienkollegInnen, die mir durch ihre Ratschläge und fachliche Kompetenz halfen, so manche Hürde aus der Welt zu schaffen und immer ein offenes Ohr für meine Fragen und Probleme hatten.

Abstract

Die bisherige Literatur zum Thema, ob verschiedene Antwortformate einen Einfluss auf die Itemschwierigkeit haben, kam zu dem Konsens, dass dieser Einfluss existiert; nämlich, dass Multiple-Choice Items meist leichter zu beantworten sind als das offene Antwortformat (vgl. z.B. Kubinger, Holocher-Ertl, Reif, Hohensinn & Frebort, 2010). Die bis zu dieser Arbeit vorliegende Literatur ging bei ihren Untersuchungen meist von Wissenstests oder speziellen Leistungstest aus. In der vorliegenden Untersuchung ging es darum, diese Einflüsse auch für die Fähigkeit zum schlussfolgernden Denken, die in den meisten Intelligenztestbatterien ebenfalls erhoben wird, nachzuweisen. Dazu wurden ein Test zum numerischen und ein Test zum verbalen logischen Schlussfolgern 208 SchülerInnen an höheren (berufsbildenden) Schulen vorgegeben. Die Ergebnisse wiesen jedoch in eine andere Richtung als die vorangegangene Literatur vermuten ließ. Es zeigten sich keine Unterschiede in der Schwierigkeit von Items mit offenem, Multiple-Choice oder Multiplen Evaluation Format.

Abstract

The previous literature on the topic if miscellaneous response formats have an effect on the item difficulty have come to the consensus on the existence of such influence; namely that multiple choice items can mostly be answered easier than constructed response items (cf. e.g. Kubinger, Holocher-Ertl, Reif, Hohensinn & Frebort, 2010). The literature so far is based for the most part on knowledge tests or special performance tests in the course of its studies. The present analysis tried to verify if these influences are true for the reasoning ability as well, which is also surveyed in most intelligence test batteries. For this purpose 208 pupils at secondary (vocational) schools were provided with a test for numeric reasoning and a test for verbal reasoning. The results, however, pointed in a different direction than the previous literature would have led one to assume. It turned out that there are no differences in the difficulty of items with constructed response format, multiple choice format or multiple evaluation format.

Inhaltsverzeichnis

	Seite
Danksagung	1
Abstract	3
I. Einleitung	9
II. Theoretischer Teil	11
1 Antwortformate	13
1.1 Offenes Antwortformat	13
1.2 Multiple-Choice Antwortformat	15
1.3 Multiple Evaluation	16
2 Einflüsse von Item- und Testgestaltung	18
2.1 Einflüsse von verschiedenen Antwortformaten.....	18
2.2 Rateeffekt	24
<i>Exkurs: Rasch-Modell</i>	<i>26</i>
2.3 Einflüsse von Positionseffekten	28
3 Reasoning	31
III. Empirischer Teil	35
4 Ziel der Untersuchung und Hypothesen	37

5	Methode	38
5.1	Untersuchungsplan	38
5.2	Erhebungsinstrumente	38
5.2.1	<i>Zahlenreihentest ZART</i>	39
5.2.2	<i>Family-Relations-Reasoning-Test FRRT</i>	40
5.2.3	<i>Testformen</i>	42
5.3	Durchführung der Untersuchung	42
5.4	Stichprobe	44
6	Ergebnisse	46
6.1	Auswertung des ZART	49
6.2	Auswertung des FRRT	59
6.3	Auswertung des Einflusses der Impulsivität	70
7	Diskussion und Ausblick	73
8	Zusammenfassung	76
IV.	Literaturverzeichnis	77
V.	Anhang	83
	Antrag Landesschulrat	83
	Elternbrief	86
	Diagramme der Regression von Impulsivität und relativen Lösungshäufigkeiten	88
VI.	Tabellenverzeichnis	94
VII.	Abbildungsverzeichnis	96
Lebenslauf		100

I. Einleitung

In der bisherigen Literatur wurde des Öfteren ein Schwierigkeitsunterschied zwischen Items mit verschiedenen Antwortformaten festgestellt (siehe Abschnitt 2.1). Diese wurden zum Beispiel für Wissenstests – wie Deutschtests (Hohensinn & Kubinger, 2011), Computerwissenschaften- und Chemietests (Thyssen, Wainer & Wang, 1994) oder Tests zum psychologisch-diagnostischen Wissen (siehe z.B.: Kubinger & Gottschall, 2007) – und Leistungstests – wie Tests zum räumlichen Vorstellungsvermögen und der Arbeitseffizienz (siehe Teubenbacher, 2009) – bereits nachgewiesen. Jedoch liegen noch keine Befunde für die in jedem Intelligenztest erfasste Fähigkeit des schlussfolgernden Denkens oder Reasoning (siehe Abschnitt 3) vor. Genau diese Lücke soll die vorliegende Arbeit abdecken. Sie versucht somit zu überprüfen, ob auch für diesen Bereich vom Einfluss verschiedener Antwortformate ausgegangen werden kann, da die bisherigen Ergebnisse nicht ohne Weiteres auf das schlussfolgernde Denken übertragbar sind.

Dazu wird im Folgenden auf die in dieser Arbeit verwendeten Antwortformate (Abschnitt 1) näher eingegangen um danach die bisherige Literatur zum Thema (Abschnitt 2) genauer beleuchten zu können. Auch wird in Abschnitt 3 auf die Fähigkeit zum schlussfolgernden Denken eingegangen.

Im Anschluss erfolgt eine Beschreibung der durchgeführten Untersuchung (Abschnitt 5) und der Auswertung (Abschnitt 6).

Zum Abschluss werden die so gewonnenen Erkenntnisse unter Bezugnahme auf die beschriebene Literatur interpretiert und eingeordnet (Abschnitt 7).

II. Theoretischer Teil

1 Antwortformate

Laut Rost (2004) besteht ein Item aus zwei Komponenten: dem Itemstamm und dem Antwortformat. Der Stamm eines Items kann aus einer Frage, einer Aussage, einem Bild und vielem anderen bestehen, das das gewünschte Testverhalten einer Person auslöst. Die Wahl des idealen Antwortformates für die Entwicklung eines Tests sollte eine wohl überlegte und nicht zu unterschätzende Entscheidung sein. Sie sollte niemals nur aufgrund wirtschaftlicher Überlegungen getroffen werden, sondern hauptsächlich aufgrund inhaltlicher Aspekte erfolgen (Kubinger, 2006). Neben der schon erwähnten Ökonomie hat die Wahl des Antwortformates auch einen beträchtlichen Einfluss auf die Objektivität eines Verfahrens (Seiwald, 2003 a), auf welche in den folgenden Abschnitten noch näher eingegangen wird. So gilt es abzuwägen, bei welchem Antwortformat in einer speziellen Testsituation die Vorzüge gegenüber den jeweiligen Nachteilen überwiegen.

Es gibt fast unendlich viele Formen von Antwortformaten, die sich jedoch alle darauf reduzieren lassen, ob sie in freier (siehe Abschnitt 1.1) oder gebundener Form (siehe Abschnitte 1.2 und 1.3) vorgegeben werden. Diese beiden Formen unterscheidet, dass im Gegensatz zum gebundenen Format beim freien Antwortformat keine Antwortmöglichkeiten vorgegeben werden. Um einen genaueren Einblick in die in dieser Arbeit verwendeten Antwortformate zu geben, werden diese in den folgenden Abschnitten genauer beschrieben.

1.1 *Offenes Antwortformat*

Das freie oder offene Antwortformat (OA) ist das intuitiv am leichtesten verständliche und aufgrund der Einfachheit der Konstruktion – neben dem Multiple-Choice Antwortformat – wohl auch das bekannteste aller Antwortformate. Hier bekommt die Testperson eine Frage gestellt, auf die sie selbständig und ohne vorgegebene Antwortmöglichkeiten antworten muss. Laut Kubinger (2006) kann mit dem OA meist mehr diagnostisch relevante Information gewonnen werden als es beispielsweise mit dem Multiple-Choice Format der Fall ist. Des Weiteren kann das OA bei Leistungstests auch dazu eingesetzt werden, um die Wahrscheinlichkeit die richtige Antwort zu erraten, zu verringern (Rost, 2004; siehe auch Abschnitt 2.2 Rateeffekt).

Auch offene Antwortformate erfordern ein Format, in dem der Testperson vorgegeben wird, in welcher Weise sie die geforderte Reaktion setzen soll. Dieses Format kann zum Beispiel das Malen eines Bildes oder das Ergänzen eines Satzes sein (Rost, 2004). Laut Rost (2004) existieren drei Arten des offenen Antwortformates:

- (1) Eine Art ist dadurch gekennzeichnet, dass – außer der Angabe des Mediums – so gut wie *keine weiteren Vorgaben* gemacht werden. D.h. die Person bekommt ein weißes Blatt Papier hingelegt mit dem Auftrag, z.B. die Mitglieder ihrer Familie als Tiere zu zeichnen (Familie-in-Tieren-Test).
- (2) Ein zweiter Typ freier Antwortformate macht eine *formale Vorgabe* für die Produktion des Verhaltens, wie z.B. ein Wort aufzuschreiben, einen Satz zu formulieren, genau drei Dinge zu nennen, so viele Antworten wie möglich zu produzieren und diese so schnell wie möglich aufzuschreiben usw.
- (3) Ein dritter Typ freier Antwortformate macht eine sogenannte *Lückenvorgabe*, d.h. die erfragte Itemantwort soll eine Leerstelle im vorgegebenen Itemstamm ausfüllen. Dies ist z.B. der Fall, wenn die Aufgabe darin besteht, ein unvollständiges Bild oder einen Satz zu ergänzen oder Geschichten oder vorgegebene Muster fortzusetzen. (S. 60)

Im Zuge der vorliegenden Arbeit wird der zweite Typ der freien Antwortformate, also die formale Vorgabe, verwendet. Wie diese im Speziellen aussieht, wird in den Abschnitten 5.2.1 und 5.2.2 genauer ausgeführt.

Ein großes Problem beim OA ist die gewünschte Verrechnungssicherheit, das heißt die Objektivität beim Auswerten eines Tests, die natürlich unter nicht ausreichend definierten Antwortkategorien leidet. Drastisch wird die Auswertungssituation, wenn neben einfachen richtig-/falsch-Entscheidungen auch richtige Teilantworten honoriert werden sollen, da es hier im Ermessen des Testleiters / der Testleiterin liegt, wie er/sie mit solchen Teilantworten umgeht (Kubinger, 2006).

Dem Problem der Bewertung der Antworten begegnet der AID 2 (Adaptives Intelligenz Diagnostikum – Version 2.1, Kubinger & Wurst, 2000), welcher ausschließlich das OA verwendet, indem er den TestleiterInnen ein umfangreiches

Kategoriensystem an die Hand gibt, um subjektiven Bewertungen entgegenzuwirken. Trotz des nicht zu unterschätzenden Problems der Verrechnungssicherheit benutzen viele Individualverfahren das OA, weil die hier erhaltenen Antworten meist diagnostisch aufschlussreicher sind als andere Formate. Daraus folgt aber auch, dass diese Verfahren weniger verrechnungssicher sind als zum Beispiel Multiple-Choice Tests und daher intensiver eingeschulte AuswerterInnen verlangen (Kubinger, 2006). Neben Individualverfahren verwenden vor allem Projektive Verfahren und Kreativitätstests das offene Antwortformat; somit ist auch hier die Auswertungsobjektivität schwieriger zu bewerkstelligen als dies bei Tests der Fall ist, die das Multiple-Choice Format verwenden (Seiwald, 2003 a).

1.2 Multiple-Choice Antwortformat

Beim Multiple-Choice Antwortformat (MC) werden dem/der TestnehmerIn mehrere Antworten vorgelegt und er/sie muss sich für eine oder auch mehrere Möglichkeiten entscheiden, die er/sie für richtig hält. Häufig findet man das MC Format „1 aus 5“, welches aus einer Lösung und vier sogenannten Distraktoren, also falschen Antwortmöglichkeiten, besteht und auch in dieser Arbeit verwendet wird (siehe Abschnitte 5.2.1 und 5.2.2). Als Vorteil dieses Formates lässt sich seine Verrechnungssicherheit verbuchen, da beim Auswerten nur danach beurteilt wird, welche Möglichkeiten von der Testperson angekreuzt wurden und daher auch kein Entscheidungsspielraum seitens der BeurteilerInnen entsteht. Weiters besteht die Möglichkeit Tests mithilfe des MC Formates als Gruppenverfahren zu konzipieren. Ein zusätzlicher Grund für die Häufigkeit dieses Formates ist wohl auch die Einfachheit in dessen Auswertung, die unter Umständen sogar mittels Computer erfolgen kann, wenn man die Antwortbögen „scannt“ (Seiwald, 2003 a), wobei bei handschriftlichen Tests auch die Auswertung mittels Schablone durchaus praktikabel sein kann.

Neben diesen Vorteilen muss jedoch darauf hingewiesen werden, dass die Entscheidung für oder gegen ein bestimmtes Antwortformat immer erst aus inhaltlicher Sicht erfolgen muss (Kubinger, 2006), da zum Beispiel bei Wissensfragen das Wiedererkennen von Inhalten, wie es beim MC Format geschieht, deutlich leichter fällt als das selbständige Reproduzieren von Wissen beim offenen Format (vgl. z.B. Zimbardo & Gerring, 1999).

Als größter Nachteil des Multiple-Choice Antwortformates muss der – im Gegensatz zum offenen Format – erhöhte Rateeffekt (siehe Abschnitt 2.2) genannt werden (Kubinger, 2006).

Ein weiterer Nachteil entsteht durch die möglichen unterschiedlichen Herangehensweisen beim Lösen von MC Items, welche die gewünschte Eindimensionalität eines Tests gefährden können. Laut Kubinger (2006) wären hier drei Herangehensweisen an die Lösung eines Items zu erwähnen: 1) die „Produktions“-Strategie, bei der erst in Gedanken die Lösung produziert und anschließend die gedachte Lösung unter den Antwortmöglichkeiten gesucht wird. Als zweite Strategie wäre die 2) „Versuch-und-Irrtum“-Strategie aufzuzeigen; hier wird jede Antwortmöglichkeit der Reihe nach auf ihre Wahrscheinlichkeit hin untersucht bis die Plausibelste entdeckt wird. Bei der letzten Strategie, der 3) „Falsifikations“-Strategie, werden die augenscheinlich unwahrscheinlichsten Möglichkeiten sofort aussortiert, um nur noch die Übrigen genauer zu untersuchen.

1.3 Multiple Evaluation

Laut Musch (2009) wird das Multiple-Choice Antwortformat (siehe Abschnitt 1.2) vor allem in der Wissens- und Leistungsdiagnostik häufig verwendet, was jedoch bei einer schlichten Auswertung in richtige und falsche Antworten zu wenigen diagnostisch nutzbaren Informationen führt. Außerdem zwingt das MC Format die Testpersonen bei nicht vorhandenem oder Teilwissen zu raten, was im in Abschnitt 2.2 beschriebenen Rateeffekt resultiert. Daher schlägt Musch (2009) als ideales Antwortformat die Multiple Evaluation vor, bei der die Testperson für jede Antwortmöglichkeit angibt, wie hoch sie die Wahrscheinlichkeit einschätzt, dass besagte Möglichkeit richtig ist. Hier drängt sich jedoch die Frage nach der besten Verrechnungsmöglichkeit auf, da die richtig-/falsch-Verrechnung in vielen Fällen nicht möglich und auch nicht erwünscht ist (Jenner, 2012). Eine Möglichkeit würde die lineare Auswertung darstellen, bei der der Item-Score dem Wert der Wahrscheinlichkeit entspricht, die die Testperson auf die Lösung gesetzt hat (Rippey, 1970). Allerdings erzielt man bei dieser Art der Verrechnung dann den höchsten Score, wenn man immer 100 Prozent auf die am wahrscheinlichsten erscheinende Möglichkeit setzt, was wiederum zu einer Art MC Format führen würde. Daher schlägt Dirkwager (2003) die logarithmische Verrechnung für die Multiple

Evaluation vor. Auch hier werden nur die Wahrscheinlichkeiten, die auf die Lösung gesetzt wurden, berücksichtigt; wenn diese jedoch zu gering ausfallen, wirkt sich dies negativ auf den Score aus. Laut Jenner (2012) haben Persönlichkeitseigenschaften einen Einfluss darauf, von welcher Verrechnungsmöglichkeit man eher profitiert. So schnitten weniger risikobereite Personen bei logarithmischer Auswertung schlechter ab, wohingegen Frauen und antwortsichere Testpersonen von dieser Auswertungsart eher profitierten. In der vorliegenden Arbeit wird jedoch lediglich bewertet, ob die TeilnehmerInnen richtig oder falsch geantwortet haben, um die Vergleichbarkeit zwischen den Antwortformaten zu gewährleisten.

Die Multiple Evaluation setzt allerdings voraus, dass die Getesteten fähig sind, ihr Wissen realistisch einzuschätzen. Dies ist jedoch nicht immer der Fall, da es Personen gibt, die ihr Wissen systematisch über- oder unterschätzen. Nach Dirkzwager (2003) kann diesem Phänomen mit ausreichender Einarbeitungszeit in das (für die meisten Personen) neue Format und der Vorgabe am Computer – mit unmittelbarer Feedbackmöglichkeit – entgegengesteuert werden.

2 Einflüsse von Item- und Testgestaltung

2.1 Einflüsse von verschiedenen Antwortformaten

Bereits im Jahr 1941 untersuchte Cronbach Unterschiede zwischen verschiedenen Antwortformaten. Er verglich ein Multiple-Choice Format (MC), bei dem nur alle richtigen Möglichkeiten angekreuzt werden mussten, mit einem Multiple-True/False Format, bei welchem für jede Antwortmöglichkeit angegeben werden musste, ob sie richtig oder falsch ist. Cronbach (1941) kam bei diesem Vergleich zu dem Schluss, dass zwischen den beiden Formaten kaum Unterschiede in der Leistung der getesteten StudentInnen bestanden, dass jedoch MC eine höhere Reliabilität aufwies und einfacher zu bewerten war.

Im Gegensatz dazu beschäftigten sich Thissen, Wainer und Wang (1994) mit der Frage, ob es zulässig sei, Scores aus Tests mit MC und offenem Antwortformat (OA) einfach zu verbinden oder, ob diese Formate doch unterschiedliche Fähigkeitsdimensionen messen. Sie überprüften diese Fragen mittels Wissenstests zur Computerwissenschaft und Chemie und kamen zu dem Schluss, dass zumindest bei diesen Tests die verglichenen Antwortformate die selbe Dimension erfassten, was auch Hohensinn und Kubinger (2011) zeigen konnten. Rodriguez (2003) beschäftigte sich ebenfalls mit der Frage, ob Items mit offenem und Multiple-Choice Format in einem Test das selbe Konstrukt erfassen. Er führte eine Metaanalyse über 29 Studien mit insgesamt 56 Korrelationen durch und kam zu dem Ergebnis, dass stammgleiche Items höher korrelieren als nicht-stammgleiche Items. Vor allem wenn sie mit der Intention erstellt wurden, das selbe Konstrukt zu erfassen, waren die Korrelationen hoch. Bei der Frage nach der Kombinierbarkeit der beiden Formate kommt Rodriguez (2003) so zu dem Schluss, dass dies für eben zu diesem Zweck designte stammgleiche Items der Fall ist.

Die Gegenüberstellung von Multiple-Choice Tests mit unterschiedlich vielen Antwortmöglichkeiten erbrachten unter anderen Delgado und Prieto (1998), da sie überprüfen wollten, ob tatsächlich drei Antwortoptionen für die meisten Getesteten am besten zu bearbeiten sind (Lord, 1944). Sie kamen nach einem Vergleich von drei und vier Antwortoptionen zu dem Schluss, dass die geringere Optionenzahl ebenso funktional sei wie die höhere und daher drei Antwortoptionen vorzuziehen seien, da diese Form weniger Zeit und Aufwand beansprucht.

Eine Meta-Analyse zu ebendiesem Thema, wie viele Antwortalternativen praktikabel wären, führte Rodriguez (2005) durch. Eine vorläufige Antwort darauf gaben Haladyna und Downing (1989), die die Meinung vertraten, dass es das Beste wäre so viele funktionale Distraktoren wie möglich zu erstellen. Um diese Frage nun genauer zu beantworten, analysierte Rodriguez (2005) 27 Studien aus einer Zeitspanne von 1925 bis 1999 zu diesem Thema genauer, die insgesamt zwei bis fünf Antwortalternativen untersuchten. Er kam, ebenso wie Delgado und Prieto (1998), zu dem Schluss, dass drei Alternativen für die meisten Settings als ideal zu bezeichnen sind. Die Reduktion von fünf auf drei Antwortalternativen ging mit nur geringen Schwierigkeitseinbußen einher, je nachdem wie sinnvoll die eliminierten Distraktoren waren. Weiters betont der Autor, dass die meisten zugrundeliegenden Studien eine geringere Alternativenzahl befürworteten und, dass der Rateeffekt (siehe Abschnitt 2.2) kein zwingendes Argument für mehr Auswahlmöglichkeiten darstellt, da in den meisten Fällen nicht blind geraten wird und Testpersonen von sich aus die unwahrscheinlicheren Alternativen ausschließen.

Zu dem selben Ergebnis kommen auch Shizuko, Takeuchi, Yashima und Yoshizawa (2006), die die Unterschiede in Studieneingangstests für Englisch an einer japanischen Universität mit drei und vier Antwortoptionen untersuchten. Der Test mit „1 aus 4“ Multiple-Choice Items wurde so administriert, dass nach Ausschluss des unattraktivsten Distraktors Items mit drei Antwortmöglichkeiten resultierten, wobei 10 Items mit vier Optionen im zweiten Test verblieben, um die Vergleichbarkeit zu gewährleisten. Nach einer Analyse basierend auf dem Rasch-Modell kamen die Autoren zu dem Ergebnis, dass in beiden Tests die Itemschwierigkeiten annähernd gleich waren, wobei sogar die Items mit einer Lösung und zwei Distraktoren tendenziell schwieriger erschienen. Dieser Umstand stützt die vorgenannten Befunde und resultiert in einem geringeren Arbeitsaufwand für die Testentwickler.

Den oben zitierten Schlüssen, dass Multiple-Choice Items mit drei Antwortmöglichkeiten zu präferieren seien, widerspricht jedoch zum Beispiel die Arbeit von Kubinger und Gottschall (2007), welche zu dem Ergebnis kamen, dass selbst Multiple-Choice „1 aus 6“ aufgrund der zu hohen Ratewahrscheinlichkeit vermieden werden sollte. Weiters erwähnten sie, dass aufgrund der noch höheren Wahrscheinlichkeit das richtige Ergebnis zu erraten, die MC Formate „1 aus 5“ und „1 aus 4“ vollkommen zu verwerfen seien. Diese Arbeit hatte das Ziel, Itemschwierigkeiten in Abhängigkeit von verschiedenen Antwortformaten zu

analysieren. Sie gaben 152 PsychologiestudentInnen und 21 praktizierenden DiagnostikerInnen den Diagnostik Info Check '05 (Kubinger, 2006) vor, der aus 105 Items bestand. Bei diesen Items wurden die Antwortformate so variiert, dass bei 45 dieser Aufgaben Itempaare mit unterschiedlichen Antwortformaten entstanden, die miteinander verglichen werden konnten. Die vorgelegten Formate waren ein freies Antwortformat (mit Antwortkatalog für die AuswerterInnen), ein MC Format „1 aus 6“ und ein MC Format „x aus 5“ (bei dem keine bis alle Antworten richtig sein konnten). Die restlichen 60 Items wurden allen TestnehmerInnen in gleicher Weise vorgelegt. Nach einigen Itemeliminierungen konnte nachgewiesen werden, dass das Rasch-Modell (siehe Abschnitt 2.2) auf die Daten passte. Die Ergebnisse ihrer Analysen waren, dass das freie Antwortformat tendenziell schwieriger war als das MC Format „1 aus 6“, sich jedoch das freie Format nicht signifikant vom MC Format „x aus 5“ in seiner Schwierigkeit unterschied. Der Vergleich der Multiple-Choice Formate „1 aus 6“ und „x aus 5“ zeigte allerdings, dass ersteres deutlich leichter war. Obwohl die Anzahl der Testpersonen und der Items eher gering war, sahen die Autoren doch einen nicht unerheblichen Einfluss von Antwortformaten auf die Itemschwierigkeit. Kubinger und Gottschall (2007) schlossen aus ihren Ergebnissen, dass, wie schon erwähnt, das MC Format „1 aus 6“ vermieden werden und, wo die Verwendung eines freien Formates nicht möglich sei, auf das MC Format „x aus 5“ zurückgegriffen werden sollte.

Auch die Ergebnisse von Hohensinn und Kubinger (2011) bestätigen einen Einfluss von verschiedenen Antwortformaten auf die Itemschwierigkeit. Die Autoren verglichen die MC Formate „1 aus 6“ und „2 aus 5“ mit einem offenen Antwortformat anhand eines Deutschtests für die achte Schulstufe, das von ExpertInnen entwickelt wurde und stammgleiche Items enthielt. Es wurden insgesamt sechs Booklets mit Fragen erstellt, wobei darauf geachtet wurde, dass die zu vergleichenden Items in besagten Booklets immer an der gleichen Stelle platziert wurden, um eventuelle Itempositionseffekte (mehr dazu siehe Abschnitt 2.3) zu vermeiden. Nach der Vorgabe an 2 285 SchülerInnen kamen sie zu dem Ergebnis, dass das offene Format und MC „2 aus 5“ schwieriger waren als das MC Format „1 aus 6“. Daraus schlossen Hohensinn und Kubinger (2011), dass das Multiple-Choice Format „2 aus 5“ eine gute Alternative für das offene Format sei.

Zum selben Ergebnis wie die vorab erwähnten Studien kommen auch Kubinger, Holocher-Ertl, Reif, Hohensinn & Frebort (2010), die in einem Large-Scale-Assessment insgesamt ca. 14 000 Schülerinnen der achte Schulstufe einen

Mathematiktest vorgaben. Auch sie verglichen die von Hohensinn und Kubinger (2011) verwendeten Formate und kamen zu dem Schluss, dass das MC Format „1 aus 6“ leichter als das MC Format „2 aus 5“ und sehr viel leichter als das freie Antwortformat zu beantworten war. Das MC Format „2 aus 5“ und das freie Format unterschieden sich jedoch nicht signifikant, was die Autoren ebenfalls zu dem Fazit veranlasste, dass, wo ein freies Format nicht realisierbar sei, auf das MC Format „2 aus 5“ zurückgegriffen werden sollte.

In diesem Sinne beschäftigte sich auch Vock (2009) in ihrer Diplomarbeit mit einem Vergleich der Itemschwierigkeiten der MC Formate „1 aus 4“, „1 aus 6“, „x aus 5“ und einem offenen Antwortformat. Es zeigte sich, dass sich das OA und das MC Format „x aus 5“ nicht signifikant unterscheiden, diese beiden Formate jedoch deutlich schwieriger waren als die MC Formate „1 aus 4“ und „1 aus 6“.

In einer ebenfalls im Jahr 2009 publizierten Meta-Analyse über insgesamt 56 Datenquellen aus 37 Studien zum Thema Testformateffekte bei Tests zum Lesen und Hören beim Erlernen einer Fremdsprache (In´nami & Koizumi, 2009), kamen die Autoren ebenso zu dem Schluss, dass die Multiple-Choice Formate durchwegs einfacher zu beantworten waren als offene Fragen.

Zu diesem Ergebnis kamen auch Currie und Chiramanee (2010), die den Einfluss von Multiple-Choice Formaten auf die Messung von Kenntnis der Sprachstruktur untersuchten. Sie gaben 186 thailändischen StudentInnen einen Englischtest erst in offenem Format vor und verwendeten die meistgenannten falschen Antworten als Distraktoren für Multiple-Choice Items. Auch diese stammgleichen Items wurden den selben StudentInnen vorgegeben, wobei Multiple-Choice „1 aus 3“, „1 aus 4“ und „1 aus 5“ zum Einsatz kamen. Abschließend wurde in einem Fragebogen erhoben, aus welchen Gründen die TeilnehmerInnen in beiden Tests unterschiedliche Antworten gaben. Es stellte sich heraus, dass die Tests mit Multiple-Choice Items höhere Scores produzierten als die Tests mit offenen Fragen. Auch zeigte sich, dass fast 74 Prozent der Antworten in beiden Tests unterschiedlich waren. Die Hauptursache hierfür ist dem Umstand zuzuschreiben, dass die StudentInnen im ersten Test mit offenen Fragen diese falsch beantwortet hatten und im zweiten Test ihre Fehler durch die gegebenen Antwortoptionen zu korrigieren vermochten. Weiters konnten die Autoren feststellen, dass die Zahl der Möglichkeiten bei den Multiple-Choice Items kaum einen Einfluss auf das Ergebnis der Tests hatte. Vor allem zwischen vier und fünf Optionen bestand kaum ein Unterschied, jedoch erschienen die „1 aus 3“ Multiple-Choice Items deutlich einfacher als die beiden vorgenannten. Die Autoren schlossen

aus diesen Ergebnissen, dass für Fremdsprachentests auf unterem Niveau Multiple-Choice Tests nicht geeignet sind, da sie nicht das tatsächliche Können der Getesteten widerspiegeln.

Einen etwas anderen Vergleich von Multiple-Choice und Items mit offenem Antwortformat traten Liu, Lee und Linn (2011) an. Sie ließen 794 SchülerInnen der sechsten und siebenten Schulstufe einen Wissenschaftstest mit Multiple-Choice Items „1 aus 4“ bearbeiten, wobei die Testpersonen in Folgeitems, die entweder in offenem Format oder Multiple-Choice „1 aus 6“ gehalten waren, ihre vorher getätigte Wahl erklären mussten. Obwohl diese Folgeitems den gleichen Itemstamm besaßen, zeigten sich Unterschiede in deren Schwierigkeit. Items mit offenem Antwortformat erschienen schwerer als solche mit dem Format Multiple-Choice. Wobei sich die Befürchtung, dass Folgeitems mit MC Format Schlüssel für die Beantwortung der eigentlichen Fragen beinhalten könnten, nicht bestätigte, da beide Gruppen bei den Testfragen gleich gut abschnitten.

Die Frage des Einflusses von ähnlichen Distraktoren auf die Itemschwierigkeit versuchten Ascalon, Meyers, Davis und Smits (2007) zu beantworten, indem sie 493 freiwilligen SchülerInnen eine Art Führerscheinprüfung vorgaben. So konnten sie belegen, dass ähnliche Distraktoren zu einer erhöhten Schwierigkeit von Items beitragen. Weiters betont Revuelta (2004), dass Items mit Distraktoren von ähnlicher Schwierigkeit mehr über die getesteten Personen aussagen als solche mit größerer Varietät.

Mit einem anderen Aspekt von Multiple-Choice Tests beschäftigten sich Huang, Trevison und Storfer (2007). Sie untersuchten den Einfluss von „all-of-the-above“-Antwortmöglichkeiten (AOTA), die implizieren, dass alle bis dahin genannten Optionen richtig sind. Dudycha und Carpenter (1973) waren der Ansicht, dass der Einsatz dieser Möglichkeit Items schwieriger macht und es somit besser zwischen guten und schlechten TestnehmerInnen zu differenzieren vermag. Jedoch geben einige Autoren zu bedenken, dass, wenn eine Antwortoption ausgeschlossen werden kann, dies auch für AOTA zutrifft, was es für erfahrene Testpersonen deutlich einfacher macht, solch ein Item zu lösen (zum Beispiel Haladyna, Downing & Rodriguez, 2002). Um diesen Zusammenhängen genauer auf den Grund zu gehen, untersuchten die Autoren 457 CollegestudentInnen anhand von drei Testformen eines Studienleistungstests. Jede dieser Testformen enthielt 10 Items, wobei in der ersten Testform AOTA die richtige Antwortmöglichkeit darstellte, in der zweiten

wurde sie als Distraktor vorgegeben und in der dritten Testform war sie entweder die Lösung oder ein Distraktor. Das Fazit von Huang, Trevison und Storfer (2007) war, dass Testpersonen mit höheren Fähigkeiten bei Items mit AOTA besser abschnitten als Personen mit niedrigen Fähigkeiten. Außerdem konnten sie beobachten, dass die Schwierigkeit für die Testpersonen höher war, wenn die AOTA Option die Lösung des Items darstellte und niedriger wurde wenn AOTA als Distraktor fungierte.

Itemschwierigkeiten können sich allerdings auch durch Eigenschaften der TeilnehmerInnen verändern. So zeigten Bielinski und Davison (2001) über Mathematikleistungstests von drei amerikanischen Studien zum Schulerfolg, die aus Multiple-Choice und offenen Fragen bestanden, dass eine Interaktion zwischen dem Geschlecht der TeilnehmerInnen und der Itemschwierigkeit besteht. Die getesteten Mädchen schnitten bei einfacheren Items besser ab als männliche Teilnehmer, hingegen zeigten Jungen bei schwierigeren Aufgaben bessere Leistungen. Taylor und Lee (2012) untersuchten anhand eines Mathematik- und Lesetests des Washington Assessment of Student Learning SchülerInnen in der vierten, siebenten und zehnten Schulstufe in drei aufeinanderfolgenden Jahren mit Gruppengrößen von 61 458 bis 75 985 Personen. Sie kamen zu dem Schluss, dass bei beiden Tests von den nicht-geschlechtsneutralen Items die Multiple-Choice Items für männliche Testpersonen einfacher waren als für ihre weiblichen Pendants, wobei es sich für Items mit offenem Antwortformat in gegensätzlicher Weise verhielt. Untersuchungen mit dem selben Datensatz konnten jedoch auch einen Zusammenhang zwischen Itemschwierigkeit bei verschiedenen Antwortformaten und der Herkunft der TeilnehmerInnen zeigen (Taylor & Lee, 2011). Die TeilnehmerInnen deklarierten selbst ihre ethnische Zugehörigkeit und es wurden die Daten aus dem Lesetest zur Analyse herangezogen. Die Ergebnisse zeigten, dass weiße Testpersonen mit Multiple-Choice Fragen besser zurechtkamen und Personen mit anderer ethnischer Zugehörigkeit bei offenen Fragen besser abschnitten.

Auch motivationale Unterschiede in den Testpersonen können sich auf die Itemschwierigkeit verschiedener Antwortformate auswirken. Diesen Beweis trat Solheim (2011) an, indem sie 217 Schülerinnen der fünften Schulstufe einen Leseverständnistest, mit in einer ersten Form 20 Multiple-Choice Items und in einer zweiten Form 20 offenen Fragen, vorgab. Sie konnte aus ihren Ergebnissen folgern, dass nur für SchülerInnen mit niedriger Lesemotivation Multiple-Choice Items schwieriger erschienen als kurze offene Fragen.

Wenn es darum geht, nicht die Fähigkeit einer Person zu erfassen, sondern mithilfe von verschiedenen Antwortformaten Lernaufgaben zu erstellen, stellt sich auch die Frage, ob dafür das Multiple-Choice oder das offene Format besser geeignet ist. Diesem Umstand widmete sich Clariana (2003) in einer Studie, an der 46 StudentInnen teilnahmen, um den Effekt von verschiedenen Formaten auf computergestütztes Lernen zu erfassen. Es zeigte sich eine größere Lernleistung bei Aufgaben mit offenem Format als bei Multiple-Choice Aufgaben, was der Autor darauf zurückführt, dass ersteres den/die SchülerIn eher anregt über die richtige Antwort nachzudenken und sich somit der Lerninhalt besser im Gedächtnis verankert.

2.2 Rateeffekt

Der Rateeffekt ist ein oft unterschätzter Einflussfaktor auf die Lösungswahrscheinlichkeit eines (Multiple-Choice) Items. Bei nicht vorhandener Fähigkeit hat ein Item eine Lösungswahrscheinlichkeit beziehungsweise eine a priori-Ratewahrscheinlichkeit von $1/k$, wobei k die Anzahl der Antwortmöglichkeiten darstellt. Daher erhöht sich die Lösungswahrscheinlichkeit eines Items, je weniger Antwortoptionen geboten werden. Liegt beispielsweise ein MC Format „1 aus 5“ vor, hat ein entsprechendes Item eine a priori-Ratewahrscheinlichkeit von $1/5 = 20$ Prozent. Im Falle von drei Antwortoptionen, so wie es zum Beispiel Delgado und Prieto (1998) forderten, hätte die Testperson sogar eine Chance von 33 Prozent die Lösung zu erraten. Das Problem des Rateeffektes verstärkt sich natürlich noch, wenn der/die TestnehmerIn über zumindest geringe Fähigkeiten verfügt und so einzelne Antwortmöglichkeiten eliminieren kann (Kubinger, 2006).

Um die a priori-Ratewahrscheinlichkeit zu verringern, könnte man die Anzahl der Distraktoren anheben. Allerdings muss man hierbei aufpassen, die Kapazitäten der Testpersonen bezüglich Konzentration, Merkfähigkeit und auch Leistungsmotivation nicht zu überreizen (Kubinger, 2006). Des Weiteren ist auch darauf zu achten, nur sinnvolle Distraktoren zu verwenden, da sehr unwahrscheinliche Antwortmöglichkeiten von den meisten Testpersonen sofort ausgeschlossen werden können (Lord, 1944).

Sollten laut Kubinger (2006) Zweifel bestehen, dem Rateeffekt in testtheoretischer Weise beikommen zu können, bleibt nur die Möglichkeit (eventuell neue) formal-

gestalterische Mittel einzusetzen. Eine Möglichkeit den Rateeffekt gering zu halten, wäre die Erhöhung der richtigen Antwortmöglichkeiten (zum Beispiel MC Format „2 aus 5“ oder aber „x aus 5“) oder die Einführung des sequentiellen Antwortformates, bei dem für jede Antwortmöglichkeit sofort bestimmt werden muss, ob sie richtig ist oder nicht, ohne jedoch die weiteren Möglichkeiten zu kennen. Außerdem besteht die Möglichkeit, die Testpersonen durch spezielle Antwortoptionen („Ich weiß die Lösung nicht“ oder „Keine Antwort ist richtig“) oder gezielte Instruktionen vom Raten abzuhalten (Kubinger, 2006).

In diesem Sinne konnten Prieto und Delgado (1999) zeigen, dass verschiedene Instruktionen einen nicht zu vernachlässigenden Beitrag zum Rateeffekt leisten. Nachdem sie vier Instruktionssituationen durch die Ergebnisse der ansonsten identischen Tests miteinander verglichen, kamen sie zu dem Ergebnis, dass es aus formal-gestalterischer und ethischer Sicht am besten wäre, Testpersonen dahingehend zu informieren, dass Raten generell nicht erwünscht ist und ausgelassene Items kleine Belohnungen nach sich ziehen. Allerdings betonen die Autoren auch, dass es leider keine Instruktion gibt, die für jede Testperson perfekt ist.

Dem Rateeffekt durch verschiedene Auswertungsmethoden für Multiple-Choice Items entgegenzuwirken versuchten Scharf und Baldwin (2007), indem sie drei Methoden verglichen. Als erstes wurden keine Strafpunkte vergeben, sondern nur die richtig markierten Lösungen berücksichtigt. Bei der zweiten Methode wurden die fälschlicher Weise als richtig gekennzeichneten Optionen bestraft und die dritte Methode bestand in maximaler Bestrafung, indem zusätzlich nicht markierte Lösungen geahndet wurden. Schlussendlich entschieden sich die Autoren für die zweite Methode als die fairste, um Raten zu bestrafen und Negativpunkte ausschließlich für als richtig markierte Distraktoren zu vergeben.

Auch die Item Response Theorie (IRT) hat einige Zugänge zu bieten, um dem Rateeffekt Herr zu werden. Zum einen kann über das 3-PL-Modell (Birnbaum, 1968) neben Personen-Fähigkeitsparameter, Item-Schwierigkeitsparameter und Item-Diskriminationsparameter auch der Item-Rateparameter geschätzt werden. Analog dazu kann auch das Difficulty-Plus-Guessing-Modell (Kubinger & Draxler, 2006) verwendet werden, das jedoch ohne einen Item-Diskriminationsparameter auskommt.

Erwähnenswert sind in diesem Zusammenhang auch die Person-Fit-Indices (Ponocny & Klauer, 2002) für Modelle der IRT, vor allem auch das 1-PL-Modell von Rasch (1960), die es ermöglichen, ratende Personen zu identifizieren.

Exkurs: Rasch-Modell

Das 1-PL-Modell, oder Rasch-Modell, beschäftigt sich grundsätzlich mit der Wahrscheinlichkeit, dass eine Testperson v ein Item i löst („+“). Diese Wahrscheinlichkeit ist abhängig von einem Personenparameter ξ_v , welcher die wahre Fähigkeit von Person v abbildet, und von einem Itemparameter σ_i , der wahren Schwierigkeit von Item i . Das heißt, dass mithilfe dieses Modells die Fähigkeit und die Schwierigkeit eindimensional gemessen werden können. Des Weiteren macht es keinen Unterschied an welcher Stelle in einem Test der Person ein Item vorgegeben wird, da ihre Leistung bei diesem Item unabhängig davon ist, welche Aufgaben sie zuvor bearbeitet hat oder noch bearbeiten wird („lokal stochastische Unabhängigkeit“). Dem Rasch-Modell liegt eine „logistische“ Wahrscheinlichkeitsfunktion mit den additiv zusammengesetzten Parametern ξ und σ zugrunde:

$$P(+ | \xi_v, \sigma_i) = \frac{e^{\xi_v - \sigma_i}}{1 + e^{\xi_v - \sigma_i}}$$

Diesen Zusammenhang kann man auch grafisch über sogenannte Item Characteristic Curves (ICC) darstellen (Abbildung 1). Aufgespannt wird diese Kurve über die Ausprägung der latenten Dimension (x-Achse) und die Lösungswahrscheinlichkeit des Items in Abhängigkeit des Personenparameters (y-Achse). Wenn die ICCs zweier Items parallel verlaufen, also nur entlang der x-Achse verschoben sind, sind sie mit dem Rasch-Modell konform (Kubinger, 2003 a), da sie sich nur über ihre Schwierigkeit voneinander unterscheiden.

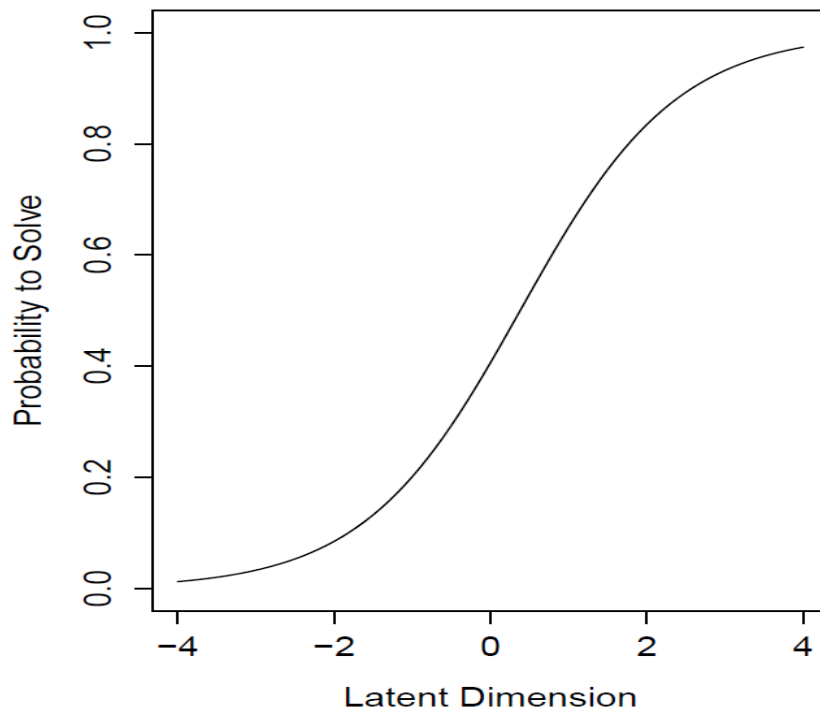


Abbildung 1: Beispiel einer Item Characteristic Curve

Da das Rasch-Modell stichprobenunabhängig ist, kann ein entsprechender Modelltest abgeleitet werden, was wiederum bedeutet, dass es nie ungeprüft vorausgesetzt werden muss (vergleiche dazu zum Beispiel Kubinger, 2003 b).

Daraus folgt auch der große Wert des 1-PL-Modells für adaptive Testverfahren, da in diesen die Reihenfolge der Vorgabe von Items nicht fixiert ist (siehe auch Abschnitt 2.3). Aber auch für konventionelle Verfahren, deren Testwert durch eine einfache Auszählung der gelösten Items ohne Gewichtung zusammengesetzt ist, muss für eine faire Verrechnung das Rasch-Modell gelten (Kubinger, 2003 b).

Das 3-PL-Modell, oder Birnbaum-Modell, erweitert das 1-PL-Modell durch einen Diskriminationsparameter α_i und einen Rateparameter. Ersterer geht auf den Umstand ein, dass die Lösungswahrscheinlichkeiten von zwei Testpersonen mit gleicher Fähigkeit bei einem Item fast gleich, bei einem anderen aber verschieden sein können. Das heißt, die Differenz $\xi_v - \sigma_i$ wird durch α_i gewichtet. Zweiterer trägt der Tatsache Rechnung, dass zum Beispiel bei Multiple-Choice Items die Lösung auch erraten werden kann, wodurch sich die Lösungswahrscheinlichkeit erhöhen würde (Kubinger, 2003 b).

Die vorhin schon erwähnten Person-Fit-Indices überprüfen nun, ob das Antwortmuster zu der erbrachten Leistung passt. So wäre es hier zum Beispiel auffällig, wenn eine Testperson leichte Aufgaben nicht lösen kann, jedoch bei schwierigen eine gute Leistung zeigt. Im Rasch-Modell besitzt jede Person eine bestimmte Wahrscheinlichkeit, ein Item mit bekannter Schwierigkeit zu lösen; diese Wahrscheinlichkeit ist abhängig von der Fähigkeit dieser Person, also von ihrem Personenparameter. Der Person-Fit-Index ist somit die Wahrscheinlichkeit, dass eine Person mit bestimmter Fähigkeit ein bestimmtes Antwortmuster produziert (Rost, 2004). Wenn nun dieses Muster nicht Rasch-Modell-konform ist, kann man davon ausgehen, dass der Score der Person keine gute Abbildung der Fähigkeiten der Testperson darstellt. Ein Grund dafür kann sein, dass die Person bei einigen Items glücklich geraten hat. Somit kann man den Person-Fit-Index zur Identifizierung von ratenden TestnehmerInnen heranziehen (Meijer, 2003). Teubenbacher (2009) konnte so mittels Person-Fit-Indices ratende Personen identifizieren und damit zeigen, dass diese ein gutes Maß für abweichende Antwortmuster sind.

Bei diesen IRT-Ansätzen handelt es sich um Möglichkeiten, die dem Rateeffekt a posteriori begegnen – im Gegensatz zu den zuvor genannten Itemgestaltungsmöglichkeiten, die den Rateeffekt a priori beeinflussen.

2.3 Einflüsse von Positionseffekten

Laut Seiwald (2003 b) sind die Einflüsse der Position eines Items innerhalb eines Tests oder Fragebogens auf dessen Beantwortung als Positionseffekte zu bezeichnen. Damit muss man vor allem zu Beginn und am Ende eines Tests rechnen, da speziell am Anfang eines Verfahrens mit mangelndem Instruktionsverständnis gerechnet werden muss und es am Ende vermehrt zu Ermüdungserscheinungen kommen kann (Rost, 2004).

Itempositionseffekte wurden vorrangig anhand von Interviews oder Fragebögen untersucht. Nur wenige Untersuchungen beschäftigen sich mit der Aufgabenreihung bei Leistungstests und es existieren sehr wenige Studien zur idealen Reihung von Untertests innerhalb von Testbatterien (Khorramdel & Frebort, 2011). Für konventionelle Tests stellt eine Nichtbeachtung dieser Effekte kein allzu großes Problem dar, da jede Testperson die Items in der selben Reihenfolge präsentiert bekommt. Jedoch ist ein Negieren von Positionseffekten bei adaptiven Testverfahren

kaum akzeptabel, denn sollte sich herausstellen, dass ein Item an verschiedenen Stellen im Test schwieriger oder leichter zu lösen ist, ist dieses Vorgehen schlicht unfair (Kubinger, Reif & Yanagida, 2011). An solchen Positionseffekten zu erwähnen wären hier vor allem Lern- und Ermüdungseffekte. Erstere führen zu höheren Testscores oder auch zu geringeren Itemschwierigkeiten zum Testende hin (siehe z.B. Kubinger, 2009), wohingegen die Zweitgenannten längere Reaktionszeiten nach sich ziehen (Földényi et al, 1999).

Kubinger, Reif und Yanagida (2011) konnten zumindest für den AID 2 (Adaptives Intelligenz Diagnostikum – Version 2.1, Kubinger & Wurst, 2000) keine dementsprechenden Itempositionseffekte feststellen. Sie folgerten trotzdem, dass jeder adaptive Test vor seiner Freigabe auf etwaige Positionseffekte getestet werden sollte, um diese im besten Fall ausschließen oder aber ihnen entgegenwirken zu können.

Zu der Schlussfolgerung, dass entsprechende Voruntersuchungen vor Veröffentlichung eines Tests unerlässlich sind, kommen auch Meyers, Miller und Way (2008). Sie untersuchten ein spezielles Testdesign anhand von Daten eines Large-Scale-Assessments, in dem zu Beginn und Ende eher leichte und in der Mitte des Tests eher schwierige Items positioniert waren. Es zeigte sich, dass sich in besagtem Design die Positionseffekte aufhoben. Weiters schlossen die Autoren aus ihren Untersuchungen, dass schwierige Items eine höhere Schwierigkeit erreichen, je näher sie zum Ende des Tests lokalisiert werden. Im Gegensatz dazu werden einfache Items leichter, je weiter sie am Beginn eines Tests stehen.

Auch für Large-Scale-Assessments stellen Itempositionseffekte ein nicht zu unterschätzendes Problem dar, da hier meist verschiedene Testzusammenstellungen mit den selben Items an verschiedenen Positionen verwendet werden. In diesen Fällen wäre es fatal, wenn sich durch Müdigkeit oder eventuelle Lerneffekte eine Veränderung der Itemschwierigkeiten ergeben würde (Hohensinn, Kubinger, Reif, Schleicher & Khorramdel, 2011). So zeigte sich bei Mathematiktests ohne Zeitlimit kein Effekt, wohingegen mit Zeitlimit ein Effekt zum Tragen kam (zum Beispiel Flaugher, Melton & Myer, 1968). Auch bei den Mathematiktests im Zuge der PISA-Studie (Programme for International Student Assessment 2000, OECD, 2002; zitiert nach Hohensinn, Kubinger & Reif et al, 2011) zeigten sich solche Positionseffekte. Es stellte sich heraus, dass Items am Ende schwieriger erschienen als auf einer Position weiter vorne im Test, was entweder dem Zeitlimit oder der zunehmenden Müdigkeit

der TestnehmerInnen zuzuschreiben ist. In der Studie zu Itempositionseffekten bei Large-Scale-Assessments von Hohensinn, Kubinger & Reif et al (2011) zeigten sich für den verwendeten Mathematiktest für die vierte Schulstufe jedoch keine nennenswerten Positionseffekte.

Eine Studie von Schweizer, Schreiner und Gold (2009), die als Erhebungsgrundlage einen Matrizen test zur Feststellung von figuralem Reasoning verwendete, kam zu dem Schluss, dass besagter Test nicht, wie postuliert eindimensional, sondern zweidimensional misst. Neben der eigentlich zu erhebenden Dimension der fluiden Intelligenz beziehungsweise des Reasoning erfasste der Test auch eine Positionsdimension, die eventuell auf einen Lerneffekt zurückzuführen ist.

Ebenfalls mithilfe eines Matrizen tests, dem Wiener Matrizen Test (WMT; Formann & Piswanger, 1979), versuchte Hahne (2008) eventuellen Itempositionseffekten auf die Spur zu kommen, konnte jedoch für diesen Test keine dementsprechenden Effekte finden.

Skinner (1999) untersuchte ebenfalls Itempositionseffekte und ging dabei von der Alltagsweisheit aus, dass, wenn zu Beginn eines Tests leichte Fragen vorgegeben werden, sich die Leistung in diesem erhöht, da die TestnehmerInnen so gleich auf Erfolgserlebnisse aufbauen können. Seine Ergebnisse ließen aber eher darauf schließen, dass es sich umgekehrt verhält, nämlich, dass vorangestellte schwere Items die Leistung bei den folgenden leichten Items erhöhen.

Itempositionseffekten bei Persönlichkeitsfragebögen versuchte Ortner (2004) auf den Grund zu gehen: Sie gab jungen Männern, die zum Militärdienst einberufen worden waren, den gleichen Fragebogen einmal in normaler und einmal in verkehrter Reihenfolge vor. Dabei stellte sich heraus, dass sich bei drei der sieben vorgegebenen Skalen die Itemschwierigkeiten bei verkehrter Vorgabereihenfolge veränderten. Daraus könnte man folgern, auch bei der Erstellung eines Persönlichkeitsfragebogens nicht auf die Betrachtung eventueller Itempositionseffekte zu verzichten.

3 Reasoning

Reasoning oder schlussfolgerndes Denken bezeichnet „die Fähigkeit, Gesetzmäßigkeiten oder logisch zwingende Zusammenhänge erkennen und zweckentsprechend verwerten zu können“ (Kubinger, 2006, S. 198).

Den Stellenwert dieser Fähigkeit für einen allgemeinen Intelligenzbegriff lässt allein schon der Umstand erkennen, dass Thurstone (1938, zitiert nach Amelang & Bartussek, 2001) sie zu einer seiner sieben „Primary Mental Abilities“ zählte:

- **verbal comprehension:** Kenntnis von Wörtern und ihrer Bedeutung sowie deren angemessene Verwendung im Gespräch.
- **word fluency:** Rasches Produzieren von Wörtern, die bestimmten strukturellen oder symbolischen Erfordernissen entsprechen.
- **number:** Geschwindigkeit und Präzision bei einfachen arithmetischen Aufgaben.
- **space:** Bewältigung von Aufgaben, die räumliches Vorstellen und Orientieren sowie das Erkennen von Objekten unter anderem Bezugswinkel erfordern [...].
- **memory (associative):** Behalten paarweise gelernter Assoziationen. ...
- **perceptual speed:** Geschwindigkeit beim Vergleich oder der Identifikation visueller Konfigurationen. ...
- **induction oder reasoning, general:** Schlussfolgerndes Denken im Sinne des Auffindens einer allgemeinen Regel in einer vorgegebenen Abfolge von Zahlen oder Symbolen und Anwendung derselben bei der Vorhersage des nächstfolgenden Elements.
(S.208-209)

Diese sieben Fähigkeiten sollten ursprünglich als gleichberechtigte Bausteine der Intelligenz nebeneinander stehen und somit unabhängig voneinander sein. Es fanden sich jedoch trotzdem Korrelationen zwischen den Gruppenfaktoren was einige Forscher (u.a. Eysenck, 1979; zitiert nach Amelang & Bartussek, 2001) zu dem Schluss kommen ließ, dass es doch einen g-Faktor beziehungsweise einen Faktor für Allgemeine Intelligenz gibt. Da sich dieser g-Faktor unmittelbar aus den „Primary Mental Abilities“ zusammensetzt, trägt Reasoning ganz entscheidend dazu bei.

Auch in der Theorie der fluiden und kristallinen Intelligenz (Horn & Cattell, 1966) spielt Reasoning eine nicht unbedeutende Rolle. So trägt es entscheidend zur fluiden Intelligenz bei, die „Basisprozesse des Denkens und anderer mentaler Aktivitäten“ (Süß, 2003 a, S. 219) betrifft und meist über Aufgaben zum induktiven oder deduktiven Denken erfasst wird. Die fluide Intelligenz dient, zusammen mit geeigneten Lerngelegenheiten, der kristallinen Intelligenz als Basis. Diese wird als Fähigkeit bezeichnet, mittels erworbenen Wissens Probleme zu lösen.

Auch die Three Stratum Theory (Carroll, 1993) beinhaltet das Konzept Reasoning. Diese Theorie resultierte aus einer Reanalyse aller bis dahin in der Literatur veröffentlichten Korrelationsmatrizen und fand auf erster Ebene Evidenzen für einen allgemeinen Intelligenzfaktor, allerdings konnten auch Faktoren auf einer zweiten Ebene bestätigt werden. Unter diesen findet sich wieder die fluide Intelligenz Cattells, die, wie oben beschrieben, Reasoning beinhaltet und sich aus generellem, verbalem, quantitativem, syllogistischem Reasoning, Induktion und der Fähigkeit zu Klassifizieren zusammensetzt, was Reasoning auch für diese Theorie zu einem unverzichtbaren Bestandteil macht.

Durch den Umstand, dass Reasoning für die meisten gängigen Intelligenztheorien eine basale Fähigkeit darstellt, die wichtig für die Intelligenz ist, wird diese in den meisten Intelligenztests oder Intelligenztestbatterien erfasst. Diese Erfassung kann unterschiedliche Formen annehmen, wie zum Beispiel die Untertests *Analogien*, *Zahlenreihen* und *Matrizen* des Intelligenz-Struktur-Test 2000 R (Amthauer, Brocke, Liepmann & Beauducel, 2001). Im Untertest *Analogien* werden der Testperson drei Begriffe vorgegeben, wobei zwischen den ersten beiden ein bestimmtes Verhältnis besteht. Zum dritten gegebenen Begriff soll aus fünf Antwortmöglichkeiten diejenige gewählt werden, die in einem ähnlichen Verhältnis zu diesem steht. Die *Zahlenreihen* werden mithilfe von Regeln gebildet, die die Testperson richtig weiterführen soll, um die nachfolgende Zahl zu bestimmen. Auch hier stehen fünf Auswahlmöglichkeiten zur Verfügung. Die *Matrizen* sind, ebenso wie die Zahlenreihen, auf Regeln aufgebaut, jedoch handelt es sich hierbei um Anordnungen von einfachen bis komplexen Figuren, welche der/die TestnehmerIn in korrekter Weise und unter Zuhilfenahme der fünf gegebenen Möglichkeiten fortsetzen muss.

Es existieren jedoch auch eigenständige Tests zum Reasoning, wie der Test *Syllogismen* (Srp, 1994), bei dem die Aufgabe darin besteht, aus zwei Prämissen auf eine Konklusion zu schließen, und der vor allem auf den Bereich des verbalen Reasoning abzielt. Weiters wäre auch *Rechnen in Symbolen* (Schmotzer, Kubinger &

Maryschka, 1994) zu erwähnen, welcher das schlussfolgernde Denken im numerischen Bereich erfasst. Hier wurden die Ziffern einer Rechnung durch Symbole ersetzt und die Testperson ist angehalten, einem dieser Symbole eine Zahl zwischen null und neun zuzuordnen. Aber auch die *Standard Progressive Matrices* (Raven, Court, Raven & Kratzmeier, 1990) müssen hier erwähnt werden. Dabei werden die Testpersonen aufgefordert, Sets von Matrizen mithilfe von Antwortvorgaben zu vervollständigen.

Bei Tests zum schlussfolgernden Denken kann leicht auf sprachliches Material verzichtet werden (siehe *Zahlenreihen*, *Rechnen in Symbolen* oder *Standard Progressive Matrices*) und daher stehen Reasoningtests oft in Zusammenhang mit sogenannten Culture fair-Tests, die den Anspruch erheben kulturunabhängig zu sein (Kubinger, 2006), da sie nicht auf früheren Lernerfahrungen beruhen (siehe zum Beispiel Süß, 2003 b).

Trotz der unbestrittenen Wichtigkeit von Tests zum Thema Reasoning verwenden die meisten und zumindest alle vorgenannten Verfahren das Multiple-Choice Antwortformat, was wahrscheinlich auch an den in Abschnitt 1.2 genannten Vorzügen dieses Formates liegen dürfte. Allerdings stellt sich die Frage, ob das MC Format wirklich die Methode der Wahl für Reasoningtests sein sollte, zumal dieses Format mit dem nicht unerheblichen Nachteil des erhöhten Rateeffektes einhergeht. Vielleicht wäre ein offenes Antwortformat (Abschnitt 1.1) oder aber die Multiple Evaluation (Abschnitt 1.3) für diese Art von Verfahren besser geeignet. Da auf diese Frage bis jetzt keine zufriedenstellende Antwort gegeben werden konnte, versucht nun die vorliegende Arbeit dieses Problem genauer zu beleuchten.

Die Schlussfolgerungen, die aus der vorliegenden Theorie gezogen werden können, sind, dass offene Formate durchweg schwieriger zu beantworten sind als Multiple-Choice Aufgaben und, dass in dieser Beziehung zur Multiplen Evaluation noch keine Aussagen gemacht werden können. Beim Multiple-Choice Format ist zusätzlich auf den erhöhten Rateeffekt zu achten und daher auch auf eine ausgewogene Kombination aus guten Distraktoren und deren Anzahl.

Eventuelle Itempositionseffekte kann man nur bei Rasch-Modell-Konformität ausschließen oder sie bei gleicher Vorgabereihenfolge der Items relativieren.

Des Weiteren kann man feststellen, dass es bezüglich des idealen Antwortformates für Reasoningtests noch keine Erkenntnisse gibt.

III. Empirischer Teil

4 Ziel der Untersuchung und Hypothese

Das vorrangige Ziel der vorliegenden Untersuchung ist es zu überprüfen, ob sich die bisherigen Ergebnisse der Literatur auch auf den Bereich des schlussfolgernden Denkens übertragen lassen und, wo sich die Itemschwierigkeiten der Multiplen Evaluation einordnen lassen.

Aufgrund der bisherigen Befunde ist von einem Schwierigkeitsunterschied zwischen Items mit verschiedenen Antwortformaten auszugehen. Daraus resultiert folgende Hypothese:

Hypothese 1: Die Schwierigkeitsgrade der Tests zum schlussfolgernden Denken unterscheiden sich signifikant bei verschiedenen Antwortformaten.

$$H_0: \mu_{\text{offen}} = \mu_{\text{me}} = \mu_{\text{mc}}$$

$$H_1: \mu_{\text{offen}} \neq \mu_{\text{me}} \neq \mu_{\text{mc}}$$

Als Nebenhypothese soll außerdem der Einfluss von Impulsivität auf die Testleistung in den beiden Tests untersucht werden, da, wie in Abschnitt 2.1 gezeigt, auch Persönlichkeitseigenschaften Auswirkungen auf die Itemschwierigkeit haben können. Um das Modell zu vervollständigen, werden das Antwortformat und auch die Interaktion von Impulsivität und Antwortformat als Prädiktoren hinzugezogen.

Hypothese 2: Die Impulsivität (Imp) einer Person hat einen linearen Einfluss auf die Testleistungen (y) in Abhängigkeit des Antwortformates (AWF).

$$H_0: Y \neq \text{Imp} + \text{AWF} + \text{Imp} * \text{AWF} + \varepsilon$$

$$H_1: Y = \text{Imp} + \text{AWF} + \text{Imp} * \text{AWF} + \varepsilon$$

5 Methode

5.1 Untersuchungsplan

Der Vergleich der Antwortformate sollte mittels dreier randomisiert zugewiesener Gruppen erfolgen, die zwar die selben Fragen, jedoch unterschiedliche Antwortformate enthalten sollten. Diese sollten ein offenes Format, ein Multiple-Choice Format „1 aus 5“ und eine Multiple Evaluation mit fünf Antwortmöglichkeiten sein. Geplant war, SchülerInnen von Gymnasien und berufsbildenden höheren Schulen zwei Tests zum schlussfolgernden Denken vorzulegen, die diese unterschiedlichen Antwortformate enthalten. Die Testung sollte in einer Unterrichtsstunde erfolgen, um den Schulalltag nicht zu sehr zu stören.

Angedacht war, im Anschluss eine Rasch-Modell Analyse (Rasch, 1960), inklusive Likelihood Ratio Test (LRT; Andersen, 1973), Graphischem Modelltest (Rasch, 1960) und Wald Test (Glas & Verhelst, 1995), durchzuführen. Darauf aufbauend sollte über die resultierenden Itemparameter eine Varianzanalyse berechnet werden.

Für diese Berechnungen sollte unter anderem das Package „eRm: extended Rasch modeling“ (Mair, Hatzinger & Maier, 2011) verwendet werden, welches im Statistikprogramm „R“ (Version 2.13.2, 2011; R Development Core Team, 2008) implementiert ist. Zusätzlich sollte mit dem Statistikprogramm „SPSS 13.0“ gearbeitet werden.

5.2 Erhebungsinstrumente

Den Testpersonen wurden der Zahlenreihentest ZART (siehe Abschnitt 5.2.1) und der Family-Relations-Reasoning-Test FRRT (siehe Abschnitt 5.2.2) vorgelegt. Zusätzlich wurden das Geschlecht, das Alter, die Schule und die Klasse erfragt. Darüber hinaus wurde mittels dreier Items, die auf einer vierstufigen Skala beantwortet werden sollten (siehe Tabelle 1), versucht, die Ausprägung der Impulsivität zu erfragen; wobei die möglichen Werte in einer Spanne von 3 bis 12 Punkten liegen konnten.

Tabelle 1: *Fragen zur Feststellung der Impulsivität*

Bitte bewerten Sie die folgenden Aussagen, je nachdem wie sehr diese auf Sie zutreffen.

Frage	Trifft nicht zu	Trifft kaum zu	Trifft meist zu	Trifft immer zu
Wenn ich Fragen beantworten muss, antworte ich meist spontan.				
Ich überlege mir immer sehr genau, was ich sagen will.				
Ich kontrolliere alle meine Arbeiten gerne noch einmal nach.				

5.2.1 Zahlenreihentest ZART

Der Zahlenreihentest ZART (Poinstingl, 2009 a) ist ein psychologisches Verfahren, das die Fähigkeit des schlussfolgernden Denkens mithilfe von Zahlenreihen zu erheben versucht (Poinstingl, 2010). Die Items des ZART wurden mittels itemgenerierender Regeln konstruiert und bestehen aus einer Serie von Zahlen, die der/die TeilnehmerIn fortsetzen soll, indem er/sie die passenden Regeln extrahiert und diese richtig weiterführt (Poinstingl, 2009 a). Der Test umfasst ursprünglich 60 Multiple-Choice Items „1 aus 8“ (inklusive „Keine Antwort ist richtig!“ und „Ich weiß die Antwort nicht!“) und wird mittels Computer ohne Zeitlimit vorgegeben. Aufgrund der Studien des Testautors kann von der Gültigkeit des Rasch-Modells (Rasch, 1960) ausgegangen werden.

Im Falle der vorliegenden Arbeit wurden den Testpersonen 22 der ZART Items vorgelegt, die in 15 Minuten zu bearbeiten waren. Des Weiteren wurde der ZART in eine Paper-Pencil-Version überführt und die Antwortformate so verändert, dass Multiple-Choice „1 aus 5“, Multiple Evaluation mit fünf Antwortmöglichkeiten und ein offenes Format entstanden.

Beispiel-Item:

x) 1 3 (?) 7 9 11 13

Multiple-Choice-Antwortformat „1 aus 5“:

x) ☐ 4 ☐ 7 ☐ 5 ☐ 8 ☐ 6

Multiple Evaluation mit 5 Antwortmöglichkeiten:

x) % 4 % 7 % 5 % 8 % 6

offenes Antwortformat:

x)

5.2.2 Family-Relations-Reasoning-Test FRRT

Bei diesem verbalen Reasoning-Test von Poinstingl, Kubinger, Skoda und Schechtner (in Vorb.) ist es die Aufgabe der Testpersonen, nach dem Lesen einer kurzen Geschichte zu Verwandtschaftsverhältnissen in einer Familie, das Verhältnis zwischen zwei Mitgliedern dieser Familie zu bestimmen (Poinstingl, 2009 b). Das im Multiple-Choice gehaltene Antwortformat besteht aus acht Antwortmöglichkeiten mit einer Lösung und sieben Distraktoren. Laut Bösch (2011) wurden die Items aufgrund von drei Konstruktionsregeln erstellt: die erste Regel legt den Komplexitätsgrad der Familienbeziehungen fest (unterteilt in vier Gruppen, von der Kern- bis zur Patchworkfamilie), die zweite Regel beschäftigt sich mit der Gesamtzahl der genannten Verwandtschaftsverhältnisse. Die nachträglich eingefügte dritte Regel beschäftigt sich mit der Art der Verwandtschaft (direkte Verbindungen wie Mutter – Kind sind einfacher zu durchschauen als Verbindungen mit dreifacher Verknüpfung wie bei Cousins / Cousinen). In einem ersten Schritt wurden auf diesem Weg

ungefähr 100 Items generiert, wovon 64 einer Rasch-Modell Analyse und diversen Überarbeitungen standhielten.

Für die vorliegende Arbeit wurden 22 Items des FRRT verwendet, welche sich aus Gruppe 1 (Kernfamilie) und Gruppe 2 (Verwandtschaft 2. Grades) zusammensetzen. Für die Bearbeitung dieses Sets hatten die Testpersonen 20 Minuten Zeit. Wie schon beim ZART (siehe Abschnitt 5.2.1.) musste auch beim FRRT das Antwortformat verändert werden, um von einem Multiple-Choice Format „1 aus 8“ zu einem MC Format „1 aus 5“, einem Multiplen Evaluation Format mit fünf Antwortmöglichkeiten und einem offenen Format zu kommen.

Beispiel-Item:

*Klara ist die Tochter von Sonja. Sonja ist die Tochter von Maria.
Maria ist die (?) von Klara.*

Multiple-Choice-Antwortformat „1 aus 5“:

☐ Tante ☐ Cousine ☐ Großmutter ☐ Nichte ☐ Enkeltochter

Multiple Evaluation mit 5 Antwortmöglichkeiten:

_____%Tante _____%Cousine _____%Großmutter _____%Nichte _____%Enkeltochter

offenes Antwortformat:

Maria ist die _____ von Klara.

5.2.3 Testformen

Um Ermüdungseffekte bei ungewohnten Antwortformaten, vor allem aber bei der Multiplen Evaluation, so gering als möglich zu halten, wurden drei Testformen mit abwechselnden Formaten entwickelt (siehe Tabelle 2).

Tabelle 2: *Zusammenstellung der Testformen*

Testform	ZART	FRRT
A	Multiple-Choice „1 aus 5“	Offenes Antwortformat
B	Multiple Evaluation	Multiple-Choice „1 aus 5“
C	Offenes Antwortformat	Multiple Evaluation

Die gewünschte Vergleichbarkeit der Testformen wurde mit Hilfe von vier Linking – Items pro Test hergestellt, welche in jeder Testform im offenen Format vorgegeben wurden. Wie schon erwähnt, wurden die Testungen zeitlich beschränkt, wobei 15 Minuten für den ZART und 20 Minuten für den FRRT an reiner Testzeit eingeplant wurden. Durch diese Zeitbeschränkung sollten die Tests in einer Unterrichtsstunde abgehalten werden können.

Eventuelle Itempositionseffekte sollten sich auf ein Minimum reduzieren lassen, da in jeder Testform die Items an der gleichen Stelle positioniert waren und des Weiteren die Tests bereits einer Rasch-Modell-Analyse unterzogen wurden.

5.3 *Durchführung der Untersuchung*

Um die im Abschnitt 4 dargelegten Hypothesen zu überprüfen, wurde eine Stichprobengröße von 200 Schülerinnen und Schülern angedacht und zu diesem Zweck 16 Schulen in Niederösterreich (Raum St. Pölten und Krems) kontaktiert. Von diesen konnte, nach Einholen des Einverständnisses der DirektorInnen (per Telefongespräch), des niederösterreichischen Landesschulrates (mittels Antrag, siehe Anhang) und der betroffenen Eltern (mittels Elternbrief, siehe Anhang), an drei Schulen getestet werden. Der Erhebungszeitraum lag zwischen 16. und 25. April 2012. An jeder der drei Schulen wurden drei Klassen in aufeinanderfolgenden

Schulstunden getestet. Es handelte sich um je zwei 9., 11. und 12. Schulstufen sowie drei 10. Schulstufen. In jeder Klasse führten zwei TestleiterInnen die Untersuchung durch; wo möglich, waren auch die jeweiligen LehrerInnen im Raum. Um Schummeln vorzubeugen, erhielten die SchülerInnen die Testformen abwechselnd (A-B-C-A-B-C usw.). Somit wurde die Form A 71 mal, die Form B 69 mal und die Form C 68 mal vorgegeben. Trotz großzügiger Einteilung wurde in manchen Klassen wegen Unruhe zu Beginn und während der Instruktionen die Zeit knapp, was in einer etwas verkürzten Testzeit resultierte (17 statt 20 Minuten für den FRRT). Da dies jedoch schon in der ersten Klasse der Fall war, wurde die verkürzte Zeit für die folgenden Testungen übernommen und war somit für alle TestnehmerInnen gleich. Während der Testungen zeigte sich, dass Testpersonen, die die Multiple Evaluation als Antwortformat erhalten hatten, trotz Instruktion vermehrt Probleme mit deren Bearbeitung aufwiesen. Jedoch konnte diese Unsicherheit nach erneuter verbaler Instruktion beseitigt werden, was eventuell darauf schließen lässt, dass die schriftliche Instruktion nicht aufmerksam gelesen, nicht verstanden oder aber nicht umfassend genug ausgeführt wurde. Trotz dieser einzelnen Fälle der Verunsicherung bei dem für die meisten SchülerInnen neuen Format, stellte es für den Großteil der Getesteten kein Problem dar. Jedoch wurde bei der Auswertung ersichtlich, dass viele Testpersonen der von ihnen favorisierten Antwort einfach 100 % gaben und somit die Multiple Evaluation wieder wie ein normales Multiple-Choice Format funktionierte. Des Weiteren fiel die großteils sehr geringe Motivation der Testpersonen auf, die sich in der schon erwähnten Unruhe und der leichten Ablenkbarkeit zeigte. Zudem konnte bei der Auswertung der Testbögen eine hohe Zahl an nicht bearbeiteten Items, sogenannten Missings, beobachtet werden (siehe Abschnitt 6).

5.4 Stichprobe

Die Stichprobe setzte sich aus 208 SchülerInnen zusammen, davon waren 127 (61%) weiblichen und 81 (39%) männlichen Geschlechts. Die Geschlechteraufteilung in den Schulen kann man der untenstehenden Tabelle 3 entnehmen.

Tabelle 3: *Geschlechteraufteilung in den Schulen*

Geschlecht	Borg St. Pölten	HAK St. Pölten	HAK Krems	Prozentuale Verteilung
weiblich	46	46	35	61%
männlich	18	24	39	39%
Gesamt	64	70	74	100%

Zum Erhebungszeitpunkt besuchten 64 SchülerInnen das Bundesoberstufenrealgymnasium (BORG) St. Pölten, 70 SchülerInnen die Bundeshandelsakademie (HAK) St. Pölten und 74 SchülerInnen die Bundeshandelsakademie (HAK) Krems (Tabelle 3). Die Klassenstärken reichten von 15 bis 31 SchülerInnen.

Die TeilnehmerInnen lagen in einer Altersspanne von 14 bis 19 Jahren, die meisten SchülerInnen waren jedoch 15 bis 17 Jahre alt, wie man anhand der Abbildung 2 sehen kann. Der Altersmittelwert der TestteilnehmerInnen lag bei 16,21 Jahren.

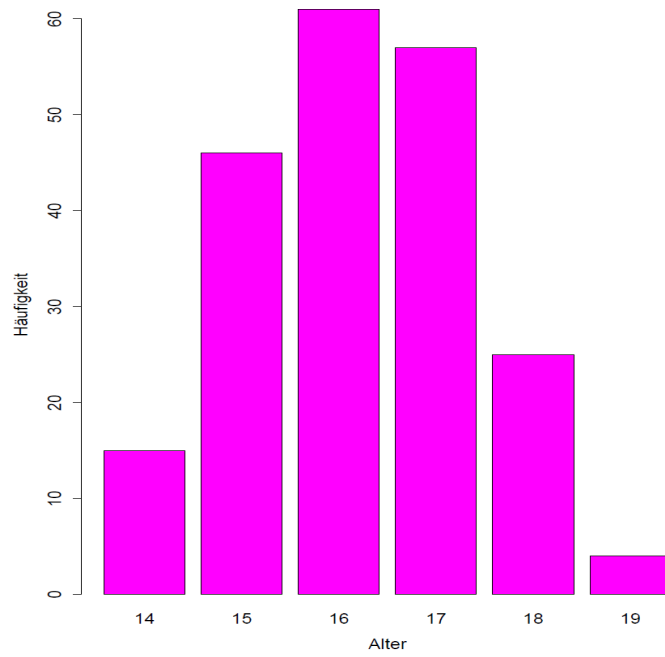


Abbildung 2: Altersverteilung der SchülerInnen

6 Ergebnisse

Hinsichtlich der Beobachtung, dass viele der Testpersonen, die die Multiple Evaluation bearbeitet hatten, einfach immer den von ihnen favorisierten Antwortalternativen 100 Prozent zuschrieben, konnte bei Betrachtung der tatsächlichen Daten dieser erste Eindruck bestätigt werden. 67,88 % (93 von 137 Personen) der TeilnehmerInnen folgten dieser Taktik insgesamt. Beim ZART, also bei Testform B, waren es geringfügig weniger (44 von 69 Personen, 63,77%) als beim FRRT (49 von 68 Personen, 72,06%), also bei Testform C, die die 100 Prozent Taktik anwandten (siehe Abbildung 3).

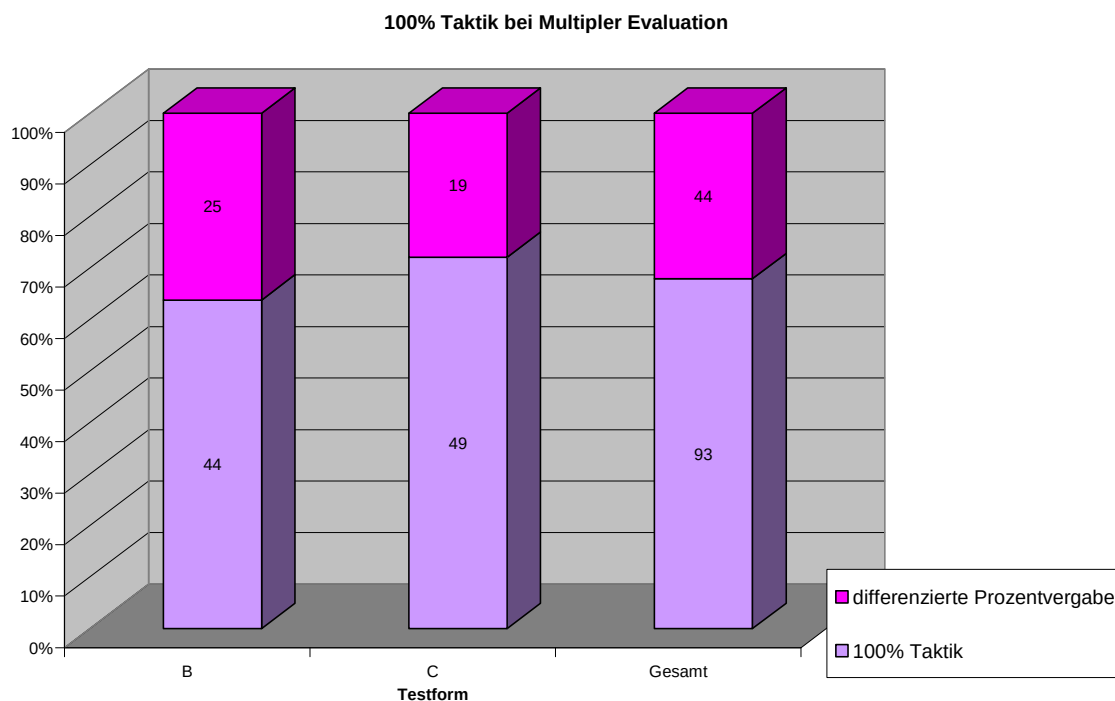


Abbildung 3: Prozentuale Verteilung der Antwortstile im Bezug auf die 100% Taktik bei Multipler Evaluation; in den Balken findet sich die Anzahl der beinhalteten Personen

Da die Zeit aufgrund der Testumstände zumindest für den FRRT gekürzt werden musste, interessierte im Folgenden, ob es die Testpersonen in der vorgegebenen Zeit schafften, die Tests zur Gänze zu absolvieren. Es zeigte sich, dass 47 % der Testpersonen im ZART auch das letzte Item bearbeiteten. Im FRRT waren es jedoch nur 43 % der Getesteten.

Die genaue prozentuale Aufteilung, wie viele Personen die beiden Tests hinsichtlich der unterschiedlichen Antwortformate zur Gänze bearbeitet haben, findet sich in Tabelle 4.

Tabelle 4: *Prozentuale Verteilung der Testpersonen hinsichtlich ihres Absolvierens der Tests und der einzelnen Antwortformate*

Antwortformat	ZART		FRRT	
	fertig	nicht fertig	fertig	nicht fertig
Offenes Format	52	48	42	58
Multiple-Choice	38	62	46	54
Multiple Evaluation	50	50	41	59
Gesamt	47	53	43	57

Die Ergebnisse der Tabelle 4 lassen darauf schließen, dass die Testpersonen unabhängig von Test und Antwortformat mehr Zeit zu ihrer Bearbeitung benötigt hätten.

In Folge genauerer Betrachtung der Missings, also der nicht bearbeiteten Items, wurde ersichtlich, dass die Testpersonen bei allen Antwortformaten gleichermaßen viele Items unbeantwortet ließen. So verzichteten sie auch beim MC Format auf die Möglichkeit des Ratens, was, wie schon in Abschnitt 5.3 erwähnt, auf die geringe Motivation der Testpersonen zurückgeführt werden muss.

Daher war es notwendig, diejenigen Personen, die weniger als die Hälfte der Items eines Tests bearbeitet hatten, also 11 Items pro Test, von den weiteren Berechnungen auszuschließen, um den Einflussfaktor Motivation etwas abzumildern. Ausgeschlossen wurden im ZART für die Form A 13, für die Form B 16 und für die Form C 12 Personen. Analog wurden im FRRT für die Form A 29, für die Form B 28 und für die Form C 30 Personen ausgeschlossen. Es resultierten demnach für die Analysen des ZART 167 und für den FRRT 121 Testpersonen.

Die für die Nebenfragestellung und als Teilungskriterium erhobenen Impulsivitätswerte der Testpersonen lagen in einer Spanne von 3 bis 11 Punkten und wiesen einen Mittelwert von 7,077 auf. Der Median lag bei 7 Punkten. Die Verteilung der Werte kann man Abbildung 4 entnehmen.

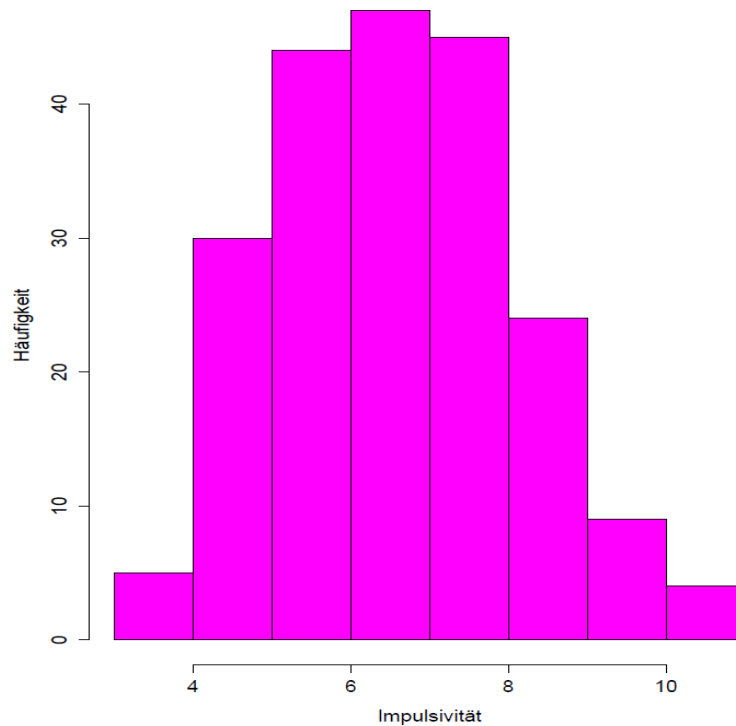


Abbildung 4: Verteilung der Impulsivitätswerte der Testpersonen

Für das offene und das MC Antwortformat wurde jeweils eine richtig-/falsch-Kodierung vorgenommen, was dazu führte, dass auch bei der Multiplen Evaluation von einer aufwendigeren Auswertung abgesehen wurde, um die Vergleichbarkeit der Formate zu ermöglichen.

Im Folgenden werden nun die Auswertungen des ZART (Abschnitt 6.1) und des FRRT (Abschnitt 6.2) genauer beleuchtet. Im Anschluss erfolgt die Betrachtung der Ergebnisse in Bezug auf den Einfluss der Impulsivität auf die Testleistung (Abschnitt 6.3).

6.1 Auswertung des ZART

Der erste Schritt in der Datenanalyse des ZART bestand darin, die Daten auf Rasch-Modell-Konformität hin zu prüfen, da für einen Vergleich der Antwortformate die Item-Schwierigkeitsparameter verwendet werden sollten, wobei jedes Item mit verschiedenem Antwortformat als separates Item in die Analysen einging. Dazu wurde das Rasch-Modell (siehe Abschnitt 2.2) für die Daten berechnet und dem Likelihood Ratio Test (LRT) nach Andersen (1973) übergeben, welcher für die Teilungskriterien Score, Geschlecht, Alter, Schule, Impulsivität und Motivation durchgeführt wurde. Dieser LRT teilt die Stichprobe nach den geforderten Kriterien und überprüft, ob alle Items für die entstandenen Gruppen die gleiche Schwierigkeit besitzen. Der ebenfalls verwendete Wald Test (Glas & Verhelst, 1995) teilt die Stichprobe ebenfalls nach bestimmten Kriterien, betrachtet jedoch für jedes einzelne Item separat, ob dessen Schwierigkeit in den Gruppen in etwa gleich groß ist. Ist dies für ein oder mehrere Items nicht der Fall, kann man zusätzlich mit Hilfe des Graphischen Modelltests (Rasch, 1960) die Daten des LRT visualisieren und somit auffällige Items erkennen und für erneute Berechnungen ausschließen.

In diesem Sinne wurde die Stichprobe für die Teilungskriterien Score, Alter, Impulsivität und Motivation bei deren Median geteilt, was zum Beispiel beim Alter in einer Gruppe resultierte, die jünger als 17 Jahre alt war, und in einer, die 17 Jahre oder älter war. Beim Geschlecht wurde nach männlichen und weiblichen Testpersonen geteilt und im Falle des Kriteriums Schule wurde die Stichprobe nach der besuchten Schule unterteilt, was in diesem speziellen Fall in drei zu vergleichenden Teilstichproben resultierte. Für das Teilungskriterium Motivation wurde die Anzahl der Missings, also der nicht bearbeiteten Items einer Person, herangezogen.

Nach diesen Berechnungen konnte die Rasch-Modellgültigkeit für die vorliegenden Daten des ZART vorerst nicht angenommen werden, da bei einem angenommenen Risiko 1. Art von 0,01 für den LRT das Teilungskriterium Motivation signifikant wurde ($p < 0,001$, siehe Tabelle 5) und für den ZART die Relationen der Itemschwierigkeiten somit nicht in allen Teilgruppen gleich zu sein schienen. Auch die Graphischen Modelltests (Abbildung 5 bis 10) zeigten zumindest für die Motivation abweichende Items, was die Ergebnisse des LRT ebenfalls bestätigten. Des Weiteren musste schon im Vorfeld ein Item aus der Multiple-Choice Gruppe ausgeschlossen werden, da es von allen TeilnehmerInnen gelöst wurde und sich

somit kein entsprechender Itemparameter berechnen ließ. Natürlich wurden auch die stammgleichen Items im offenen Format und der Multiplen Evaluation von den weiteren Berechnungen ausgeschlossen.

Tabelle 5: *Ergebnisse des Andersen Likelihood Ratio Tests für den ZART vor den Ausschlüssen*

Teilungskriterium	LR - Wert	df	p - Wert
Score (2)	41,125	50	0,810
Geschlecht (2)	65,355	55	0,160
Alter (2)	37,367	54	0,959
Schule (3)	118,428	108	0,232
Impulsivität (2)	57,549	54	0,345
Motivation (2)	104,587	55	< 0,001

Anmerkungen. In den Klammern findet sich die Teilstichprobenanzahl, $\alpha=0,01$.

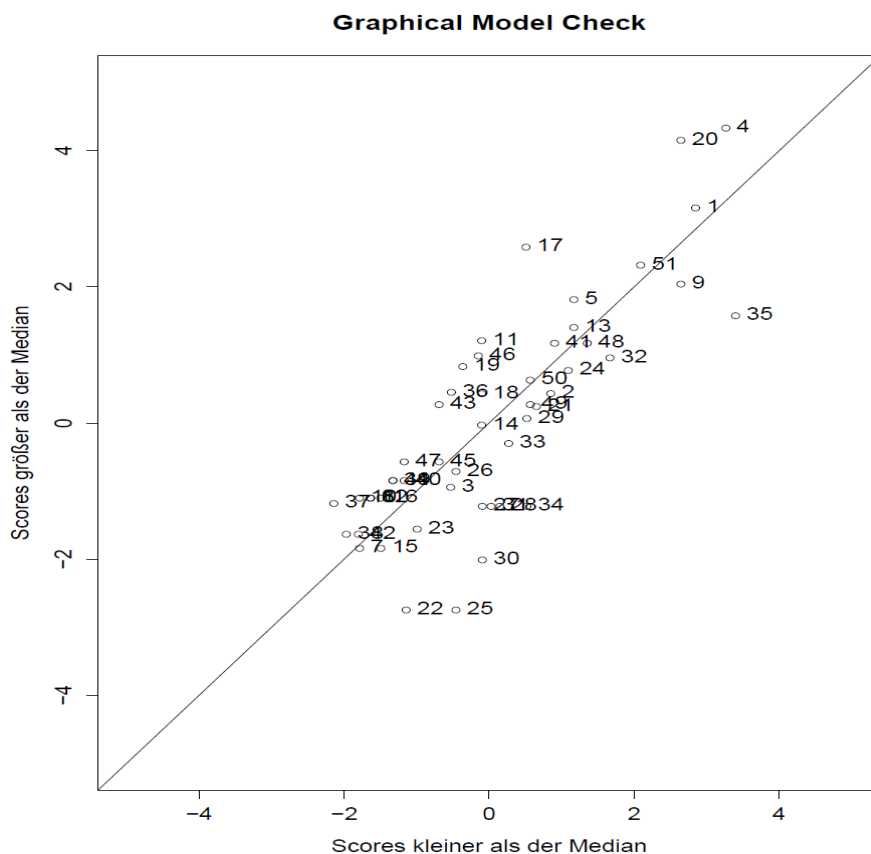


Abbildung 5: Graphischer Modell-Test für das Teilungskriterium Score (ZART) vor den Ausschlüssen, geteilt beim Median

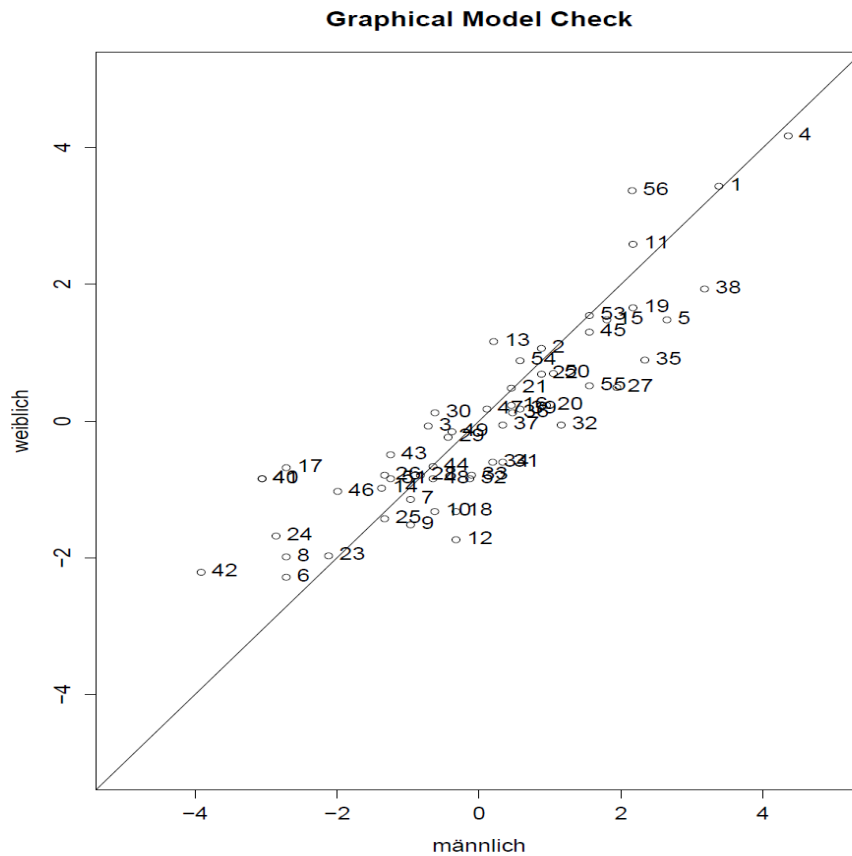


Abbildung 6: Graphischer Modell-Test für das Teilungskriterium Geschlecht (ZART) vor den Ausschlüssen, geteilt nach männlich und weiblich

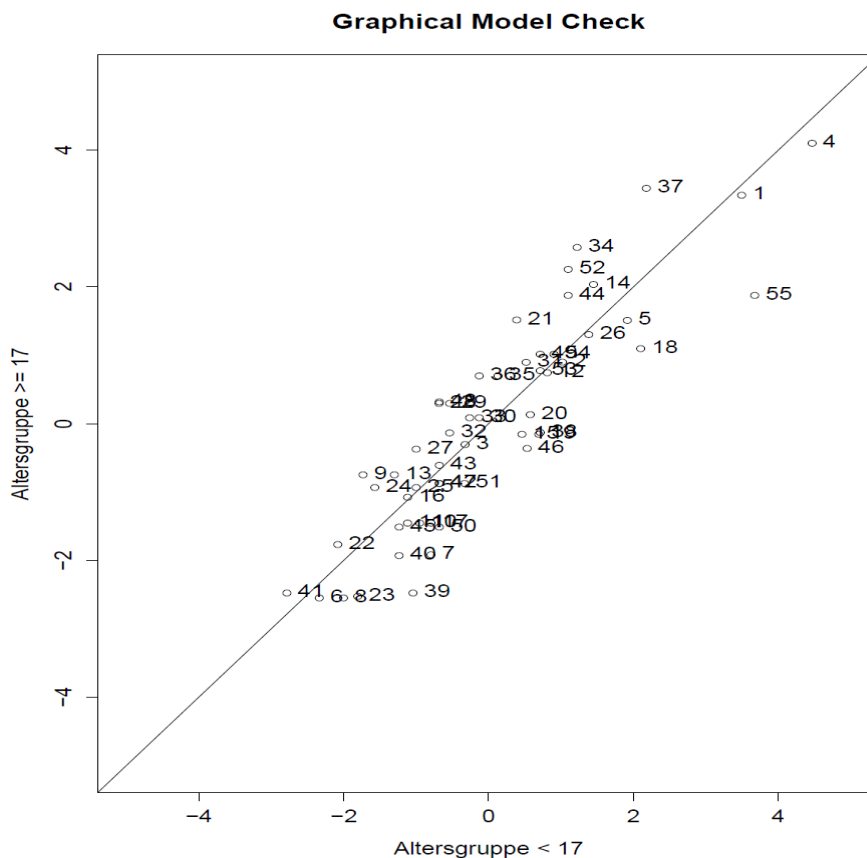


Abbildung 7: Graphischer Modell-Test für das Teilungskriterium Alter (ZART) vor den Ausschlüssen, geteilt beim Median (17)

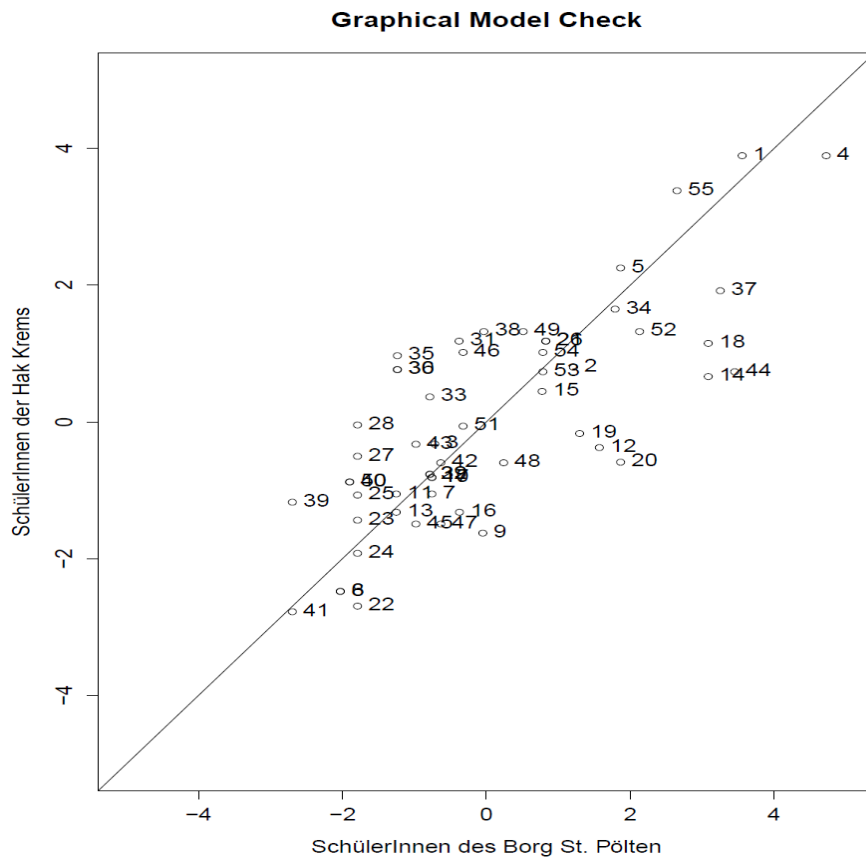


Abbildung 8: Graphischer Modell-Test für das Teilungskriterium Schule (ZART) vor den Ausschlüssen, geteilt nach Borg St. Pölten und Hak Krems

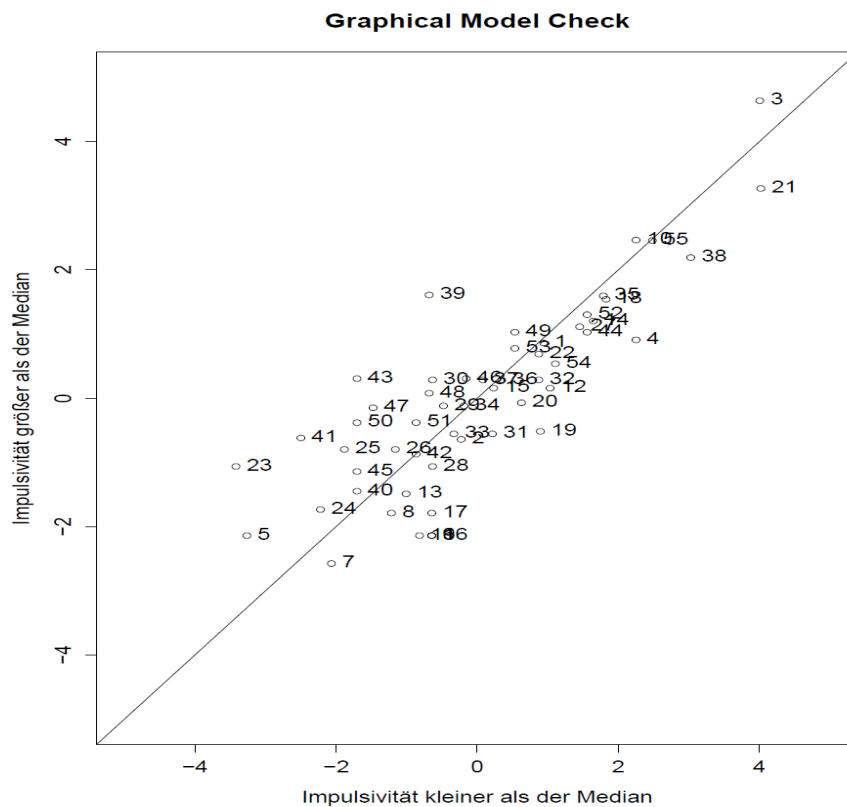


Abbildung 9: Graphischer Modell-Test für das Teilungskriterium Impulsivität (ZART) vor den Ausschlüssen, geteilt beim Median (7)

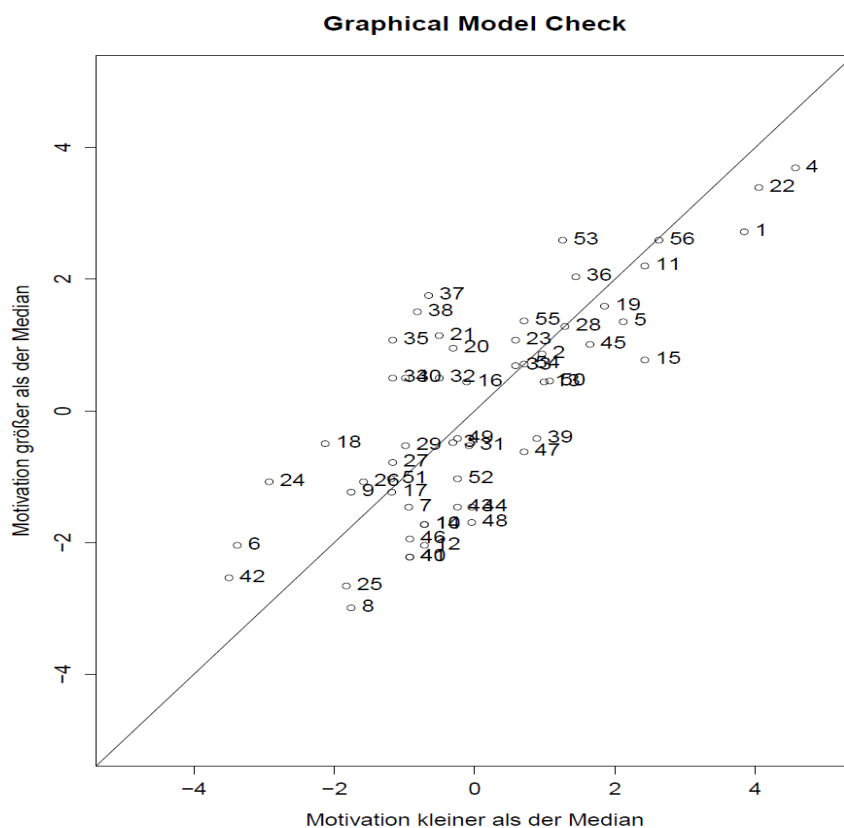


Abbildung 10: Graphischer Modell-Test für das Teilungskriterium Motivation (ZART), geteilt beim Median (4)

Mit dem Ziel eine Passung des Rasch-Modells auf die Daten zu ermöglichen, wurden der Reihe nach vier Multiple-Choice Items ausgeschlossen. Natürlich wurden für die weiteren Berechnungen auch die entsprechenden Items im offenen Format und der Multiplen Evaluation ausgeschlossen.

Nach den Ausschlüssen konnte die Geltung des Rasch-Modells für die Daten angenommen werden, da die Andersen Likelihood Ratio Tests (siehe Tabelle 6) für die jeweiligen Teilungskriterien nicht signifikant (bei $\alpha=0,01$) wurden und diese Annahme auch von den Graphischen Modelltests (siehe Abbildungen 11 bis 16) und den Wald Tests gestützt wurde.

Tabelle 6: *Ergebnisse des Andersen Likelihood Ratio Tests für den ZART nach den Ausschlüssen*

Teilungskriterium	LR - Wert	df	p - Wert
Score (2)	33,881	45	0,888
Geschlecht (2)	62,470	51	0,130
Alter (2)	36,464	50	0,924
Schule (3)	108,749	100	0,258
Impulsivität (2)	56,138	50	0,256
Motivation (2)	76,313	51	0,012

Anmerkungen. In den Klammern findet sich die Teilstichprobenanzahl, $\alpha=0,01$.

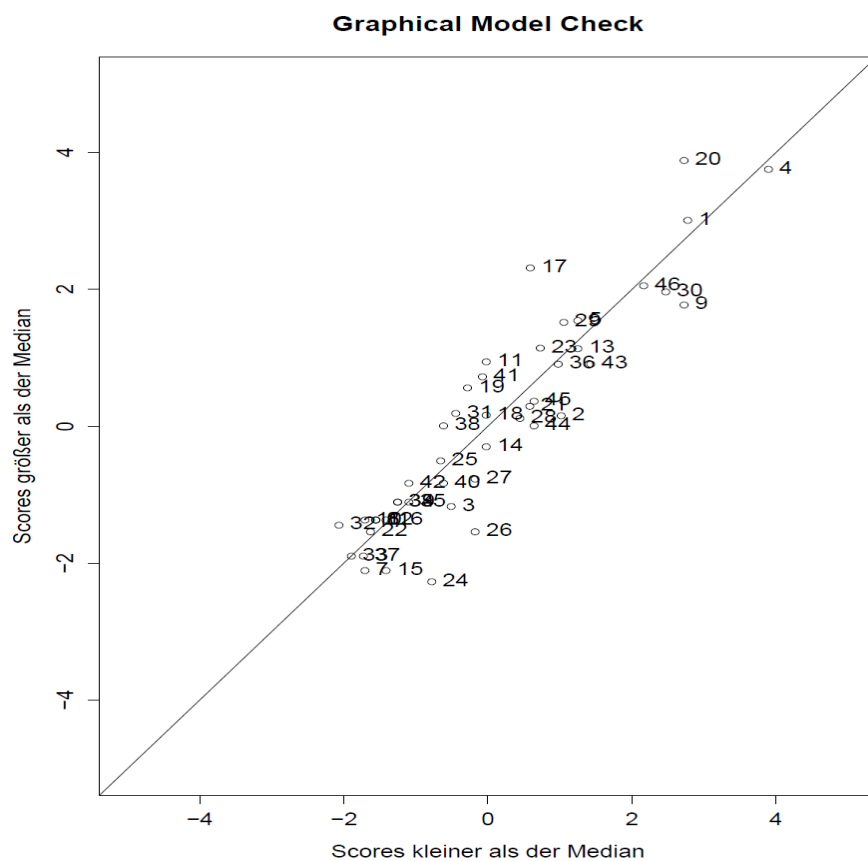


Abbildung 11: Graphischer Modell-Test für das Teilungskriterium Score (ZART) nach den vier MC Itemausschlüssen, geteilt beim Median

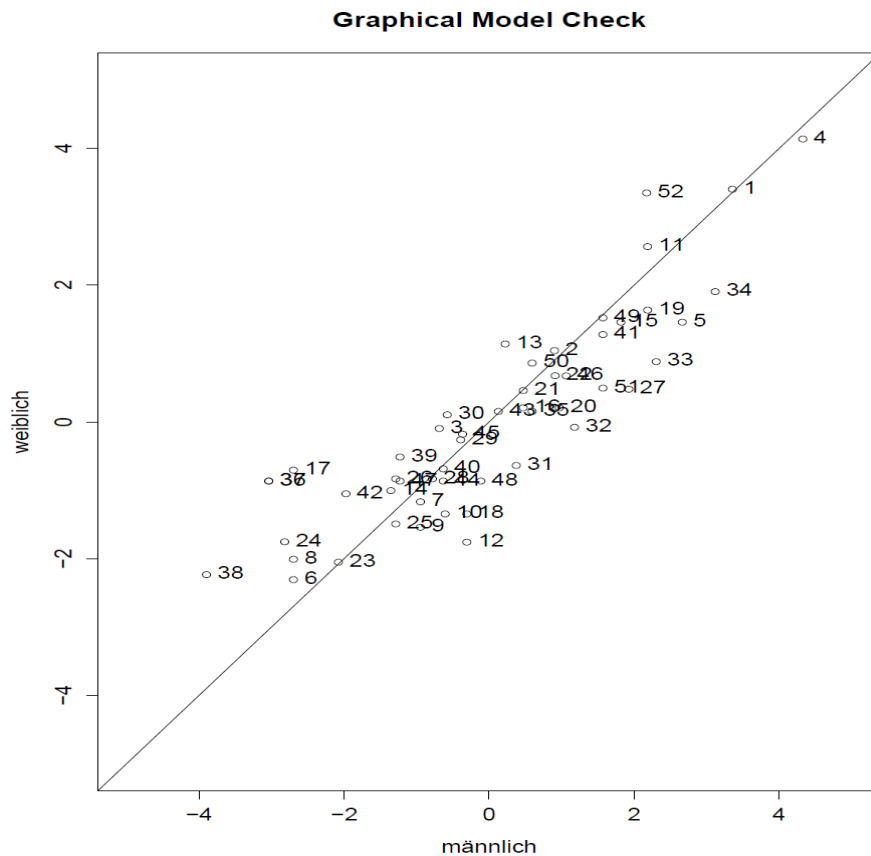


Abbildung 12: Graphischer Modell-Test für das Teilungskriterium Geschlecht (ZART) nach den vier MC Itemausschlüssen, geteilt nach männlich und weiblich

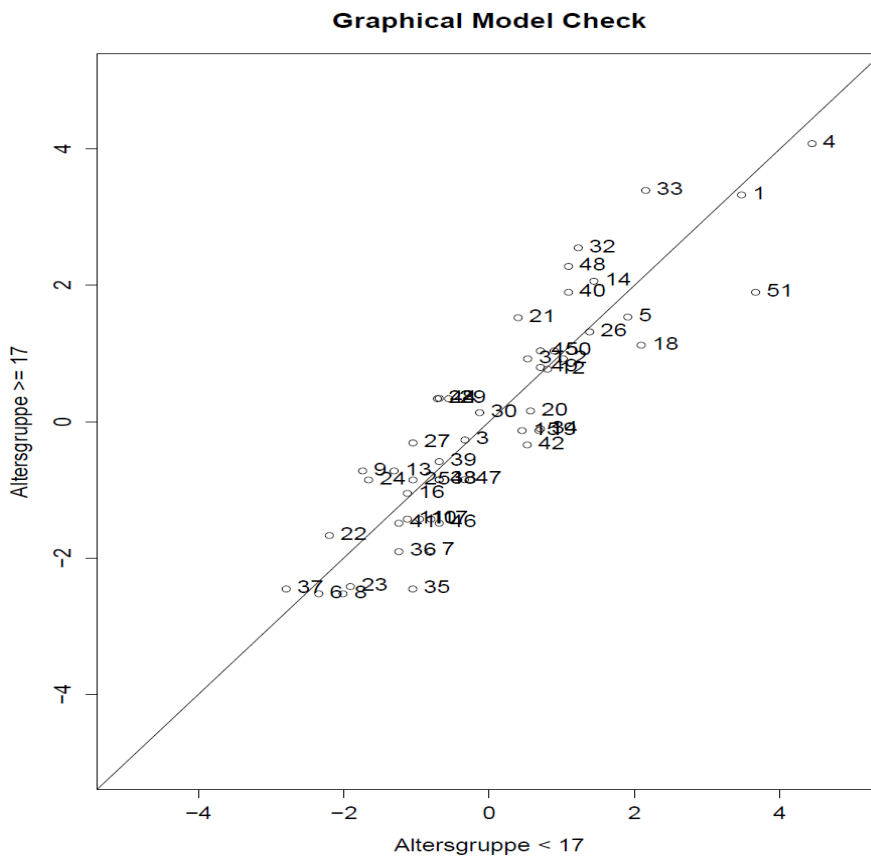


Abbildung 13: Graphischer Modell-Test für das Teilungskriterium Alter (ZART) nach den vier MC Itemausschlüssen, geteilt beim Median (17)

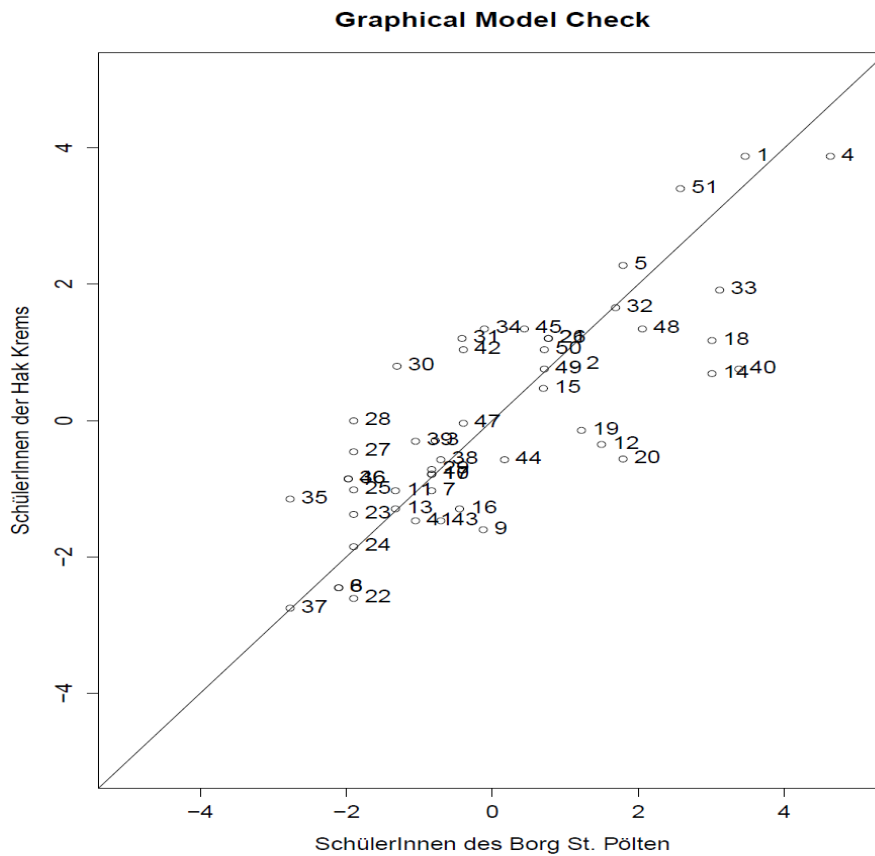


Abbildung 14: Graphischer Modell-Test für das Teilungskriterium Schule (ZART) nach den vier MC Itemausschlüssen, geteilt nach Borg St. Pölten und Hak Krems

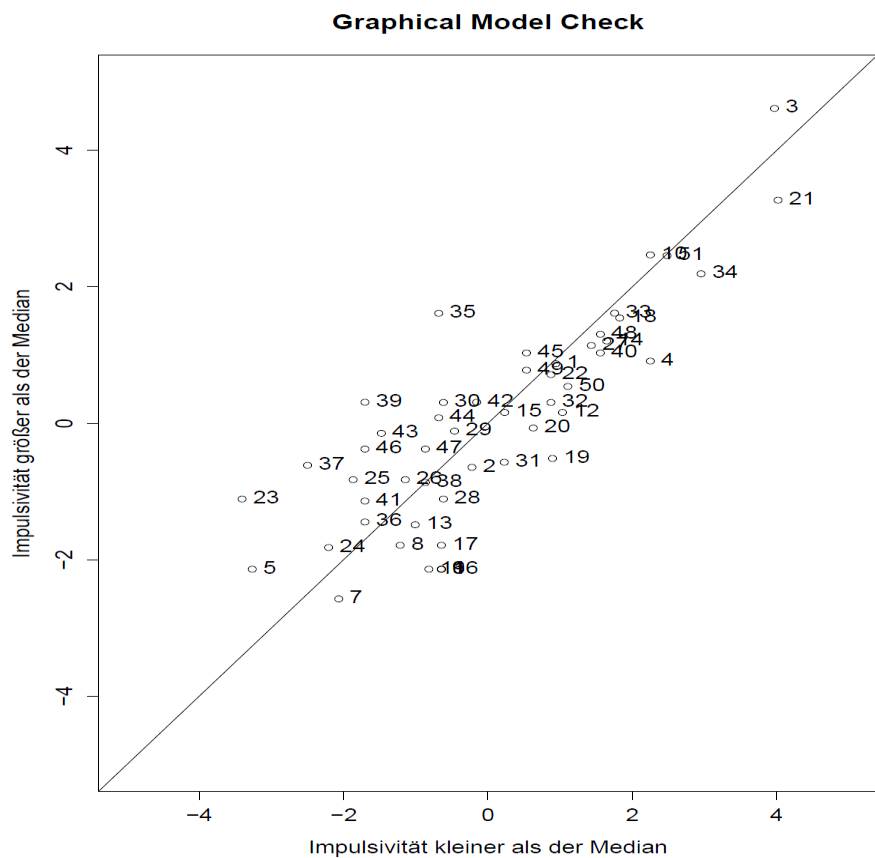


Abbildung 15: Graphischer Modell-Test für das Teilungskriterium Impulsivität (ZART) nach den vier MC Itemausschlüssen, geteilt beim Median (7)

Die anschließende Varianzanalyse der verschiedenen Antwortformate über die Itemparameter des ZART zeigte, dass sich diese nicht signifikant ($p = 0,124$) voneinander unterscheiden (siehe Tabelle 8 sowie die grafische Darstellung in Abbildung 17).

Tabelle 8: *Varianzanalyse der Antwortformate über die Itemparameter des ZART*

Varianzanalyse Antwortformate ZART					
	Df	Sum. Sq.	Mean Sq.	F – Wert	p – Wert (>F)
Antwortformate	2	0,992	0,496	2,231	0,124
Error	32	1,115	0,222		

Anmerkung. $\alpha=0,01$.

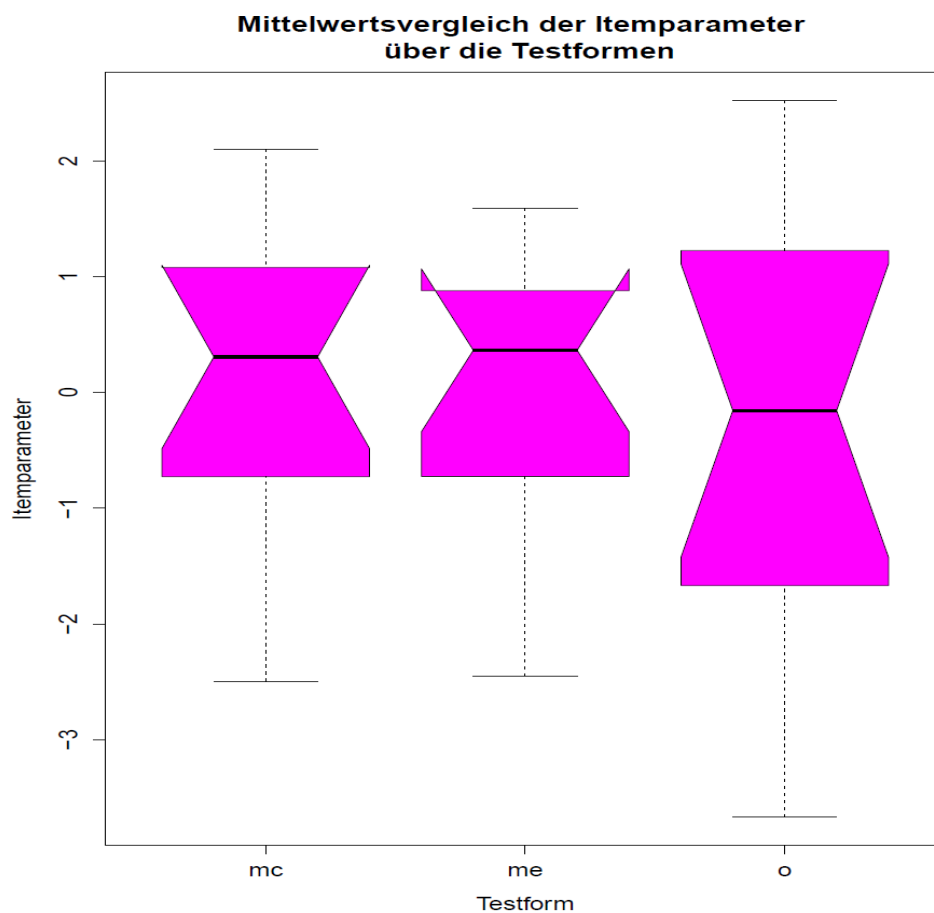


Abbildung 17: Mittelwertsvergleich der Itemparameter des ZART über die Testformen

6.2 Auswertung des FRRT

Analog zum ZART wurde auch für die Daten des FRRT eine Rasch-Modell Analyse durchgeführt, jedoch mit dem Ergebnis, dass die Daten in Bezug auf den Score und die Motivation (beide $p < 0,001$) nicht konform mit dem Rasch-Modell waren, wie Tabelle 9 und die Abbildungen 18 bis 23 zeigen.

Tabelle 9: *Ergebnisse des Andersen Likelihood Ratio Tests für den FRRT vor den Ausschlüssen*

Teilungskriterium	LR – Wert	df	p – Wert
Score (2)	116,459	57	<0,001
Geschlecht (2)	71,809	56	0,076
Alter (2)	35,738	57	0,988
Schule (3)	98,720	114	0,845
Impulsivität (2)	81,417	57	0,019
Motivation (2)	134,129	56	<0,001

Anmerkungen. Den Klammern ist die Anzahl der Teilstichproben zu entnehmen, $\alpha = 0,01$.

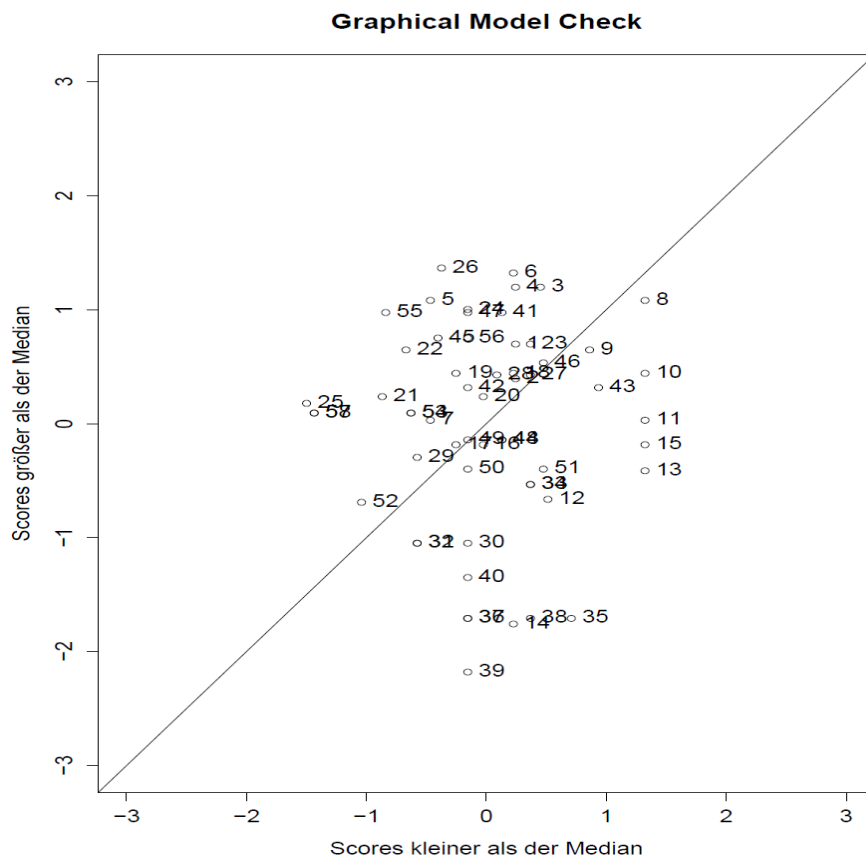


Abbildung 18: Graphischer Modell-Test für das Teilungskriterium Score (FRRT) vor den Ausschlüssen, geteilt beim Median

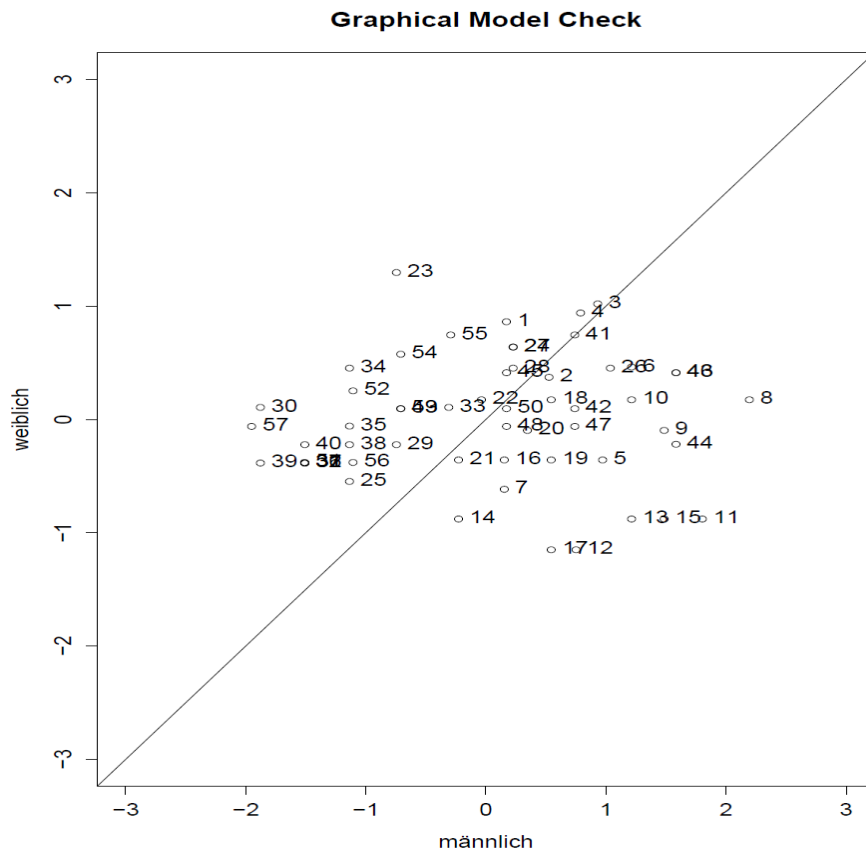


Abbildung 19: Graphischer Modell-Test für das Teilungskriterium Geschlecht (FRRT) vor den Ausschlüssen, geteilt nach männlich und weiblich

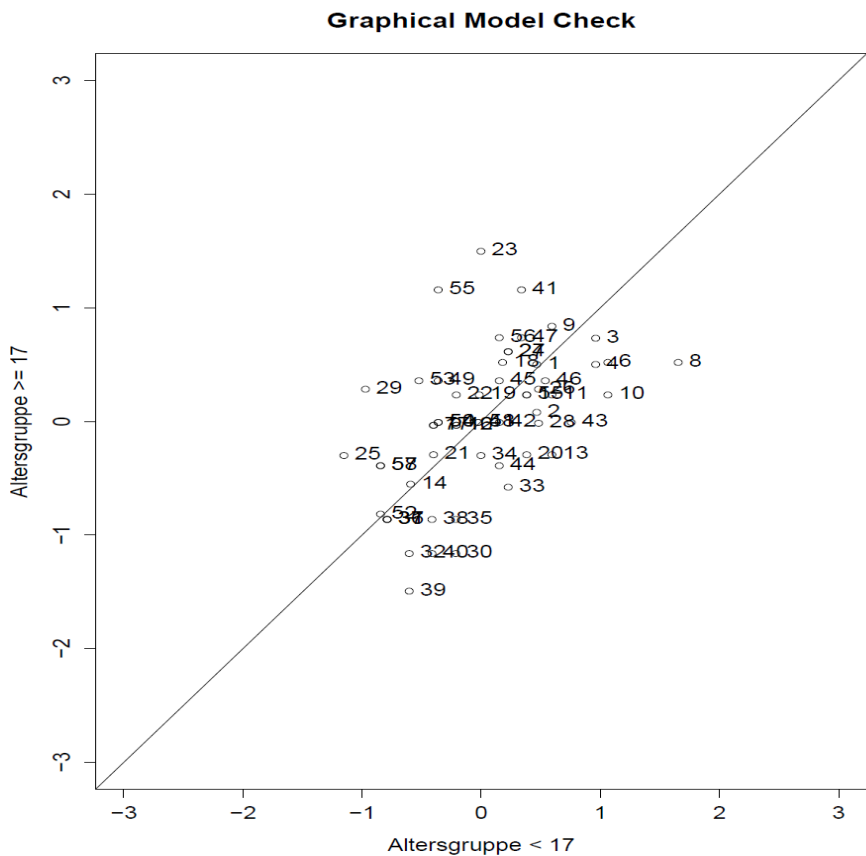


Abbildung 20: Graphischer Modell-Test für das Teilungskriterium Alter (FRRT) vor den Ausschlüssen, geteilt beim Median (17)

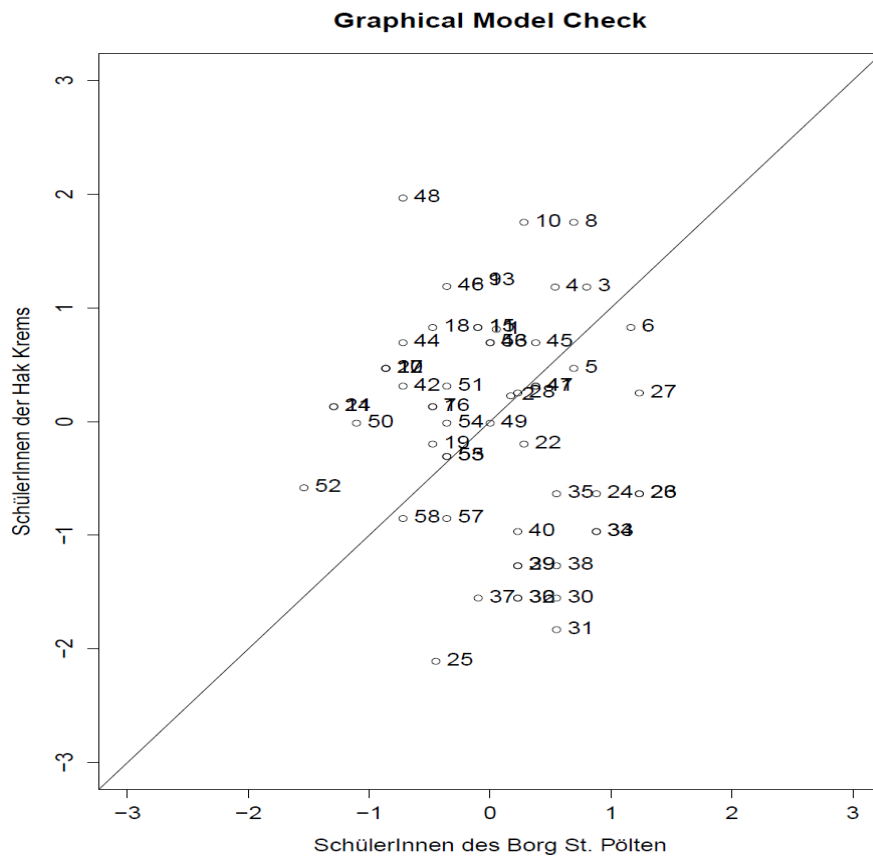


Abbildung 21: Graphischer Modell-Test für das Teilungskriterium Schule (FRRT) vor den Ausschlüssen, geteilt nach Borg St. Pölten und Hak Krems

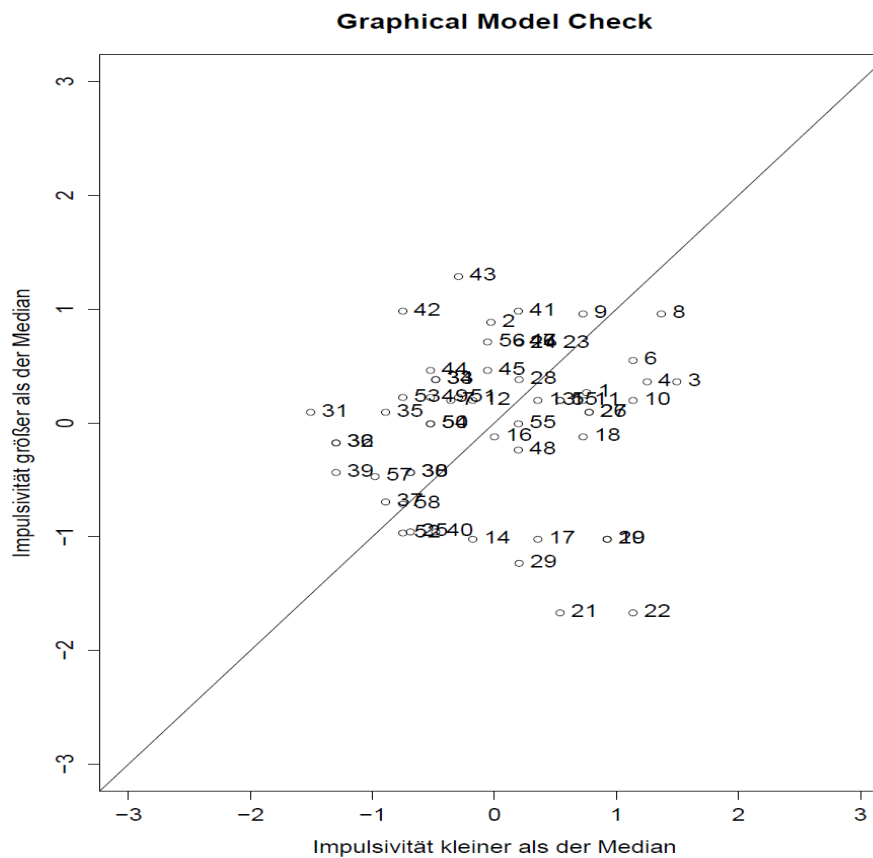


Abbildung 22: Graphischer Modell-Test für das Teilungskriterium Impulsivität (FRRT) vor den Ausschlüssen, geteilt beim Median (7)

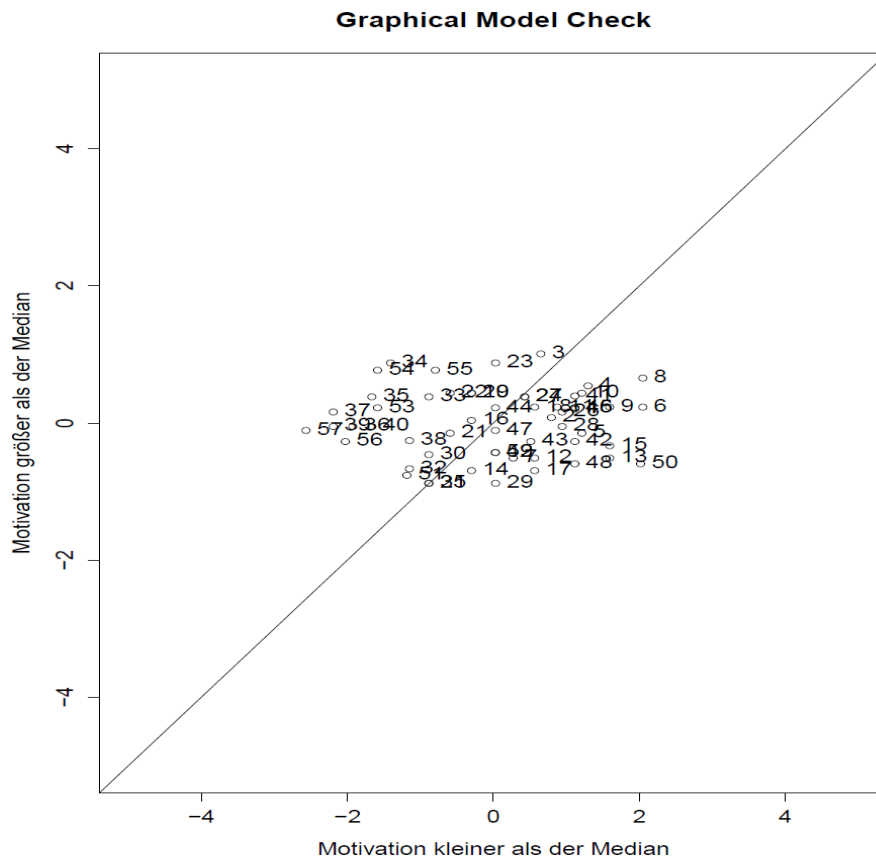


Abbildung 23: Graphischer Modell-Test für das Teilungskriterium Motivation (FRRT) vor den Ausschlüssen, geteilt beim Median (4)

Leider besserte sich die Modellpassung in Bezug auf die Motivation auch nach sechs Itemausschlüssen (22,7% der Itemtrios) nicht (siehe Tabelle 10 und Abbildungen 24 bis 29). Es wurden vier Items mit Multiple-Choice Format und zwei Items der Multiplen Evaluation ausgeschlossen, wobei je ein MC und ein ME Item den gleichen Itemstamm aufwiesen. Insofern musste diese Auswertungsstrategie aufgegeben werden, da hier keine Rasch-Modell Gültigkeit angenommen werden konnte und die Itemparameter für weitere Analysen nicht verwendet werden durften.

Tabelle 10: *Ergebnisse des Andersen Likelihood Ratio Tests für den FRRT nach den Ausschlüssen*

Teilungskriterium	LR - Wert	df	p - Wert
Score (2)	66,630	51	0,070
Geschlecht (2)	63,549	50	0,094
Alter (2)	30,512	51	0,990
Schule (3)	83,525	102	0,909
Impulsivität (2)	77,227	51	0,010
Motivation (2)	89,832	50	<0,001

Anmerkungen. Den Klammern ist die Teilstichprobenzahl zu entnehmen, $\alpha=0,01$.

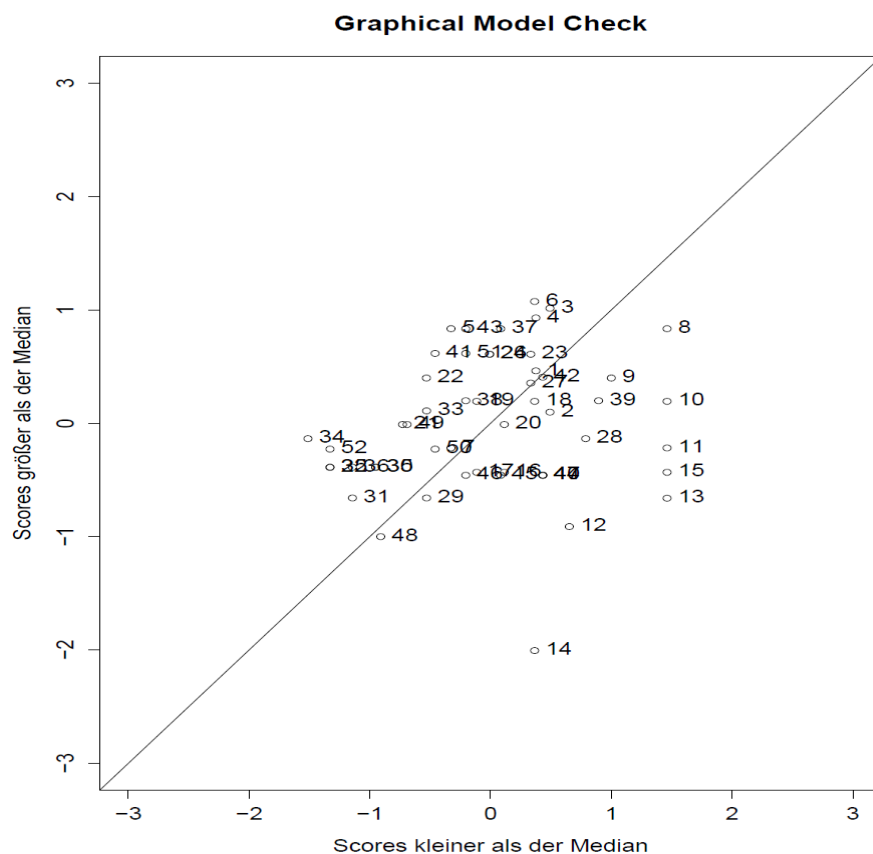


Abbildung 24: Grafischer Model-Test für das Teilungskriterium Score (FRRT) nach den Ausschlüssen von vier MC und zwei ME Items, geteilt beim Median

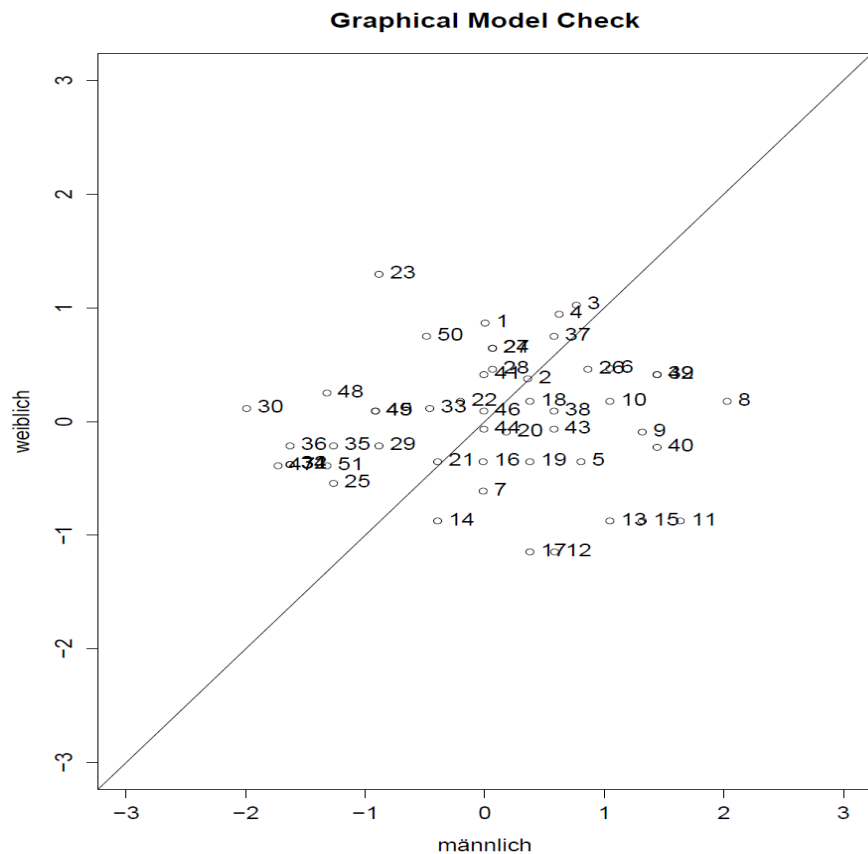


Abbildung 25: Grafischer Model-Test für das Teilungskriterium Geschlecht (FRRT) nach den Ausschlüssen von vier MC und zwei ME Items, geteilt nach männlich und weiblich

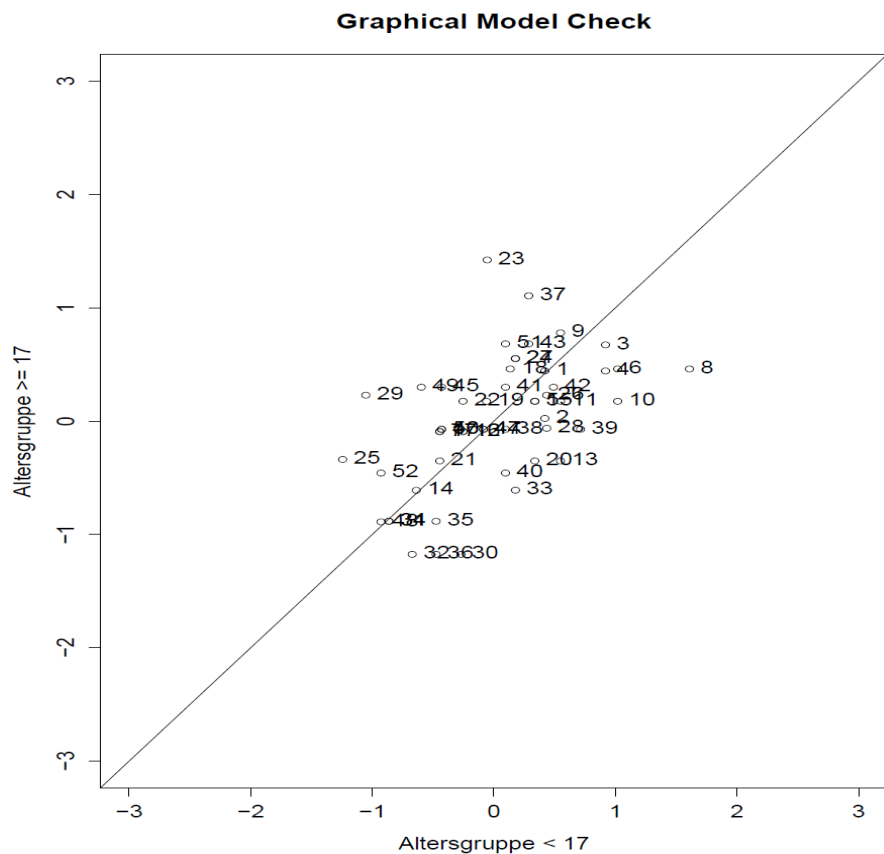


Abbildung 26: Grafischer Model-Test für das Teilungskriterium Alter (FRRT) nach den Ausschlüssen von vier MC und zwei ME Items, geteilt beim Median (17)

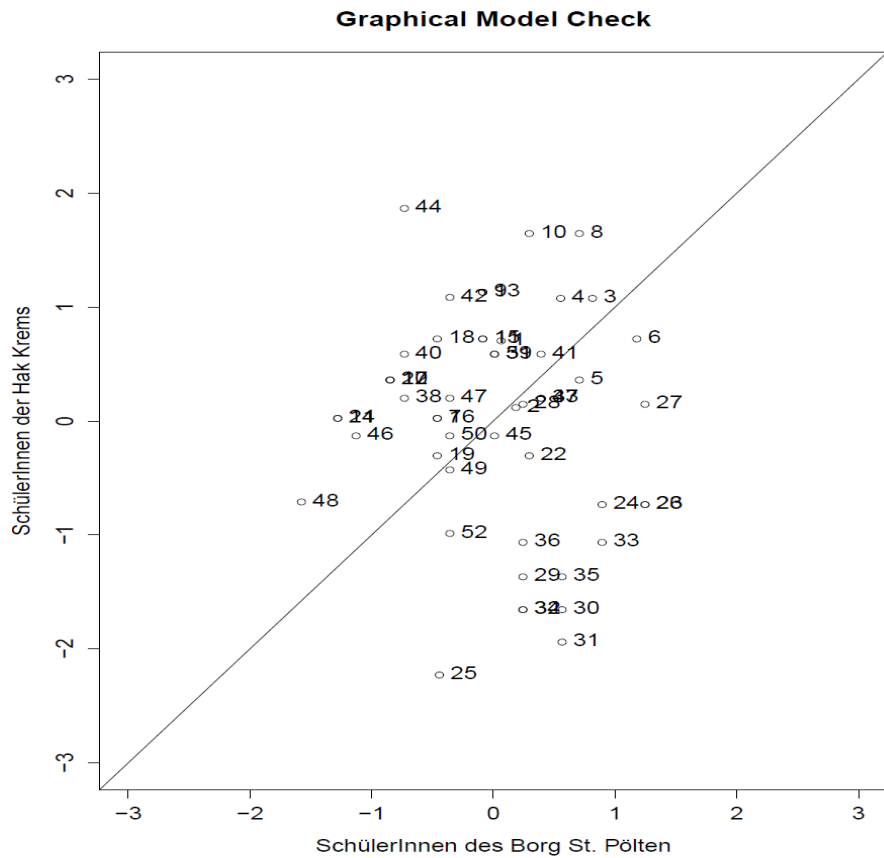


Abbildung 27: Grafischer Model-Test für das Teilungskriterium Schule (FRRT) nach den Ausschlüssen von vier MC und zwei ME Items, geteilt nach Borg St. Pölten und Hak Krems

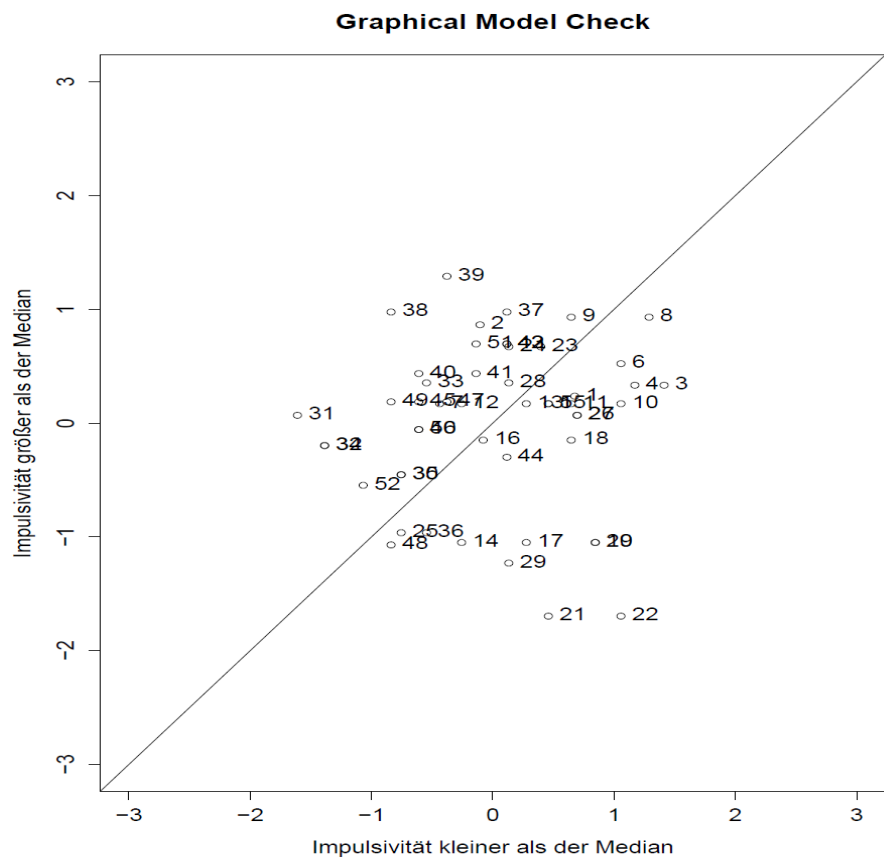


Abbildung 28: Grafischer Model-Test für das Teilungskriterium Impulsivität (FRRT) nach den Ausschlüssen von vier MC und zwei ME Items, geteilt beim Median (7)

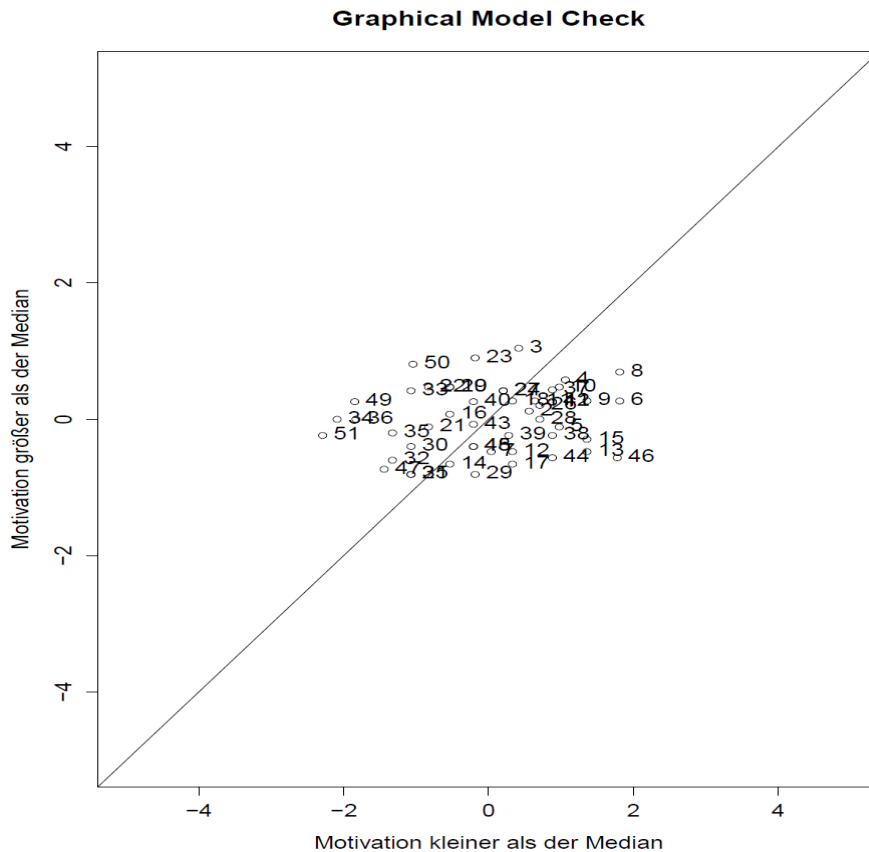


Abbildung 29: Grafischer Model-Test für das Teilungskriterium Motivation (FRRT) nach den Ausschlüssen von vier MC und zwei ME Items, geteilt beim Median (4)

Daher wurde eine alternative Vorgangsweise angedacht, nämlich eine Varianzanalyse für abhängige Stichproben über die relativen Lösungshäufigkeiten der Items des FRRT zu berechnen. Als Voraussetzung dafür musste jedoch wiederum durch eine Varianzanalyse für unabhängige Stichproben über die Personenparameter aus dem ZART festgestellt werden, dass sich die Personenfähigkeiten zwischen den vorgegebenen Testformen nicht signifikant voneinander unterscheiden. Weil in diesem Fall keine Itemparameter verwendet werden konnten, ist diese Prüfung notwendig um zu zeigen, dass nicht etwaige Gruppenunterschiede der Fähigkeiten zu Unterschieden zwischen den relativen Lösungshäufigkeiten der Items führen. Da die Varianzhomogenität der Daten angenommen werden konnte (siehe Tabelle 11, $p=0,374$), durfte die Varianzanalyse für unabhängige Stichproben über die Personenparameter durchgeführt werden.

Tabelle 11: *Levene's Test über die Homogenität der Varianzen der Personenparameter des ZART*

Levene's Test zur Prüfung der Varianzhomogenität			
	Df	F – Wert	p – Wert
Antwortformate	2	0,990	0,374
	161		

Anmerkung. $\alpha=0,01$.

Tabelle 12: *Varianzanalyse für unabhängige Stichproben über die Personenparameter des ZART*

Varianzanalyse Personenparameter					
	Df	Sum. Sq.	Mean Sq.	F – Wert	p – Wert (>F)
Antwortformate	2	6,136	3,081	2,441	0,090
Residuen	161	203,270	1,263		

Anmerkung. $\alpha=0,01$.

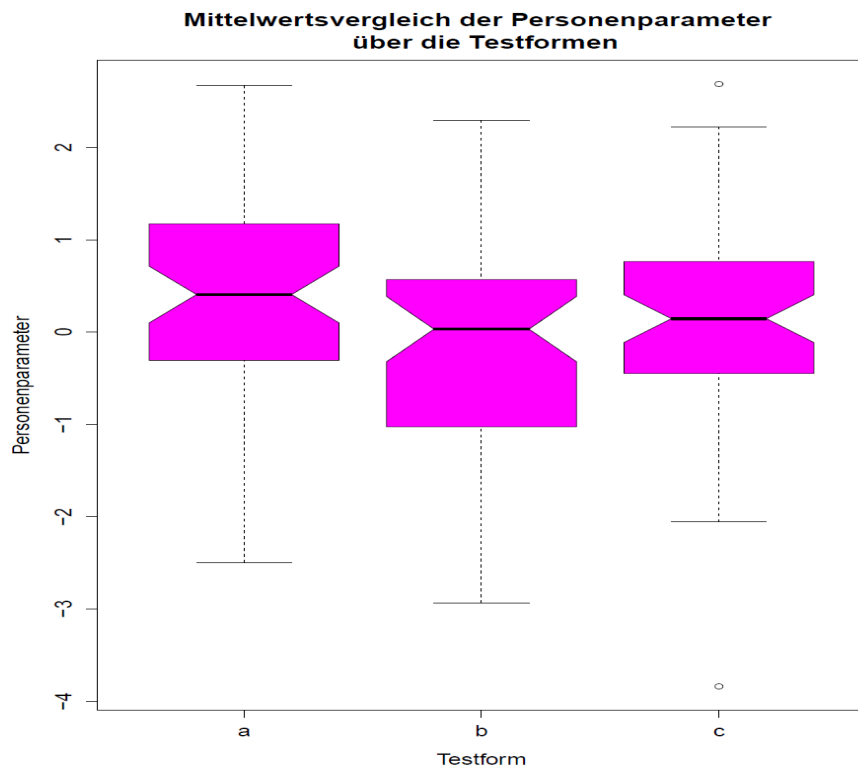


Abbildung 30: Mittelwertsvergleich der Personenparameter des ZART über die Testformen

Wie Tabelle 12 zeigt, konnten keine signifikanten Unterschiede der Personenparameter des ZART über die Testformen festgestellt werden ($p=0,090$, $\alpha=0,01$, siehe auch Abbildung 30).

Zusätzlich wurde der Ausgang dieser Varianzanalyse über einen graphischen Mittelwertsvergleich über die relative Lösungshäufigkeit der Linking-Items des FRRT (Abbildung 31) verifiziert. Leider zeigte sich hier augenscheinlich eine ungleiche Verteilung und so wurde auch über die relative Lösungshäufigkeit der Linking-Items des FRRT eine Varianzanalyse für abhängige Stichproben durchgeführt, um zu prüfen, ob sich die drei Personengruppen hinsichtlich ihrer zugrundeliegenden Fähigkeit nicht grundsätzlich unterscheiden.

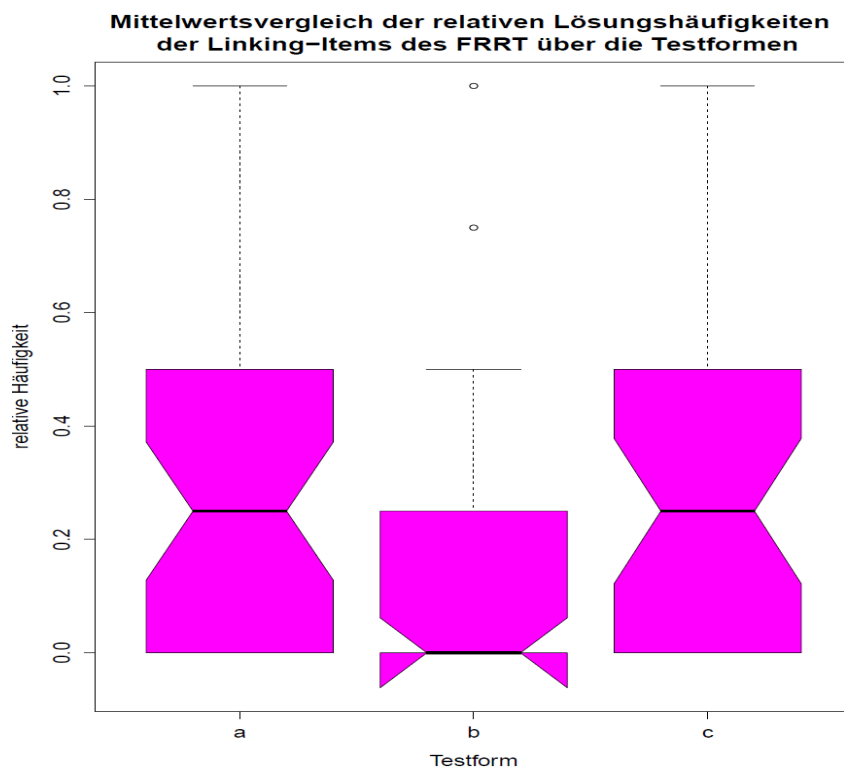


Abbildung 31: Mittelwertsvergleich der relativen Lösungshäufigkeiten der Linking-Items des FRRT

Bezüglich der Voraussetzungen können Intervallskalenniveau und Sphärizität für die verwendeten Daten angenommen werden (siehe Tabelle 13, $p=0,423$). Hier zeigte die Varianzanalyse, dass der beobachtete Unterschied im graphischen Mittelwertsvergleich ebenfalls nicht signifikant ist (Tabelle 14, $p=0,361$).

Tabelle 13: *Mauchly's Test auf Sphärizität der relativen Lösungshäufigkeiten der Linking-Items des FRRT*

Mauchly's Test der Sphärizität				
	Mauchly's W	Chi ²	Df	p – Wert
Relative Häufigkeit Linking-Items	0,953	1,720	2	0,423

Anmerkung. $\alpha=0,01$.

Tabelle 14: *Varianzanalyse für abhängige Stichproben über die relative Lösungshäufigkeit der Linking-Items des FRRT*

Varianzanalyse Linking-Items					
	Df	Sum. Sq.	Mean Sq.	F – Wert	p – Wert (>F)
Relative Häufigkeit Linking-Items	2	0,171	0,086	1,034	0,361
Error	74	6,121	0,083		

Anmerkung. $\alpha=0,01$.

Da durch die beiden Varianzanalysen gezeigt werden konnte, dass sich die Testpersonen in den einzelnen Gruppen nicht signifikant unterschieden, konnte die Varianzanalyse für abhängige Stichproben über die relativen Lösungshäufigkeiten der gelösten Items des FRRT durchgeführt werden. Hierfür konnten die Voraussetzungen nur teilweise erfüllt werden, da die Intervallskalierung zwar gegeben war, jedoch die Sphärizität der Daten nicht angenommen werden konnte ($p<0,001$, siehe Tabelle 15). Daher wurde für die durchgeführte Varianzanalyse auf die Greenhouse-Geisser Korrektur zurückgegriffen.

Tabelle 15: *Mauchly's Test auf Sphärizität der relativen Lösungshäufigkeiten der gelösten Items des FRRT*

Mauchly's Test der Sphärizität				
	Mauchly's W	Chi ²	Df	p – Wert
Relative Häufigkeit gelöste Items	0,125	33,293	2	<0,001

Anmerkung. $\alpha=0,01$.

Tabelle 16: *Varianzanalyse für abhängige Stichproben über die relative Lösungshäufigkeit der gelösten Items des FRRT mit Greenhouse-Geisser Korrektur*

Varianzanalyse gelöste Items des FRRT					
	Df	Sum. Sq.	Mean Sq.	F – Wert	p – Wert (>F)
Relative Häufigkeit gelöste Items	2	0,074	0,070	1,352	0,263
Error	18,132	0,935	0,052		

Anmerkung. $\alpha=0,01$.

Wie in Tabelle 16 zu sehen ist, konnten auch beim FRRT zwischen den verschiedenen Antwortformaten keine signifikanten ($p=0,263$ bei $\alpha=0,01$) Unterschiede in den Lösungshäufigkeiten gefunden werden.

6.3 Auswertung des Einflusses der Impulsivität

Um die Frage nach einem Einfluss der Impulsivität einer Person auf ihre Leistung in den einzelnen Antwortformaten des ZART und des FRRT zu beantworten, wurden für die beiden verwendeten Tests jeweils hierarchische lineare Regressionsanalysen berechnet. In einem ersten Schritt fungierten die Impulsivitätswerte der Personen als Prädiktoren, die relativen Lösungshäufigkeiten der Testpersonen in den Tests stellten die Kriterien dar. In Schritt zwei wurde die Art des Antwortformates als Prädiktor zum Modell hinzugefügt. Schritt drei bestand aus der Implementierung einer zentrierten Interaktionsvariable aus Impulsivität und Antwortformat. Da die verwendeten Daten intervallskaliert waren und auch in den hinzugezogenen Streudiagrammen (siehe Anhang) keine Anhaltspunkte für einen anderen als einen linearen Zusammenhang gefunden wurde, konnten die geplanten Analysen durchgeführt werden. Im Zuge dessen konnte in keinem Fall die Multikollinearität der Prädiktoren festgestellt werden.

Die Ergebnisse dieser Berechnungen sind in den Tabellen 17 und 18 dargestellt.

Tabelle 17: *Zusammenfassung der hierarchischen Regressionsanalyse für den ZART*

Variable ZART		B	SE	β
1. Schritt	Impulsivität	0,005	0,009	0,035
2. Schritt	Impulsivität	0,004	0,009	0,032
	Antwortformat	-0,040	0,018	-0,155
3. Schritt	Impulsivität	-0,033	0,026	-0,240
	Antwortformat	-0,040	0,018	-0,155
	Impulsivität * Antwortformat	0,017	0,011	0,292

Anmerkungen. $R^2 = 0,001$ für Schritt 1; $\Delta R^2 = 0,024$ für Schritt 2; $\Delta R^2 = 0,011$ für Schritt 3; Annahme der H_0 . $\alpha=0,01$.

Aus Tabelle 17 ist ersichtlich, dass die Impulsivität allein kaum Variation (0,1 %) in den relativen Lösungshäufigkeiten des ZART zu erklären vermag, was sich auch in den nicht signifikanten Koeffizienten ($p=0,612$) niederschlägt. Bei Integration des Prädiktors Antwortformat erhöht sich die erklärbare Variation in den Daten auf 2,5 Prozent, jedoch sind in diesem Fall die berechneten Koeffizienten ebenfalls nicht signifikant (Impulsivität: $p=0,644$; Antwortformat: $p=0,026$). Auch unter Einbezug der gebildeten Interaktionsvariablen aus Impulsivität und Antwortformat kann dieses Modell nur 3,7 Prozent der Varianz der Daten erklären und die Koeffizienten liegen noch immer im nicht signifikanten Bereich (Impulsivität: $p=0,203$; Antwortformat: $p=0,026$; Impulsivität*Antwortformat: $p=0,121$). Daraus kann geschlossen werden, dass die Impulsivität keinen signifikanten Einfluss auf die Leistungen im ZART hat.

Tabelle 18: *Zusammenfassung der hierarchischen Regressionsanalyse für den FRRT*

Variable FRRT		B	SE	β
1. Schritt	Impulsivität	0,003	0,009	0,025
2. Schritt	Impulsivität	0,003	0,009	0,025
	Antwortformat	0,005	0,018	0,020
3. Schritt	Impulsivität	-0,025	0,026	-0,181
	Antwortformat	0,005	0,018	0,020
	Impulsivität * Antwortformat	0,013	0,011	0,221

Anmerkungen. $R^2 = 0,001$ für Schritt 1; $\Delta R^2 = 0,000$ für Schritt 2; $\Delta R^2 = 0,007$ für Schritt 3; Annahme der H_0 . $\alpha=0,01$.

Ähnliches kann man bei der hierarchischen Regressionsanalyse für den ZART beobachten (siehe Tabelle 18). Der Anteil der Impulsivität an erklärter Varianz der Daten zum FRRT beträgt nur 0,1 Prozent und auch hier konnten keine signifikanten Koeffizienten beobachtet werden ($p=0,724$). Auch unter Einbezug des Antwortformates ändert sich an diesem Bild nichts, da der Anteil der erklärten Varianz auf gleichem Niveau bleibt und auch die Koeffizienten nicht signifikant werden (Impulsivität: $p=0,720$; Antwortformat: $p=0,774$). Nach Integration der Interaktionsvariablen vermag das Modell 0,8 Prozent der Streuung der Daten zu erklären, doch auch hier bleiben die Koeffizienten auf nicht signifikantem Niveau (Impulsivität: $p=0,344$; Antwortformat: $p=0,775$; Impulsivität*Antwortformat: $p=0,247$).

Aus der Gesamtheit der Ergebnisse lässt sich der Eindruck gewinnen, dass weder die Impulsivität noch das Antwortformat einen linearen Einfluss auf die relativen Lösungshäufigkeiten in ZART und FRRT haben. Auch scheint die Impulsivität in Abhängigkeit vom Antwortformat keinen linearen Einfluss auf die Leistung in den beiden Verfahren zu haben.

7 Diskussion und Ausblick

Die in Abschnitt 2 erläuterten Untersuchungen hinsichtlich des Einflusses von Antwortformaten auf die Itemschwierigkeit hatten eine Kernaussage gemein: nämlich, dass sich zumindest in den verwendeten Verfahren die Itemschwierigkeiten bei unterschiedlichen Formaten unterschieden. Meist zeigte sich, dass Multiple-Choice Items leichter zu beantworten waren als solche mit offenem Antwortformat. Eine Ausnahme bildeten die Multiple-Choice Formate „2 aus 5“ und „x aus 5“, die in ihrer Schwierigkeit mit dem offenen Format vergleichbar waren. Auch konnte gezeigt werden, dass die Güte der Distraktoren bei Multiple-Choice Items ganz entscheidend zu deren Schwierigkeit beiträgt (vergleiche zum Beispiel Ascalon, Meyers, Davis & Smits, 2007). Zusätzlich wurden Persönlichkeitseigenschaften wie etwa die Motivation als ein Prädiktor der Itemschwierigkeit identifiziert (Solheim, 2011). Positionseffekte sollten ebenfalls nicht unbeachtet bleiben, da auch diese einen Einfluss auf die Schwierigkeit von Items nehmen können. Für diese Arbeit können sie jedoch außer Acht gelassen werden, da sich stammgleiche Items in den verschiedenen Testformen an den gleichen Stellen befanden und eventuell auftretende Effekte für alle TestnehmerInnen in der selben Weise wirkten.

Natürlich lassen sich die vorliegenden Ergebnisse aufgrund des Motivationsmangels der Testpersonen nur sehr eingeschränkt interpretieren, dennoch wird im Folgenden versucht, sie mit Bezug auf die bisherigen Aussagen der Literatur einzuordnen.

In das schon bestehende Bild lassen sich die Ergebnisse dieser Arbeit nicht so einfach einfügen, da sie, im Gegensatz zur bisherigen Literatur, keine Unterschiede in den Schwierigkeiten von offenen, Multiple-Choice „1 aus 5“ und Multiplen Evaluation Antwortformaten aufweisen, was sowohl die Daten des ZART über die Itemschwierigkeiten als auch die Befunde aus dem FRRT über die relativen Lösungshäufigkeiten zeigten. Eine Ursache dafür kann man in der erfassten Fähigkeit vermuten, da bei dieser Art von Aufgaben zum schlussfolgernden Denken auch für Multiple-Choice und Multiple Evaluation ein Wiedererkennen von Antworten, wie dies Zimbardo und Gerring (1999) zumindest für Multiple-Choice Items postulieren, ausgeschlossen werden kann. Man kann vermuten, dass bei den vorgelegten Items bei allen drei Antwortformaten mit der selben Strategie an deren Bearbeitung herangegangen wird, da zum Beispiel bei den Zahlenreihen des ZART der Weg zur Lösung nur über ein Verständnis der zugrundeliegenden Rechenregeln führen kann. Natürlich könnte dies auch darin begründet sein, dass die Distraktoren der

gebundenen Formate gut gewählt waren, was, wie oben erwähnt, ganz entscheidend zur Schwierigkeit eines Items beitragen kann. Das heißt, dass die Testpersonen diese nicht von vornherein ausschließen konnten und sich so intensiver mit der Fortführung der Reihe beschäftigen mussten.

Diese Schlussfolgerungen legen nahe, dass es für das Ergebnis eines Verfahrens zur Erfassung des schlussfolgernden Denkens kaum einen Unterschied macht, welches Antwortformat verwendet wird. Daher ist das Multiple-Choice Format „1 aus 5“ vorzuziehen, da es in der Auswertung ökonomischer und reliabler ist als das offene Antwortformat und für die meisten Testpersonen bekannter ist als die Multiple Evaluation. Weiters spricht gegen die Verwendung der Multiplen Evaluation, dass annähernd 70 Prozent der Testpersonen den von ihnen favorisierten Alternativen die vollen 100 Prozent gaben, was im Grunde wieder einem Multiple-Choice Format entspricht.

Die Frage, ob die Impulsivität einer Person einen Einfluss auf ihre Leistung in den verwendeten Tests in Abhängigkeit von deren Antwortformat hat, muss aufgrund der Ergebnisse ebenfalls verneint werden. Im vorliegenden Fall war diese Persönlichkeitseigenschaft weder ein guter Prädiktor für die Fähigkeit zum schlussfolgernden Denken, noch zogen Personen mit höheren oder niedrigeren Impulsivitätswerten einen Vorteil aus einem der Antwortformate.

Natürlich lassen sich die Schlussfolgerungen dieser Arbeit nicht einfach auf alle Reasoningtests oder gar Intelligenztests übertragen, da es noch zu wenig Forschung gibt, die zu den selben Ergebnissen kommt, und die bisherigen Vergleiche der Itemschwierigkeit bei verschiedenen Antwortformaten aufzeigen, dass durchaus Unterschiede existieren. Möglicherweise ließen sich im vorliegenden Fall keine Unterschiede nachweisen, da bei den untersuchten Verfahren zum logischen Schlussfolgern Vorwissen keine Rolle spielt, mit Ausnahme der Rechenfähigkeit bei den Zahlenreihen des ZART und einem gewissen Grundwissen von Familienstrukturen beim FRRT. Somit ist es bei optimal gewählten Distraktoren nicht möglich, aus den gegebenen Antwortmöglichkeiten einen Vorteil zu ziehen.

Kritischerweise muss bei dieser Untersuchung auf einige Limitationen hingewiesen werden. Zum Ersten kann über den großen Mangel an Motivation der Testpersonen nicht hinweggegangen werden, ohne darauf hinzuweisen, dass dieser Umstand die Güte der aus dieser Untersuchung gezogenen Schlussfolgerungen massiv

beeinträchtigt und auf alle Fälle eine neuerliche Untersuchung nach sich ziehen sollte.

Zum Zweiten muss auf die bereits erwähnte und begründete Einschränkung auf die verwendeten Verfahren hingewiesen werden. Zum Dritten wurde in der vorliegenden Arbeit die Muttersprache der Testpersonen nicht miterhoben, was sicherlich einen interessanten Aspekt für die Interpretationen aus dem sprachlichen FRRT mit sich gebracht hätte. Zum Vierten spielt für die Ergebnisse sicherlich auch die knapp bemessene Testzeit eine Rolle, die die Schwierigkeiten der Items ebenfalls beeinflusst haben könnte, jedoch im vorliegenden Fall leider aufgrund organisatorischer Gegebenheiten nicht vermeidbar war. Der Zeitfaktor ist vor allem deshalb von Bedeutung, da so die Schnelligkeit einer Person nicht von ihrer wahren Fähigkeit separiert werden kann. Die fünfte Limitation besteht in der Erfassung der Impulsivität, die nur aus drei neukonstruierten Fragen bestand. Hierfür existieren natürlich keine Vergleichswerte und des Weiteren ist nicht gesichert, dass diese Fragen das Konstrukt der Impulsivität zu messen im Stande sind.

In weiterer Folge wäre es angebracht, fortführende Untersuchungen mit Einbezug der Muttersprache und ohne Zeitlimitation zum Einfluss von Antwortformaten bei Tests zum logischen Schlussfolgern beziehungsweise Intelligenztests durchzuführen. Dabei ist besonders auf eine hohe Testmotivation zu achten, um die Ergebnisse der vorliegenden Arbeit zu untermauern. Da die Motivation von Freiwilligen jedoch schwer zu erhöhen ist, wäre es sinnvoll, zur Testung Personen heranzuziehen, die Konsequenzen aus ihrer Leistung erwarten, wie es zum Beispiel bei der Bewerberauswahl der Fall ist. Weiters wäre es interessant, den wahren Einfluss von Impulsivität auf die Leistung im schlussfolgernden Denken beziehungsweise auf die einzelnen Antwortformate zu überprüfen, indem man zur Feststellung dieser Persönlichkeitseigenschaft einen erprobten Test heranzieht.

8 Zusammenfassung

Bis zur vorliegenden Arbeit ging die einschlägige Literatur davon aus, dass zwischen den Schwierigkeiten von Items mit verschiedenen Antwortformaten Unterschiede bestehen. Meist kamen die AutorInnen zu dem Ergebnis, dass Multiple-Choice Tests einfacher zu bearbeiten sind als solche mit offenen Fragen. Allerdings waren die Grundlage dieser Untersuchungen Wissenstests, wie im Falle von Thissen, Wainer und Wang (1994) oder Kubinger und Gottschall (2007). Vereinzelt wurden die Arbeiten auch auf Leistungstests aufgebaut, wie bei Teubenbacher (2009). Leider lassen sich diese Ergebnisse und Schlussfolgerungen nicht auf alle Verfahren übertragen ohne deren Gültigkeit vorher zu überprüfen. Daher war das Ziel dieser Arbeit eine Überprüfung der Hypothese, dass Multiple-Choice Tests auch bei Verfahren zum logischen Schlussfolgern als Grundlage eine geringere Schwierigkeit aufweisen als Tests mit offenem Antwortformat. Zusätzlich wurde auch das Antwortformat der Multiplen Evaluation vorgegeben, das bezüglich seiner Schwierigkeit noch nicht eingeordnet wurde. Zu diesem Zweck wurden für den Zahlenreihentest ZART (Poinstingl, 2009 a) und den Family-Relations-Reasoning-Test FRRT (Poinstingl, Kubinger, Skoda & Schechtner, in Vorb.) die bestehenden Multiple-Choice Antwortformate so verändert, dass für jedes Item ein Multiple-Choice Format „1 aus 5“, eine Multiple Evaluation mit fünf Antwortmöglichkeiten und ein offenes Antwortformat entstand. Diese Tests wurden in drei verschiedene Formen gebracht und 208 SchülerInnen an drei höheren beziehungsweise berufsbildenden Schulen in Niederösterreich vorgegeben. Die Ergebnisse einer Varianzanalyse über die Itemparameter des ZART zeigten keine Unterschiede zwischen den Antwortformaten, was auch die Varianzanalyse über die relativen Lösungshäufigkeiten aus dem FRRT bestätigte. Daraus kann man nun folgern, dass es für diese Art von Tests, also Verfahren zum logischen Schlussfolgern, keinen Unterschied für deren Schwierigkeit macht, welches Antwortformat verwendet wird, da davon ausgegangen werden kann, dass sich die Herangehensweise an die Lösung nicht ändert. Da die vorliegenden Ergebnisse jedoch nicht mit den bisherigen Schlussfolgerungen konform gehen und die geringe Motivation der Testpersonen die Güte der Aussagen dieser Arbeit stark in Frage stellt, sollte eine erneute Untersuchung angedacht werden, möglichst mit TeilnehmerInnen, bei denen die Testteilnahme mit Konsequenzen verbunden sind um eine hohe Testmotivation zu erreichen.

IV. Literaturverzeichnis

Amelang, M & Bartussek, D. (2001). *Differentielle Psychologie und Persönlichkeitsforschung*. Stuttgart: Kohlhammer.

Amthauer, R., Brocke, B., Liepmann, D. & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)*. Göttingen: Hogrefe.

Andersen, E. B. (1973). A Goodness of Fit Test for the Rasch model. *Psychometrika*, 38, 123-140.

Ascalon, M. E., Meyers, L. S., Davis, B. W. & Smits, N. (2007). Distractor Similarity and Item-Stem Structure : Effects on Item Difficulty. *Applied Measurement in Education*, 20 (2), 153-170.

Bielinski, J. & Davison M. L. (2001). A Sex Difference by Item Difficulty Interaction in Multiple-Choice Mathematics Items Administered to National Probability Samples. *Journal of Educational Measurement*, 38 (1), 51-77.

Bösch, A. R. (2011). *Konstruktionsartefakte des Family-Relations-Reasoning-Test?*. Unveröff. Dipl. Arbeit, Universität, Wien.

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies*. New York: Cambridge University Press.

Clariana, R. B. (2003). The Effectiveness of Constructed-response and Multiple-choice Study Tasks in Computer Aided Learning. *Journal of Educational Computing Research*, 28 (4), 395-406.

Cronbach, L. J. (1941). An experimental Comparison of the Multiple True-False and Multiple Multiple-Choice Tests. *Journal of Educational Psychology*, 32 (7), 533-543.

Currie, M. & Chiramanee, T. (2010). The effect of multiple-choice item format on the measurement of knowledge of language structure. *Language Testing*, 27 (4), 471-491.

Delgado, A. R. & Prieto, G. (1998). Further Evidence Favoring Three-Option Items in Multiple-Choice Tests. *European Journal of Psychological Assessment*, 14 (3), 197-201.

Dirkzwager, A. (2003). Multiple Evaluation: A New Testing Paradigm That Exorcizes Guessing. *International Journal of Testing*, 3(4), 333-352.

Dudycha, A.L., & Carpenter, J. B. (1973). Effects of item formats on item discrimination and difficulty. *Journal of Applied Psychology*, 58, 116 - 121.

Flaugher, R.L., Melton, R.S., & Myer, C.T. (1968). Item rearrangement under typical test conditions. *Educational and Psychological Measurement*, 28, 813-824.

- Földényi, M., Tagwerker-Neuenschwander, F., Giovanoli, A., Schallberger, U., & Steinhausen, H.-C. (1999). Die Aufmerksamkeitsleistungen von 6–10-jährigen Kindern in der TAP [Attentional performance of 6–10-year-old children on the TAP]. *Zeitschrift für Neuropsychologie*, 10, 87–102.
- Formann, A. K., & Piswanger, K. (1979). *Viennese Matrizen test. Ein Rasch-skaliertes sprachfreier Intelligenztest*. Weinheim: Beltz Test.
- Glas, C. A. W. & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (eds.), *Rasch models* (pp. 69–95). New York: Springer.
- Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly*, 50 (3), 379 – 390.
- Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, 2, 51–78.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15 (3), 309–334.
- Hohensinn, C. & Kubinger, K. D.. (2011). Applying Item Response Theory Methods to Examine the Impact of Different Response Formats. *Educational and Psychological Measurement*, 71, 732 – 746.
- Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E. & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation: An international Journal on Theory and Practice*, 17 (6), 497 – 509.
- Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, 57 (5), 253 – 270.
- Huang, Y.-M., Trevisan, M. & Storfer, A. (2007). The Impact of the “all-of-the-above” Option and Student Ability on Multiple Choice Tests. *International Journal for the Scholarship of Teaching and Learning*, 1 (2), 1–13.
- In’ nami, Y. & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, 26 (2), 219–244.
- Jenner, L. (2012). *Multiple Evaluation. Pilotstudie zu einem neuen Antwortformat*. Unveröff. Dipl. Arbeit, Universität, Wien.
- Khorramdel, L. & Frebort, M. (2011). Context Effects on Test Performance – What About Test Order?. *European Journal of Psychological Assessment*, 27 (2), 103 – 110.
- Kubinger, K. D. (2003 a). Adaptives Testen. In Kubinger, K. D. & Jäger, R. S. (Hrsg.). *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 1–9). Berlin: Beltz.

Kubinger, K. D. (2003 b). Testtheorie, Probabilistische. In Kubinger, K. D. & Jäger, R. S. (Hrsg.). *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 415-423). Berlin: Beltz.

Kubinger, K. D. (2006). *Psychologische Diagnostik. Theorie und Praxis psychologischen Diagnostizierens*. Göttingen: Hogrefe.

Kubinger, K.D. (2009a). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement*, 69, 232–244.

Kubinger, K.D. & Draxler, C. (2006). A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In M. von Davier & C.H. Carstensen (Hrsg.), *Multivariate and Mixture Distribution Rasch Models - Extensions and Applications* (S. 295-312). New York: Springer.

Kubinger, K. D. & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on different item response formats – an experiment in fundamental research on psychological assessment. *Psychology Science*, 49 (4), 361-374.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C. & Frebort, M. (2010). On Minimizing Guessing Effects on Multiple-Choice Items: Superiority of a two solutions and three distractors item format to a one solution and five distractors format. *International Journal of Selection and Assessment*, 18 (1), 111 – 115.

Kubinger, K. D., Reif, M. & Yanagida, T. (2011). Branched adaptive testing with a Rasch-model-calibrated test: analysing item presentation's sequence effects using the Rasch-model-based LLTM. *Educational Research and Evaluation: An international Journal on Theory and Practice*, 17 (5), 373 – 385.

Kubinger, K. D. & Wurst, E. (2000). *Adaptives Intelligenz Diagnostikum – Version 2.1 (AID 2)*. Göttingen: Beltz.

Liu, O. L., Lee, H.-S. & Linn, M. C. (2011). An Investigation of Explanation Multiple-Choice Items in Science Assessment. *Educational Assessment*, 16 (3), 164-184.

Lord, F. M. (1944). Reliability of multiple choice tests as a function of number of choices per item. *Journal of Educational Psychology*, 35, 175–180.

Mair, P., Hatzinger, R. & Maier, M. (2011). eRm: Extended Rasch Modeling [online]. <http://r-forge.r-project.org/projects/erm/> [04.04.2013]

Meijer, R. R. (2003). Diagnosing Item Score Patterns on a Test Using Item Response Theory-Based Person-Fit Statistics. *Psychological Methods*, 8 (1), 72 – 87.

Meyers, J. L., Miller, G. E. & Way, W. D. (2008). Item Position and Item Difficulty Change in an IRT-Based Common Item Equating Design. *Applied Measurement in Education*, 22 (1), 38 – 60.

Musch, J. (2009). *Multiple Evaluation statt Multiple Choice: Zur Verbesserung der Wissensdiagnostik durch die Nutzung der Antwortkonfidenz*. Positionsreferat zur 10. Arbeitstagung der Fachgruppe Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik.

Ortner, T. M. (2004). On changing the position of items in personality questionnaires: Analysing effects of item sequence using IRT. *Psychology Science*, 46 (4), 466-476.

Poinstingl, H. (2009 a). *Remote Testing*. Referat anlässlich der General Online Research 2009 – Conference.

Poinstingl, H. (2009 b). The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychology Science Quarterly*, 51 (2), 123 – 134.

Poinstingl, H. (2010). *Der Remote Testing Ansatz am Beispiel des Zahlenreihentests ZART*. Referat anlässlich der 9. Tagung der österreichischen Gesellschaft für Psychologie (Salzburg).

Ponocny, I. & Klauer, K. C. (2002). Towards identification of unscalable personality questionnaire respondents: The use of person fit indices. *Psychologische Beiträge*, 44 (1), 94 – 107.

Prieto, G. & Delgado, A. R. (1999). The Effect of Instructions on Multiple-Choice Test Scores. *European Journal of Psychological Assessment*, 15 (2), 143 – 150.

R Development Core Team (2008). *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing (URL <http://www.R-project.org>).

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Raven, J. C., Court, C., Raven, J. & Kratzmeier, H. (1990). *Standard Progressive Matrices: [SPM]*. Weinheim: Beltz Test.

Revuelta, J. (2004). Analysis of Distractor Difficulty in Multiple-Choice Items. *Psychometrika*, 69 (2), 217-234.

Rippey, R. M. (1970). A Comparison of Five Different Scoring Functions for Confidence Tests. *Journal of Educational Measurement*, 7 (3), 165-170.

Rodriguez, M. C. (2003). Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations. *Journal of Educational Measurement*, 40 (2), 163-184.

Rodriguez, M. C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice*, 24 (2), 3-13.

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

Scharf, E. M. & Baldwin, L. P. (2007). Assessing multiple choice question (MCQ) tests – a mathematical perspective. *Active Learning in Higher Education*, 8 (1), 31-47.

- Schmotzer, C., Kubinger, K. D. & Maryschka, C. (1994). *Rechnen in Symbolen. Manual*. Frankfurt/M.: Swets Test Service.
- Schweizer, K. , Schreiner, M. & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: A two-dimensional Model of APM. *Psychology Science Quarterly*, 51 (1), 47 – 64.
- Seiwald, B. B. (2003 a). Antwortformat. In Kubinger, K. D. & Jäger, R. S. (Hrsg.). *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 23-28). Berlin: Beltz.
- Seiwald, B. B. (2003 b). Antworttendenzen. In Kubinger, K. D. & Jäger, R. S. (Hrsg.). *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 29-32). Berlin: Beltz.
- Shizuka, T., Takeuchi, O., Yashima, T. & Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing*, 23 (1), 35-57.
- Skinner, N. F. (1999). When the Going Gets Tough, the Tough Get Going: Effects of Order of Item Difficulty on Multiple-Choice Test Performance. *The North American Journal of Psychology*, 1 (1), 79 – 82.
- Solheim, O. J. (2011). The Impact of Reading Self-Efficacy and Task Value on Reading Comprehension Scores in Different Item Formats. *Reading Psychology*, 32 (1), 1-27.
- Srp, G. (1994). *Syllogismen. Manual*. Frankfurt/M.: Swets Test Services.
- Süß, H.-M. (2003 a). Intelligenztheorien. In Kubinger, K. D. & Jäger, R. S. (Hrsg.). *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 216-224). Berlin: Beltz.
- Süß, H.-M. (2003 b). Culture fair. In Kubinger, K. D. & Jäger, R. S. (Hrsg.). *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 82-86). Berlin: Beltz.
- Taylor, C. S. & Lee, T. (2011). Ethic DIF in Reading Tests With Mixed Item Formats. *Educational Assessment*, 16 (1), 35-68.
- Taylor, C. S. & Lee, T. (2012). Gender DIF in Reading and Mathematics Test With Mixed Item Formats. *Applied Measurement in Education*, 25 (3), 246-280.
- Teubenbacher, N. (2009). *Korreliert ein höherer Rateeffekt mit einem niedrigeren Person-fit Index?*. Unveröff. Dipl. Arbeit, Universität, Wien.
- Thissen, D., Wainer, H. & Wang, X. (1994). Are Tests Comprising Both Multiple-Choice and Free-Response Items Necessarily Less Unidimensional than Multiple-Choice Tests? An Analysis of Two Tests. *Journal of Educational Measurement*, 31 (2), 113-123.

Vock, K. (2009). *Eine experimentelle Studie zur Itemschwierigkeit von Freiem Antwortformat sowie "1 aus 4"-, "1 aus 6"- und "x aus 5"-MC-Antwortformat.* Unveröff. Dipl. Arbeit, Universität, Wien.

Zimbardo, P. G. & Gerring, R. J. (1999). *Psychologie* (7. Auflage). Berlin: Springer.

V. Anhang

Antrag Landesschulrat

Carmen Pils
Diplomandin am Arbeitsbereich Psychologische Diagnostik
Institut für Angewandte Psychologie: Gesundheit, Entwicklung und Förderung
Fakultät für Psychologie, Universität Wien
Hans Handl-Gasse 3/2/2
3100, St. Pölten
Tel.: 0650/4588112
c.pils@gmx.net

An den
Landesschulrat für Niederösterreich
Rennbahnstraße 29
3100 St. Pölten

St. Pölten, am 15. Februar 2012

Betreff: Antragstellung zur Genehmigung einer Schüler(innen)-Testung zum Einfluss verschiedener Antwortformate bei Intelligenztests

Sehr geehrte Damen und Herren,

Zur Zeit schreibe ich meine Diplomarbeit zum Thema “Zum Einfluss verschiedener Antwortformate bei Intelligenztests“ (Betreuerin: Mag. Dr. Christine Hohensinn) an der Universität Wien. Um diese Frage zu beantworten, soll im Zentrum meiner Arbeit eine Erhebung an Schülerinnen und Schülern stehen.

Schon frühere Forschungen zu diesem Thema konnten Schwierigkeitsunterschiede zwischen verschiedenen Antwortformaten in den verschiedensten Bereichen, wie mathematischen Tests oder Leistungsüberprüfungen, feststellen. Eine Prüfung im Bereich des numerischen und verbalen logischen Schlussfolgerns steht allerdings noch aus. Diese Intelligenzbereiche sind unter anderem auch für den Mathematikunterricht und das logische Argumentieren im Fach Deutsch wichtig.

In meiner Diplomarbeit wird auf drei Antwortformate näher eingegangen:

Für das offene Antwortformat sprechen die große Flexibilität und die Einfachheit der Erstellung. Nachteilig an diesem Format ist jedoch der verhältnismäßig große Aufwand, der bei der Auswertung beziehungsweise der Verbesserung der Fragen anfällt. Im Gegensatz dazu ist die Auswertung beim Multiple Choice Format sehr viel ökonomischer und unter Umständen auch mittels Computer zu realisieren. Allerdings steht dieses Format im Verdacht zu leicht zu sein, da die getesteten Personen die richtige Antwort nur wiedererkennen müssen. In dieser Diplomarbeit soll auch noch ein drittes Antwortformat überprüft werden, nämlich das Format der Multiplen Evaluation. Dies ist ein noch kaum beforschtes Format, das von den Testpersonen verlangt, 100 % auf alle Antwortalternativen aufzuteilen, je nachdem wie sicher sie sich sind, dass die Alternativen zutreffen. Dieses Format bietet denselben

Herstellungsaufwand wie das Multiple Choice Format, ist allerdings in der Lage, auch teilweises Wissen zu honorieren.

Neue Forschung zu diesem Bereich ist sicherlich auch für den Schulalltag relevant, da auch hier im Zuge jedes Tests und jeder Schularbeit die Frage beantwortet werden muss, welches Antwortformat adäquat ist. Zudem kann die Teilnahme an einer psychologischen Testung horizontweiternd sein und auch aufzeigen, womit sich Psychologen, unter anderem im Arbeitsalltag beschäftigen.

Daher bitte ich Sie, mir eine Testung an folgenden Schulen zu ermöglichen:

Bezirk St. Pölten – Stadt:

- Bundeshandelsakademie und Bundeshandelsschule St. Pölten
- Bundesrealgymnasium und Bundesoberstufenrealgymnasium St. Pölten, Schulring

Bezirk Krems – Stadt:

- Bundesgymnasium und Bundesrealgymnasium Krems, Piaristengasse
- Bundesrealgymnasium Krems, Ringstrasse
- MARY WARD Oberstufenrealgymnasium der Vereinigung der Ordensschulen Österreichs, Krems
- Bundeshandelsakademie und Bundeshandelsschule Krems
- Bundesgymnasium Krems, Rechte Kremszeile

In der Hoffnung auf Ihre Unterstützung verbleibe ich mit freundlichen Grüßen,

Carmen Pils

**Anhang: 1. Untersuchungsplan und Testbeschreibung
 2. Elternbrief**

Untersuchungsplan und Testbeschreibung

Numerisches und verbales logisches Schlussfolgern werden mit 2 Tests überprüft:

1. Zahlenreihentest (ZART; Poinstingl, H., unpubliziert)

Hier wird das schlussfolgernde Denken in Bezug auf numerisches Material getestet. Es werden Zahlenreihen vorgegeben, die nach bestimmten Regeln gebildet wurden. Der Testteilnehmer muss nun diese Regeln herausfinden und richtig weiter anwenden. Die Vorgabe des ZART dauert 15 Minuten und es sollen maximal 22 Fragen beantwortet werden.

Ein Beispiel:

1 3 (?) 7 9 11 13

Die Regel lautet **Addition mit +2**. Die Zahl vor dem Fragezeichen ist 3. Damit ist die gesuchte Zahl: $3 + 2 = 5$

2. Family Relations Reasoning Test (FRRT; Poinstingl, Kubinger, Skoda & Schechtner, in Vorb.)

In diesem Test wird das schlussfolgernde Denken in Bezug auf verbales Material überprüft. Es werden Familienbeziehungen vorgegeben und die Testperson muss aufgrund dieser Vorinformation das Verwandtschaftsverhältnis zweier Personen schlussfolgern. Hier dauert die Vorgabe 20 Minuten und es werden wiederum 22 Fragen vorgegeben.

Ein Beispiel:

Klara ist die Tochter von Sonja. Sonja ist die Tochter von Maria.

Maria ist die _____ von Klara.

Hier wäre die richtige Lösung: Maria ist die **Großmutter** von Klara.

Um einen etwaigen Unterschied zwischen verschiedenen Formaten feststellen zu können, wird jeder der beiden Tests mit drei verschiedenen Antwortformaten vorgegeben. Zu diesem Zweck wurden drei Testformen erstellt, die dieselben Fragen aufweisen, sich jedoch in den Antwortformaten unterscheiden.

Folgende Antwortformate werden vorgegeben:

- a) *offenes Format*: Hier muss die Testperson die Antworten frei produzieren.
- b) *Multiple Choice*: In diesem Fall werden 5 Antwortmöglichkeiten vorgegeben, aus denen die richtige herauszufinden ist.
- c) *Multiple Evaluation*: Hier muss jeder der 5 Antwortmöglichkeiten ein Prozentwert zugeordnet werden, je nachdem wie sicher sich die Testperson ist, dass die fragliche Antwort richtig ist.

Zusätzlich wird mittels dreier Fragen die Ausprägung auf der Skala Impulsivität/Reflexivität grob erhoben, um etwaige Zusammenhänge mit den unterschiedlichen Antwortformaten feststellen zu können.

Durch die Zeitbeschränkungen der beiden Tests kann die gesamte Erhebung **in einer Unterrichtsstunde abgehalten** werden und findet **während der Unterrichtszeit** statt. Die Testung kann in **Gruppenvorgabe** abgehalten werden, was bedeutet, dass es auch möglich ist ganze Klassen zu testen. Die Erhebung findet in von der Schule zur Verfügung gestellten Räumen statt, diese benötigen an Ausstattung lediglich Tische und Bänke. Die teilnehmenden Schüler(innen) sollten die Oberstufe eines Gymnasiums oder eine Berufsbildende Höhere Schule besuchen und deren **Eltern müssen das Einverständnis zur Testung gegeben haben**. Im Sinne der Analyse der Ergebnisse, ist geplant ca. **200 Schüler** zu testen.

Fragen zur Person der Schüler beschränken sich auf das Geschlecht, das Alter, die Schule und Klasse, daher kann auch die erforderliche **Anonymität gewährleistet** werden.

Die Termine werden abgesprochen, sobald feststeht, welche Jugendlichen das Einverständnis ihrer Eltern erhalten haben.

Als Testleiter(innen) werden Diplomand(inn)en des Arbeitsbereiches Psychologische Diagnostik der Universität Wien fungieren, die in der Vorgabe der Tests geschult sind.

Elternbrief

Liebe Eltern!

Bei sämtlichen Tests, die Kinder, Jugendliche und schließlich auch Erwachsene bearbeiten, steht am Beginn die Konstruktion dieser Tests. Ein Aspekt, der im Zuge dessen beachtet werden muss, ist die Art der Antwortformate. Hiervon gibt es zum Beispiel das offene Antwortformat, bei dem auf eine Frage mit selbst produziertem Wissen geantwortet werden muss. Des Weiteren wird das Multiple Choice Format angewendet, bei dem aus einer bestimmten Anzahl von Antwortalternativen die richtige herausgefunden werden muss. Ein drittes Format stellt die Multiple Evaluation dar, bei der auf alle Antwortalternativen 100 % vergeben werden, je nachdem wie sicher man sich bei der Richtigkeit der jeweiligen Alternative ist. Alle diese Antwortformate haben ihre Vor- und auch Nachteile, allerdings ist schon in vielen Bereichen erwiesen worden, dass sie sich in ihrer Schwierigkeit unterscheiden, da sie unterschiedliche Anforderungen stellen. Eine solche Prüfung steht für das verbale und numerische logische Schlussfolgern, das zum Beispiel in den Fächern Mathematik und Deutsch höchst wichtig ist, jedoch noch aus.

Daher möchte ich **im Rahmen meiner Diplomarbeit an der Universität Wien** (Betreuung: Mag. Dr. Christine Hohensinn) genau diese Frage klären. Im Zuge dessen findet eine **Schüler(innen)-Testung zum Einfluss verschiedener Antwortformate auf die Leistung im logischen Schlussfolgern** statt.

Hierzu würde ich Sie gerne um Unterstützung und daher um **Ihre Einwilligung** bitten, dass Ihre Tochter / Ihr Sohn an der Schüler(innen)-Testung zum Einfluss verschiedener Antwortformate teilnehmen darf.

Die Teilnahme an der Testung ist selbstverständlich **freiwillig** und auch bei einer Nichtteilnahme entstehen keine Nachteile. Natürlich kann eine einmal gegebene Einwilligung jederzeit widerrufen werden.

Die Untersuchung wird **während der Schulzeit** durchgeführt und wird **eine Unterrichtsstunde** in Anspruch nehmen.

Für Rückfragen stehe ich jederzeit zur Verfügung:

Carmen Pils

Tel.: 0650/4588112

e-Mail: c.pils@gmx.net

Die Ergebnisse der Testung werden im Hinblick auf den Datenschutz **anonymisiert** und ausschließlich zu wissenschaftlichen Zwecken verwendet. Des Weiteren werden sie in keinem Fall an die Direktion oder die Lehrer(innen) weitergegeben.

Ich würde Ihre Tochter / Ihren Sohn sehr gerne in die Testung mit einbeziehen und wende mich daher mit der Bitte an Sie Ihr Kind teilnehmen zu lassen, vorausgesetzt Ihre Tochter / Ihr Sohn ist einverstanden.

In diesem Sinne möchte ich Sie höflichst ersuchen, mit Ihrer Unterschrift auf dem beiliegenden Formular Ihr Einverständnis zur Teilnahme Ihrer Tochter / Ihres Sohnes an der oben beschriebenen wissenschaftlichen Untersuchung zu erteilen.

Herzlichen Dank im Voraus!

Carmen Pils

Diplomandin am Arbeitsbereich Psychologische Diagnostik

Institut für Entwicklungspsychologie und Psychologische Diagnostik, Fakultät für Psychologie

Universität Wien

Ich erkläre mich mit der Teilnahme meiner Tochter / meines Sohnes

_____, geboren am _____, an der
Name des Schülers Geburtsdatum

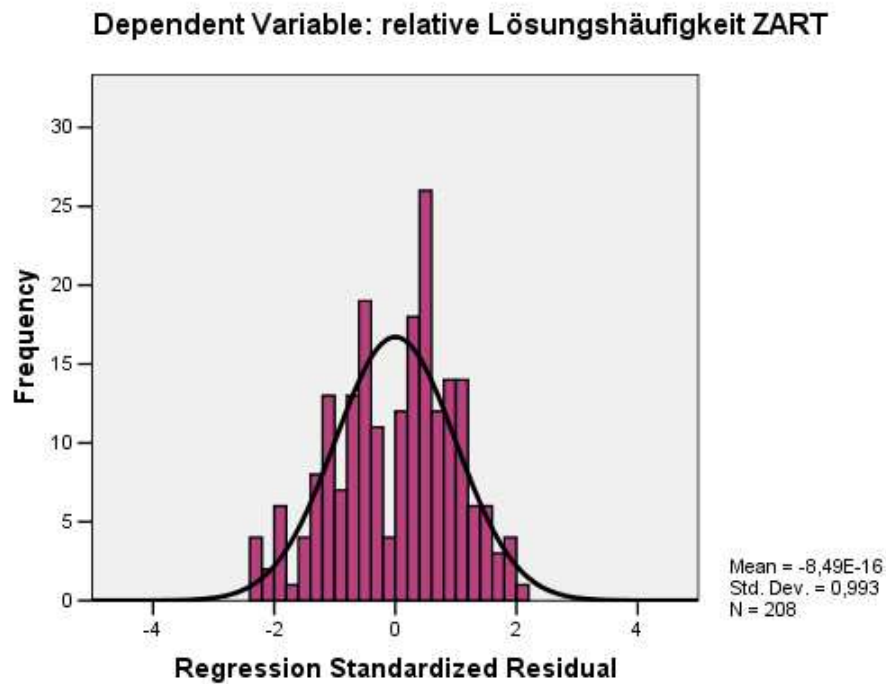
Schüler(innen)-Testung zum Einfluss von Antwortformaten auf die Leistung im logische Schlussfolgern einverstanden.

Datum

Unterschrift der/des Erziehungsberechtigten

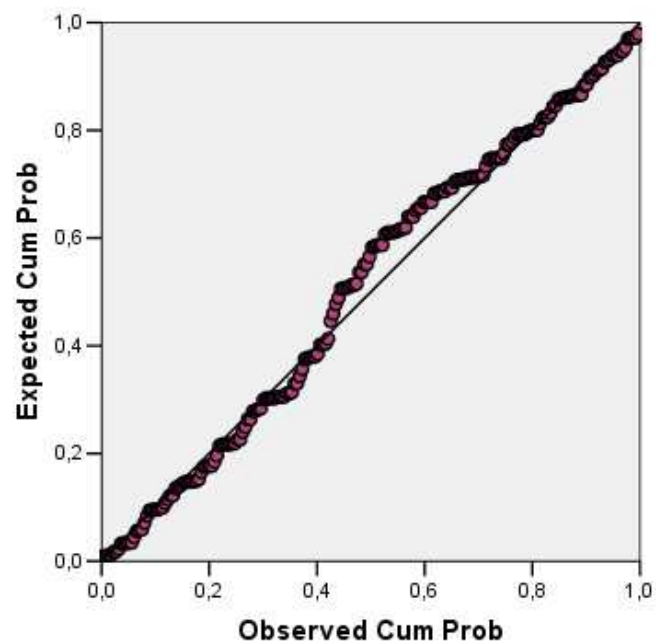
Diagramme der Regression von Impulsivität und relativen Lösungshäufigkeiten

Histogramm

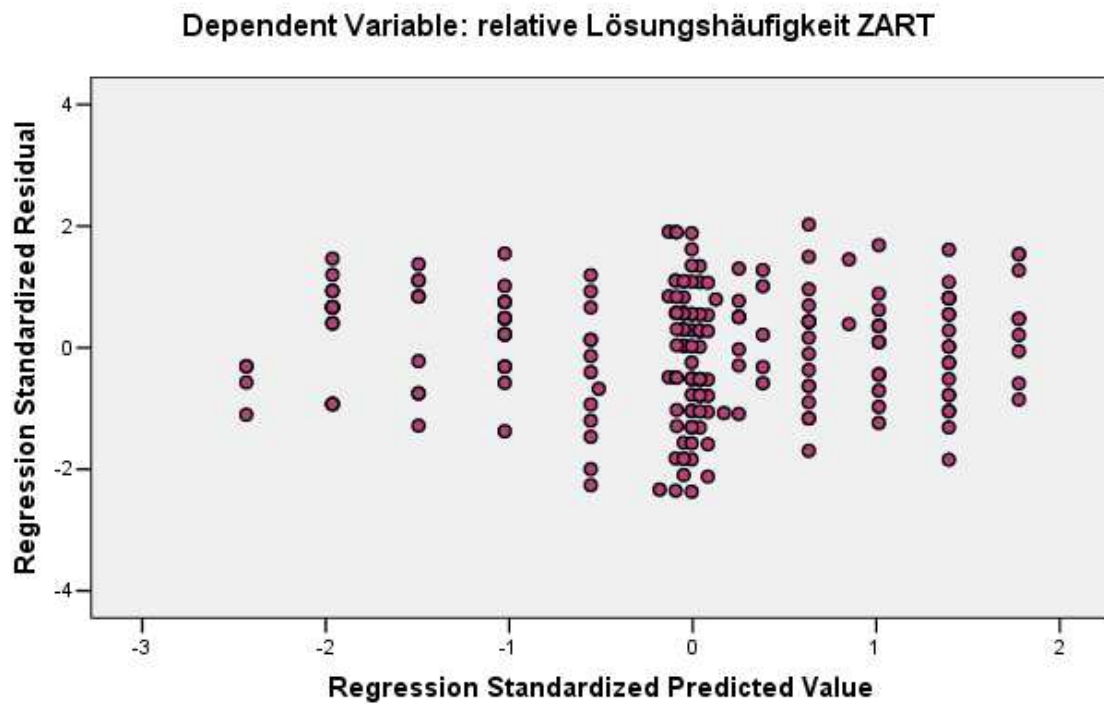


Normal P-P Plot of Regression Standardized Residual

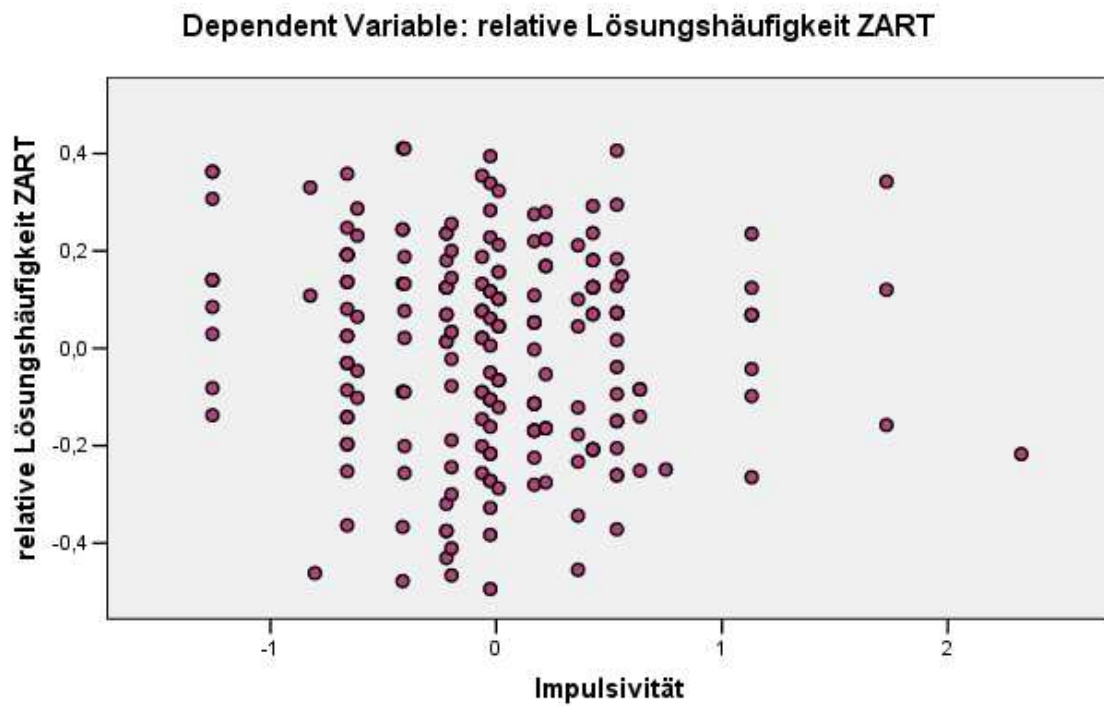
Dependent Variable: relative Lösungshäufigkeit ZART



Scatterplot

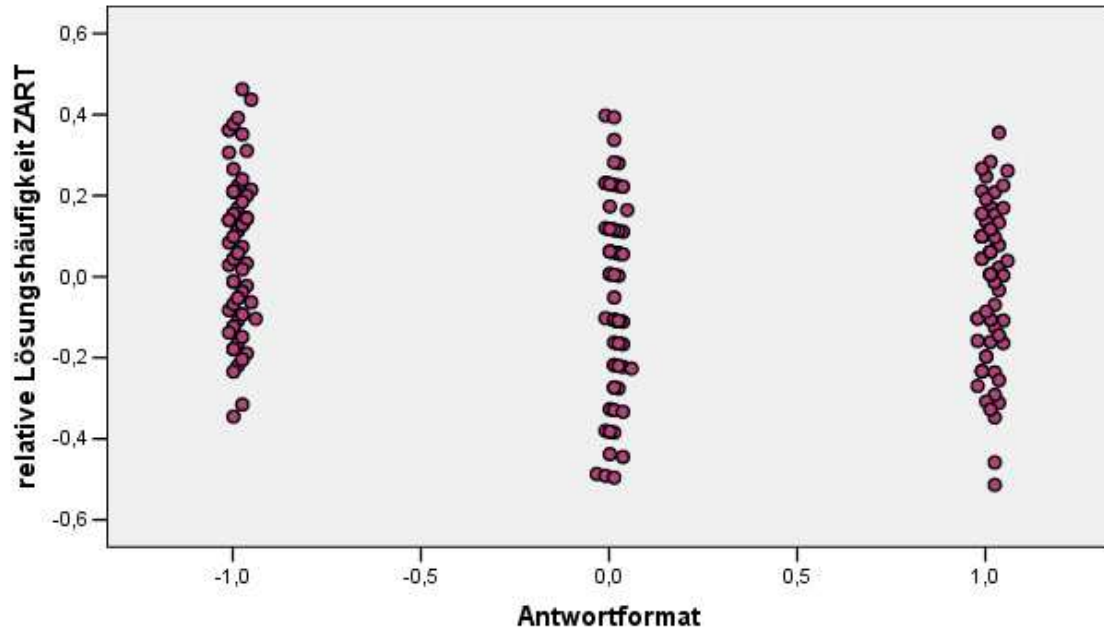


Partial Regression Plot



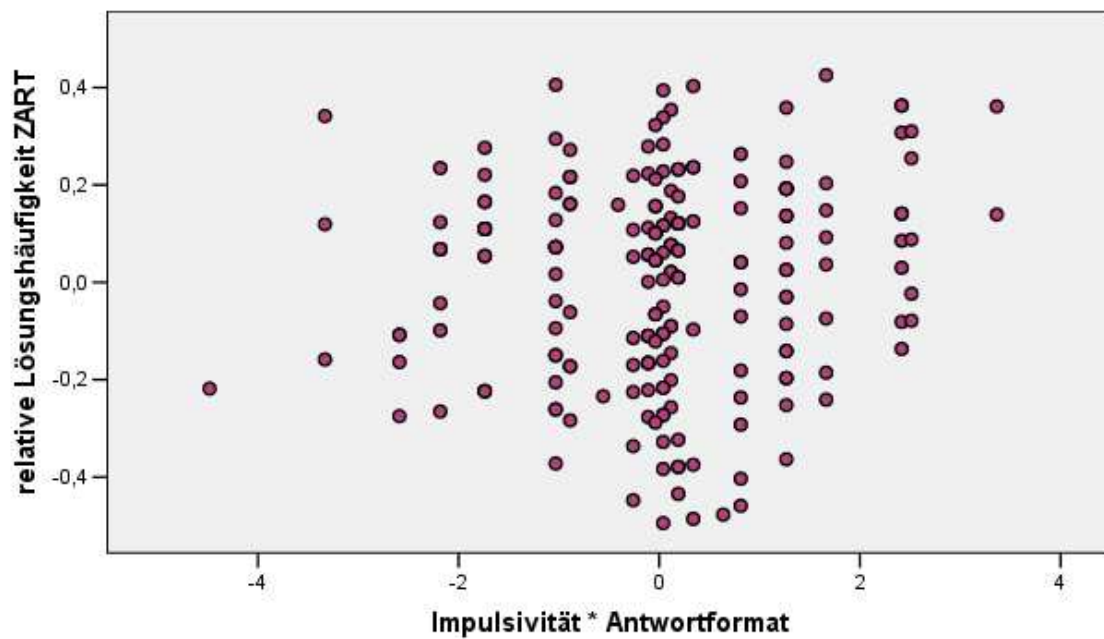
Partial Regression Plot

Dependent Variable: relative Lösungshäufigkeit ZART

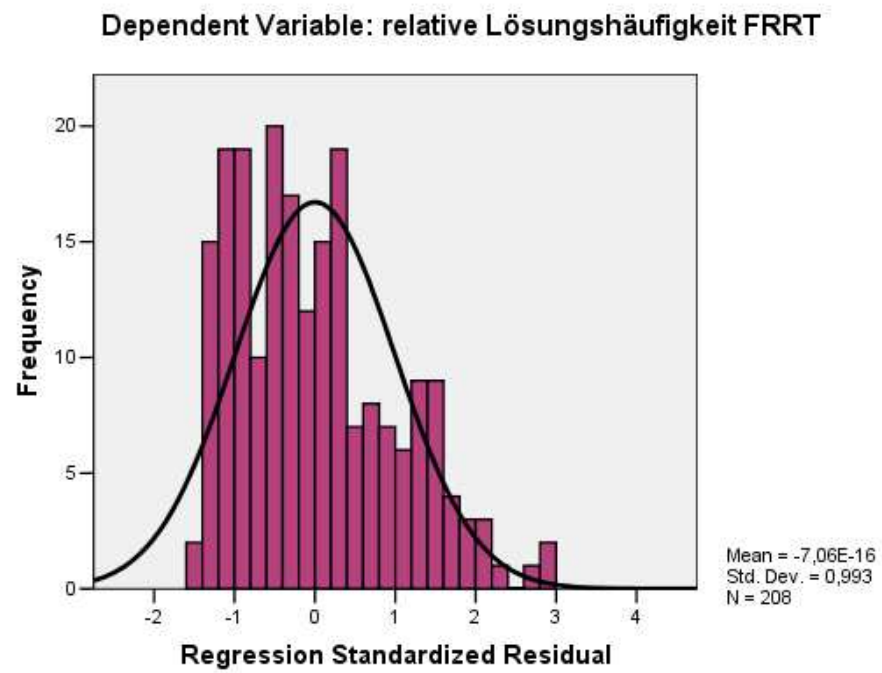


Partial Regression Plot

Dependent Variable: relative Lösungshäufigkeit ZART

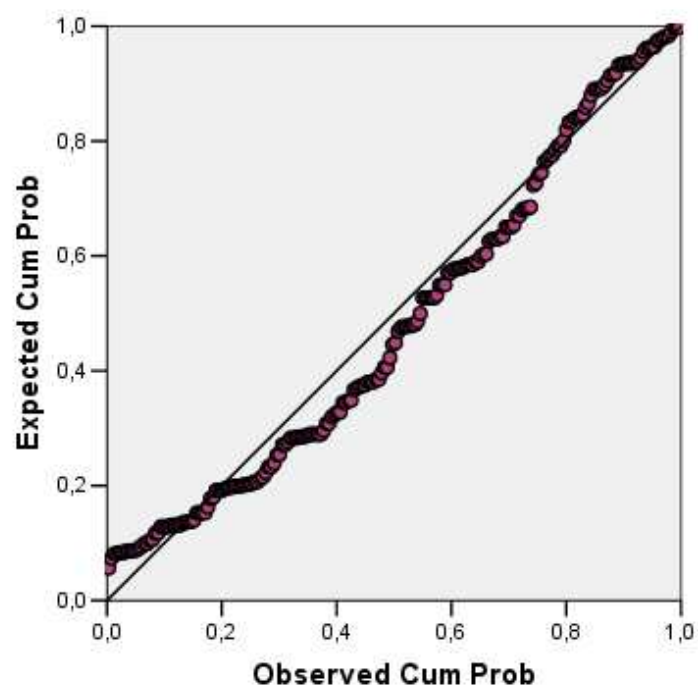


Histogram



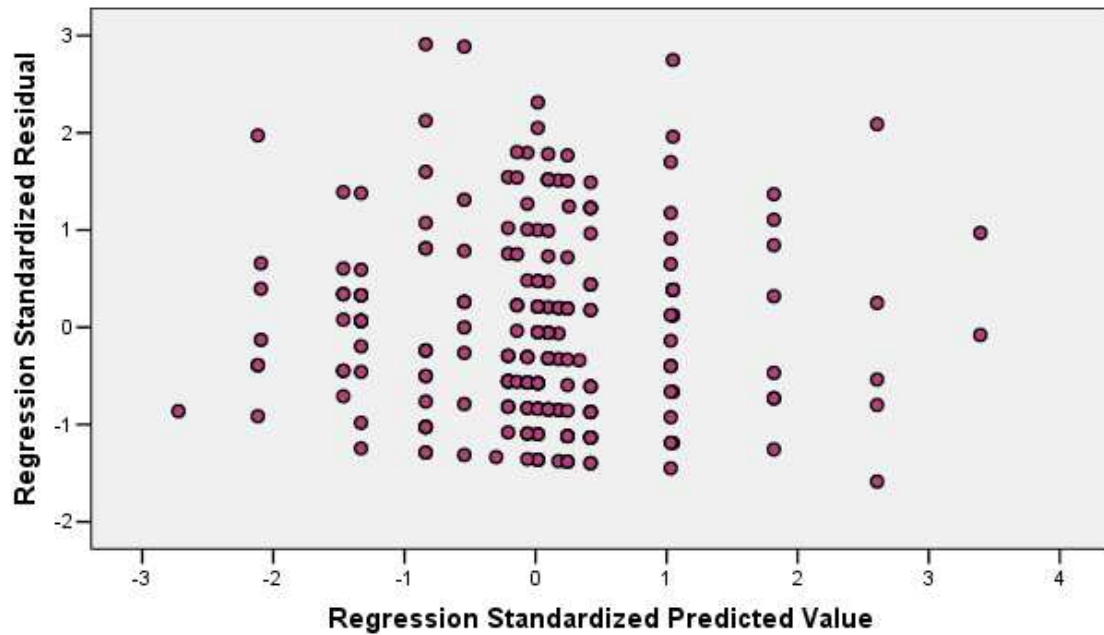
Normal P-P Plot of Regression Standardized Residual

Dependent Variable: relative Lösungshäufigkeit FRRT



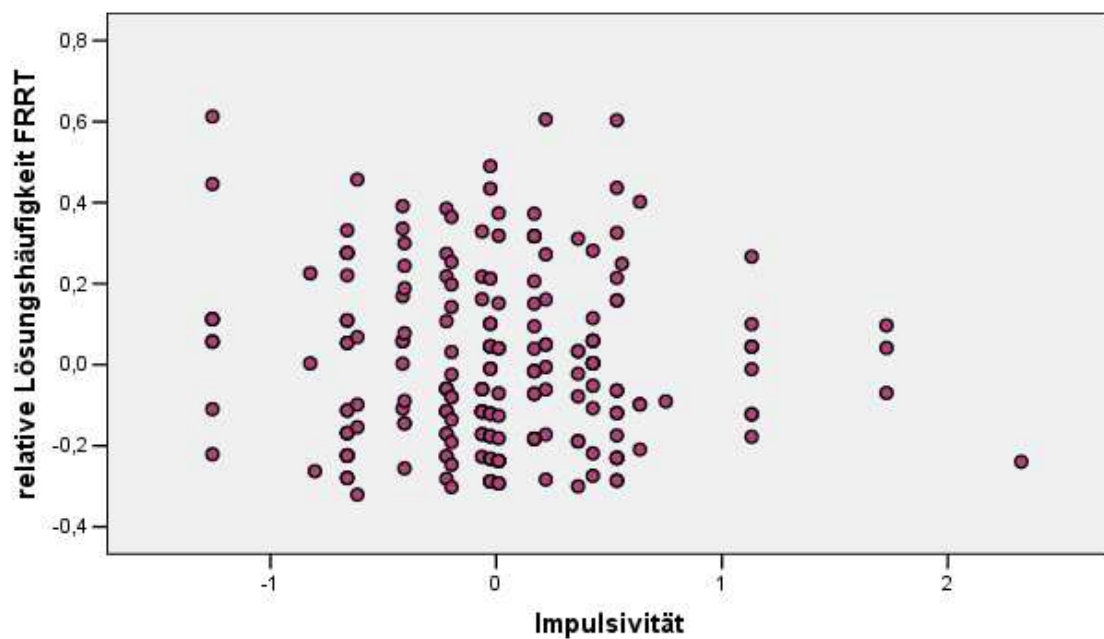
Scatterplot

Dependent Variable: relative Lösungshäufigkeit FRRT



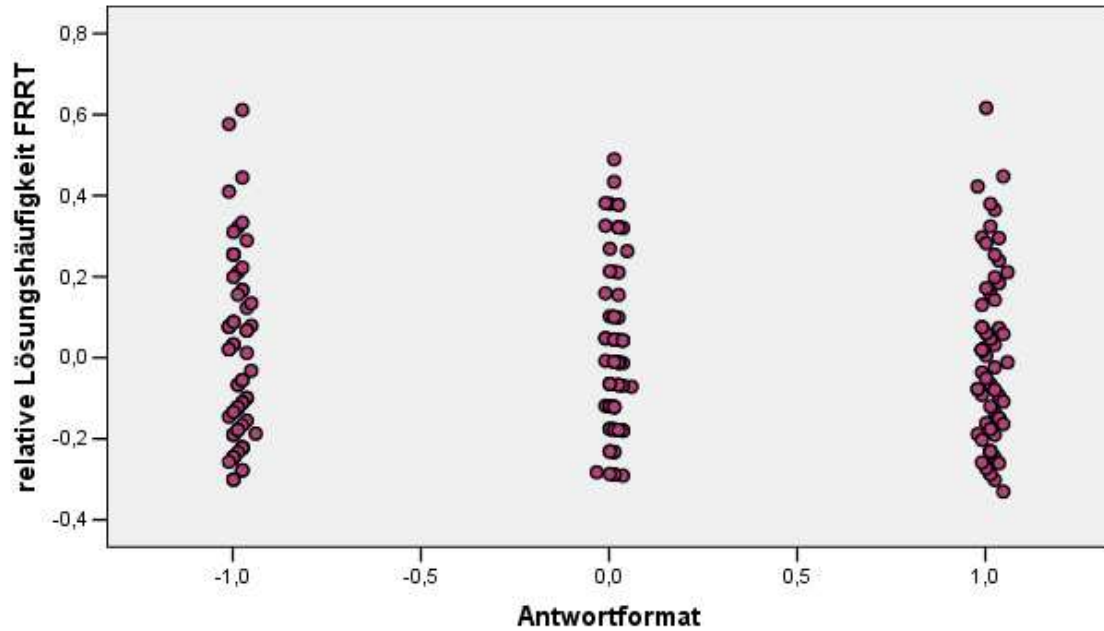
Partial Regression Plot

Dependent Variable: relative Lösungshäufigkeit FRRT



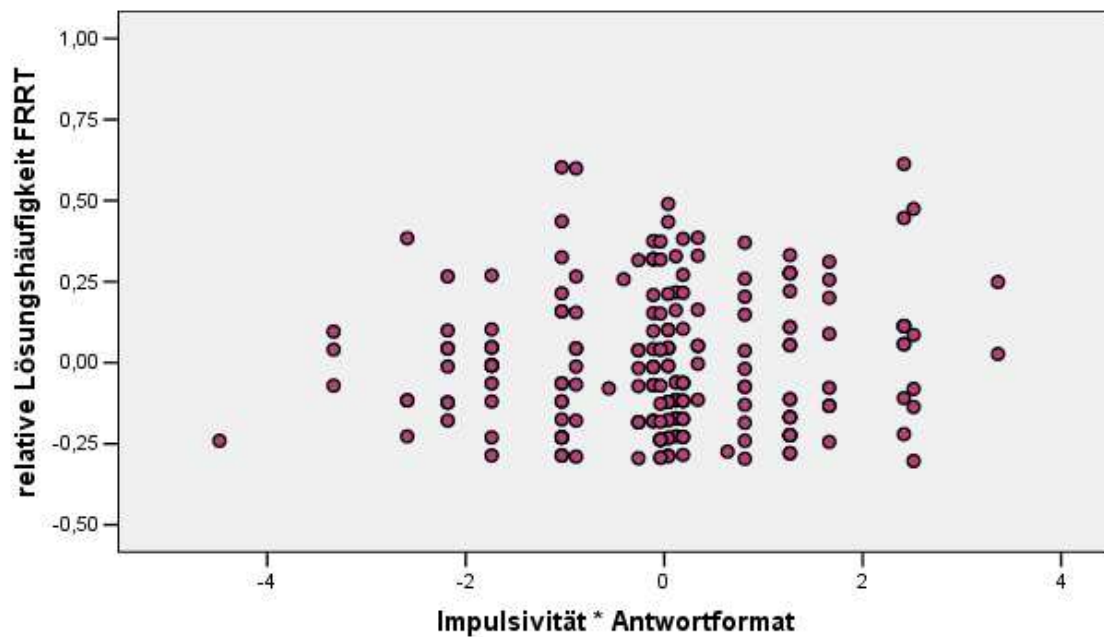
Partial Regression Plot

Dependent Variable: relative Lösungshäufigkeit FRRT



Partial Regression Plot

Dependent Variable: relative Lösungshäufigkeit FRRT



VI. Tabellenverzeichnis

	Seite
Tabelle 1: Fragen zur Feststellung der Impulsivität	39
Tabelle 2: Zusammenstellung der Testformen	42
Tabelle 3: Geschlechteraufteilung in den Schulen	44
Tabelle 4: Prozentuale Verteilung der Testpersonen hinsichtlich ihres Absolvierens der Tests und der einzelnen Antwortformate	47
Tabelle 5: Ergebnisse des Andersen Likelihood Ratio Tests für den ZART vor den Ausschlüssen	50
Tabelle 6: Ergebnisse des Andersen Likelihood Ratio Tests für den ZART nach den Ausschlüssen	54
Tabelle 7: Mauchly's Test auf Sphärizität der Itemparameter des ZART	57
Tabelle 8: Varianzanalyse der Antwortformate über die Itemparameter des ZART	58
Tabelle 9: Ergebnisse des Andersen Likelihood Ratio Tests für den FRRT vor den Ausschlüssen	59
Tabelle 10: Ergebnisse des Andersen Likelihood Ratio Tests für den FRRT nach den Ausschlüssen	63
Tabelle 11: Levene's Test über die Homogenität der Varianzen der Personenparameter des ZART	67
Tabelle 12: Varianzanalyse für unabhängige Stichproben über die Personenparameter des ZART	67

Tabelle 13:	Mauchly's Test auf Sphärizität der relativen Lösungshäufigkeiten der Linking-Items des FRRT	69
Tabelle 14:	Varianzanalyse für abhängige Stichproben über die relative Lösungshäufigkeit der Linking-Items des FRRT	69
Tabelle 15:	Mauchly's Test auf Sphärizität der relativen Lösungshäufigkeiten der gelösten Items des FRRT	69
Tabelle 16:	Varianzanalyse für abhängige Stichproben über die relative Lösungshäufigkeit der gelösten Items des FRRT mit Greenhouse-Geisser Korrektur	70
Tabelle 17:	Zusammenfassung der hierarchischen Regressionsanalyse für den ZART	71
Tabelle 17:	Zusammenfassung der hierarchischen Regressionsanalyse für den FRRT	72

VII. Abbildungsverzeichnis

	Seite
Abbildung 1: Beispiel einer Item Characteristic Curve	27
Abbildung 2: Altersverteilung der SchülerInnen	45
Abbildung 3: Prozentuale Verteilung der Antwortstile in Bezug auf die 100% Taktik bei Multipler Evaluation	46
Abbildung 4: Verteilung der Impulsivitätswerte der Testpersonen	48
Abbildung 5: Graphischer Modell-Test für das Teilungskriterium Score (ZART) vor den Ausschlüssen	50
Abbildung 6: Graphischer Modell-Test für das Teilungskriterium Geschlecht (ZART) vor den Ausschlüssen	51
Abbildung 7: Graphischer Modell-Test für das Teilungskriterium Alter (ZART) vor den Ausschlüssen	51
Abbildung 8: Graphischer Modell-Test für das Teilungskriterium Schule (ZART) vor den Ausschlüssen	52
Abbildung 9: Graphischer Modell-Test für das Teilungskriterium Impulsivität (ZART) vor den Ausschlüssen	52
Abbildung 10: Graphischer Modell-Test für das Teilungskriterium Motivation (ZART) vor den Ausschlüssen	53
Abbildung 11: Graphischer Modell-Test für das Teilungskriterium Score (ZART) nach den vier MC Itemausschlüssen	54

Abbildung 12:	Graphischer Modell-Test für das Teilungskriterium Geschlecht (ZART) nach den vier MC Itemausschlüssen	55
Abbildung 13:	Graphischer Modell-Test für das Teilungskriterium Alter (ZART) nach den vier MC Itemausschlüssen	55
Abbildung 14:	Graphischer Modell-Test für das Teilungskriterium Schule (ZART) nach den vier MC Itemausschlüssen	56
Abbildung 15:	Graphischer Modell-Test für das Teilungskriterium Impulsivität (ZART) nach den vier MC Itemausschlüssen	56
Abbildung 16:	Graphischer Modell-Test für das Teilungskriterium Motivation (ZART) nach den vier MC Itemausschlüssen	57
Abbildung 17:	Mittelwertsvergleich der Itemparameter des ZART über die Testformen	58
Abbildung 18:	Graphischer Modell-Test für das Teilungskriterium Score (FRRT) vor den Ausschlüssen	59
Abbildung 19:	Graphischer Modell-Test für das Teilungskriterium Geschlecht (FRRT) vor den Ausschlüssen	60
Abbildung 20:	Graphischer Modell-Test für das Teilungskriterium Alter (FRRT) vor den Ausschlüssen	60
Abbildung 21:	Graphischer Modell-Test für das Teilungskriterium Schule (FRRT) vor den Ausschlüssen	61

Abbildung 22:	Graphischer Modell-Test für das Teilungskriterium Impulsivität (FRRT) vor den Ausschlüssen	61
Abbildung 23:	Graphischer Modell-Test für das Teilungskriterium Motivation (FRRT) vor den Ausschlüssen	62
Abbildung 24:	Grafischer Model-Test für das Teilungskriterium Score (FRRT) nach den Ausschlüssen von vier MC und zwei ME Items	63
Abbildung 25:	Grafischer Model-Test für das Teilungskriterium Geschlecht (FRRT) nach den Ausschlüssen von vier MC und zwei ME Items	64
Abbildung 26:	Grafischer Model-Test für das Teilungskriterium Alter (FRRT) nach den Ausschlüssen von vier MC und zwei ME Items	64
Abbildung 27:	Grafischer Model-Test für das Teilungskriterium Schule (FRRT) nach den Ausschlüssen von vier MC und zwei ME Items	65
Abbildung 28:	Grafischer Model-Test für das Teilungskriterium Impulsivität (FRRT) nach den Ausschlüssen von vier MC und zwei ME Items	65
Abbildung 29:	Grafischer Model-Test für das Teilungskriterium Motivation (FRRT) nach den Ausschlüssen von vier MC und zwei ME Items	66

Abbildung 30:	Mittelwertsvergleich der Personenparameter des ZART über die Testformen	67
Abbildung 31:	Mittelwertsvergleich der relativen Lösungshäufigkeiten der Linking-Items des FRRT	68

Lebenslauf

1. Persönliche Daten

Name: Carmen Pils
Anschrift: Hans Handl-Gasse 3/2/2,
A-3100 St. Pölten
Telefon: +43/650/4588112
E-Mail: c.pils@gmx.net

Staatsangehörigkeit: Österreich
Geburtsdatum und -ort: 11.12.1985, St. Pölten
Familienstand: ledig



2. Ausbildung

Oktober 2008	Erste Diplomprüfung Psychologie
seit Oktober 2004	Diplomstudium der Psychologie an der Universität Wien
2000 – 2004	BORG St. Pölten (Abschluss: Matura)
1996 – 2000	BRG St. Pölten
1992 – 1996	Volksschule der Englischen Fräulein St. Pölten

3. Berufserfahrung

Juni 2011	Anwesenheit bei Nachschulungskurs für Alkoholauffällige Lenker des Kuratoriums für Verkehrssicherheit
Februar 2011	Anwesenheit bei diversen Verkehrspsychologischen Untersuchungen des Kuratoriums für Verkehrssicherheit
Juli, August 2010	6-Wochen-Praktikum WIFI Berufsinformationszentrum St. Pölten
seit November 2005	Feinkost Interspar St. Pölten (geringfügige Beschäftigung)
August 2004/2005/2006 Juli 2002/2003	Ferialpraktikum Allgemein Gemeinnützige Wohnungsgenossenschaft St. Pölten