



universität
wien

Diplomarbeit

Titel der Arbeit

Entwicklung des Tests „Gleichungen“
zur Erfassung von Reasoning

Verfasserin

Claudia Gamsjäger

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, im Juli 2012

Studienkennzahl: 298

Studienrichtung: Psychologie

Betreuer: Univ.-Prof. Dr. Mag. Klaus D. Kubinger

Danksagung

An erster Stelle möchte ich mich bei Herrn Univ.-Prof. Dr. Mag. Klaus D. Kubinger für die Betreuung dieser Arbeit und das entgegengebrachte Vertrauen bedanken.

Ein großer Dank gebührt auch meiner Subbetreuerin, Fr. Mag. Bettina Hagenmüller, die mich das letzte Jahr über in jeder Hinsicht und zu jedem Zeitpunkt unterstützt hat.

Bei meinen Eltern, Josef und Christiane, als auch bei meinem Bruder Christoph möchte ich mich für die jahrelange Unterstützung, nicht nur während des Studiums, von ganzem Herzen bedanken.

An dieser Stelle möchte ich auch Brigadier Reis und Oberst Liebmann sowie dem Lehrpersonal der SIAK Traiskirchen und SIAK Ybbs danken, die mir die Durchführung der Testungen ermöglichten.

Rafaela Lorenz danke ich für das Korrekturlesen dieser Arbeit, die moralische und tatkräftige Unterstützung während meines Studiums und insbesondere während der Entstehung dieser Arbeit.

Danken möchte ich auch Angelika Hackl für ihre moralische Unterstützung und ihre Freundschaft.

Abstract (Deutsch)

Der Test *Gleichungen* soll die Fähigkeit zum logisch-schlussfolgernden Denken (*Reasoning*) mit numerischem Aufgabenmaterial bei Personen ab 18 Jahren erfassen. Die Personen müssen dabei mathematische Gleichungen, die zwei Variablen enthalten, mithilfe eines Multiple-Choice Antwortformats lösen. Auf Basis eines eigens entwickelten Konstruktionsrationalis wurden 44 Items generiert. Die Verfahren *Gleichungen*, *AN-TOP* (Kubinger & Heuberger, in Vorb.) und *figurale Analogien* (Ünal, in Vorb.) wurden insgesamt 311 Auszubildenden der Sicherheitsakademien der Polizei in Niederösterreich vorgegeben. Der Test *AN-TOP* beansprucht ebenfalls *Reasoning* zu erfassen, jedoch unter Verwendung eines unterschiedlichen Aufgabenmaterials, und dient in vorliegender Arbeit zur Validierung der *Gleichungen*. Die anschließenden *Rasch*-Modell-Analysen zeigten, dass für den Itempool *Gleichungen*, nach sukzessivem Ausschließen von 11% der Items, *a-posteriori* Modellgültigkeit angenommen werden kann. Der Itempool *AN-TOP* erwies sich nach Ausscheiden von 7% der Items als *Rasch*-Modell-konform. Die Übereinstimmungsvalidität ist mit $r = 0,43$, zwischen *Gleichungen* und *AN-TOP* nicht gegeben.

Schlüsselwörter: *Reasoning*, *Rasch*-Modell, regelgeleitete Itemkonstruktion, Gleichungen

Abstract (English)

The test *Gleichungen* was developed to measure numeric reasoning in adults. The testees have to solve equations, including two (unknown) variables, with the aid of multiple choice response options. 44 items were generated by using a rule-based item construction. In sum 311 individuals, trainees of the Federal Security Academy in Lower Austria, were tested with the tests *Gleichungen*, *AN-TOP* (Kubinger & Heuberger, in process) and *figurale Analogien* (Ünal, in process). *AN-TOP* is a test to measure reasoning as well, using different item material, and serves in the context of the current paper for validation of *Gleichungen*. Rasch model analyses show that after excluding 11 percent of all items together, the items of *Gleichungen* fit *a-posteriori* to the Rasch model. The items of *AN-TOP* fit to the Rasch model after 4 items were removed. Concurrent validity ($r = 0.43$) between *Gleichungen* and *AN-TOP* does not exist.

Keywords: reasoning, Rasch model, rule-based item construction, equations

Inhaltsverzeichnis

I.	EINLEITUNG	15
II.	THEORETISCHER TEIL	17
1.	Intelligenztheoretische Grundlage	17
1.1.	Primärfaktoren-Modell der Intelligenz von Thurstone.....	17
1.2.	Theorie der fluiden und kristallisierten Intelligenz von Cattell.....	18
2.	Das Konstrukt <i>Reasoning</i>	19
3.	Gleichungen.....	20
4.	Antwortformat	21
4.1.	Verschiedene Antwortformate.....	21
4.2.	Ratewahrscheinlichkeit beim Multiple-Choice-Format	21
5.	Regelgeleitete Itemkonstruktion	24
6.	Testtheoretische Grundlage	25
6.1.	Item-Response-Theorie	25
6.2.	Das <i>Rasch</i> -Modell	25

III. EMPIRISCHER TEIL.....	29
7. Ziel der Untersuchung und Hypothesen	29
8. Entwicklung des Verfahrens <i>Gleichungen</i>	30
8.1. Definition der zu erfassenden Fähigkeit	30
8.2. Geltungsbereich und Testformat.....	30
8.3. Aufgabenmaterial und Antwortformat.....	30
8.4. Itemgenerierende Regeln.....	31
8.5. Itempool und Testdesign.....	32
8.6. Instruktion und Testvorgabe.....	33
9. Test und Testdesign <i>AN-TOP</i>	36
10. Datenerhebung	38
11. Stichprobenbeschreibung.....	40
12. Darstellung der Ergebnisse	42
12.1. Ergebnisse der <i>Rasch</i> -Modell-Analysen des Tests <i>Gleichungen</i>	42
12.2. Ergebnisse der <i>Rasch</i> -Modell-Analysen des Tests <i>AN-TOP</i>	50
12.3. Übereinstimmungsvalidität zwischen <i>Gleichungen</i> und <i>AN-TOP</i>	56
13. Diskussion und Ausblick.....	57
14. Zusammenfassung	59
IV. LITERATURVERZEICHNIS.....	61
V. ANHANG	65

Abbildungsverzeichnis

Abbildung 1:	Das Instruktionsitem der <i>Gleichungen</i>	31
Abbildung 2:	Testdesign (Ausschnitt) des Verfahrens <i>Gleichungen</i> (TF=Testform)....	33
Abbildung 3:	Testinstruktion <i>Gleichungen</i>	34
Abbildung 4:	Itembeispiel des Tests <i>AN-TOP</i> (Kubinger & Heuberger, in Vorb.). Die Lösung ist 5.....	36
Abbildung 5:	Testdesign (Ausschnitt) des Verfahrens <i>AN-TOP</i> (TF=Testform).....	37
Abbildung 6:	Itembeispiel des Tests <i>AN-TOP</i> (Kubinger & Heuberger, in Vorb.). Die Lösung ist ○.	37
Abbildung 7:	Grafischer Modelltest <i>Gleichungen</i> , Teilungskriterium <i>Score</i> (alle Items).....	44
Abbildung 8:	Grafischer Modelltest <i>Gleichungen</i> , Teilungskriterium <i>Geschlecht</i> (alle Items).....	44
Abbildung 9:	Grafischer Modelltest <i>Gleichungen</i> , Teilungskriterium <i>Alter</i> (alle Items).....	45
Abbildung 10:	Das <i>Item 1</i> , des Tests <i>Gleichungen</i> , das sich als nicht <i>Rasch</i> -Modell- konform erwies.....	45
Abbildung 11:	Grafischer Modelltest <i>Gleichungen</i> , Teilungskriterium <i>Score</i> (nach dem Ausschluss von 5 Items)	48
Abbildung 12:	Grafischer Modelltest <i>Gleichungen</i> , Teilungskriterium <i>Geschlecht</i> (nach dem Ausschluss von 5 Items)	48
Abbildung 13:	Grafischer Modelltest <i>Gleichungen</i> , Teilungskriterium <i>Alter</i> (nach dem Ausschluss von 5 Items)	49
Abbildung 14:	Grafischer Modelltest <i>AN-TOP</i> , Teilungskriterium <i>Score</i> (alle Items).....	51
Abbildung 15:	Grafischer Modelltest <i>AN-TOP</i> , Teilungskriterium <i>Geschlecht</i> (alle Items).....	51
Abbildung 16:	Grafischer Modelltest <i>AN-TOP</i> , Teilungskriterium <i>Alter</i> (alle Items).....	52
Abbildung 17:	Grafischer Modelltest <i>AN-TOP</i> , Teilungskriterium <i>Score</i> (nach dem Ausschluss von 4 Items)	54
Abbildung 18:	Grafischer Modelltest <i>AN-TOP</i> , Teilungskriterium <i>Geschlecht</i> (nach dem Ausschluss von 4 Items)	55
Abbildung 19:	Grafischer Modelltest <i>AN-TOP</i> , Teilungskriterium <i>Alter</i> (nach dem Ausschluss von 4 Items)	55

Tabellenverzeichnis

Tabelle 1:	Häufigkeit der Position der Lösung von x innerhalb der Antwortmöglichkeiten (Test <i>Gleichungen</i>)	35
Tabelle 2:	Die Verteilung der getesteten Personen in Bezug auf das Alter in Jahren ($n = 331$).....	40
Tabelle 3:	Die Verteilung der getesteten Personen auf die höchste abgeschlossene Ausbildung ($n = 331$).....	41
Tabelle 4:	Ergebnisse der LRTs <i>Gleichungen</i> (alle Items).....	43
Tabelle 5:	z -Test <i>Gleichungen</i> nach Fischer und Scheiblechner, Teilungskriterium <i>Score</i>	46
Tabelle 6:	Ergebnisse der LRTs <i>Gleichungen</i> (nach Ausschluss von 5 Items).....	47
Tabelle 7:	Ergebnisse der LRTs <i>AN-TOP</i> (alle Items)	50
Tabelle 8:	z -Test <i>AN-TOP</i> nach Fischer und Scheiblechner, Teilungskriterium <i>Score</i>	52
Tabelle 9:	Ergebnisse der LRTs <i>AN-TOP</i> (nach Ausschluss von 4 Items)	54

I. EINLEITUNG

Die vorliegende Diplomarbeit beinhaltet den Ansatz logisch-schlussfolgerndes Denken anhand von mathematischen Gleichungen zu messen und basiert auf einer Idee von Herrn Univ.-Prof. Dr. Mag. Klaus D. Kubinger.

Gleichungen wurden bisher hauptsächlich in verschiedenen Rechentests und Mathematiktests zur Erfassung der Rechenfähigkeit bzw. -fertigkeit eingesetzt. *Reasoning*¹ andererseits wird mithilfe verschiedener Verfahren, klassischerweise Matrizentests als Vertreter „sprachfreier“ *Reasoning*-Tests sowie andererseits diverser sprachlicher *Reasoning*-Tests (z.B. Syllogismenaufgaben), erfasst. Die Möglichkeit *Reasoning* mit mathematischen Gleichungen als Aufgabenmaterial zu messen, wurde beispielsweise im Untertest *Rechenaufgaben* der Intelligenztestbatterie IST 2000-R (Intelligenz-Struktur-Test 2000-R, Liepmann, Beauducel, Brocke & Amthauer, 2001) verwirklicht. Dieser Untertest unterscheidet sich vom vorliegenden Verfahren jedoch durch sein Aufgabenmaterial, genauer gesagt durch die Anzahl der verwendeten Variablen, als auch durch die Art der Konstruktion und der Überprüfung der Items.

Der Ansatz logisch-schlussfolgerndes Denken mit Hilfe von Gleichungen zu erfassen, bringt es mit sich, dass *Reasoning sensu* Cattell (siehe Abschnitt 1.2.) nicht im Bereich der fluiden, sondern der kristallisierten Intelligenz erfasst wird, da zum Lösen von mathematischen Gleichungen vorangegangene Lernerfahrungen notwendig sind.

Gleichzeitig soll das neu entwickelte Verfahren hohe psychometrische Standards erfüllen. Insbesondere wird hier auf die Verrechnungsfairness² hingewiesen, welche durch Geltung des dichotomen logistischen Testmodells nach Georg Rasch als erfüllt betrachtet werden kann. Die Geltung des Modells bedeutet auch, dass das Verfahren eindimensional misst, es wird also nur eine einzige Fähigkeit erfasst.

¹ In vorliegender Arbeit werden die Begriffe *Reasoning* und logisch-schlussfolgerndes Denken synonym verwendet.

² Verrechnungsfairness bedeutet, dass „die gegebenen Verrechnungsvorschriften eines Tests zu Testwerten führen, die verhaltensadäquate Relationen wiedergeben“ (Kubinger, 2009, S. 87).

Zu Beginn dieser Arbeit werden zunächst zwei Intelligenztheorien sowie das Konstrukt *Reasoning* vorgestellt, im Anschluss daran erfolgt eine Definition von Gleichungen. Anschließend werden verschiedene Antwortformate und das damit einhergehende Problem der Ratewahrscheinlichkeit besprochen. Den Schluss des empirischen Teils dieser Arbeit bilden Grundlagen über die regelgeleitete Itemkonstruktion und über die Item-Response-Theorie, insbesondere des dichotomen logistischen Testmodell nach Georg Rasch. Der empirische Teil beginnt mit einer kurzen Darstellung des Ziels der Untersuchung und den abgeleiteten Hypothesen. Danach erfolgt eine Beschreibung der Entwicklung der *Gleichungen*, des Tests *AN-TOP* sowie der Datenerhebung und der Stichprobe. Zuletzt werden die Ergebnisse der *Rasch*-Modell Analysen hinsichtlich der Items der *Gleichungen* und des *AN-TOP* sowie das Ergebnis der Validierung dargestellt.

II. THEORETISCHER TEIL

Im theoretischen Teil dieser Arbeit werden zuerst die zugrundeliegenden Intelligenztheorien sowie das Konstrukt *Reasoning* vorgestellt. Danach erfolgt eine Beschreibung von Gleichungen. Des Weiteren wird kurz auf das Antwortformat sowie die regelgeleitete Itemkonstruktion eingegangen. Der letzte Abschnitt beinhaltet testtheoretische Grundlagen, genauer die Item-Response-Theorie sowie das dichotome logistische Testmodell nach Georg Rasch.

1. Intelligenztheoretische Grundlage

Grundlage der vorliegenden Arbeit bilden zwei bekannte Intelligenztheorien. Diese werden im Folgenden kurz dargestellt:

1.1. Primärfaktoren-Modell der Intelligenz von Thurstone

L. L. Thurstone postulierte ein Modell, welches hauptsächlich durch Multiple-Faktorenanalyse gewonnene Gruppenfaktoren beinhaltet. Zusätzlich zu seinen mittlerweile berühmten Primary Mental Abilities akzeptierte er allerdings einen allgemeinen Intelligenzfaktor („g“) auf höchster Stufe. Den Primärfähigkeiten auf zweiter Ebene galten jedoch seine Hauptbemühungen. Diese sind namentlich: Space, Number, Verbal Comprehension, Verbal Fluency, Memory, Reasoning und Perceptual Speed (vgl. dazu Jäger, 1973; Stemmler et al., 2011; Süß, 2003).

Thurstone betont, dass es sich beim Faktor Number nicht „um die Befähigung für höhere Mathematik oder allgemein für formallogisches Denken im Zahlenbereich“ (Jäger, 1973, S.79) handelt, sondern um die Fähigkeit zur Ausführung von relativ einfachen Rechenoperationen (Jäger, 1973). Die Primärfähigkeit *Reasoning* dagegen beinhaltet die Fähigkeit zur Induktion, zur Deduktion sowie zum logischen Schließen und soll nicht an eine bestimmte Art von Material gebunden sein (Jäger, 1973).

1.2. Theorie der fluiden und kristallisierten Intelligenz von Cattell

Trotz der Annahme Cattells von mehreren Faktoren menschlicher Intelligenz, wird im Folgenden nur auf die beiden Faktoren fluide und kristallisierte Intelligenz näher eingegangen (vgl. dazu Cattell, 1961; Stemmler et al., 2011; Süß, 2003).

Der Faktor fluide Intelligenz beinhaltet die Fähigkeit, sich neuen Situationen und/oder Problemen anzupassen. Er betrifft die Basisprozesse des Denkens und ist weitgehend unabhängig von früheren Lernerfahrungen, weshalb die fluide Intelligenz auch als vorrangig genetisch determiniert angesehen wird. Erfasst wird der Faktor häufig mit Aufgaben, die das Erkennen von Regeln bzw. deren Anwendung erfordern, wie zum Beispiel Matrizentests (vgl. dazu Cattell, 1961; Stemmler et al., 2011; Süß, 2003). Das Aufgabenmaterial zur Messung des logisch-schlussfolgernden Denkens ist häufig figural-abstrakt (Horn & Cattell, 1966), weshalb fluide Intelligenz auch oft mit sogenannten „kulturfreien“ Tests in Verbindung gebracht wird (siehe dazu z.B. Süß, 2003; Kubinger, 2009).

Dem gegenüber steht der Faktor der kristallisierten Intelligenz. Dieser erfasst kognitive Fertigkeiten, die sich durch vorangegangenes Lernen gebildet haben und in Folge kulturabhängig sind (Stemmler et al., 2011; Süß, 2003). Kristallisierte Intelligenz wird deshalb überwiegend durch Wortverständnis, Satzergänzung sowie Satzbildung erfasst (Stemmler et al., 2011).

2. Das Konstrukt *Reasoning*

Wie bereits in Abschnitt 1.1 ausgeführt, stellt *Reasoning* einen der sieben Primärfaktoren nach Thurstone dar. Für vorliegende Arbeit soll als Definition von *Reasoning* jedoch Folgende gelten: Kubinger (2009) leitet aus dem Verfahrensinventar zum logisch-schlussfolgernden Denken ab, dass *Reasoning* die Fähigkeit ist, „Gesetzmäßigkeiten oder logisch zwingende Zusammenhänge zu erkennen und zweckentsprechend verwerten zu können“ (S. 206).

Reasoning kann mit einer Vielzahl von psychologisch-diagnostischen Verfahren erfasst werden.

An erster Stelle seien die Matrizen Tests genannt, die wohl als typischer Repräsentant sprachfreier *Reasoning*-Tests gelten. Publierte Verfahren sind beispielsweise der Wiener Matrizen-Test (WMT, Formann & Piswanger, 1979) oder der Adaptive Matrizen Test (AMT, Hornke, Etzel & Rettig, 1997), die aufgrund des verwendeten Testmaterials, welches unabhängig von der kulturellen Lerngeschichte sein soll, den Anspruch erheben „kulturfreie“ Tests zu sein (Kubinger, 2009).

Des Weiteren sind auch sprachliche *Reasoning*-Tests im Verfahrensinventar der Psychologischen Diagnostik zu finden. Beispiele hierfür sind der Untertest *Analogien* im IST 2000-R (Intelligenz-Struktur-Test 2000-R, Liepmann, Beauducel, Brocke & Amthauer, 2001) als auch das Verfahren Syllogismen (Spr, 1994).

3. Gleichungen

„Eine Gleichung ist im Sinne der Mathematischen Logik eine Aussage über die Gleichheit zweier Dinge. Meist handelt es sich dabei um Größen, die man durch Zahlen ausdrücken kann“ (Kerr, 1999, S. 142).

Bei Gleichungen zwischen Zahlen, Variablen und Rechenausdrücken fungiert die Variable als Platzhalter bzw. als Symbol für eine noch nicht näher bezeichnete Zahl. Die Anzahl der verschiedenen Variablen innerhalb einer Gleichung ist unterschiedlich. Im Allgemeinen haben Gleichungen mit zwei Variablen als Lösungen viele Wertepaare (Kerr, 1999). Um das richtige Wertepaar zu bestimmen und somit zu einer Lösung zu kommen, steht meist eine zweite Gleichung mit denselben Variablen zur Verfügung.

Gleichungen als numerisches Aufgabenmaterial finden bereits in diversen Mathematik- und Schulleistungstests Verwendung (vgl. dazu Rettig, Hornke & Schiff, 1989). Zur Erfassung von *Reasoning* wurden sie beispielsweise im Untertest *Rechenaufgaben* des IST 2000-R (Intelligenz-Struktur-Test 2000-R, Liepmann, Beauducel, Brocke & Amthauer, 2001) eingesetzt. Die dort zu findenden Gleichungen enthalten jeweils nur eine Variable, wodurch die Gleichungen durch einfaches Ausführen der Rechenoperationen zu lösen sind. Zur Beantwortung steht ein freies Antwortformat zur Verfügung. Es wird versucht den sprachlichen Anteil durch nicht verbales Präsentieren weitgehend auszuschalten.

4. Antwortformat

Nachstehend werden die verschiedenen Antwortformate kurz beschrieben um dann auf die Ratewahrscheinlichkeit bei Multiple-Choice-Aufgaben einzugehen.

4.1. Verschiedene Antwortformate

Eine große Rolle im Zuge der Testentwicklung spielt die Wahl des Antwortformats, welches festlegt auf welchem Weg die Reaktionen der Testpersonen auf die Items festgehalten werden. Grundlegend wird zwischen einem freien und einem gebundenen Antwortformat unterschieden (Seiwald, 2003).

Das freie Antwortformat ist gekennzeichnet, durch die Möglichkeit der Testperson die Antwort auf die Aufgabe selbst zu formulieren, in Form von Sprache, Schrift oder zeichnerisch (Seiwald, 2003).

Das gebundene Antwortformat hingegen bietet der Testperson zwei oder auch mehr Möglichkeiten zur Beantwortung der Aufgaben an. Innerhalb des gebundenen Antwortformats unterscheidet man zwischen dem Richtig-Falsch-Antworttyp und der Multiple-Choice-Aufgabe (Lienert & Raatz, 1998).

Derzeit findet das Multiple-Choice-Format (MC-Format) die häufigste Verwendung, was vermutlich in der einfacheren Auswertung, der höheren Verrechnungssicherheit³ als auch in einer höheren Testökonomie aufgrund der Möglichkeit von Gruppentestungen liegt (vgl. dazu Seiwald, 2003; Kubinger, 2009). Manche Testkonzepte benötigen sogar das Multiple-Choice-Format, da erst die Antwortmöglichkeiten die Aufgabenstellung überhaupt definieren (Kubinger, 2009).

4.2. Ratewahrscheinlichkeit beim Multiple-Choice-Format

Ein Nachteil des MC-Formats bei Leistungstest liegt in möglichen Rateeffekten, die sowohl die Reliabilität als auch die Validität eines Verfahrens beeinträchtigen. Gemeint ist hiermit, dass eine Testperson mit einem extrem niedrigen Fähigkeitsniveau die Entscheidung, welche Antwortmöglichkeit sie wählt,

³ Verrechnungssicherheit bedeutet, dass die Reglementierung, wie die Testleistungen zu Testwerten zu verrechnen sind, „derart exakt festgelegt ist, dass jeder Auswerter zu denselben Ergebnissen kommt“ (Kubinger, 2009, S. 43).

ausschließlich nach dem Zufall fällt. Die Wahrscheinlichkeit für das Lösen einer Aufgabe ist somit nur von der Anzahl der Antwortmöglichkeiten abhängig und wird als „*a-priori* Ratewahrscheinlichkeit“ bezeichnet. Diese beträgt für eine Aufgabe mit fünf Antwortmöglichkeiten, also vier Distraktoren⁴ und die Lösung, $\frac{1}{5} = 0,20$. Besitzt nun eine Testperson ein mittleres Fähigkeitsniveau und kann dadurch bestimmte Distraktoren ausschließen, so wird für sie die Ratewahrscheinlichkeit bedeutend höher. Keine Seltenheit ist das Schwanken zwischen zwei Antwortmöglichkeiten, was einer Wahrscheinlichkeit von $\frac{1}{2}$ entspricht (Kubinger, 2009).

Bemühungen, die Rateeffekte zu verringern bestehen seitens der Testautoren und werden nachstehend kurz angeführt.

Die Möglichkeit der Erhöhung der Distraktorenanzahl ist nur bedingt möglich, da zu viele Antwortmöglichkeiten die Konzentration und die Merkfähigkeit der Testperson beeinflussen.

Die Hinzunahme der Antwortmöglichkeiten „Keine Lösung ist richtig“ und/oder „Ich weiß die Antwort nicht“ muss nicht zwingend die Ratewahrscheinlichkeit reduzieren und bringt auch weitere Probleme mit sich (siehe dazu Kubinger, 2009).

Ein durchaus praktikabler Ansatz besteht darin, die Anzahl der Lösungen unter den Antwortmöglichkeiten zu erhöhen. Bei einer Aufgabe mit 5 Antwortmöglichkeiten, 2 Lösungen und 3 Distraktoren, beträgt die Ratewahrscheinlichkeit nur noch $\binom{5}{2} = \frac{1}{10}$, wobei das Item nur als gelöst gilt, wenn beide Lösungen und kein Distraktor von der Testperson markiert wurden (Kubinger, 2009).

Zuletzt sei auch noch die Möglichkeit der sequentiellen Vorgabe erwähnt. Dabei werden der Testperson die Antwortmöglichkeiten hintereinander präsentiert. Die Testperson muss sich bei jeder einzeln dargebotenen Antwortmöglichkeit entscheiden, ob sie diese mit richtig oder falsch bewertet. Dies geschieht solange, bis die Testperson entweder alle Antwortmöglichkeiten ausgeschlagen hat oder sich für eine entschieden hat. Diese aufeinanderfolgenden

⁴ Distraktoren bedeuten wörtlich übersetzt „Ablenker“ und stellen „die vorgegebenen Falsch-Alternativen in einer Mehrfachwahl-Aufgabe“ (Häcker & Stapf, 2009, S. 223) dar.

Entscheidungen stellen unabhängige Ereignisse dar, weshalb sich die Ratewahrscheinlichkeit je nach Platzierung der Lösung verringert (Seiwald, 2003).

Für Tests mit Multiple-Choice-Format kann trotz möglicher Rateeffekte das dichotome logistische Testmodell nach Rasch gelten, „die befürchteten Rateeffekte sind dann für alle Aufgaben gleich, im günstigsten Fall sogar alle (nahezu) null“ (Kubinger, 2009, S. 136).

5. Regelgeleitete Itemkonstruktion

Für die Konstruktion von Items eines Tests bestehen zwei Möglichkeiten:

Auf der einen Seite kann die Konstruktion, wie schon lange Zeit praktiziert, intuitiv, erfahrungsgemäß und dem Glück überlassen erfolgen. Erst im Nachhinein wird das Konstruktionsziel des Verfahrens durch Datenerhebungen und Itemanalysen abgesichert, obwohl dieses bereits im Vorhinein feststeht (vgl. Hornke & Habon, 1984; Klauer, 1978; Schott & Wieberg, 1984).

Dem gegenüber steht die rationale bzw. regelgeleitete Itemkonstruktion. Feger (1984) versteht unter rationaler Itemgenerierung die Entwicklung von so weitgehenden Vorschriften zu einem bestimmten Bereich für die Erstellung von Items, „daß voneinander unabhängig arbeitende Testautoren zu gleichartigen, d.h. weitgehend übereinstimmenden Items gelangen“ (S. 24). Dies soll nicht anschließend an diverse Analysen, sondern bereits in einem ersten Schritt erfolgen. Durch solch eine Itemkonstruktion kann auch die inhaltliche Validität⁵ sichergestellt werden und die Konstruktvalidität⁶ wäre bereits im Vorfeld gegeben (siehe dazu Embretson, 1999; Schott & Wieberg, 1984). Klauer (1978) schlägt ein dreischrittiges Verfahren zur Objektivierung der Konstruktion von Items vor. In einem ersten Schritt soll der Inhalt, welcher später in Aufgaben transformiert wird, vollständig und präzise dargestellt werden. Der nächste Schritt besteht in der Wahl einer geeigneten Aufgabenform bzw. einem Itemformat und schlussendlich sind Regeln zu formulieren, die die Umwandlung des Inhalts in das Itemformat festlegen. Solch itemgenerierende Regeln (dargestellt in einem sog. Konstruktionsrational) machen es wahrscheinlich, „daß jedes konstruierte Item zu der durch den Regelsatz inhaltlich definierten Grundmenge (Itemuniversum) gehört“ (Hornke & Habon, 1984, S. 204).

Welche der beiden Möglichkeiten zur Konstruktion von Items ein Testautor wählt, bleibt jedem selbst überlassen. Und schlussendlich muss sich auch jeder selbst fragen, ob sein Test „tatsächlich jenes psychische Merkmal misst, welches er zu messen behauptet“ (Kubinger, 2009, S. 55).

⁵ Inhaltliche Validität, auch inhaltliche Gültigkeit genannt, bedeutet, dass der Test selbst, „quasi definitionsgemäß, das optimale Kriterium des interessierenden Merkmals darstellt (Kubinger, 2009, S. 55).

⁶ Konstruktvalidität ist einem Verfahren zuzusprechen, wenn dieser „bestimmte theoriegeleitete Ansätze (nachweislich) umsetzt“ (Kubinger, 2003b, S. 199).

6. Testtheoretische Grundlage

Im nachfolgenden Abschnitt werden die Grundlagen der Item-Response-Theorie und speziell des dichotomen logistischen Testmodells nach Georg Rasch (1-PL-Modell) dargestellt.

6.1. Item-Response-Theorie

Der klassischen Testtheorie gegenüber steht die Item-Response-Theorie (IRT) oder latent-trait-Theorie genannt, welche auch als probabilistische Testtheorie bezeichnet wird (Kubinger, 2003a).

Die Item-Response-Theorie geht davon aus, dass für beobachtbare (manifeste) Reaktionen einer Testperson auf Items latente, nicht beobachtbare Eigenschaften eben dieser Testperson verantwortlich sind. Die manifesten Reaktionen stehen mit diesen latenten Eigenschaften in einem wahrscheinlichkeitsfunktionalem Zusammenhang, der je nach Modell spezifisch ist (Kubinger, 1989).

Die bereits angesprochene klassische Testtheorie stellt ein nicht prüfbares, deterministisches Modell dar, welches hauptsächlich auf korrelationsstatistischen Betrachtungen aufbaut. Im Gegensatz dazu strebt die probabilistische Testtheorie nach einem wahrscheinlichkeitstheoretischen Ansatz und gilt als prüfbares Modell (Kubinger, 2003a).

Innerhalb der psychologischen Diagnostik kommt dem *Rasch-Modell*⁷ die größte Bedeutung zu. Der Großteil aller weiteren probabilistischen Modelle ist entweder als eine Verallgemeinerung oder als ein Spezialfall des *Rasch-Modells* zu sehen (Kubinger, 2003a).

6.2. Das *Rasch-Modell*

Das *Rasch-Modell* kann nur bei dichotomen Daten angewendet werden, was bedeutet, dass die Reaktionen auf Items in einem psychologisch-diagnostischen Verfahren mit „0“ oder „1“ bzw. „nicht gelöst“ oder „gelöst“ kodiert werden müssen

⁷ Mit *Rasch-Modell* ist selbstverständlich das dichotome logistische Testmodell von Georg Rasch gemeint, der Einfachheit halber wird jedoch die Bezeichnung *Rasch-Modell* für vorliegende Arbeit verwendet.

(Kubinger, 1989). Die Summe der gelösten Items entspricht dann dem Testwert. Sieht ein Test diesen Verrechnungsmodus vor, so muss *notwendigerweise* das *Rasch-Modell* gelten, damit Verrechnungsfairness vorliegt, die Testleistungen müssen die Verhaltensreaktionen also adäquat repräsentieren. (Kubinger, 2009).

Die Wahrscheinlichkeit, „dass Tp v Item i löst („+“), in Abhängigkeit des Personenparameters ξ_v , das ist die (wahre) Fähigkeit von v , und des Itemparameters σ_i , das ist die (wahre) Schwierigkeit von i “ (Kubinger, 2009, S. 89) stellt das *Rasch-Modell* folgendermaßen dar:

$$P(+|\xi_v, \sigma_i) = \frac{e^{\xi_v - \sigma_i}}{1 + e^{\xi_v - \sigma_i}} \quad (6.2.1.)$$

Die Wahrscheinlichkeit, dass Item i von der Tp v nicht gelöst wird („-“), beschreibt die Komplementärwahrscheinlichkeit (Kubinger, 2003a).

Die verschiedenen Annahmen, welche dem *Rasch-Modell* zu Grunde liegen werden im Folgenden kurz dargestellt:

Erstens wird eine *erschöpfende Statistik* postuliert, welche aus der Zahl der gelösten Aufgaben besteht. Der Rohwert enthält also bereits die gesamte Information über die latente Fähigkeit der Testperson (Fischer, 1974).

Des Weiteren sind die Reaktionen der Testperson auf alle Items *lokal stochastisch unabhängig*, ob eine Testperson ein Item löst oder nicht, hängt nur von der Schwierigkeit des Items und der Fähigkeit der Testperson ab. Welche Aufgaben bereits gelöst wurden oder noch gelöst werden ist davon unabhängig (Kubinger, 1989, 2003a).

Eine Besonderheit des Modells besteht darin, dass es *spezifisch objektive Vergleiche* erlaubt. Unabhängig davon welche Items bearbeitet wurden, kann die Testleistung von Testpersonen miteinander verglichen werden.

Umgekehrt zeigt sich auch die Itemschwierigkeit σ_i und σ_j zweier Items i und j unabhängig in Bezug auf die Stichprobe, die dafür verwendet wird, wodurch dem *Rasch-Modell statistische Stichprobenunabhängigkeit* zuzuschreiben ist (Kubinger, 2009).

Alle Aufgaben eines Tests, der konform mit dem *Rasch-Modell* ist, messen *eindimensional*. Erfasst wird demnach von allen Items nur eine einzige latente Fähigkeit (Kubinger, 2003a).

Der Vorteil des *Rasch*-Modells besteht auch darin, dass es sich um ein prüfbares Modell handelt. Zur Prüfung der Modellgültigkeit wird häufig der *Likelihood-Ratio*-Test nach Andersen (1973) eingesetzt. Dabei werden die Itemparameter in verschiedenen Teilstichproben separat geschätzt und anschließend wird geprüft, ob diese die beobachteten Daten besser erklären als die Parameterschätzungen in der gesamten Stichprobe. Zumeist erfolgt die Teilung der Stichprobe in nur zwei Teile, wobei der Score als internes Kriterium, sowie Alter oder Geschlecht als externes Kriterium, häufig verwendet werden (Kubinger 1989, 2003a).

Die Schätzungen der Itemparameter in den zwei Teilstichproben lassen sich auch durch den grafischen Modelltest (grafische Modellkontrolle) darstellen. Dabei werden die Parameterschätzungen in einem rechtwinkligen Koordinatensystem gegenübergestellt. Liegen alle Punkte auf oder in der Nähe von der durch den Ursprung gehenden 45°-Geraden, so gilt für den Test das *Rasch*-Modell (Kubinger, 2003a, 2009).

Der z-Test nach Fischer und Scheiblechner (1970, zitiert nach Kubinger 1989) ist ein ebenfalls häufig eingesetzter Modelltest, der als Entscheidungshilfe dient, wenn es darum geht Items aus dem Itempool zu entfernen (Kubinger, 1989).

III. EMPIRISCHER TEIL

Im empirischen Teil werden zu Beginn das Ziel der Untersuchung und die Hypothesen kurz vorgestellt. Danach erfolgt die Beschreibung der Entwicklung des Verfahrens *Gleichungen* sowie der durchgeführten Untersuchung. Zuletzt erfolgt die Darstellung der Ergebnisse.

Auf eine vollständige Darstellung der entwickelten Items muss aus urheberrechtlichen Gründen verzichtet werden, dem nachfolgenden Abschnitt sind jedoch Beispielitems zu entnehmen.

7. Ziel der Untersuchung und Hypothesen

Das Ziel dieser Untersuchung war vorrangig die Konstruktion eines neuen Verfahrens zur Messung von *Reasoning*, welches als Aufgabenmaterial mathematische Gleichungen vorsieht, und dessen Überprüfung auf Geltung des *Rasch*-Modells. Gleichzeitig sollte der Test *AN-TOP* auf *Rasch*-Modell-Konformität untersucht werden. Die Übereinstimmungsvalidität⁸ beider Verfahren soll durch das Korrelieren der Personenparameter beider Verfahren bestimmt werden.

Daraus leiten sich folgende Hypothesen ab:

H₀(1)

Die Items des Verfahrens *Gleichungen* sind konform mit dem *Rasch*-Modell.

H₀(2)

Die Items des Verfahrens *AN-TOP* sind konform mit dem *Rasch*-Modell.

H₀(3)

Zwischen den Leistungen in *Gleichungen* und *AN-TOP* besteht kein Zusammenhang.

⁸ Als Übereinstimmungsvalidität bezeichnet man die Korrelation mit einem Test, welcher beansprucht das gleiche Konstrukt zu messen (Kubinger, 2009).

8. Entwicklung des Verfahrens *Gleichungen*

8.1. Definition der zu erfassenden Fähigkeit

Der Test *Gleichungen* soll logisch-schlussfolgerndes Denken (*Reasoning*) mit numerischem Aufgabenmaterial messen. Das numerische Aufgabenmaterial, welches aus Gleichungen besteht, setzt das Wissen um Rechenzeichen und -operationen voraus, so dass *Reasoning sensu* Cattell (vgl. Abschnitt 1.2.) im Faktor der kristallisierten Intelligenz erfasst wird, da vorangegangene Lernerfahrung notwendig ist.

8.2. Geltungsbereich und Testformat

In einem ersten Schritt wurde entschieden, dass das Verfahren bei Personen ab 18 Jahren eingesetzt werden soll. Um den Test auch möglichst ökonomisch zu gestalten, fiel die Wahl auf ein Paper-Pencil Verfahren, welches in der Gruppe administriert werden sollte.

8.3. Aufgabenmaterial und Antwortformat

Wie bereits aus dem Namen des Verfahrens ersichtlich, stellen mathematische Gleichungen das Aufgabenmaterial und gleichzeitig das Itemformat dar. Jede Aufgabe besteht aus einer einzelnen Gleichung, die zwei unterschiedliche Variablen (x und y) enthält. Wie bereits in Abschnitt 3 erwähnt, haben solche Gleichungen meist mehrere Wertepaare als Lösung. Aus diesem Grund wird die Lösung durch das Antwortformat, dem MC-Format definiert. Wie in Abbildung 1 zu erkennen ist, sind für diese Gleichung viele Lösungen möglich, solange x und y in Summe den Wert 10 ergeben. Betrachtet man nun die Antwortmöglichkeiten, wird ersichtlich, dass es nur eine Lösung, also eine richtige x - y -Kombination, gibt. Prinzipiell sind alle x -Werte (obere Zeile) bei entsprechendem y möglich, ebenso sind alle y -Werte (untere Zeile) bei entsprechendem x möglich. Pro Item gibt es aber immer nur eine richtige x - y -Kombination.

$$x + y = 10$$

x	2	8	5	10
y	9	5	1	3

Abbildung 1: Das Instruktionsitem der *Gleichungen*.

Eine Aufgabe wird nur dann als richtig gewertet, wenn sowohl der richtige Wert für x als auch für y , sowie kein Distraktor, markiert werden.

Es wurde versucht, die Distraktoren so plausibel wie möglich zu gestalten. Da ein Großteil der konstruierten Gleichungen mehrere Wertepaare als Lösung haben, bestehen die Distraktoren oftmals aus einer Teillösung.

Da die Testpersonen zweimal aus je vier Antwortmöglichkeiten die Lösung auswählen müssen, beträgt die Ratewahrscheinlichkeit $\frac{1}{4} * \frac{1}{4} = \frac{1}{16} = 6,25 \%$.

8.4. Itemgenerierende Regeln

Zur Generierung der Items wurden Regeln konstruiert, die hauptsächlich die Schwierigkeit einer Gleichung beeinflussen. Diese werden im Folgenden dargestellt.

- **Wert der Variablen:** Jede Gleichung enthält zwei Variablen, x und y . Diese können sowohl für positive als auch für negative Zahlen stehen. Die Verwendung einer positiven und einer negativen als auch zweier negativer Zahlen als Variable soll die Schwierigkeit einer Gleichung erhöhen.
- **Zahlenraum:** Die verwendeten Zahlen innerhalb einer Gleichung befinden sich entweder im Einer- oder Zehnerzahlenbereich. Zahlen aus dem Zehnerbereich sollen die Schwierigkeit einer Gleichung erhöhen.
- **Rechenzeichen und -operationen:** Durch die verwendeten Rechenzeichen und -operationen lässt sich die Schwierigkeit einer Gleichung am besten variieren. Man gelangt hier jedoch sehr rasch an die Grenze des

Zumutbaren, da zur Lösung keine Hilfsmittel verwendet werden dürfen.

Diese Regel wird unterteilt in:

addieren (+)

subtrahieren (−)

multiplizieren (×)

dividieren (÷)

Vorhandensein eines Bruchs

potenzieren zweiten Grades (x^2)

Wurzel ziehen ($\sqrt{\quad}$)

x im Nenner

x^2 im Nenner

Zahl hoch x

potenzieren dritten Grades (x^3)

Verwendung von Klammern

- **Umformbarkeit:** Diese Regel berücksichtigt die Möglichkeit der Umformung bestimmter Gleichungen in eine einfachere Gleichung.
- **Häufigkeit einer Variable:** In Abhängigkeit der Schwierigkeit werden in einer Gleichung die zwei Variablen ein- oder mehrmals verwendet, wobei die mehrmalige Verwendung eine Erhöhung der Schwierigkeit bedeutet.

Das daraus entstandene Konstruktionsrational ist im Anhang zu finden.

8.5. Itempool und Testdesign

Zu Beginn wurde durch Anwendung der itemgenerierenden Regeln versucht, einen möglichst umfangreichen Itempool zu erstellen. Die Schwierigkeit der Items sollte dabei hauptsächlich durch die verwendeten Rechenzeichen und -operationen variiert werden. Nach der ersten Konstruktionsphase wurde jedoch ersichtlich, dass dieses Vorgehen Probleme bereitet, da die Aufgaben bald zu schwierig wurden und ohne Hilfsmittel kaum zu lösen waren. Aus diesem Grund wurden die als zu schwierig eingestuften Rechenoperationen (beispielsweise $\sqrt[3]{\quad}$) aus den

itemgenerierenden Regeln entfernt. Ein weiterer Versuch bestand darin, die Antwortmöglichkeiten sprachlich zu formulieren und die Gleichungen mit so wenig Zahlen wie möglich zu gestalten. Die entwickelten Distraktoren stellten sich jedoch als unzufriedenstellend dar, da die Variationsbreite sehr eingeschränkt war.

Schlussendlich gelang es durch Kombination aller oben beschriebenen itemgenerierenden Regeln einen hinreichend großen Itempool zu erstellen. Für die Testgestaltung standen 44 Items zur Verfügung, wobei 16 davon als Parallelitems konstruiert wurden. Auf die Entwicklung weiterer Parallelitems wurde, obwohl durch Vorhandensein der itemgenerierenden Regeln möglich, im Rahmen dieser Arbeit verzichtet.

Im Rahmen einer Vorstudie bearbeiteten zwei erwachsene Personen zwei Mal zwanzig Items und wurden anschließend gebeten, die Aufgaben in Bezug auf ihre Schwierigkeit zu rangreihen. Zusätzlich wurde die Bearbeitungszeit gestoppt, um eine Richtlinie für die künftigen Testungen zu haben.

Der Itempool von 44 Aufgaben diene als Grundlage zur Erstellung von 3 Testformen mit je 20 Items. Wie in Abbildung 2 ersichtlich, wurde ein *Booklet-Design* verwendet, bei dem einige Items, wie beispielsweise *Item 3* oder *Item 10* in allen Testformen vorkommen, und dadurch sogenannte *Linking-Items* darstellen. Auch Items, die in nur zwei Testformen vorkommen, sind *Linking-Items*. Die Zahlen geben die Positionen der Items im jeweiligen Testheft wieder.

Item Nr.	1	2	3	4	5	6	7	8	9	10	11
TF 1	1		2	3			4	5		6	7
TF 2		1	2		3	4			5	6	
TF 3			1	2		3	4		5	6	

Abbildung 2: Testdesign (Ausschnitt) des Verfahrens *Gleichungen* (TF=Testform)

8.6. Instruktion und Testvorgabe

Die Testvorgabe erfolgte konventionell. Jede Testperson erhielt eine Testform mit 20 Items. Nach dem Lesen der Instruktion (siehe Abbildung 3) hatten die Testpersonen 18 Minuten Zeit, die Gleichungen zu lösen.

Auf den nächsten Seiten finden Sie insgesamt 20 Gleichungen.

Beispiel:

$$x + y = 10$$

x	2	8	5	10
y	9	5	1	3

Die Lösung ist $x = 5$ und $y = 5$. Nur diese beiden Werte ergeben in Summe 10.

Wie bei diesem Beispiel gibt es auch bei allen weiteren Aufgaben immer nur eine Lösung. Bitte markieren Sie diese eindeutig durch Ankreuzen der zutreffenden Zahl. Sollten Sie eine Auswahl korrigieren wollen, streichen Sie bitte die angekreuzte Zahl durch und markieren Sie Ihre Lösung indem Sie einen Kreis um die Zahl zeichnen.

Zur Lösung der Aufgaben dürfen Sie keine Hilfsmittel verwenden und sich auch keine Notizen machen.

Bearbeiten Sie alle Aufgaben in der gegebenen Reihenfolge. Wenn Sie bei einer Aufgabe nicht weiterkommen, gehen Sie zur Nächsten über.

Für die Testbearbeitung stehen Ihnen 18 Minuten zur Verfügung.

Falls Sie vor Ablauf der Zeit mit der Bearbeitung der Aufgaben fertig sind, warten Sie bitte auf ein Zeichen des Testleiters.

Gibt es von Ihrer Seite noch Fragen?

Abbildung 3: Testinstruktion *Gleichungen*

Wie in Abbildung 3 ersichtlich, dürfen zur Bearbeitung der Aufgaben keine Hilfsmittel wie beispielsweise ein Taschenrechner verwendet werden, ebenfalls dürfen keine Notizen gemacht werden.

Die Bearbeitungszeit von 18 Minuten wurde auf Basis der benötigten Zeit der zwei Testpersonen in der Vorstudie festgelegt. Die Zeitbeschränkung ist aus zweierlei Gründen notwendig. Erstens wurde zusätzlich der *AN-TOP* sowie das Verfahren *Figurale Analogien* (siehe Punkt 11) vorgegeben, weshalb eine Standardisierung notwendig war. Auf der anderen Seite sollte damit erreicht werden, dass Personen, im Fall der Anwendung der „Versuch-Irrtum“-Strategie, möglichst nicht mit der Bearbeitung der gesamten Items fertig werden. „Versuch-Irrtum“-Strategie bedeutet, dass die Testperson „alle Antwortmöglichkeiten sofort und der Reihe nach auf ihre Angemessenheit“ (Kubinger, 2009, S. 133) hin prüft, bis sie die Lösung gefunden hat. Es soll also damit erreicht werden, dass nur

Testpersonen, die bereits durch logisch-schlussfolgerndes Denken bestimmte Antwortmöglichkeiten ausschließen können oder bereits in Gedanken eine Gesetzmäßigkeit zur Lösungsfindung konstruiert haben, die Möglichkeit haben alle Items zu bearbeiten. Zusätzlich wurde die Lösung von x nie an erster Stelle positioniert (siehe Tabelle 1), um das Prüfen der Reihe nach zu erschweren.

Tabelle 1: Häufigkeit der Position der Lösung von x innerhalb der Antwortmöglichkeiten
(Test Gleichungen)

Testform	1. Stelle	2. Stelle	3. Stelle	4. Stelle
1	0	8	6	6
2	0	7	7	6
3	0	6	9	5

Obwohl das Verfahren somit einen *Speed-and-Power-Test* darstellt, kann das *Rasch-Modell* gelten. Bestenfalls bedeutet dies, „dass die gegebene Bearbeitungszeit tatsächlich angemessen gewählt worden ist“ (Kubinger, 2009, S. 144).

9. Test und Testdesign *AN-TOP*

Das psychologisch-diagnostische Verfahren Alpha Numeric Topologies (kurz: AN-TOP) (Kubinger & Heuberger, in Vorb.) ist ein Test zur Erfassung von *Reasoning* im figuralen Bereich.

Die Testpersonen müssen hierbei Gesetzmäßigkeiten herausfinden, die das jeweils fehlende Symbol eindeutig festlegen. *Sensu* Cattell (vergleiche Abschnitt 1.2.) wird mit diesem Verfahren *Reasoning* im Bereich der fluiden Intelligenz erfasst, da wie bei Matrizentests auf sprachliches Material verzichtet wurde und die Testpersonen die Aufgaben ohne vorrangegangene Lernerfahrung lösen können sollten.

Folgende Instruktion wurde von Heuberger (in Vorb.) übernommen und vor der Bearbeitungszeit von 18 Minuten von den Testpersonen gelesen:

„In diesem Test geht es darum, Gesetzmäßigkeiten zu erkennen und entsprechend anzuwenden. Das dabei jeweils gesuchte Symbol ist in das vorgegeben Kästchen einzutragen.“

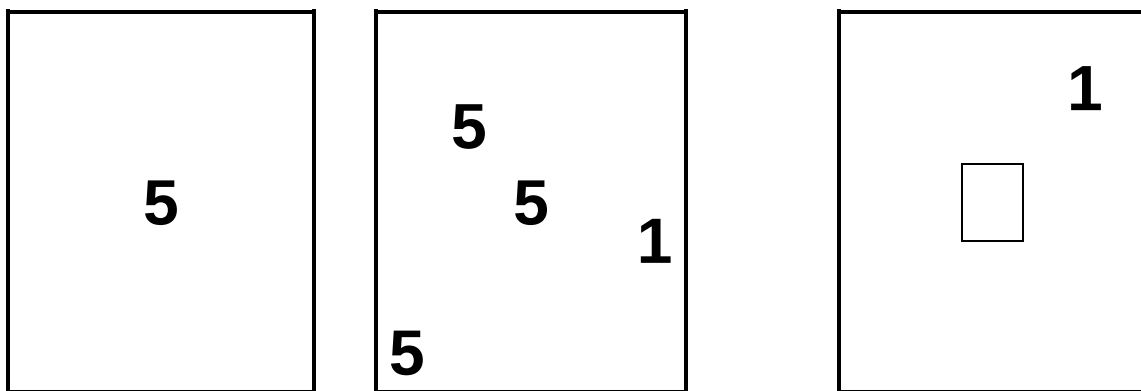


Abbildung 4: Itembeispiel des Tests *AN-TOP* (Kubinger & Heuberger, in Vorb.). Die Lösung ist 5.

Der *AN-TOP* besteht aus insgesamt 56 Items, wobei exakt die Hälfte Parallelitems darstellen. Um auch hier 3 Testformen zu erhalten, wurde wie bereits bei den *Gleichungen* ein *Booklet-Design* erstellt (siehe Abbildung 5). Jede Testform besteht aus 28 Items.

Item Nr.	1a	1b	2a	2b	3a	3b	4a	4b	5a	5b	6a	6b
TF 1	1			2	3		4		5			6
TF 2		1		2		3		4	5		6	
TF 3	1		2			3	4			5	6	

Abbildung 5: Testdesign (Ausschnitt) des Verfahrens *AN-TOP* (TF=Testform)

Auch hier muss auf eine vollständige Darstellung der Items aus urheberrechtlichen Gründen verzichtet werden. Zwei Instruktionsitems können den Abbildungen 4 und 6 entnommen werden.

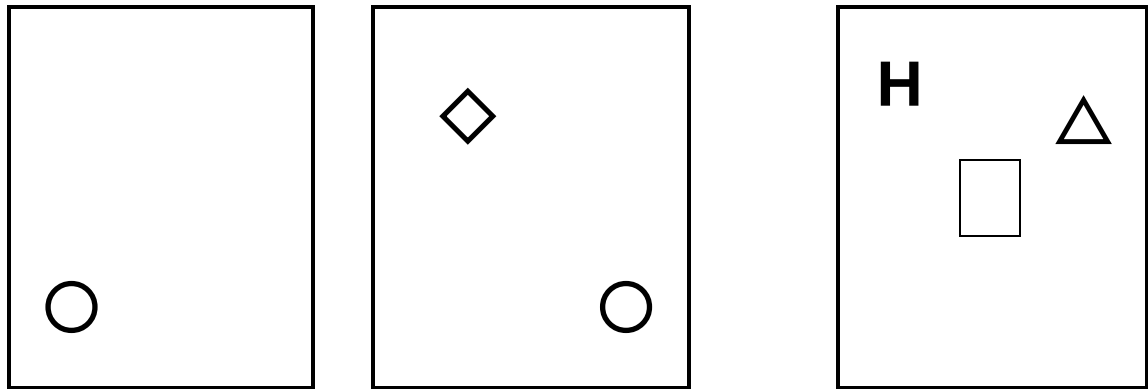


Abbildung 6: Itembeispiel des Tests *AN-TOP* (Kubinger & Heuberger, in Vorb.). Die Lösung ist O.

10. Datenerhebung

Da das Verfahren für Personen ab 18 Jahren entwickelt wurde und die Stichprobe nicht nur aus StudentInnen bestehen sollte, fand die Datenerhebung in den Sicherheitsakademien der Polizei statt. Um entsprechende Analysen durchführen zu können, sollten in etwa 300 Personen getestet werden.

Nach Vorgesprächen mit dem Schulkommandanten der Sicherheitsakademie (SIAK) Traiskirchen, Brigadier Arthur Reis, wurde im März ein entsprechender Antrag an das Bundesministerium für Inneres (siehe Anhang) gestellt. Nach offizieller Genehmigung der Testung erfolgte ein weiteres Gespräch mit den Verantwortlichen der SIAK Traiskirchen. Dabei stellte sich heraus, dass ca. 200 Auszubildende getestet werden können. Die Termine wurden für Mitte April festgelegt und es wurden zwei Unterrichtsstunden fix in den Stundenplan eingeplant. Um die angestrebte Stichprobengröße zu erreichen, wurde nachträglich um eine Genehmigung für Testungen in der Sicherheitsakademie Ybbs gebeten. Nach Zustimmung des Ministeriums sowie des Schulkommandanten Oberst Liebmann wurden die Termine in Ybbs für Ende April und Anfang Mai fixiert. Die Schulkommandanten der Sicherheitsakademien sowie die zuständigen LehrerInnen bekamen vorab ein Informationsschreiben (siehe Anhang), um die Klassen über die Testung zu informieren.

In der SIAK Traiskirchen wurden an zwei Vormittagen jeweils 4 Klassen parallel, in der SIAK Ybbs an 4 Terminen insgesamt 6 Klassen während der Unterrichtszeit getestet. Für die Testungen standen jeweils 2 Schulstunden à 50 Minuten zur Verfügung, benötigt wurden im Durchschnitt jedoch nur 55 Minuten, da die Testpersonen für die Bearbeitung des Verfahrens *Figurale Analogien* nicht die gesamte geplante Bearbeitungszeit benötigten. Die Klassengröße variierte zwischen 21 und 27 Personen und die Testung fand im jeweiligen Unterrichtsraum der Klasse statt. Als Testleiterinnen fungierten Diplomandinnen des Arbeitsbereichs Psychologische Diagnostik der Fakultät für Psychologie, Universität Wien.

Die Testungen fanden zusammen mit Frau Ünal statt, deren neu entwickelter Test *Figurale Analogien* zusätzlich vorgegeben wurde, wobei es hier nur eine Testform gibt. Jede Testperson erhielt somit ein Blatt zur Erhebung der demografischen Daten, je eine Testform der *Gleichungen* und des *AN-TOP* sowie

das Testheft *Figurale Analogien*. Es wurde versucht jede der 9 Kombinationen von Testheften in gleicher Anzahl zu verwenden.

Zu Beginn der Testung wurde den TeilnehmerInnen versichert, dass die Daten anonymisiert und nur für wissenschaftliche Zwecke verwendet werden und dass die Ergebnisse der Testung keinerlei Einfluss auf ihre schulischen Bewertungen in der SIAK haben. Nach Ausfüllen der demografischen Daten konnten die Testpersonen die Instruktion der *Gleichungen* selbstständig lesen und hatten danach 18 Minuten für die Aufgabenbearbeitung Zeit. Nach Ablauf der Zeit folgte das Verfahren *Figurale Analogien*. Auch hier hatten die Personen 18 Minuten Zeit, die jedoch wie bereits erwähnt, selten zur Gänze in Anspruch genommen wurden. Der letzte Teil der Testung bestand aus dem Verfahren *AN-TOP*. Wie bereits in Abschnitt 10 erwähnt gab es bei diesem Test ebenfalls die Zeitbeschränkung von 18 Minuten für die Bearbeitung der Aufgaben.

11. Stichprobenbeschreibung

Wie in Tabelle 2 ersichtlich, wurden insgesamt 331 Auszubildende an den Sicherheitsakademien im Alter von 18 bis 32 Jahren ($Md_{Alter}=22$ Jahre) getestet. Die Stichprobe setzt sich aus 79,2 % männlichen und 20,8 % weiblichen TeilnehmerInnen zusammen, wobei davon ausgegangen werden kann, dass diese Verteilung repräsentativ für die Ausbildungskandidaten der Polizei ist.

Tabelle 2: Die Verteilung der getesteten Personen in Bezug auf das Alter in Jahren ($n = 331$)

Alter in Jahren	Absolute Häufigkeit	%
18	2	0,6
19	15	4,5
20	52	15,7
21	58	17,5
22	39	11,8
23	32	9,7
24	39	11,8
25	19	5,7
26	18	5,4
27	23	6,9
28	12	3,6
29	6	1,8
30	12	3,6
31	3	0,9
32	1	0,3

72,5 % der Testpersonen wohnen in Niederösterreich, 15,7 % in Wien, die restlichen Prozent verteilen sich auf 5 weitere österreichische Bundesländer. Deutsch als Muttersprache gaben 97 % der Testpersonen an, wobei hier zu beachten ist, dass die österreichische Staatsbürgerschaft eine Voraussetzung der Aufnahme in die Polizeischule ist. Die am häufigsten höchste abgeschlossene Ausbildung stellt eine BHS (berufsbildende höhere Schule) dar, die genaue Verteilung ist der Tabelle 3 zu entnehmen.

Tabelle 3: Die Verteilung der getesteten Personen auf die höchste abgeschlossene Ausbildung
($n = 331$)

höchste abgeschlossene Ausbildung	absolute Häufigkeit	%
Pflichtschule	9	2,7
Lehre	88	26,6
AHS	94	28,4
Berufsbildende Höhere Schule (z.B. BHS, HTL, HAK)	135	40,8
Hochschule (Universität, FH, PH)	5	1,5

12. Darstellung der Ergebnisse

Im Folgenden werden zuerst die Ergebnisse der *Rasch*-Modell-Analysen in Bezug auf *Gleichungen* und *AN-TOP* dargestellt, danach das Ergebnis zur Übereinstimmungsvalidität beider Verfahren.

Für die statistischen Analysen sowie die Grafiken wurde die frei verfügbare Statistik-Software R unter Verwendung des Programmpakts *eRm* (Mair, Hatzinger & Meier, 2010) genutzt. Alle deskriptiven Auswertungen erfolgten mit Hilfe der Statistik-Software SPSS (13.0).

12.1. Ergebnisse der *Rasch*-Modell-Analysen des Tests *Gleichungen*

Der *Likelihood-Ratio*-Test (LRT) nach Andersen (1973) diente zur Prüfung der Modellgültigkeit der 44 Items, während die grafische Modellkontrolle zusätzlich zur Veranschaulichung durchgeführt wurde. Im Falle eines signifikanten LRTs sollte der z-Test nach Fischer und Scheiblechner (1970, zitiert nach Kubinger, 1989) durchgeführt werden.

Nachstehend werden die 3 Teilungskriterien des *Likelihood-Ratio*-Test (LRT) erörtert:

Internes Teilungskriterium:

- „**Score**“: Rohwert $\leq$ Median vs. Rohwert $>$ Median

Externe Teilungskriterien:

- **Geschlecht**: männlich vs. weiblich
- **Alter**: 18- bis 22-Jährige vs. 23- bis 32-Jährige

Das Risiko 1. Art (Irrtumswahrscheinlichkeit) wurde mit $\alpha = 0.05$ festgelegt. Aufgrund der Durchführung von drei Signifikanztests, was zu einer Kumulierung der Irrtumswahrscheinlichkeit führt, wurde eine α -Anpassung auf $\alpha = 0.01$ vorgenommen. Im Falle signifikanter LRTs, die eine Modellabweichung bedeuten, sollten schrittweise Items ausgeschieden werden, bis *a-posteriori* Modellgültigkeit gegeben ist.

Abweichungen in der grafischen Modellkontrolle werden als relevant angesehen, wenn der Unterschied der Itemparameterschätzungen zwischen 2 Teilstichproben größer als ein Zehntel der Spannweite aller Itemparameterschätzungen ist.

Die Anzahl der schätzbaren Items ($df + 1$) ist zumindest in zwei Teilungskriterien verringert (siehe Tabelle 4). Items sind nicht schätzbar, wenn entweder alle Testpersonen oder keine Testperson einer Teilstichprobe das bzw. die in der Berechnung nicht berücksichtigten Items lösen konnten. Des Weiteren ist der Tabelle 4 zu entnehmen, dass es nur im Teilungskriterium *Score* zu signifikanten Modellabweichungen gekommen ist.

Tabelle 4: Ergebnisse der LRTs *Gleichungen* (alle Items)

Teilungskriterium	χ^2	df	χ^2_{krit}	p
Score	64,516	38	61,162	0,005
Geschlecht	50,919	43	67,459	0,19
Alter	48,579	42	66,206	0,225

Die grafischen Modelltests der Teilungskriterien sind in den Abbildungen 7 bis 9 zu finden.

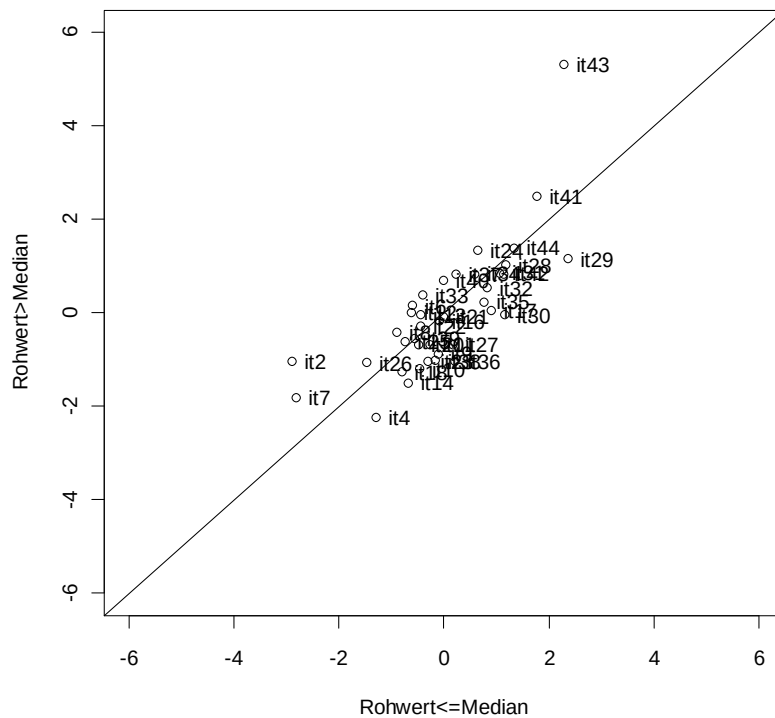


Abbildung 7: Grafischer Modelltest *Gleichungen*, Teilungskriterium *Score* (alle Items)

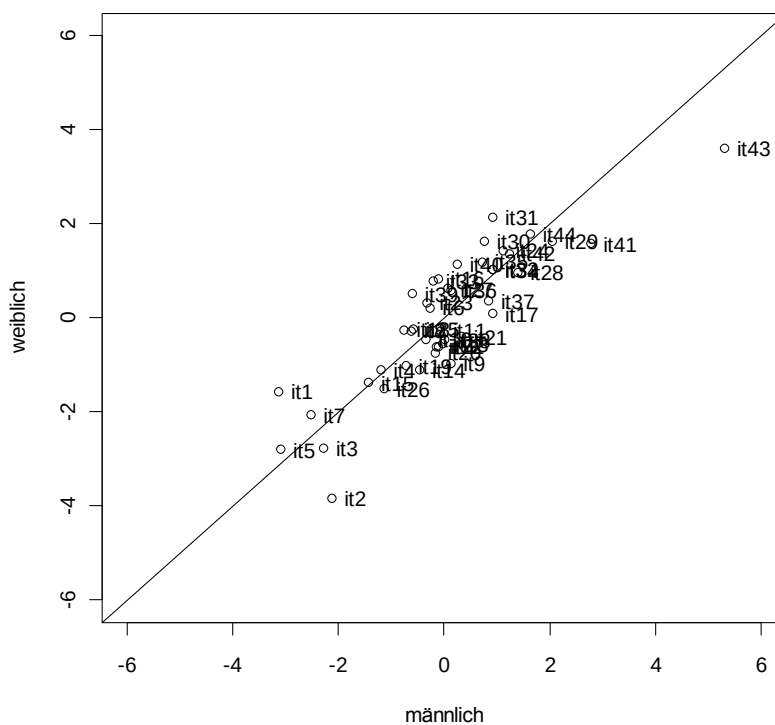


Abbildung 8: Grafischer Modelltest *Gleichungen*, Teilungskriterium *Geschlecht* (alle Items)

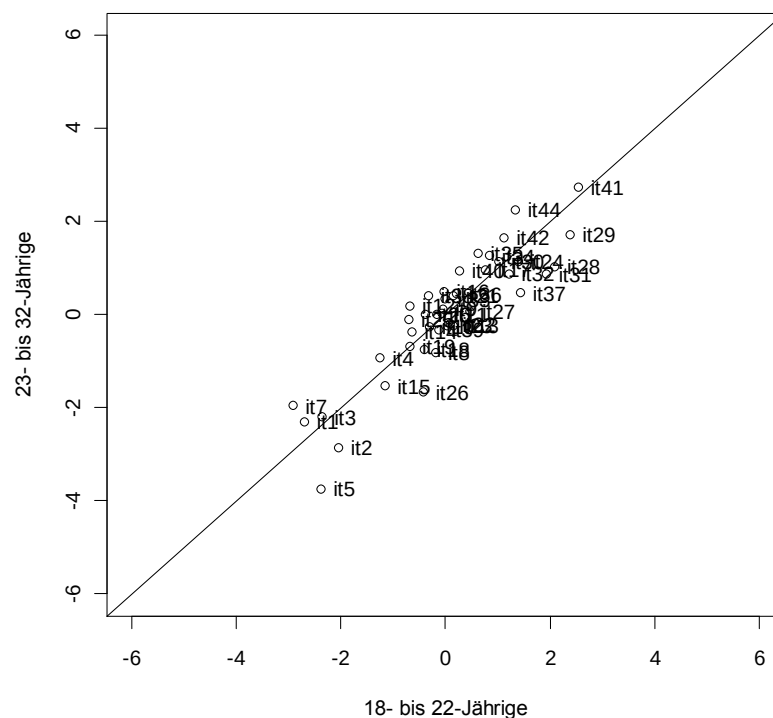


Abbildung 9: Grafischer Modelltest *Gleichungen*, Teilungskriterium *Alter* (alle Items)

Auffällig in den grafischen Modelltests sind vor allem die *Items 2* und *43* in den Teilungskriterien *Score* und *Geschlecht*. *Item 1* weicht im Teilungskriterium *Geschlecht* deutlich von der 45°-Geraden ab, ebenso wie *Item 5* im Teilungskriterium *Alter*. Inhaltliche Gründe für die auffälligen Items konnten nicht gefunden werden. Möglich ist jedoch, dass *Item 1* (siehe Abbildung 10) und *Item 2* aufgrund ihrer geringen Schwierigkeit nur mit Rechenfähigkeit gelöst wurden und somit *Reasoning* nicht erfassen.

1	$x = 15 - y$	x	5	12	2	7
		y	8	4	11	9

Abbildung 10: Das *Item 1*, des Tests *Gleichungen*, das sich als nicht *Rasch*-Modell-konform erwies.

Aufgrund des signifikanten LRT im Teilungskriterium *Score* wurde zusätzlich der z-Test nach Fischer und Scheiblechner berechnet. Dabei wurde kein Item auffällig (siehe Tabelle 5).

Tabelle 5: z-Test *Gleichungen* nach Fischer und Scheiblechner, Teilungskriterium *Score*

Item	z-Wert	p-Wert
it2	2,065	0,039
it4	-1,260	0,208
t6	2,072	0,038
it7	1,531	0,126
it8	0,821	0,412
it9	-1,591	0,112
it10	-1,712	0,087
it11	-0,632	0,527
it12	1,251	0,211
it13	0,693	0,488
it14	-1,479	0,139
it16	-0,149	0,881
it17	-1,555	0,120
t18	-0,823	0,411
it20	-0,292	0,770
it21	-0,151	0,880
it22	0,276	0,783
it23	-0,940	0,347
it24	1,439	0,150
it25	0,214	0,830
it26	0,583	0,560
it27	-1,323	0,186
it28	-0,283	0,777
t29	-2,097	0,036
it30	-2,041	0,041
t31	-0,358	0,720
it32	-0,974	0,330
it33	1,422	0,155

it34	0,707	0,479
it35	-0,914	0,361
it36	-1,579	0,114
it37	1,167	0,243
it38	-1,080	0,280
it39	0,055	0,956
it40	1,292	0,196
it41	1,404	0,160
it42	-0,502	0,615
t43	2,454	0,014
it44	0,112	0,911

Auf Basis der grafischen Modellkontrolle wurden sukzessive die *Items 43, 2, 5, 1* und *7*, das sind ungefähr 11 % der Items, ausgeschlossen. Den verbleibenden 39 Items kann *a-posteriori* Modellgültigkeit zugeschrieben werden.

Die drei LRTs fallen nach Ausscheiden der fünf Items nicht signifikant aus (siehe Tabelle 6). Die grafischen Modelltests können den Abbildungen 11 bis 13 entnommen werden und sind für alle drei Teilungskriterien zufriedenstellend. Der Test *Gleichungen* misst somit das Konstrukt *Reasoning* eindimensional.

Tabelle 6: Ergebnisse der LRTs *Gleichungen* (nach Ausschluss von 5 Items)

Teilungskriterium	χ^2	<i>df</i>	χ^2_{krit}	<i>p</i>
Score	55,167	36	58,619	0,021
Geschlecht	43,434	38	61,162	0,251
Alter	44,528	38	61,162	0,216

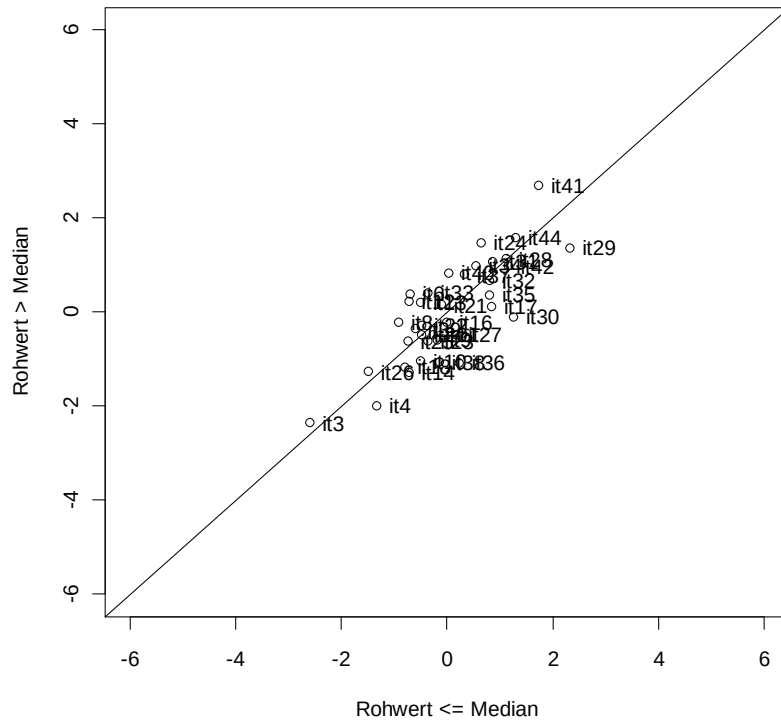


Abbildung 11: Grafischer Modelltest *Gleichungen*, Teilungskriterium *Score*
(nach dem Ausschluss von 5 Items)

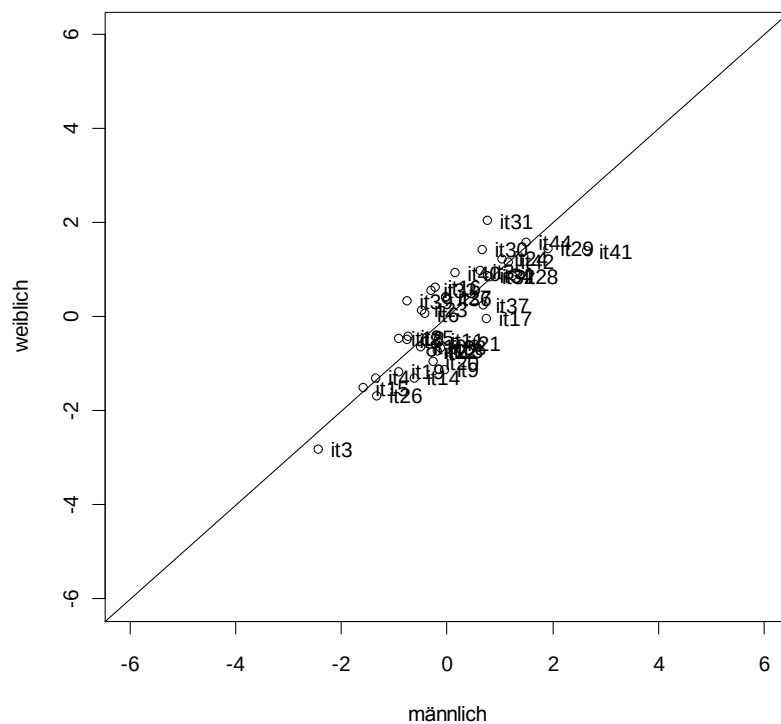


Abbildung 12: Grafischer Modelltest *Gleichungen*, Teilungskriterium *Geschlecht*
(nach dem Ausschluss von 5 Items)

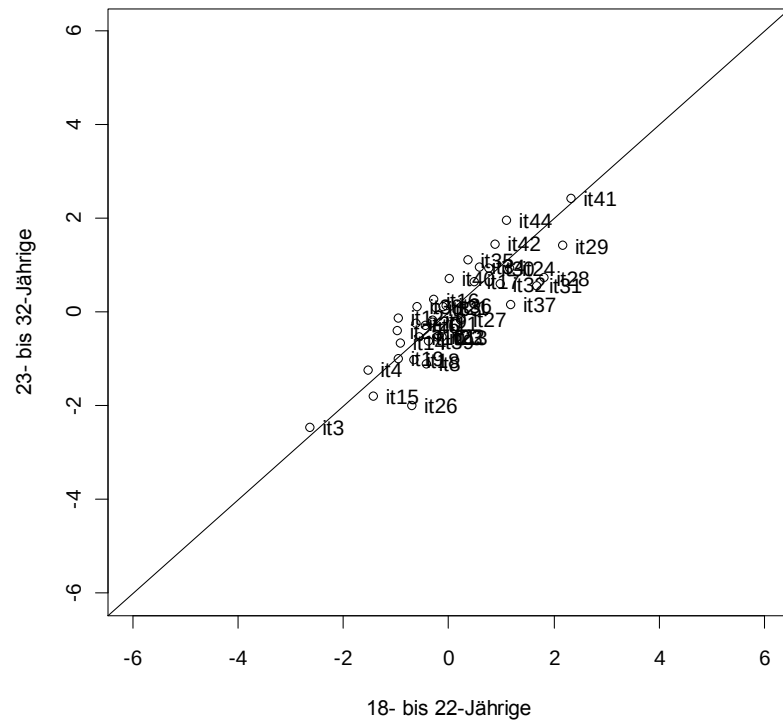


Abbildung 13: Grafischer Modelltest *Gleichungen*, Teilungskriterium *Alter*
(nach dem Ausschluss von 5 Items)

12.2. Ergebnisse der Rasch-Modell-Analysen des Tests AN-TOP

Die Prüfung auf Geltung des *Rasch*-Modells für die 56 Items des *AN-TOP* erfolgte ebenfalls mit dem *Likelihood-Ratio*-Test (LRT) nach Andersen (1973). Zusätzlich wird der grafische Modelltest als auch der *z*-Test nach Fischer und Scheiblechner (1970, zitiert nach Kubinger 1989) eingesetzt. Die Teilungskriterien sind analog zu den *Gleichungen Score, Geschlecht* und *Alter*. Das Signifikanzniveau wurde gleicherweise auf 1 % adjustiert.

Da *Item 1a* und *Item 1b* von allen Testpersonen gelöst wurden, sind diese durch das *Rasch*-Modell nicht schätzbar. Der Itempool zur Prüfung der Modellgültigkeit verkleinert sich somit auf 54 Items.

Wie in Tabelle 7 ersichtlich sind, bis zu 14 Items im Teilungskriterium *Score* nicht schätzbar (schätzbare Items = $df + 1$), da diese entweder von keiner oder von allen Testpersonen einer Teilstichprobe gelöst wurden.

Ebenfalls der Tabelle 7 ist zu entnehmen, dass es nur im Teilungskriterium *Score* zu signifikanten Modellabweichungen gekommen ist.

Tabelle 7: Ergebnisse der LRTs *AN-TOP* (alle Items)

Teilungskriterium	χ^2	df	χ^2_{krit}	p
Score	80,432	39	62,428	0
Geschlecht	35,921	45	69,956	0,831
Alter	49,974	50	76,154	0,474

Die Abbildungen 14 bis 16 zeigen die grafischen Modelltests der drei Teilungskriterien. Obwohl keine inhaltlichen Gründe gefunden werden konnten, wurden die *Items 12b* und *20b* im Teilungskriterium *Score* und *Item 23b* im Teilungskriterium *Geschlecht* grafisch auffällig. Auch *Item13a* fiel grafisch in den Teilungskriterien *Geschlecht* und *Score* auf, *Item 4a* im Teilungskriterium *Alter*.

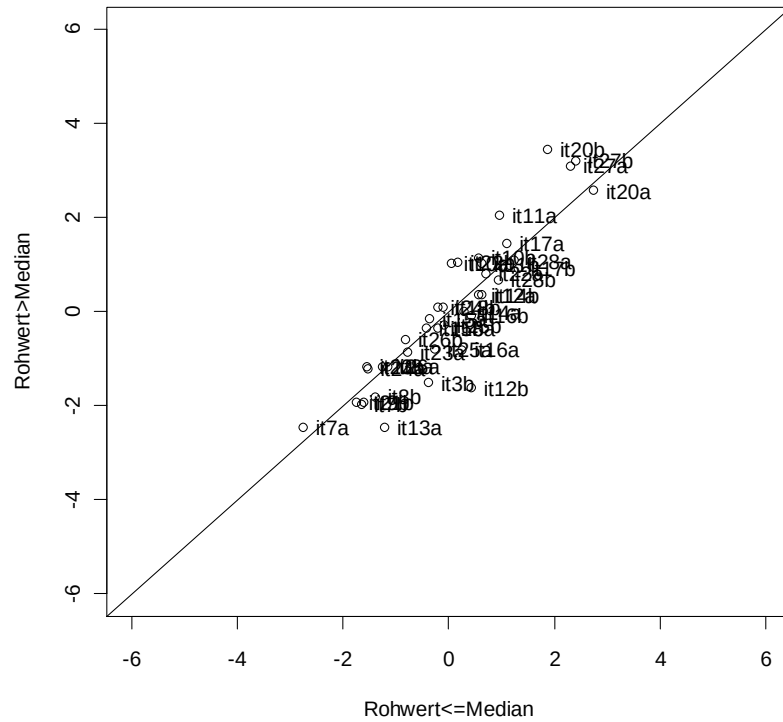


Abbildung 14: Grafischer Modelltest AN-TOP, Teilungskriterium *Score* (alle Items)

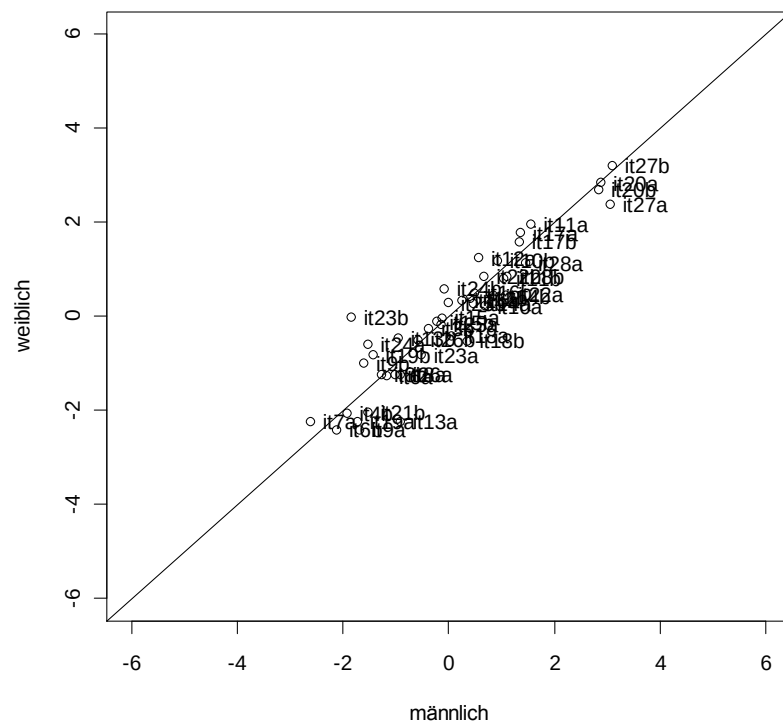


Abbildung 15: Grafischer Modelltest AN-TOP, Teilungskriterium *Geschlecht* (alle Items)

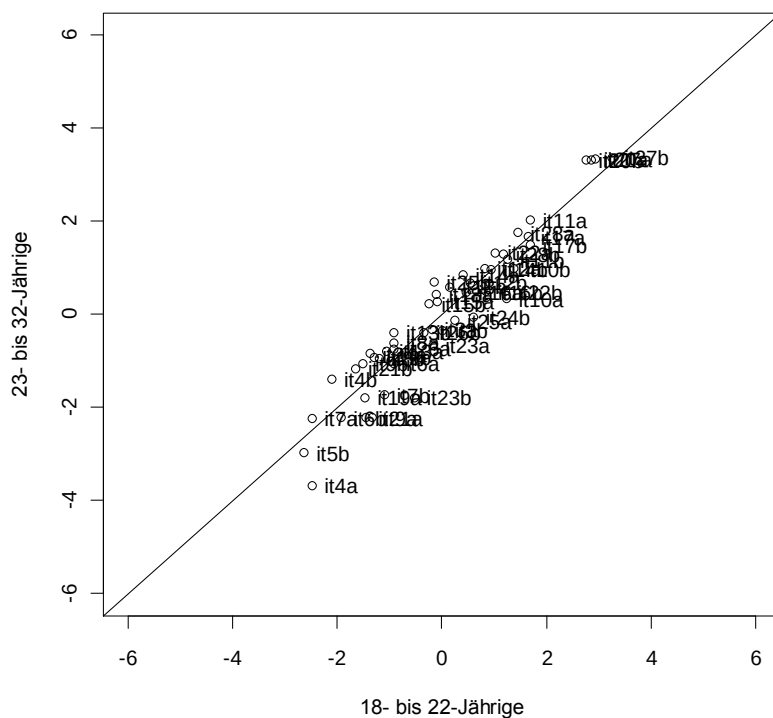


Abbildung 16: Grafischer Modelltest *AN-TOP*, Teilungskriterium *Alter* (alle Items)

Vor dem Ausschluss von Items wurde der z-Test nach Fischer und Scheiblechner berechnet. Dabei wurden die *Items 10a, 11a* und *20b* auffällig (siehe Tabelle 8).

Tabelle 8: z-Test *AN-TOP* nach Fischer und Scheiblechner, Teilungskriterium *Score*

Item	z-Wert	p-Wert
it3b	-1,839	0,066
it7a	0,279	0,780
it7b	-0,300	0,764
it8a	-0,005	0,996
it8b	-0,545	0,586
it9b	-0,424	0,671
it10a	3,176	0,001
it10b	1,424	0,155
it11a	3,628	0,000

it11b	0,943	0,345
it12a	-0,585	0,558
it12b	-1,988	0,047
it13a	-1,220	0,222
it14a	-0,562	0,574
it14b	-0,744	0,457
it15a	0,378	0,706
it15b	0,170	0,865
it16a	-2,122	0,034
it16b	-0,982	0,326
it17a	1,175	0,240
it17b	-0,884	0,377
it18a	-0,374	0,708
it18b	0,387	0,699
it20a	-0,287	0,774
it20b	4,092	0,000
it21b	-0,224	0,822
it22a	0,319	0,750
it22b	2,095	0,036
it23a	-0,159	0,874
it23b	0,476	0,634
it24a	0,569	0,569
it24b	0,555	0,579
it25a	-1,049	0,294
it25b	-0,356	0,722
it26a	0,116	0,907
it26b	0,503	0,615
it27a	1,463	0,143
it27b	2,092	0,036
it28a	-0,424	0,672
it28b	-0,554	0,580

Im Anschluss wurden schrittweise die Items *4a*, *11a*, *20b* und *23b* ausgeschieden. Nach Ausschluss dieser 4 Items, das sind in etwa 7 % der Items, ist

a-posteriori Modellgültigkeit gegeben. Wie in Tabelle 9 ersichtlich, fallen die LRTs der drei Teilungskriterien nicht signifikant aus. Die zufriedenstellenden grafischen Modelltests können den Abbildungen 16 bis 18 entnommen werden. Das Konstrukt *Reasoning* wird somit eindimensional mit dem Test *AN-TOP* gemessen.

Tabelle 9: Ergebnisse der LRTs *AN-TOP* (nach Ausschluss von 4 Items)

Teilungskriterium	χ^2	<i>df</i>	χ^2_{krit}	<i>p</i>
Score	57,356	38	61,162	0,023
Geschlecht	26,557	42	66,206	0,97
Alter	44,135	46	71,201	0.551

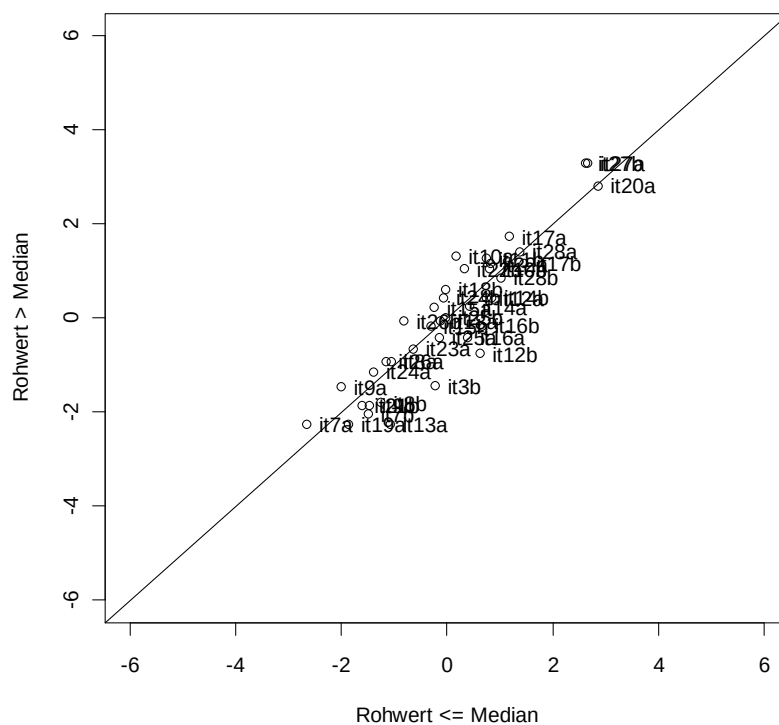


Abbildung 17: Grafischer Modelltest *AN-TOP*, Teilungskriterium *Score* (nach dem Ausschluss von 4 Items)

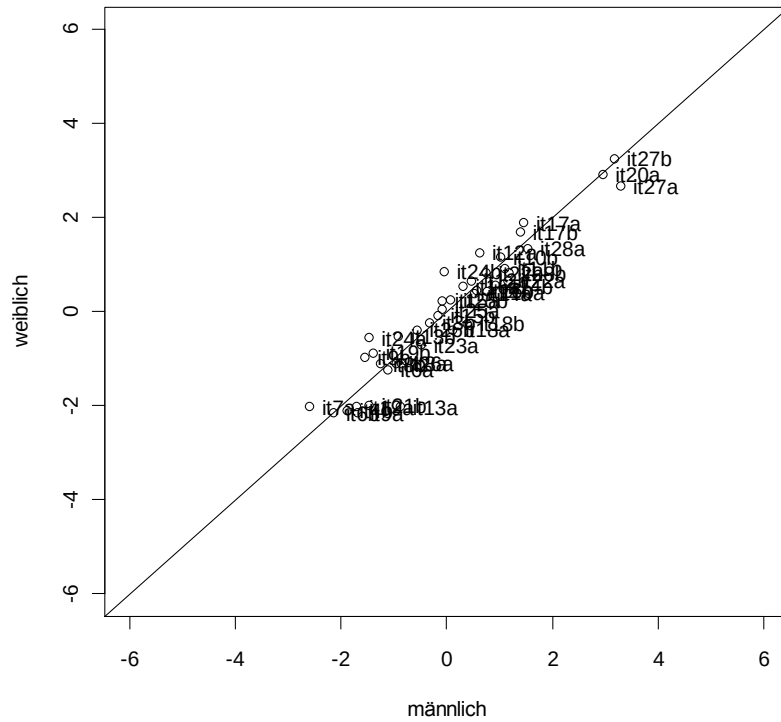


Abbildung 18: Grafischer Modelltest AN-TOP, Teilungskriterium *Geschlecht* (nach dem Ausschluss von 4 Items)

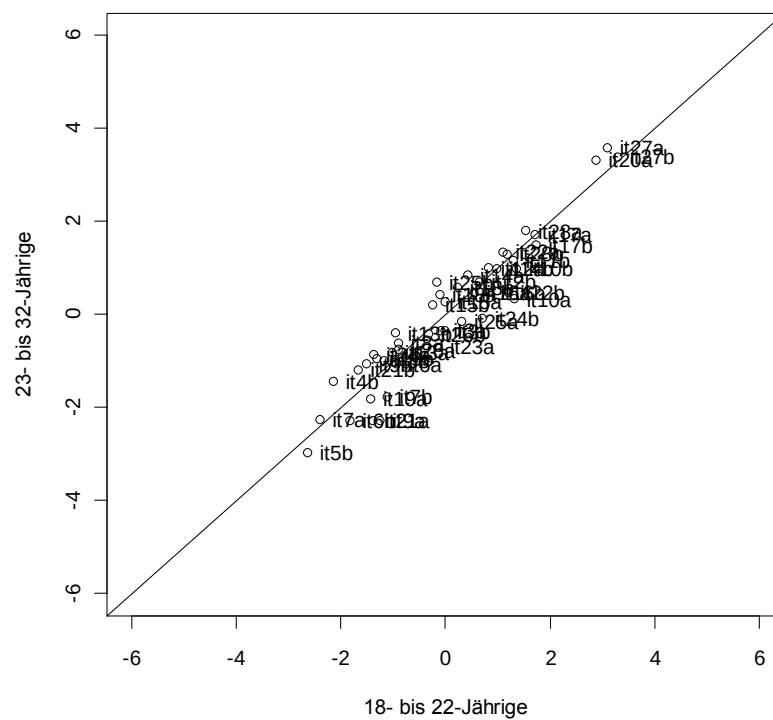


Abbildung 19: Grafischer Modelltest AN-TOP, Teilungskriterium *Alter* (nach dem Ausschluss von 4 Items)

12.3. Übereinstimmungsvalidität zwischen *Gleichungen* und *AN-TOP*

Die Übereinstimmungsvalidität zwischen *Gleichungen* und *AN-TOP* wurde aus zweierlei Gründen berechnet. Erstens um zu überprüfen, ob der Test *Gleichungen* auch tatsächlich *Reasoning* misst. Zum Zweiten ob dies in unterschiedlichen Faktoren der Intelligenz nach Cattell geschieht. Durch den Umstand, dass zum Lösen von Gleichungen vorangegangene Lernerfahrungen notwendig sind, soll *Reasoning* im Faktor der *kristallisierten* Intelligenz erfasst werden, wohingegen der Test *AN-TOP Reasoning* im Faktor der *fluiden* Intelligenz messen soll. Des Weiteren wird *Reasoning* bei den *Gleichungen* durch numerisches Aufgabenmaterial erfasst, beim *AN-TOP* mithilfe von figuralem Aufgabenmaterial. Folglich wird eine Übereinstimmungsvalidität erwartet, die zwar als vorhanden, jedoch nicht als groß bezeichnet werden kann.

Um die Übereinstimmungsvalidität zu bestimmen, wurden die Personenparameter herangezogen, welche intervallskaliert sind. Deshalb ist eine Berechnung des Zusammenhangs der Personenparameter mittels Pearson-Korrelationskoeffizienten möglich. Die Berechnung erfolgte unter Verwendung der Open-Source Statistik Software R.

Vorab wurden unter Verwendung des Programmpakets „PP“ (Person Parameter estimation) (Reif, 2012) alle Personenparameter extrapoliert, die zuvor nicht geschätzt werden konnten. Ein Personenparameter kann dann nicht geschätzt werden, wenn diese Person entweder alle Items oder kein einziges Item gelöst hat.

Zwischen den Personenparametern der *Gleichungen* und des *AN-TOP* besteht ein positiver Zusammenhang. Der Pearson-Korrelationskoeffizient beträgt $r = 0,43$, das Bestimmtheitsmaß $B = 0,19$. Das heißt 19 % der Varianz der Leistung in einem Test werden durch das Ergebnis des anderen Tests erklärt.

Die Übereinstimmungsvalidität kann nicht als gegeben angesehen werden. Das Ergebnis entspricht weitgehend den oben angeführten Erwartungen. Zu beachten ist jedoch noch, dass die „wahre“ Validität aufgrund von Messfehlern stets unterschätzt wird (vgl. dazu Rost, 2004).

13. Diskussion und Ausblick

Im Rahmen dieser Arbeit ist es gelungen einen *Rasch*-Modell-konformen Test zur Erfassung von *Reasoning* für Personen ab 18 Jahren zu entwickeln. Die Testpersonen müssen dabei mathematische Gleichungen, welche das Aufgabenmaterial darstellen und zwei Variablen enthalten, lösen. Zum Lösen der Aufgabe stehen für jede Variable je 4 Antwortmöglichkeiten zur Verfügung. Das Multiple-Choice Format als Antwortformat ist für den Test *Gleichungen* aus folgendem Grund notwendig: Eine Gleichung mit zwei Variablen hat viele Wertpaare als Lösung, weshalb meistens eine zweite Gleichung zum Lösen zur Verfügung steht. Da dies im Test *Gleichungen* nicht der Fall ist, können die Gleichungen nur mit Hilfe der dargebotenen Antwortmöglichkeiten gelöst werden. Durch den Umstand, dass die Testpersonen zweimal aus vier Antwortmöglichkeiten die Lösung wählen müssen, beträgt die *a-priori*-Ratewahrscheinlichkeit nur 6,25 %.

In Bezug auf die Stichprobe, welche aus Auszubildenden der Sicherheitsakademien der Polizei besteht, ist anzumerken, dass die ältesten Testpersonen 32 Jahre alt waren. Eine Stichprobe, deren Altersbereich breiter gefächert ist, wäre für künftige Studien wünschenswert. Kritisch zu betrachten ist ebenfalls die Anzahl weiblicher Testpersonen innerhalb der Stichprobe. Ein hoher Anteil männlicher Testpersonen (annähernd 80%) ist sicherlich repräsentativ für Ausbildungskandidaten der Polizei, jedoch sollte in künftigen Untersuchungen eine ausgewogenere Verteilung der männlichen und weiblichen Testpersonen innerhalb der Stichprobe angestrebt werden.

Nach dem Ausscheiden von 5 Items (etwa 11 % des Itempools) konnte dem Test *Gleichungen a-posteriori* Modellgeltung des *Rasch*-Modells zugeschrieben werden. Die *a-posteriori* Modellgeltung war beim Test *AN-TOP* bereits nach Ausschluss von 4 Items gegeben.

Zwei auszuschneidende Items der *Gleichungen* könnten möglicherweise alleine durch Rechenfähigkeit gelöst werden, weshalb diese *Reasoning* vermutlich nicht erfassen. Inhaltliche Gründe für die 3 weiteren ausgeschlossenen Items der *Gleichungen* als auch für die ausgeschiedenen Items des *AN-TOP* konnten nicht gefunden werden.

Durch die *a-posteriori* erreichte Modellgeltung des *Rasch*-Modells kann sowohl dem Test *Gleichungen* als auch dem Test *AN-TOP* attestiert werden, dass diese das Gütekriterium der Skalierung hinreichend erfüllen, genauer gesagt, dass sie verrechnungsfair sind. Dies bedeutet, dass der Rohwert in einem Test ein faires Maß der Verrechnung der Testleistungen ist. Des Weiteren erfassen beide Tests eindimensional, unter Vorbehalt weiterer Validierungsstudien, das Konstrukt *Reasoning*.

Die Übereinstimmungsvalidität, welche durch Korrelieren der Personenparameter beider Tests berechnet wurde, kann als gering eingestuft werden. Dies ist vermeintlich darauf zurückzuführen, dass beide Verfahren zwar eindimensional *Reasoning*, jedoch zwei verschiedenen Faktoren der Intelligenz *sensu* Cattell (kristallisierte vs. fluide) erfassen und zur Erfassung unterschiedliches Aufgabenmaterial (numerisch vs. figural) verwendet wird. Eine Validierung anhand weiterer Verfahren ist in zukünftigen Untersuchungen wünschenswert, um sichergehen zu können, dass der Test *Gleichungen* tatsächlich *Reasoning* erfasst.

Ein weiteres Ziel ist die Überprüfung des Konstruktionsrational, mit dessen Hilfe die vorliegenden Items des Tests *Gleichungen* erstellt wurden, um zu untersuchen, ob dieses die Itemschwierigkeiten adäquat abbildet und dem Test somit Konstruktvalidität zugesprochen werden kann.

Ferner ist es notwendig den verbleibenden Itempool, dem *a-posteriori* Modellgeltung des *Rasch*-Modells zugeschrieben werden kann, erneut an einer neuen, unabhängigen Stichprobe auf Geltung des *Rasch*-Modells zu prüfen (vgl. dazu Kubinger, 2003a; Kubinger, 2005).

Zusammenfassend ist es in dieser Arbeit gelungen einen neuen Test zur Erfassung von *Reasoning* zu entwickeln, der aus *Rasch*-Modell-konformen Items besteht. Weiters stellt der Test *Gleichungen* auch das erste psychologisch-diagnostische Verfahren dar, mit welchem es möglich ist, *Reasoning* im kristallisierten Bereich *sensu* Cattell zu erfassen. Auch konnte gezeigt werden, dass sich mathematische Gleichungen durchaus als numerisches Aufgabenmaterial eignen.

14. Zusammenfassung

Die vorliegende Diplomarbeit hatte das Ziel einen neuen Test zur Erfassung von logisch-schlussfolgerndem Denken im numerischen Bereich für Personen ab 18 Jahren zu entwickeln, der dem *Rasch*-Modell entspricht.

Dafür wurden auf Basis eines selbst erstellten Konstruktionsrationalis insgesamt 44 Items generiert, deren Aufgabengabematerial aus Gleichungen besteht. Das Multiple-Choice-Antwortformat wurde aus inhaltlichen Gründen gewählt, da jede Gleichung im Test zwei Variablen enthält und die Antwortmöglichkeiten somit zum Lösen notwendig sind. Das Multiple-Choice-Format weist jedoch in diesem Fall eine geringe *a-priori* Ratewahrscheinlichkeit von 6,25 % auf.

Das im Rahmen dieser Arbeit entwickelte Verfahren *Gleichungen* wurde gemeinsam mit dem Test *AN-TOP* (Kubinger & Heuberger, in Vorb.) und dem Test *Figurale Analogien* (Ünal, in Vorb.) 311 Auszubildenden der Sicherheitsakademien der Polizei in Niederösterreich vorgegeben. Die so gewonnenen Daten wurden nach dem *Rasch*-Modell analysiert. Neben dem grafischen Modelltest wurden der *Likelihood-Ratio-Test* (LRT) nach Andersen (1973) sowie der z-Test nach Fischer und Scheiblechner (1970, zitiert nach Kubinger, 1989) zur Modellkontrolle verwendet. Als Teilungskriterien fungierten *Score*, *Geschlecht* sowie *Alter*. Bei den *Gleichungen* mussten 5 Items sukzessive ausgeschieden werden um *a-posteriori* Modellgeltung des *Rasch*-Modells zu erhalten. Der *AN-TOP* Itempool erwies sich nach Ausscheiden von 4 Items als *Rasch*-Modell-konform. Für die fehlende Modellgeltung in Bezug auf die ausgeschiedenen Items konnte bislang keine inhaltliche Erklärung gefunden werden.

Die Korrelation der Personenparameter der *Gleichungen* und des *AN-TOP* ergab mit 19 % erklärter Varianz eine geringe Übereinstimmungsvalidität. Durch Geltung des *Rasch*-Modells kann folglich beiden Verfahren Verrechnungsfairness und Eindimensionalität zugesprochen werden. Eine Kreuzvalidierung mit dem verbliebenen Itempool an einer neuen Stichprobe ist notwendig, um die Modellgeltung abzusichern.

IV. LITERATURVERZEICHNIS

- Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, 38, 123-140.
- Cattell, R. B. (1961). Fluid and Crystallized Intelligence. In J. J. Jenkins & D. G. Paterson (Eds.), *Studies in individual differences. The search for intelligence* (pp. 738-746). London: Methuen & Co.
- Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, 64, 407-433.
- Feger, B. (1984). Die Generierung von Testitems zu Lehrtexten. *Diagnostica*, 30, 24-46.
- Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen*. Bern: Hans Huber.
- Formann, A. K. & Piswanger, K. (1979). *Wiener Matrizen-Test. WMT*. Weinheim: Beltz-Test.
- Häcker, H. & Stapf, K. H. (2009). *Dorsch Psychologisches Wörterbuch* (15. überarbeitete und erweiterte Aufl.). Bern: Hans Huber.
- Heuberger, N. (in Arbeit). *Item-generierende Regeln bei AN-TOP*. Wien.
- Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, 57, 253-270.
- Hornke, L. & Habon, M. (1984). Erfahrungen zur rationalen Konstruktion von Testaufgaben. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 5, 203-212.
- Hornke, L. F., Etzel, S. & Rettig, K. (1997). *Adaptiver Matrizentest (AMT)*. Mödling: Schuhfried.
- Jäger, A. O. (1973). *Dimensionen der Intelligenz* (3. Aufl.). Göttingen: Hogrefe.
- Kerr, R. (1999). *Goldmann-Lexikon Mathematik* (überarbeitete Taschenbuchausgabe). Gütersloh: Bertelsmann Lexikographisches Institut.
- Klauer, K. J. (1978). Kontentvalidität. In K. J. Klauer (Hrsg.), *Handbuch der Pädagogischen Diagnostik* (S. 225-256). Düsseldorf: Schwan.

- Kubinger, K. D. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie. In K. D. Kubinger (Hrsg.), *Moderne Testtheorie* (S. 19-83). Weinheim: Beltz.
- Kubinger, K. D. (2003a). Probabilistische Testtheorie. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 415-423). Weinheim: Beltz.
- Kubinger, K. D. (2003b). Gütekriterien. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 195-204). Weinheim: Beltz.
- Kubinger, K. D. (2005). Psychological test calibration using the Rasch model – Some critical suggestions on traditional approaches. *International Journal of Testing*, 5, 377-394.
- Kubinger, K. D. (2009). *Psychologische Diagnostik*. Göttingen: Hogrefe.
- Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Beltz.
- Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2001). *Intelligenz-Struktur-Test 2000 R. Intelligenz-Struktur-Test zweitausend*. Göttingen: Hogrefe.
- Mair, P., Hatzinger, R. & Maier, M. (2010). *Extended Rasch Modeling. Package eRm* [Manual]. Verfügbar unter:
<http://cran.r-project.org/web/packages/eRm/eRm.pdf> [07.07.2012]
- Reif, M. (2012). *Person Parameter estimation*. Verfügbar unter:
<http://cran.r-project.org/web/packages/PP/PP.pdf> [09.07.2012]
- Rettig, K., Hornke, L. F. & Schiff, B. (1989). Konstruktion eines Rechentests. In Bundesministerium der Verteidigung – P II 4 (Hrsg.), *Untersuchungen des Psychologischen Diensts der Bundeswehr (24/89)* (S. 139-201). Bonn: Bundesministerium der Verteidigung
- Rost, J. (2004). *Lehrbuch Testtheorie-Testkonstruktion* (2. überarbeitete und erweiterte Aufl.). Bern: Hans Huber.
- Schott, F. & Wieberg, H.-J. W. (1984). Regelgeleitete Itemkonstruktion. Ein Verfahren zur Definition von Itemuniversa und deren kontextvalider Abbildung in Itemmengen für Tests und Treatments. *Diagnostica*, 30 (1), 47-66.

- Seiwald, B. B. (2003). Antwortformat. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe Psychologischer Diagnostik* (S. 23-28). Weinheim: Beltz.
- Srp, G. (1994). *Syllogismen*. Frankfurt am Main: Swets Test Service.
- Stemmler, G., Hagemann, D., Amelang, M. & Bartussek, D. (2011). *Differentielle Psychologie und Persönlichkeitsforschung* (7. vollständig überarbeitete Aufl.). Stuttgart: Kohlhammer.
- Süß, H.-M. (2003). Intelligenztheorien. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 217-224). Weinheim: Beltz.
- Ünal, S. (in Vorbereitung). *Figurale Analogien* [vorläufiger Arbeitstitel]. Unveröffentlichte Diplomarbeit, Universität Wien, Wien.

V. ANHANG

Testdesign des Tests *Gleichungen*:

Item Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
TF1	1		2	3			4	5		6	7			8		9		10		11				12			13	
TF2		1	2		3	4			5	6			7		8			9			10				12			
TF3			1	2		3	4		5	6		7		8			9		10		11	12				13		14

Item Nr.	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
TF 1		14		15	16		17				18	19		20		
TF 2	13			14		15		16		17	18		19			20
TF 3			15	16		17			18		19				20	

4 Gesamtverlinkungen; 3x TF 1 und TF 3; 1x TF 1 und TF 2; 4x TF 2 und TF 3

Itemparameter des Tests *Gleichungen*:

Itemnummer	Itemparameter
3	-2,53
4	-1,35
6	-0,35
8	-0,70
9	-0,23
10	-0,54
11	-0,27
12	-0,40
13	-0,29
14	-0,76
15	-1,57
16	-0,01
17	0,59
18	-0,81
19	-0,97
20	-0,43
21	0,02
22	-0,38
23	-0,35
24	1,07
25	-0,67
26	-1,40
27	0,07
28	1,15
29	1,81
30	0,85
31	1,00
32	0,80
33	-0,09
34	0,78
35	0,70
36	0,06
37	0,58
38	-0,24
39	-0,49
40	0,35
41	2,38
42	1,13
44	1,50

Anmerkung: Itemparameter sind Itemschwierigkeitsparameter, wobei hohe (positive) Werte für eine hohe Schwierigkeit und niedrige (negative) Werte für eine geringe Schwierigkeit stehen.

Konstruktionsrational des Tests *Gleichungen*:

Item Nr	Wert der Variable			verwendete Zahlen		Rechenzeichen- operationen											Häufigkeit			
	+/+	+/-	-/-	Einer	Zehner	+	-	x	:	Bruch	x ² 1-15	√x 1-15	x im Nenner	x ² im Nenner	hoch x	x ³ 1-10	()	umform bar	xy	xy
1,2	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1
3	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1
4,5	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1
6	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1
7	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1
8,9	1	0	0	2	0	0	1	1	0	0	0	0	0	0	0	0	1	1	1	1
10	1	0	0	4	0	1	1	2	0	0	0	0	0	0	0	0	1	1	1	1
11,12	1	0	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	1	1
13	1	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	1	1	1
14,15	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	1	1
16,17	0	1	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	1
18,19	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	1	1	1
20,21	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1
22,23	1	0	0	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1
24,25	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	1	1	1	1
26	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	1	1	1
27,28	1	0	0	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0	2	1
29	0	1	0	2	0	0	1	2	0	0	1	0	0	0	0	0	0	0	2	1
30,31	1	0	0	2	0	0	1	1	0	1	1	0	0	0	0	0	0	1	2	1
32	1	0	0	1	0	1	1	0	0	0	0	0	2	0	0	0	0	0	2	2
33,34	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	1
35,36	1	0	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0	1	1
37,38	1	0	0	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	1	1
39	1	0	0	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	2	1
40,41	0	1	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	1
42	1	0	0	2	0	1	1	1	0	0	2	0	0	0	0	0	0	0	2	1
43	0	1	0	2	0	1	1	2	0	0	2	0	0	0	0	0	0	0	2	2
44	0	0	1	2	0	2	0	2	0	0	0	0	0	0	0	0	0	0	2	2

Testdesign des Tests *AN-TOP*:

ItemNr.	1a	1b	2a	2b	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b	8a	8b	9a	9b	10a	10b	11a	11b	12a	12b	13a	13b
TF 1	1				2	3	4	4	5	5		6	7	7		8	9		10		11			12	13	
TF 2		1			2			4	5		6				7	8		9		10	11		12			13
TF 3	1		2				4			5	6		7		8			9	10			11	12		13	

14a	14b	15a	15b	16a	16b	17a	17b	18a	18b	19a	19b	20a	20b	21a	21b	22a	22b	23a	23b	24a	24b	25a	25b	26a	26b	27a	27b	28a	28b
	14	15		16		17			18	19		20	20	21		22		23				24	25			26	27	28	
	14		15		16	17		18			19	20	20		21		22	23		24			25		26		27		28
14			15	16				17	18	19		20			21	22			23	24		25		26			27	28	

Itemparameter des Tests AN-TOP:

Itemnummer	Itemparameter	Itemnummer	Itemparameter
2a	-3,29	15b	0,23
2b	-4,25	16a	0,75
3a	0,03	16b	0,92
3b	0,08	17a	1,96
4b	-1,54	17b	1,83
5a	-4,25	18a	0,41
5b	-2,57	18b	0,65
6a	-0,75	19a	-1,38
6b	-1,79	19b	-0,89
7a	-2,09	20a	3,33
7b	-1,12	21a	-1,53
8a	-0,50	21b	-1,17
8b	-0,84	22a	1,46
9a	-1,42	22b	1,14
9b	-1,04	23a	-0,14
10a	1,09	24a	-0,86
10b	1,44	24b	0,58
11b	1,46	25a	0,33
12a	1,15	25b	0,49
12b	0,91	it26a	-0,58
13a	-0,66	26b	-0,13
13b	-0,44	27a	3,54
14a	0,89	27b	3,58
14b	1,23	28a	1,89
15a	0,38	28b	1,48

Anmerkung: Itemparameter sind Itemschwierigkeitsparameter, wobei hohe (positive) Werte für eine hohe Schwierigkeit und niedrige (negative) Werte für eine geringe Schwierigkeit stehen.

Claudia Gamsjäger

Kleinhain 30
3107 St. Pölten

Tel: +43 650 99 50 277

E-Mail:

a0447130@unet.univie.ac.at

Bundesministerium für Inneres

Abt I/9 (SIAK)

zH. Sg. Hr. Direktor Dr. Norbert Leitner

Bildungszentrum Wien
Marokkanergasse 4
1030 Wien

St. Pölten, am 05.03.2012

Betreff: Ansuchen um eine Bewilligung für die Datenerhebung im Rahmen einer Diplomarbeit an den Ausbildungszentren der Bundespolizei

Ich studiere Psychologie an der Universität Wien und schreibe meine Diplomarbeit am Institut für Angewandte Psychologie: Gesundheit, Entwicklung und Förderung, Fakultät für Psychologie, unter der Leitung von Univ.-Prof. Dr. Mag. Klaus D. Kubinger.

In meiner Diplomarbeit beschäftige ich mich mit einem neu konstruierten psychologisch-diagnostischen Verfahren mit dem Arbeitstitel „Gleichungen“ zur Erfassung der kristallinen, numerischen Intelligenz (*Reasoning*). Die Testpersonen sollen dabei mathematische Gleichungen lösen, um Rückschlüsse über ihr logisch-schlussfolgerndes Denken im numerischen Bereich zu erhalten.

Untersuchungsablauf:

Es ist geplant das psychologisch-diagnostische Verfahren in der Gruppe zu administrieren. Somit ist es möglich mehrere Auszubildende zu einem Zeitpunkt zu testen. Die Auszubildenden müssten sich in einem von der Bundespolizei zur Verfügung gestellten Raum einfinden, in dem die Testungen stattfinden können (zB. Klassenzimmer oder Schulungsraum). Der Raum muss, abgesehen von ausreichend Tischen und Stühlen, über keine weitere Ausstattung verfügen.

Zusätzlich zu dem neu konstruierten Verfahren „Gleichungen“ wird der AN-TOP (*Alpha Numeric Topologies*) (Kubinger & Heuberger, in Vorb.) vorgegeben. Dieser erfasst ebenso das logisch-schlussfolgernde Denken, dabei im verbalen, figuralen und numerischen Bereich. Die Vorgabe dieses zweiten Verfahrens soll zur Validierung der „Gleichungen“ dienen.

Der dritte Teil der Testung besteht aus dem von Fr. Ünal im Rahmen ihrer Diplomarbeit ebenfalls neu konstruiertem Verfahren „Figurale Analogien“. Dieses erfasst die kristalline Intelligenz (*Allgemeinwissen*).

Die Testung dauert insgesamt maximal sechzig Minuten.

Sobald eine offizielle Genehmigung für die Testungen vorliegt, werden Termine fixiert, an denen Testleiter(innen) zum Testen kommen dürfen. Bei den Terminvereinbarungen legen wir großen Wert darauf, den Ablauf des regulären Unterrichts so wenig wie möglich zu unterbrechen.

Als Testleiter(innen) fungieren neben mir selbst, Diplomanden(innen) des Arbeitsbereichs Psychologische Diagnostik der Fakultät für Psychologie, die sowohl über theoretische als auch praktische Erfahrung beim Testen verfügen.

Insgesamt sollen die Daten von 250 – 300 Personen erhoben werden, um entsprechende statistische Analysen vornehmen zu können.

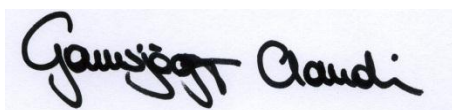
Die gewonnen Daten werden im Sinne des Datenschutzes ausschließlich für wissenschaftliche Zwecke verwendet. Jedem Auszubildenden wird zu Beginn der Testung ein Probanden-Code zugewiesen, um größtmögliche Anonymität zu gewährleisten.

Folgende Ausbildungszentren haben sich bereits zur Teilnahme an der Studie, vorbehaltlich der Bewilligung durch das Ministerium, bereit erklärt:

- Bildungszentrum Traiskirchen (GenMjr Mag. Arthur Reis)

Über eine positive Entscheidung Ihrerseits würde ich mich sehr freuen.

Mit der Bitte um Unterstützung und freundlichen Grüßen,

A handwritten signature in black ink on a light blue background. The signature is written in a cursive, flowing style and reads "Gamsjäger Claudia".

Claudia Gamsjäger

INFORMATION ZU DER TESTUNG AM 20. APRIL 2012

Liebe LehrerInnen und liebe Auszubildende!

Am 20. April finden in Ihrer Klasse in der 5. Stunde Testungen statt. Zu diesem Zweck kommen Psychologie-Studentinnen der Universität Wien in die Schule.

Worum geht's?

Die Studentinnen haben im Rahmen ihrer Diplomarbeit neue psychologisch-diagnostische Verfahren konstruiert, die sie nun an ca. 300 Personen erproben müssen.

Die Testung besteht aus drei aufeinanderfolgenden Teilen, mit denen logisch-schlussfolgerndes Denken in verschiedenen Bereichen sowie Allgemeinwissen erfasst werden soll und dauert in etwa 60 Minuten.

Bringen Sie bitte ein Schreibgerät (zB. Kugelschreiber) zur Testung mit.

Was passiert damit?

Die Ergebnisse der Testungen haben keinen Einfluss auf Ihre schulischen Bewertungen. Die gewonnen Daten werden ausschließlich für wissenschaftliche Zwecke benutzt und sind anonymisiert.

Für weitere Fragen stehen Ihnen die Testleiterinnen gerne zur Verfügung.

Lebenslauf

CLAUDIA GAMSJÄGER

Allgemeine Angaben

Adresse: Kleinhain 30
3107 St. Pölten
E-Mail: claudiagams@gmx.net
Geboren am: 29.10.1982
Staatsbürgerschaft: Österreich

Ausbildung

ab SS 2005 Psychologiestudium an der Universität Wien
1997-2002 Bundeslehranstalt für Tourismus, Krems
1993-1997 Hauptschule, Oberwölbling

Berufserfahrung / Praktika

seit 2010 Erzieherin im Schülerinternat der Kolpingsfamilie St. Pölten
2009 6-Wochen-Praktikum in der Psychiatrischen Abteilung für
Kinder und Jugendliche im Krankenhaus Hietzing, Wien
2002-2009 Wirtschaftskammer Österreich, Aussenwirtschaft Österreich
Assistentin im Bereich Märkte und Auslandsveranstaltungen