universität wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Item-generierende Regeln bei AN-TOP“

verfasst von / submitted by

Nina Heuberger

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magistra der Psychologie (Mag.rer.nat)

Wien, 2016 / Vienna, 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 298
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Diplomstudium Psychologie
Betreut von / Supervisor:	Univ.-Prof. i.R. Mag. Dr. Klaus Kubinger

Danksagung

An erster Stelle möchte ich Herrn Univ.- Prof. Dr. Mag. Klaus D. Kubinger danken, einerseits für Ihre fachlichen Ratschläge, Ihre Unterstützung und die kritische Auseinandersetzung mit meinen Ideen, andererseits würde diese Arbeit nicht ohne Ihr Vertrauen, Ihre Zuversicht und vor allem Ihre Geduld existieren. Sie haben mich für die Psychologische Diagnostik begeistert und mir ermöglicht tiefe Einblicke in die Arbeit eines Wissenschaftlers zu bekommen. Vielen Dank!

Ein weiterer Dank geht an Bettina Hagenmüller und Jan Steinfeld, für viele fachliche Diskussionen, statistische Ratschläge und die gemeinsam verbrachte Zeit.

Großer Dank geht an die Schulen, die sich bereit erklärten an dieser Untersuchung teilzunehmen.

Gleich anschließend geht ein großes Dankeschön an meine Tante Barbara, die mich schon von klein auf für die Psychologie begeisterte und mir in vielen fachlichen und persönlichen Dingen Vorbild und Stütze ist.

Ohne die vielfältige Unterstützung meiner Eltern und Großeltern, die mir ein Studium ermöglichten, mir da weiter halfen, wo ich alleine nicht weiter gekonnt hätte, und mir stets das Gefühl gaben auf dem richtigen Weg zu sein, wäre diese Arbeit nie entstanden. Danke für euer Vertrauen und eure Unterstützung.

Schließlich möchte ich meinen Freunden und Studienkollegen für die vielen offenen Ohren, kritischen Auseinandersetzungen, hilfreichen Ratschläge und Momente der Entspannung danken.

Und zuletzt: Merci Thierry, merci pour tout.

Abstract - Deutsch

Die vorliegende Arbeit thematisiert die Konstruktion eines nonverbalen Tests zum schlussfolgernden Denken basierend auf einem neuen Testkonzept, unter Zuhilfenahme der regelgeleiteten Itemkonstruktion als optimale Variante zur Sicherung der Konstruktvalidität auf Itemebene. Ziel der Arbeit war es, auf Basis von itemgenerierenden Regeln, einen Rasch-Modell-konformen Test zur Messung des Konstrukts Reasoning bei Kindern zwischen 6 und 15;11 Jahren zu entwickeln. Der theoretische Teil der Arbeit gibt einen Überblick über die Literatur zur regelgeleiteten Testkonstruktion sowie zum Konstrukt Reasoning. Außerdem wird auf die probabilistische Testtheorie als Basis zur regelgeleiteten Testkonstruktion eingegangen. Auf Basis gängiger Taxonomien zum Lösen von Aufgaben zum schlussfolgernden Denken wurde zunächst ein Konstruktionsrational zur regelgeleiteten Itemgenerierung entwickelt und auf die Aufgaben der AN-TOP (*A*lpha-*N*umerical *Top*ologies) angewendet. Anschließend wurde der Test an einer Stichprobe von 423 Schülerinnen und Schülern erprobt. Für die Items der AN-TOP konnte nach dem sukzessiven Ausscheiden von etwa 8% der Items a posteriori Modellgeltung für den verbleibenden Itempool erreicht werden. Um die Geltung des Regelsystems zu überprüfen, wurde im nächsten Schritt der Rasch-Modell-konforme Itempool unter Verwendung des Linear-Logistischen Testmodells analysiert, wobei die Gewichtsmatrix entsprechend dem Konstruktionsrational definiert wurde. Die Schätzung hielt dem Vergleich mit dem Rasch-Modell stand. Dies bedeutet inhaltlich, dass die Itemschwierigkeiten, ebenso gut durch das für die Itemkonstruktion verwendete Regelsystem erklärt werden können wie durch die Schätzung durch das Rasch-Modell. Die vorliegende Studie stellt somit einen Rasch-Modell-konformen Test zur Messung von Reasoning bei Kindern von 6 bis 15;11 Jahren zur Verfügung, deren Validität bereits auf Itemebene gegeben ist.

Schlagwörter: regelgeleitete Itemkonstruktion, LLTM, Reasoning,

Abstract - English

The following thesis is about the construction of a nonverbal reasoning test for children aged 6 to 15;11 based on a new test concept. It underlines rule-based item construction as an ideal construction strategy to ensure construct validity on item level. The goal of the study was to develop a Rasch model conform reasoning test on the basis of an item construction rational. The theoretical section of this thesis provides an overview of reasoning and rule-based test construction. It also introduces probabilistic test theory as its methodological basis. In the study, a construction principle for rule-based item generation was developed on the basis of common taxonomies for reasoning tests. The test AN-TOP (*A*lpha-*N*umerical *Top*ologies) was then tested on a sample of 423 children and adolescents. After roughly 8% of the items were successively eliminated, Rasch model analyses of the test showed a posteriori model fit for the remaining items. In order to test the construction principle, the Rasch model conform item pool was analyzed using the linear logistic test model (LLTM), with the weight matrix defined according to the construction principle. The estimation according to the LLTM was comparable to the Rasch model, showing that item difficulties are adequately explained by the construction rational. The following thesis thus provides readers with a Rasch model conform reasoning test for children aged 6 to 15;11, whose validity is ensured on item level.

Key Words: rule-based item construction, LLTM, reasoning,

Inhaltsverzeichnis

I. Einführung

Die Fähigkeit, Probleme zu lösen und sich durch Schlussfolgerungen im Alltag zu Recht zu finden, wird in unserer Gesellschaft als zentraler Indikator für Intelligenz angesehen. Auch in der psychometrischen Tradition wird dem schlussfolgernden Denken seit jeher große Bedeutung beigemessen. Verfahren zur Erfassung dieses Merkmals, unabhängig von verbalen Fähigkeiten, wurden schon früh im Rahmen der Intelligenzstrukturforschung konstruiert und werden auch heute noch rege neu und weiter entwickelt (siehe beispielsweise zu diesem Thema Stemmler, Hagemann, Amelang & Bartussek, 2011 sowie Suen & French, 2003).

Ein besonders häufig verwendetes Testkonzept zur nonverbalen Erfassung von schlussfolgerndem Denken, beziehungsweise *Reasoning*, stellen geometrische Testaufgaben dar. Beispielhaft zu nennen und besonders weit verbreitet sind die Testreihen von Raven (für deutschsprachige Manuale siehe Heller, Kratzmeier & Lengfelder, 1998 a, b) sowie die Reihe der Culture Fair Tests von Cattell und Horn (für die deutsche Version siehe Weiß, 2006). In Kapitel 1 wird auf die Grundlagen dieser Tests, die Theorie zum logischen Denken und Schlussfolgern sowie deren Umsetzung in psychologisch-diagnostischen Verfahren näher eingegangen.

In Anlehnung an Cattells Theorie der Intelligenz (Cattell, 1963 und Cattell & Horn, 1978) und kognitionspsychologischen Überlegungen, die sich darauf stützen (Carpenter, Just & Shell, 1990; Embretson 1995, 1998; Kunda, McGreggor & Goel, 2013; Sternberg, 1977), können die Anforderungen der Aufgaben an die Testperson in drei Schritte aufgeteilt werden. Nach der Wahrnehmung eines Sachverhalts, bei Matrizenaufgaben sind dies zumeist mehr oder weniger regelmäßig gestaltete geometrische Formen, folgt das Erkennen von Regeln zwischen den wahrgenommenen

Elementen der Aufgabe durch die Suche von Korrespondenzen und schließlich die analoge Anwendung der gefundenen Regeln auf neue Elemente.

Aufgrund der hohen Formalisierbarkeit dieser Anforderungen und der Möglichkeit, sowohl die zugrundeliegenden Regeln und Elemente der Aufgaben detailgenau zu definieren, wurden schon vor einigen Jahrzehnten Versuche unternommen Matrizentestaufgaben unter Zuhilfenahme von kognitionspsychologischen Modellen regelgeleitet zu konstruieren. Als wesentliche Ausgangspunkte sind in diesem Zusammenhang die Studien von Formann (1973), Hornke und Habon (1986), sowie Embretson (1995, 1998) zu erwähnen. Als Gemeinsamkeiten liegen den erwähnten Studien die Formulierung von Materialeigenschaften beziehungsweise systematischen Item-Konstruktionsregeln, die die Schwierigkeit der Aufgaben bedingen sollen, sowie das Streben nach Aufgaben mit einem hohen psychometrischen Niveau zugrunde.

Während die eben genannten Ansätze verwendet wurden um Tests mit einem vor der Testvorgabe definierten Itempool zu konstruieren, der sowohl klassisch mit festgelegter Itemreihenfolge als auch adaptiv vorgegeben werden kann (siehe beispielsweise Kubinger, 2009a, zu den Vor- und Nachteilen des adaptiven Testens), so ermöglichen heutzutage technologische Fortschritte, insbesondere die Existenz von leistungsstarken Computern, Weiterentwicklungen in der Testkonstruktion. So konnten etwa Holling, Bertlings und Zeuch (2009) psychometrisch hochqualitative Items „on the fly“ erzeugen, also während der Testvorgabe und individuell auf das Leistungsniveau der Testperson zugeschnitten.

Kapitel 2 der vorliegenden Arbeit geht näher auf kognitionspsychologische Grundlagen sowie deren Einfluss auf die Komplexität eines Items ein. Während sich in Kapitel 3 Erläuterungen zur Vorgehensweise bei der regelgeleiteten Itemkonstruktion finden, beschreibt Kapitel 4 die zugrundeliegende Methodik.

In Zusammenhang mit regelgeleiteter Itemkonstruktion ist die empirische Gültigkeit des dichotomen logistischen Modells von Rasch (1960), welches das Grundmodell der Item-Response-Theorie darstellt (IRT; siehe z. B.: Hambleton, Swaminathan & Rogers,

1991; Kubinger, 1989), von zentraler Bedeutung. Darauf aufbauend stellt die Gültigkeit des Linear Logistischen Testmodells (LLTM; Scheiblechner 1972; Fischer 1973) eine methodische Grundlage für diesen Zugang zur Testkonstruktion dar (Kubinger, 2008; 2009b).

Der in der vorliegenden Studie behandelte Test AN-TOP ist ein sprachfreier Leistungstest mit dichotomem Antwortformat, der das Konstrukt Reasoning unter Anwendung logischer Operationen zwischen figuralen Elementen zu messen beansprucht. Wie bereits eingangs erwähnt, stellen geometrische Testaufgaben einen großen Teil der verwendeten Verfahren zur Messung schlussfolgernden Denkens dar und sind auch häufig in Intelligenztestbatterien enthalten. Ziel der vorliegenden Studie ist es, einen Test zum schlussfolgernden Denken zu entwickeln, der einem völlig neuen Testkonzept folgt und somit sowohl für Personen, die regelmäßig Reasoningtests absolvieren als auch für Testnovizen eine faire Messung ermöglicht. Dies folgt der Annahme, dass die wiederholte Teilnahme an Testungen mittels gleichartiger Aufgaben zu höheren Testwerten führen kann (Mackey, Hill, Stone & Bunge, 2011; Salthouse, Schroeder & Ferrer, 2004). Wie in Kapitel 6 detailreicher umrissen wird, beschreibt diese Studie den Versuch der Validierung eines Konstruktionsrationals zur Erstellung eines Itempools für einen Test, der einem neuen und bisher nicht publizierten Konzept folgt und den Faktor Reasoning bei Kindern und Jugendlichen zwischen 6;0 und 15;11 Jahren messen soll. Eine Beschreibung des Testmaterials sowie des verwendeten Konstruktionsrationals, das auf den in Kapitel 2 und 3 vorgestellten theoretischen Grundlagen basiert, findet sich in Kapitel 6.

Kapitel 7 beschreibt das Studiendesign und die Vorgehensweise bei der Datenerhebung sowie die verwendete Stichprobe. In Kapitel 8 werden die Ergebnisse der statistischen Berechnungen vorgestellt. In der Diskussion im darauffolgenden Kapitel 9 werden schließlich die positiven Aspekte und Mängel der empirischen Arbeit beleuchtet. In diesem Kapitel findet sich auch ein Ausblick auf zukünftige Forschungsfelder. Eine Zusammenfassung der Arbeit wird schließlich in Kapitel 10 gegeben.

II. Theoretischer Hintergrund

1. Das Konstrukt Reasoning

Kubinger (2009a) definiert das Konstrukt Reasoning als „die Fähigkeit, Gesetzmäßigkeiten oder logisch zwingende Zusammenhänge erkennen und zweckentsprechend verwerten zu können“ (S. 206).

Reasoning, oder auch logisches beziehungsweise schlussfolgerndes Denken ist dementsprechend allgegenwärtig. Dies zeigt sich auch dadurch, dass das Lösen logischer Probleme sowie der beteiligten Prozesse Gegenstand der Forschung in den verschiedensten Disziplinen ist, wie etwa der Psychologie, der Philosophie, der Neurowissenschaften aber auch der Forschung im Bereich künstlicher Intelligenz und Linguistik. Speziell in der psychometrischen Forschungstradition wurden diesen Denkprozessen - nicht zuletzt aufgrund der beobachteten hohen Korrelation des Merkmals mit allgemeiner Intelligenz - seit jeher hohe Bedeutung beigemessen. Verfahren zur Erfassung dieses Merkmals wurden schon früh im Rahmen der Intelligenzstrukturforschung konstruiert (siehe beispielsweise zu diesem Thema Stemmler et al., 2011, und Suen & French, 2003).

Eine etwas psychometrisch orientiertere Definition findet sich bei Thurstone & Thurstone (1941; zitiert nach Stemmler et al., 2011), die den Primärfaktor Reasoning als „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 Elementes“ (S. 149) definieren.

Im folgenden Kapitel sollen einerseits die Grundlagen des schlussfolgernden Denkens aus Sicht der Logik und der kognitiven Psychologie erläutert werden. Andererseits wird näher auf Reasoning als Teil gängiger Intelligenztheorien sowie die Operationalisierung von schlussfolgerndem Denken in der psychologischen Diagnostik eingegangen.

1.1 Schlussfolgern aus Sicht der Logik

In der formalen Logik besteht ein logischer Schluss bzw. Syllogismus aus drei Aussagen: zwei Prämissen und einer Konklusion. Um zur richtigen Konklusion zu kommen, muss die Information aus den beiden Prämissen logisch richtig verwendet werden.

Laut Fröhlich (2010) sowie Beckmann und Guthke (1999) können logische Schlüsse in drei Kategorien unterteilt werden, nämlich deduktive, induktive und analoge Schlüsse. Im Folgenden wird auf diese drei Arten näher eingegangen.

Induktiver Schluss

Beim induktiven Schließen wird durch das Erkennen von Regelmäßigkeiten und Gesetzen sowie durch Wahrscheinlichkeiten und Erfahrungen vom Speziellen auf eine unendliche Menge von gleichartigen und bisher unbeobachteten Fällen, beziehungsweise das Allgemeine geschlossen (Fröhlich, 2010 S.257).

Da von einer endlichen Zahl an Einzelfallbeobachtungen auf alle zukünftigen Fälle geschlossen wird, ist die Konklusion mit Unsicherheit verbunden. Die Konklusionen können nicht eindeutig behauptet werden, da nicht genügend Information in den Prämissen steckt (vgl. Beckmann und Guthke, 1999, Fröhlich, 2010). Man könnte einen induktiven Schluss daher durch eine einzige widersprüchliche Beobachtung leicht widerlegen. Beckmann und Guthke (1999) betonen allerdings, dass sich bei dieser Art des logischen Schließens neue Erkenntnisse ergeben können, da das Wissen durch die gegebenen Prämissen in der Konklusion erweitert wird.

Induktives Schlussfolgern wird vor allem in den Erfahrungswissenschaften angewandt, man findet es aber auch in unserem alltäglichen Leben wieder. So lehren uns beispielsweise unsere Erfahrungen, dass Zitronen sauer und Bonbons süß sind.

Obwohl in Zusammenhang mit allgemeiner Intelligenz zumeist der Aspekt des induktiven Schlussfolgerns untersucht wurde und auch Verfahren zur Erfassung von schlussfolgerndem Denken in der Literatur häufig mit dieser Form des logischen Schlusses gleichgesetzt werden, betonen jedoch besonders Beckmann und Guthke (1999), dass induktives Schließen nur ein wesentlicher Bestandteil von Reasoning ist.

Deduktiver Schluss

Deduktive Schlüsse stützen sich auf Prämissen, die das bereits vorhandene Wissen repräsentieren. Das Schließen erfolgt somit vom Allgemeinen auf das Spezielle und der Schluss, der sich zwingend ergibt, ist somit gesichert (Fröhlich, 2010, S.128).

Bei dieser Art des logischen Schließens werden demnach Sachverhalte erkannt, die durch die Prämissen impliziert werden. Die in den Prämissen enthaltene Information muss dafür nicht inhaltlich richtig sein um eine Deduktion zu erlauben, wichtig ist laut Beckmann und Guthke (1999) viel mehr, dass es bei der Konklusion zu keiner Informationserweiterung kommt. Diese Art des Schließens findet sich vor Allem bei der Anwendung von generellen Prinzipien und Regeln (Beckmann & Guthke, 1999; Fröhlich, 2010).

Logisches Schließen in psychologisch-diagnostischen Verfahren erfordert oft eine erfolgreiche Anwendung von Deduktion und Induktion, eine strenge Unterscheidung ist daher laut Beckmann und Guthke (1999) artifiziell und unzureichend.

Analoger Schluss

Der analoge Schluss bezieht sich auf die Gleichwertigkeit von Gegenständen beziehungsweise Vorgängen und ist dem induktiven Schluss damit ähnlich (Fröhlich, 2010 S. 58).

Beim analogen Schluss wird zum Beispiel eine Ähnlichkeit zwischen den Verhältnissen A:B sowie C:D erkannt. Äquivalent zum induktiven Schluss sind analoge Schlüsse ebenfalls nur wahrscheinlich (Beckmann & Guthke, 1999).

1.2 Reasoning als Kernfaktor vieler Intelligenztheorien

„Reasoning is a thinking activity that is of crucial importance throughout our lives. Consequently, the ability to reason is of central importance in all major theories of intelligence structure.“ (Wilhelm, 2005, S. 373).

In fast allen gängigen Intelligenztheorien, die im Rahmen der Intelligenzstrukturforschung der letzten hundert Jahre postuliert wurden, findet sich der Faktor Reasoning an mehr oder weniger prominenter Stelle wieder. Das folgende Kapitel soll einen kurzen Überblick hierzu geben.

Aufgrund positiver Korrelationen zwischen verschiedenen kognitiven Leistungen ging Charles Spearman (1904, zitiert nach Stemmler et al. 2011) von der Annahme aus, dass allen kognitiven Leistungen ein gemeinsamer Faktor zu Grunde liegt, welchen er als Generalfaktor *g* bezeichnete. Die verbleibende, nicht durch *g* erklärbare Varianz des Tests erklärte Spearman durch zusätzliche spezifische Faktoren *s*, wobei es so viele spezifische Faktoren geben kann wie Aufgabentypen bzw. Tests existieren. Seiner Annahme folgend, erfordern einzelne Aufgaben jeweils unterschiedlich hohe *g*-Anteile, so dass der jeweilige spezifische Anteil entsprechend variiert (Stemmler et al., 2011).

Spearman postulierte in seiner Zweifaktorentheorie der Intelligenz keinen speziellen Faktor für schlussfolgerndes Denken, berichtete aber bereits 1927 (vgl. Sternberg & Williams, 1998) über Korrelationen von r = .68 bis r = .84 zwischen Tests zum induktiven Denken und dem *g*-Faktor. Zwei Jahrzehnte später räumte er dem Erkennen von Relationen und dem Finden von Regeln, neben der Fähigkeit zum räumlichen

Vorstellungsvermögen, die wichtigste Bedeutung beim Erfolg in allen Tests ein, die auf dem *g*-Faktor hoch laden (vgl. Sternberg & Williams, 1998).

Auf der Grundlage der multiplen Faktorenanalyse entwickelte Louis Thurstone (1938, zitiert nach Stemmler et al. 2011) ein Modell, das die Existenz mehrerer unabhängiger Einzelfaktoren (Primärfaktorenmodell) annimmt und im Gegensatz zum Modell von Spearman nicht von der Existenz eines (berechenbaren) Generalfaktors ausgeht. Zunächst ging Thurstone (1938, zitiert nach Stemmler et al., 2011) von neun primären Intelligenzfaktoren aus, später sprach er nur noch von sieben (Thurstone & Thurstone, 1941, zitiert nach Stemmler et al., 2011). Bei diesen sieben bereichsspezifischen Intelligenzfaktoren handelt es sich um räumliches Vorstellungsvermögen (Space, S), Wahrnehmungs- und Auffassungsgeschwindigkeit (Perceptual Speed, P), Rechenfähigkeit (Number, N), Merkfähigkeit bzw. Kurzzeitgedächtnis (Memory, M), Wortflüssigkeit bzw. Leichtigkeit der Wortfindung (Word Fluency, W), verbales Verständnis (Verbal Comprehension, V), und Regelerkennen bzw. schlussfolgerndes Denken (Reasoning, R). R stellt einen relativ heterogenen Faktor dar, der sich nach Thurstone aus den Unterfaktoren I (Induktion) und D (Deduktion) zusammensetzt, wobei I für R bedeutsamer ist (Stemmler et al., 2011).

Cattell (1963; siehe auch: Horn & Cattell 1966 sowie Cattell & Horn 1978) geht in seinem Modell der Intelligenz zunächst von zwei übergeordneten Generalfaktoren aus, dem General Fluid Ability-Faktor g_f und dem General Crystallized Ability-Faktor g_c. Unter fluider Intelligenz versteht Cattell die angeborene und überwiegend von gesellschaftlichen und kulturellen Einflüssen unabhängige Fähigkeit, sich neuartigen Situationen anzupassen und bisher unbekannte Probleme zu lösen ohne auf zuvor angeeignetes Wissen zurückzugreifen. Dies entspricht beispielsweise den messbaren psychologischen Konstrukten der Verarbeitungsgeschwindigkeit, dem schlussfolgernden Denken und der Fähigkeit zum Erkennen von Mustern. Demgegenüber steht die kristalline Intelligenz, die Cattell als die Fähigkeit beschreibt, angeeignetes Wissen, Fähigkeiten und Erfahrung zu benutzen. Fluide Intelligenz wird vor allem über Aufgaben zum schlussfolgernden Denken, zur intellektuellen Geschwindigkeit und zum Erkennen von figuralen Beziehungen erfasst. Kristalline

Intelligenz wird hingegen meistens durch allgemeine Wissenstests sowie Wortschatztests gemessen (Cattell, 1963 und Cattell & Horn, 1978).

Erwähnenswert ist auch, dass Intelligenztests in allen Altersstufen eine Kombination des kristallinen und des fluiden Faktors messen. Während der fluide Faktor g_f im Kindes- und Jugendalter besonders stark ausgeprägt zu sein scheint, ist der kristalline Faktor g_c vor allem bei älteren Personen für Leistungsspitzen verantwortlich (Cattell, 1941, Horn & Cattell, 1967, Schaie, 2005). Die Korrelation der beiden Faktoren g_f und g_c variiert je nach Population, deren Unabhängigkeit wurde aber mehrfach wiederlegt (siehe z. B. Carroll, 1993, Cattell, 1963 und Cattell & Horn, 1978).

Auch im Modell von Carroll (1993, 2005) spielt schlussfolgerndes Denken eine wichtige Rolle. Die „Three Stratum Theory of Cognitive Abilities“ weist drei verschiedene Hierarchieebenen auf, die von einem Generalfaktor ähnlich wie bei Spearman dominiert werden. Dieser *g*-Faktor (Stratum III) ist psychologisch wie Spearmans *g* zu interpretieren und stellt eine übergeordnete allgemeine Intelligenz dar, welche eine hohe Generalität aufweist und allen intellektuellen Aktivitäten zu Grunde liegt. Auf der darunter liegenden Schicht (Stratum II) finden sich acht breite Faktoren: fluide Intelligenz g_f, kristalline Intelligenz g_c, Gedächtnis- und Lernfähigkeit, visuelle und auditive Wahrnehmung, Abruffähigkeit, kognitive Schnelligkeit und Verarbeitungsgeschwindigkeit. Dass allgemeine Intelligenz eine sehr enge Beziehung zu Aufgaben aufweist, die schlussfolgerndes Denken verlangen, spiegelt sich empirisch in besonders hohen *g*-Ladungen wieder (vgl. Carroll, 1993).

Die unterste Schicht (Stratum I) besteht aus 69 spezifischen Fähigkeiten, die den verschiedenen Faktoren der zweiten Schicht zugeordnet werden können. Fluide Intelligenz g_f wird dabei durch drei verschiedene Fähigkeiten zum schlussfolgernden Denken definiert: induktives Denken, quantitatives Schlussfolgern und sequentielles bzw. deduktives Schlussfolgern. Methodisch sind vor allem Tests zu Zahlenfolgen (quantitatives Schlussfolgern), sprachliche Analogien (deduktives Schlussfolgern) und Figurenfolgen (induktives Schlussfolgern) in diesem Bereich anzusiedeln (vgl. Beckmann & Guthke, 1999).

1.3 Diagnostik des schlussfolgernden Denkens

Schlussfolgerndes Denken ist also eine weithin anerkannte und vielfach untersuchte Fähigkeit, die es uns ermöglicht, uns in unserer Umwelt zu Recht zu finden. Da sich bei diesem psychologischen Konstrukt interindividuelle Unterschiede in der Ausprägung und Entwicklung beobachten lassen, ist es der Psychologie, wie bereits eingangs erwähnt, seit Beginn der Intelligenzforschung ein besonderes Anliegen diese Fähigkeit zu messen.

Eine mögliche Charakterisierung von Tests zum schlussfolgernden Denken kann über die verwendeten Aufgabentypen erfolgen. Beckmann und Guthke (1999) unterscheiden zum Beispiel Matrizenaufgaben, Vervollständigung von Reihen (zum Beispiel Buchstaben- und Zahlenreihen), Klassifikationsaufgaben, Analogien, und Syllogismen. Je nach Aufgabentyp ist für die Bearbeitung eine unterschiedliche Art des schlussfolgernden Denkens oder eine Kombination aus induktiven, deduktiven und analogen Schlüssen notwendig. Eine weitere Kategorisierung kann bezüglich des verwendeten Aufgabenmaterials getroffen werden. Die meisten Verfahren erfassen die Fähigkeit Reasoning über numerisches, verbales oder figurales Aufgabenmaterial beziehungsweise durch eine Kombination dieser Darbietungsweisen.

2. Kognitive Prozesse als Grundlage von Reasoning

Die Forschung im Randgebiet zwischen der psychologischen Diagnostik und der kognitiven Psychologie beschäftigt sich mit den grundlegenden geistigen Prozessen, die der Bearbeitung von Aufgaben zum schlussfolgernden Denken, zugrundeliegen.
In diesem Kapitel sollen zuerst allgemeine Problemlösungsprozesse erläutert werden. Anschließend findet sich ein Abschnitt zu Faktoren, die die Komplexität von Reasoningaufgaben beeinflussen.

2.1 Problemlösungsprozesse nach Sternberg

Sternberg (1977) zerlegt den Problemlösungsprozess bei Analogieaufgaben und identifiziert fünf Komponenten beziehungsweise Denkschritte, die während der Lösung dieser Probleme durchlaufen werden. Diese fünf Schritte sollen anhand des folgenden Beispiels beschrieben werden (vgl. Sternberg, 1977):

Kuh : Kalb = Stute : ?

Frosch – Fohlen – Küken – Welpe – Stier

1. Encoding: Der Schritt des Enkodierens beschreibt die Übersetzung der Aufgabe und ihrer Bestandteile in eine interne Repräsentation, die dazu dient weitere Denkprozesse zu ermöglichen. (Welche charakteristischen Merkmale haben die Elemente Kuh, Kalb und Stute?)
2. Inference: In diesem Schritt wird die Beziehung zwischen den ersten beiden Elementen der Analogie erkannt. (In welcher Beziehung stehen Kuh und Kalb?)

3. Mapping: Beim Mapping steht die Analyse der Beziehung zwischen dem ersten und dem dritten Element der Analogie im Vordergrund. (In welcher Beziehung stehen Kuh und Stute?)
4. Application: Anschließend wird die Regel (auf die Antwortmöglichkeiten) angewandt und die am besten passende ausgewählt. (Wie nennt man das Kind einer Stute? Ist ein Frosch ein Kind einer Stute?)
5. Responding: Die im vorherigen Prozessschritt gefundene Antwort wird gegeben. (Die Lösung lautet Fohlen.)

Laut Sternberg (1977) unterscheiden sich intelligentere von weniger intelligenten Versuchspersonen in der Zeit, die sie auf die einzelnen Schritte verwenden. Weniger intelligente Personen verwenden tendenziell weniger Zeit auf den Schritt des Enkodierens und mehr, beziehungsweise zumindest gleich viel Zeit, auf die anderen Schritte wie intelligentere Personen.

Diese Einteilung in fünf Schritte ist in der Literatur nicht unumstritten. Es finden sich einige weitere Prozessmodelle mit mehr oder weniger Schritten, da in manchen Modellen einige Schritte genauer expliziert oder stärker zusammengefasst sind. Gemein ist den Modellen jedoch der Ablauf der einzelnen Denkprozesse (Primi, 2001).

2.2 Komplexitätsfaktoren bei Reasoningaufgaben

Marshalek, Lohman und Snow (1983) betonen die Wichtigkeit der Unterscheidung zwischen der Komplexität und der Schwierigkeit einer Aufgabe. Ihrer Definition zufolge ist die Schwierigkeit eines Items mit dessen Lösungswahrscheinlichkeit in einer definierten Population gleichzusetzen. Die Komplexität eines Items bezieht sich jedoch auf dessen Korrelation mit *g*.

Primi (2001) identifiziert zwei Hauptfaktoren, die einen signifikanten Einfluss auf die Komplexität von Aufgaben zum schlussfolgernden Denken haben, nämlich die im Item

enthaltene Informationsmenge einerseits sowie die Art der auftretenden Regeln und die perzeptive Organisation der Aufgabe andererseits. Zusätzlich stellt er die Frage nach differentiellen Effekten dieser Faktoren in den Raum. Diese Frage konnte jedoch bisher noch nicht in zufriedenstellendem Ausmaß beantwortet werden. Die beiden folgenden Kapitel gehen näher auf die beiden erwähnten Hauptfaktoren ein.

2.2.1 Informationsmenge

Der Einfluss, der im Item enthaltenen Informationsmenge, also die Anzahl der auftretenden Elemente und der sie verbindenden Beziehungen beziehungsweise Regeln, wurde speziell von Mulholland, Pellegrino und Glaser (1980) untersucht. Durch die systematische Variation der Anzahl der Elemente und Regeln in einem geometrischen Reasoning-Test konnten sie zeigen, dass die Bearbeitungszeit bei Zunahme der Anzahl der Elemente und Regeln mehr steigt als durch eine einfache additive Funktion vorhergesagt werden kann. Sie schlossen daraus, dass bei komplexeren Items eine substantielle externe Gedächtniskomponente benötigt wird, die aber nicht vorhanden ist. Daher müssen zeitintensive alternative Lösungsstrategien benützt werden. Dies trifft speziell bei multiplen Transformationen innerhalb einzelner Items zu, da Zwischenergebnisse sowie die Reihenfolge der Operationen gespeichert und weiterverwendet werden müssen (Carpenter et al. 1990).

Primi (2001) assoziiert diese beiden Variablen mit der Informationsmenge, die im Arbeitsgedächtnis bearbeitet werden muss. Ein höheres Ausmaß an Informationen stellt einen höheren Anspruch an das Arbeitsgedächtnis um die ganze präsentierte prozedurale und deklarative Information zu behalten. Gleichzeitig wird eine bessere Organisation der Ziele und enkodierten Elemente benötigt (Primi, 2001). Der Zusammenhang zwischen Arbeitsgedächtnisleistung und *g* konnte in vielen Studien gezeigt werden. Allerdings gibt es große Schwankungen zwischen den gefundenen Ergebnissen. In einer Metaanalyse von Ackerman, Beier und Boyle (2005) konnte nur etwa 25% der gemeinsamen Varianz erklärt werden, wohingegen die erklärte gemeinsame Varianz bei Bühner, Krumm und Pick (2005) bei etwa 95% liegt.

Chuderski (2013) erklärt die hohen Unterschiede durch die Art der Vorgabeweise: bei reinen Power-Tests scheint der Einfluss der Arbeitsgedächtnisleistung geringer zu sein als bei Speed-and-Power-Tests. Er schließt daraus, dass obwohl Zeitdruck kaum Einfluss auf die Reliabilität eines Tests hat, er sehr wohl die Konstruktvalidität eines Tests substantiell beeinflusst. Speziell Personen mit niedriger Fähigkeit im Bereich Arbeitsgedächtnis hätten ohne Zeitdruck die Möglichkeit, diese Schwäche durch effektivere und komplexere Problemlösestrategien, wie beispielsweise die Unterteilung in eine höhere Anzahl an Subproblemen, Lernen und bessere Abstraktion, zu kompensieren. (siehe hierzu auch Kubose, Holyoak & Hummel, 2002, und Carpenter et al., 1990).

Eine weitere Möglichkeit das Konstrukt Reasoning besonders valide zu messen wird bei Klein Entink, Kuhn, Hornke und Fox (2009) beschrieben. In ihrem Modell werden die Antworten und Antwortzeiten parallel modelliert.

2.2.2 Art der auftretenden Regeln und Perzeptive Organisation

Die Art der Relationen beziehungsweise Regeln zwischen den auftretenden Elementen, ist ein in der Literatur ausführlich behandelter Faktor für Itemkomplexität. Besonders für geometrische Reasoningaufgaben, wie beispielsweise Matrizenaufgaben, liegen sehr ausführlich beschriebene Taxonomien vor. Auch konnte der Zusammenhang zwischen der Art, der im Item auftretenden Regeln und dessen Komplexität vielfach belegt werden (siehe u. a. Carpenter et al., 1990; Dimitrov & Raykov, 2003; Embretson, 1984; Embretson & Daniel, 2008; Forman & Piswanger, 1979; Hornke & Habon, 1986; Mulholland et al., 1980).

Die genannten Studien beschäftigen sich mit unterschiedlichen geometrischen Aufgabentypen und sind daher nicht ohne Weiteres übertragbar. Gemein ist ihnen jedoch häufig eine Klassifikation nach dem geforderten Abstraktionslevel (Primi, 2001). Das in dieser Arbeit verwendete Regelsystem wird in Kapitel 6.3 vorgestellt.

Ein weit weniger häufig behandelter Faktor für Itemkomplexität ist die perzeptive Organisation eines Items. Unter perzeptiver Organisation versteht man die das Item charakterisierenden Gestalt-Prinzipien beziehungsweise visuellen Itemmerkmale. Primi (2001) konnte zeigen, dass die perzeptive Organisation eines Items einen signifikanten Effekt auf die Itemkomplexität hat und etwa 53% der Gesamtvarianz erklärt.

Gestalt-Prinzipien können die Itemkomplexität sowohl erhöhen als auch senken. In perzeptiv komplexen Items ist die Wahrscheinlichkeit höher, dass irrelevante Gruppen von Elementen aufgrund perzeptiver Details geformt werden. Es wird angenommen, dass diese Items besonders die visuelle Komponente des Arbeitsgedächtnisses beanspruchen (Primi, 2001).

Auch Arendasy und Sommer (2005) betonen, dass die perzeptive Organisation eines Items, obwohl in der Forschung bisher wenig beachtet, einen wichtigen Komplexitätsfaktor darstellt. Sie konnten zeigen, dass die Manipulation von Gestalt-Faktoren zwar die Eindimensionalität eines Tests nicht beeinflusst, jedoch einen Einfluss auf die Itemschwierigkeit hat. Speziell in Bezug auf Paralleltestversionen und bei der automatischen Itemkonstruktion müssen also Gestalt-Prinzipien beachtet werden.

3. Regelgeleitete Itemkonstruktion

Vor allem die in den letzten Jahrzehnten stark zunehmende Entwicklung des (computerisierten) adaptiven Testens bedingt das Vorliegen (theoretisch) unbegrenzter oder zumindest sehr großer Itempools um die Vorteile dieser Methode gänzlich nutzen zu können (Arendasy & Sommer, 2012; Hornke 1999; Reckase, 2010; ein Überblick über Vor- und Nachteile des adaptiven Testens findet sich bei Kubinger, 2009a). Regelbasierte Itemkonstruktion ist eine Methode, die es Testautoren ermöglicht, auf ökonomische Art und Weise eine große Menge an Testaufgaben zu konstruieren und gleichzeitig die Konstruktvalidität dieser bereits auf Itembasis zu gewährleisten (vgl. Embretson, 1983).

Bei der regelgeleiteten Itemkonstruktion werden die Faktoren, die die Itemschwierigkeit determinieren, a priori festgelegt (Kyllonen, 2002). Dafür wird oft der *Radicals-Incidentals* Ansatz herangezogen. Der Terminus *Radicals* bezieht sich auf strukturelle Charakteristika der Aufgabe, die einen signifikanten Einfluss auf die Itemschwierigkeit oder Itemdiskrimination haben. Der Begriff *Incidentals* hingegen bezeichnet jene oberflächlichen Merkmale der Aufgabe, die dieser nur ein unterschiedliches Aussehen, jedoch keine systematisch unterschiedliche Schwierigkeit verleihen (Irvine, 2002; Irvine, Dann & Anderson, 1990).

Ein erster Schritt im regelgeleiteten Itemkonstruktionsprozess besteht nun darin, Radicals und Incidentals entweder mit Hilfe von Expertenratings oder durch auf kognitiver Theorie basierenden Überlegungen zu definieren. Dann wird unter Zuhilfenahme dieser gestalterischen Merkmale entweder per Hand oder mit dem Computer ein Itempool konstruiert, der möglichst vielfältige Kombinationen von Radicals und Incidentals realisiert und dadurch ein breit gestreutes Schwierigkeitsniveau der Aufgaben anstrebt (vgl. Freund, Hofer und Holling, 2008; Hornke & Habon, 1986; Irvine, 2002). Manche der konstruierten Items sind *isomorph*,

das bedeutet, dass sie aufgrund derselben Kombination von Radicals psychometrisch ident sind, aufgrund der Verwendung verschiedener Incidentals aber anders aussehen (Bejar, 2002; Hornke & Habon, 1986).

Anschließend wird dieser konstruierte Itempool einer Stichprobe vorgegeben, die der interessierenden Population entstammt. Mit Hilfe der so gewonnenen Daten kann das zugrundeliegende kognitive Modell überprüft und die Schwierigkeiten der Radicals festgestellt werden. Nun kann der Itempool mithilfe des etablierten Konstruktionsrationals um jede sinnvolle Kombination aus Radicals und Incidentals erweitert werden. Die Schwierigkeit eines Items kann nach Anwendung dieser Methode berechnet werden, ohne das spezielle Item vortesten zu müssen. Sie ergibt sich aus einer additiven Funktion der Schwierigkeiten der zur Konstruktion der Testaufgabe verwendeten Teiloperationen, d. h. Radicals (vgl. Kubinger, 2008).

Beispiele für auf diese Art konstruierte psychologisch-diagnostische Verfahren zur Leistungsdiagnostik stellen etwa der *Adaptive Matrizentest* (*AMT*, Hornke, Etzel & Rettig, 1997; Hornke & Habon, 1986), der *Abstract Reasoning Test* (Embretson, 1998) oder der *Wiener Matrizentest* (*WMT*; Formann, 1973; Formann & Piswanger, 1979; *WMT-2*; Waldherr, Formann & Piswanger, 2011) dar.

Die bei diesen Verfahren und in zahlreichen experimentellen Studien (siehe z. B.: Arendasy & Sommer, 2012; Embretson & Daniel, 2008; Freund et al., 2008; Sonnleitner, 2008) eingesetzte Methode zur Schätzung der Schwierigkeiten der Radicals stellt das im nächsten Kapitel näher beschriebene *Lineare Logistische Testmodell* (LLTM; Fischer, 1972) dar. In der Literatur finden sich auch weitere methodische Zugänge, etwa über andere Modelle der probabilistischen Testtheorie (Geerlings, Glas & van der Linden, 2011; Holling et al., 2009), über die Regressionanalyse (Embretson & Daniel, 2008; Gorin & Embretson, 2006) sowie über Strukturgleichungsmodelle (Dimitrov und Raykov, 2003).

Die regelbasierte Itemkonstruktion kommt großteils bei der Konstruktion von Leistungstests zur Anwendung. Eine neuere Entwicklung in der psychologisch-

diagnostischen Forschung stellt die Übertragung des Prinzips auf Persönlichkeitsfragebogen dar (siehe zu diesem Thema beispielsweise Reif, 2008).

Das Bestreben, die Validität eines Tests zu untersuchen geht auf die Anfänge der Testkonstruktion zurück. Wie oben bereits erwähnt, gibt die Methode der regelgeleiteten Testkonstruktion einen Hinweis auf die Validität eines Tests auf Itemebene. Embretson (1983) und Messick (1995) betonen, dass die Modellierung der kognitiven Prozesse, die der Itembearbeitung zugrunde liegen, einer der möglichen Wege ist, die Validität eines Tests nachzuweisen. Embretson (1983) unterscheidet in ihrem wegweisenden Artikel zwischen zwei Ansätzen zum Informationsgewinn in Bezug auf die Validität eines Verfahrens. Einerseits spricht sie von der nomothetischen Spanne eines Verfahrens. Hiermit sind Korrelationen des Verfahrens mit externen Faktoren gemeint, wie etwa Testwerte in anderen Tests, die das gleiche Konstrukt zu messen beanspruchen. Als zweiten Ansatz beschreibt sie die Konstruktrepräsentation. Hiermit ist vielmehr der Bezug eines Verfahrens auf die ihm zugrundeliegende Theorie gemeint. Die beschriebene Vorgehensweise der regelgeleiteten Itemkonstruktion ist ein möglicher Zugang um die Konstruktrepräsentation eines Verfahrens zu überprüfen (vgl. Embretson, 1983; 1995; 1998; Embretson & Gorin, 2001).

Diese Methode der Testkonstruktion birgt noch weitere Vorteile, beispielsweise die Möglichkeit, sich beim Testdesign auf kognitive Komplexitätsfaktoren zu stützen, die nachweislich die Itemschwierigkeit beeinflussen, effiziente und gezielte Itemkalibrierung mit kleineren Stichprobengrößen sowie die Konstruktion von Items mit vorhersagbaren psychometrischen Charakteristika. Nach mehrfacher erfolgreicher Bestätigung des Konstruktionsrationals kann das computerisierte adaptive Testen auch auf eine neue Ebene gebracht werden. Anstatt Items aus einem vorgegebenen (endlichen) Pool auszuwählen, können Items „on the fly“ generiert werden, die einen optimalen Informationsgewinn über die Fähigkeit der Testperson ermöglichen (Embretson, 1983, 1999; Embretson & Daniel, 2008; Holling et al., 2009).

4. Methodik

Im folgenden Kapitel sollen die zur statistischen Analyse der Fragestellung herangezogenen Modelle der probabilistischen Testtheorie vorgestellt werden.

Generell kann man bei der Konstruktion psychologisch-diagnostischer Verfahren zwei methodische Ansätze unterscheiden, die *klassische Testtheorie* sowie die *moderne bzw. probabilistische Testtheorie,* die im englischsprachigen Raum auch Item-Response Theorie (IRT) genannt wird und auf die Arbeiten von Georg Rasch (vgl. Rasch, 1960) zurückgeht. Historisch gesehen wurde ein Großteil der heute verfügbaren Tests anhand von in der der klassischen Testtheorie verwendeten Methoden und Mitteln konstruiert, seit den letzten Jahrzehnten kommen aber immer mehr Verfahren auf den Markt, bei deren Entwicklung auf Modelle der probabilistischen Testtheorie zurückgegriffen wurde.

Während die Qualität eines Verfahrens in der klassischen Testtheorie anhand der festgelegten (Haupt-)Gütekriterien *Reliabilität*, *Validität* und *Objektivität* beurteilt wird, legt die probabilistische Testtheorie hingegen besonderes Augenmerk auf das Gütekriterium *Skalierung*, das Kubinger (2009a) wie folgt definiert: „Ein Test erfüllt das Gütekriterium Skalierung, wenn die laut Verrechnungsvorschriften resultierenden Testwerte die empirischen Verhaltensrelationen adäquat abbilden“ (S.82). Hinter dieser Definition verbirgt sich die Frage ob die Verrechnung von Testleistungen (in diesem Fall Rohwerten) zu numerischen Testwerten empirisch angemessen ist. Dieses Gütekriterium ist einerseits von der klassischen Testtheorie weitgehend ungeachtet andererseits kann es mit dem in der klassischen Testtheorie zur Verfügung stehenden Methodeninventar auch gar nicht ausreichend überprüft werden. Eine Beweisführung hierzu findet sich beispielsweise bei Kubinger (2009a). Wenn ein psychologisch-diagnostisches Verfahren das Gütekriterium Skalierung nicht erfüllt erübrigt sich im Grunde die Überprüfung der Hauptgütekriterien (siehe hierzu speziell Kubinger &

Proyer, 2004). Ein allgemeiner Überblick zu den Problemen der klassischen Testtheorie findet sich beispielsweise bei Kubinger (2009a) sowie bei Hambleton und Kollegen (1991).

Im Rahmen der probabilistischen Testtheorie werden Annahmen über das Zustandekommen von Antworten auf Items getroffen, daher stammt auch der englische Begriff der Item-Response Theorie. Kubinger (2009a) schreibt über die Gemeinsamkeiten der Modelle der probabilistischen Testtheorie:

> Ausgehend von einer Stichprobe *S*, mit den Testpersonen {1,2, . . . *v* . . . *n*} und einem Itempool *I*, mit den Items {1,2, . . . *i* . . . *k*}, ist ihnen allen eine wesentliche Idee gemeinsam: Für die beobachteten Reaktionen, also die manifest werdenden Verhaltensweisen der Testpersonen aus *S* auf die Items aus *I* sind weiter nicht näher beobachtbare, sog. latente „Eigenschaften" der Person verantwortlich zu machen. Und zwar stehen erstere mit letzteren modellspezifisch in wahrscheinlichkeitsfunktionalem Zusammenhang. (S. 21)

Die angesprochene Lösungswahrscheinlichkeit ist pro Item eine mathematische Funktion aus Item- und Personenparametern und hängt laut der zentralen Annahme der IRT von der Fähigkeit der Testperson sowie von Parametern ab, die das Item charakterisieren. Das können beispielsweise Itemschwierigkeit, Trennschärfe und Ratewahrscheinlichkeit sein (Hambleton et al., 1991).

Im einfachsten Fall, der durch das im Folgenden vorgestellte Rasch-Modell, auch 1-PL-Modell (1-parametrisches logistisches Testmodell) genannt, beschrieben wird, wird genau eine latente Dimension angenommen und nur zwei mögliche Reaktionen der Testperson (richtig/falsch bzw. ja/nein) unterschieden. Das 2-PL-Modell nimmt zusätzlich zur Itemschwierigkeit einen weiteren Parameter an, mithilfe dessen eine unterschiedliche Gewichtung der Items möglich ist. Beim 3-PL-Modell wird zudem ein Rateparameter geschätzt (siehe hierzu Hambleton et al., 1991; Kubinger 1989; Rost, 2004). In der Item Response Theorie finden sich allerdings auch Modelle für mehrere

latente Dimensionen sowie weitere unterschiedliche Reaktionen der Testperson (vgl. hierzu z. B.: Hambleton et al., 1991; Kubinger, 1989). Da die Daten bei der testtheoretischen Auswertung des hier untersuchten Verfahrens AN-TOP mithilfe des Rasch-Modells geschätzt wurden, soll im Kapitel 4.1 näher darauf eingegangen werden.

Die bisher erwähnten Modelle analysieren Daten auf Itemebene. Um jedoch die den Items zugrundeliegenden schwierigkeitsrelevanten Teilaspekte schätzen zu können, müssen Daten auf Itemkomponentenebene analysiert werden. Rost (2004) definiert Itemkomponenten wie folgt: „Itemkomponenten können etwa verschiedene Elemente im Prozess der Aufgabenbearbeitung sein, die *gemeinsam* die *Schwierigkeit* eines Items ausmachen. Von der Itemkonstruktion her betrachtet, können solche Itemkomponenten auch Elemente sein, aus denen man die Items ‚zusammensetzt' und *somit systematisch* konstruiert" (S. 253).

Ein testtheoretisches Modell, das diesem Anspruch genügt ist das linear-logistische Testmodell (Scheiblechner, 1972; Fischer, 1973). Es schätzt anstatt der oben erwähnten Itemparameter Schwierigkeitsparameter der einzelnen kognitiven Schritte, die zur Lösung eines Items führen. Kapitel 4.2 stellt dieses Modell der probabilistischen Testtheorie vor.

4.1 Das dichotome logistische Modell von Rasch (1-PL-Modell)

Das dichotome logistische Modell von Rasch, kurz: Rasch-Modell ist eines der am häufigsten verwendeten Modelle der IRT. Die in dieser Arbeit verwendete Darstellung des Modells orientiert sich an Fischer (1974) und Kubinger (1989).
Neben der bereits erwähnten Überprüfbarkeit des Modells bietet das Rasch-Modell gegenüber der klassischen Testtheorie weitere Vorteile. Im Folgenden soll auf die

erschöpfenden Statistiken, die Stichprobenunabhängigkeit der Parameter-schätzungen sowie die Möglichkeit spezifisch objektiver Vergleiche eingegangen werden.

Das Rasch-Modell geht vom Vorliegen dichotomer Antwortkategorien aus, das heißt es wird nur zwischen Aufgabe gelöst (+) und Aufgabe nicht gelöst (-) unterschieden. Werden die Antworten aller *n* Personen auf alle *k* Items nun in eine *n* x *k* Datenmatrix eingegeben, wobei die Zeilen den Antworten einer bestimmten Person *v* und die Spalten den Lösungen des Items *j* entsprechen, so kann aus den Zeilensummen der *Personen*- oder *Rohscore* und aus den Spaltensummen der *Itemscore* abgelesen werden. Je höher nun dabei der Rohscore ist, desto mehr Aufgaben hat Person *v* gelöst und desto höher ist ihre Fähigkeit. Je höher der Itemscore ist, desto öfter wurde das Item *j* gelöst und desto einfacher ist es.
Diese Annahme besagt, dass der Rohscore als Summe der gelösten Aufgaben der Person *v*, unabhängig von deren individuellem Lösungsmuster, eine „*erschöpfende Statistik*“ für den Personenparameter dieser Person darstellt (Fischer, 1974; Kubinger, 1989). Diese Überlegung basiert auf der *lokalen stochastischen Unabhängigkeit*, einer zentralen Annahme des Rasch-Modells. Infolge dieser Annahme, hängt die Lösung eines Items durch eine Testperson nicht davon ab, welche anderen Aufgaben sie bereits gelöst hat oder noch lösen wird, sondern nur von ihrer Fähigkeit und von der Schwierigkeit des Items (Kubinger, 1989, 2009a). Zurückkehrend zum bereits erwähnten Gütekriterium Skalierung muss also das Rasch-Modell (bzw. eine monotone Transformation davon) gelten wenn die Anzahl der gelösten Items in einem Test ein faires Maß für die Fähigkeit einer Person sein soll. Dies gilt auch im Kehrschluss: wenn das Raschmodell gilt, kann davon ausgegangen werden, dass der Rohscore die Leistung der Testperson hinreichend beschreibt (Kubinger, 1989; 2005). Ein Beweis hierzu findet sich bei Fischer (1974).

Wie bereits oben erwähnt, hängt die Lösungswahrscheinlichkeit für ein bestimmtes Item, also abgesehen vom Zufall, nur von der Fähigkeit einer Person in der untersuchten Dimension und der Schwierigkeit der Aufgabe ab. Die Fähigkeit einer Person in dieser einzigen angenommen Dimension kann durch eine einzige Zahl, den Personen(fähigkeits-)parameter ξ dargestellt werden. Die Schwierigkeit des Items kann

gleichfalls durch eine einzige Zahl, nämlich durch den Item(schwierigkeits)parameter σ dargestellt werden (Kubinger, 1989).

Das Rasch-Modell beschreibt also formal betrachtet die Wahrscheinlichkeit, dass eine Person *v* ein Item *i,* abhängig vom Schwierigkeitsparameter des Items σ_i und dem Personenparameter ξ_v löst. Diese Annahme wird durch die folgende logistische Wahrscheinlichkeitsfunktion veranschaulicht:

$$P\left(+\middle|\xi_{\mathrm{v}},\sigma_{\mathrm{i}}\right)=\frac{\mathrm{e}^{\xi_{\mathrm{v}}-\sigma_{\mathrm{i}}}}{1+\mathrm{e}^{\xi_{\mathrm{v}}-\sigma_{\mathrm{i}}}}$$

Formel 1: Rasch-Modell

Man kann jedoch bei gegebener Fähigkeit ξ einer Person *v* und bekannter Itemschwierigkeit σ_i nicht sicher vorhersagen, ob die Person diese Aufgabe lösen wird. Allenfalls kann vorhergesagt werden wie wahrscheinlich sie zur Lösung kommen wird. Bei konstant gehaltener Schwierigkeit des Items ist die Wahrscheinlichkeit ein Item zu lösen höher je größer die Fähigkeit ξ ist. Abbildung 1 zeigt für vier Items der AN-TOP die sogenannte Item Characteristic Curve (ICC). Die ICC eines Items ist die graphische Darstellung der Lösungswahrscheinlichkeit als Funktion der latenten Fähigkeitsdimension ξ (Molenaar, 1995).

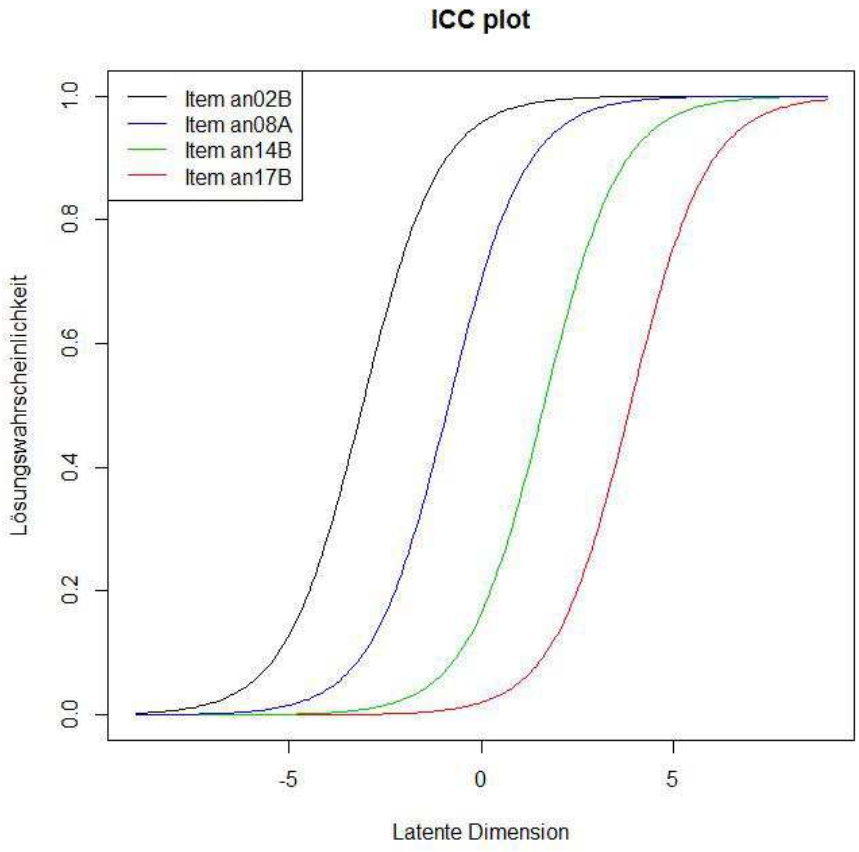

Abbildung 1: Item Characteristic Curve für vier Items der AN-TOP

Auf der x-Achse ist der Wertebereich der latenten Fähigkeitsdimension ξ aufgetragen, die y-Achse gibt die Lösungswahrscheinlichkeit an. Obwohl der Personenparameter ξ theoretisch zwischen $-\infty$ und $+\infty$ liegen kann, liegt dieser praktisch meistens zwischen -5 und +5 (Kubinger, 1989). Die ICC Kurven haben bei Rasch-Modell-Konformität einen ogivenförmigen Kurvenverlauf. Der Itemschwierigkeitsparameter σ definiert den Wendepunkt der ICC Kurve, also jene Position, bei der die Lösungswahrscheinlichkeit einer Person mit der Fähigkeit ξ_v 50 % für das Item i beträgt. Ist der Itemparameter positiv, ist das Item schwierig und die ICC-Kurve liegt in der Grafik weiter rechts. Ist der Itemparameter negativ, ist das Item leichter und die Kurve befindet sich weiter links (Rost, 2004). Anhand dieser Grafik kann auch demonstriert werden, dass eine Testperson umso fähiger sein muss, um ein Item mit einer 50%-Wahrscheinlichkeit zu lösen, je schwieriger das Item ist (Fischer, 1995a; Hambleton et al., 1991).

Ein besonderer Vorzug durch das Modell ergibt sich durch die Möglichkeit *spezifisch objektive Vergleiche* anzustellen (Kubinger, 1989). Das heißt, dass der Unterschied in der Eigenschaftsdimension ξv und ξw zweier Personen v und w unabhängig davon bestimmt werden kann welche Items eines spezifischen Itempools sie bearbeitet haben. Auch der Kehrschluss gilt. Das bedeutet, dass zwei Items i und j bezüglich ihrer Itemschwierigkeitsparameter σi und σj unabhängig von der dafür verwendeten Stichprobe verglichen werden können. Die Stichprobenunabhängigkeit der Parameterschätzungen ist gegeben (Kubinger, 1989).

4.1.1 Prüfung der Gültigkeit des Rasch-Modells

Das Rasch-Modell und auch generell die Modelle der IRT müssen auf Modellgültigkeit geprüft werden um sicherzustellen, dass ein Itempool auch tatsächlich die oben genannten Eigenschaften hat. Neben sogenannten parameterfreien Modelltests, deren praktische Umsetzung allerdings zu ungelösten mathematischen Problemen führt, stehen in der Praxis hauptsächlich parametrische Modelltest zur Verfügung. Diese bedienen sich der Eigenschaft der Stichprobenunabhängigkeit. Das bedeutet, dass die Schätzungen für einen Parameter σ bei Modellgültigkeit in verschiedenen Teilstichproben statistisch gleich sein müssen. Am anschaulichsten kann dies mithilfe des grafischen Modelltests dargestellt werden (Kubinger, 1989; siehe beispielsweise Kapitel 8.1.3).

Dabei werden die Parameterschätzungen in einem rechtwinkeligen Koordinatensystem aufgetragen. Bei Modellgültigkeit müssen sich die Punkte, die jeweils ein Item geschätzt in den beiden Teilstichproben darstellen, der 45° Geraden annähern. Im Idealfall wären die Schätzungen in beiden Teilstichproben gleich und lägen genau darauf. Weichen manche Punkte signifikant von dieser Geraden ab, so ist das Kriterium der Stichprobenunabhängigkeit verletzt und das Verfahren misst nicht fair (Kubinger, 1989).

Inferenzstatistisch kann das Rasch-Modell beispielsweise mit dem (Conditional) Maximum Likelihood-Ratio-Test (LRT) nach Andersen (1973) überprüft werden. Hierbei

wird die Gesamtstichprobe nach internen (Score; verglichen werden Teilstichproben von Personen mit hoher bzw. niedriger Testleistung) oder externen Kriterien (z. B. Alter, Muttersprache, Geschlecht, etc.) aufgeteilt und anschließend überprüft, ob die Daten durch die jeweils in den Teilstichproben geschätzten Parameter der Items besser erklärt werden können als durch jene der Schätzung der Gesamtstichprobe (Glas & Verhelst, 1995, Kubinger, 1989). Weitere Modelltests finden sich ebenfalls bei Kubinger (1989) und Glas & Verhelst (1995).

Der LRT beantwortet allerdings nur die Frage nach der globalen Modellgültigkeit, er kann nicht anzeigen welche Items nicht modellkonform sind. Ein statistisches Verfahren, das dies ermöglicht ist der *z*-Test nach Fischer und Scheiblechner (Fischer & Scheiblechner, 1970) mit der Schätzfunktion nach Wald (Wald, 1943). Bei diesem Verfahren handelt es sich um einen Test auf Itemebene, der ähnlich dem LRT auch Schätzungen in nach Teilungskriterien aufgeteilten Teilstichproben vergleicht (Glas & Verhelst, 1995).

Nach Ausschluss nicht modellkonformer Items kann somit zumindest a-posteriori Modellgültigkeit erzielt werden (Kubinger, 1989). Kubinger und Draxler (2007) betonen allerdings, dass der Anteil an auszuscheidenden Items nicht über 10% liegen sollte.

4.2 Das Linear-Logistische Test-Modell (LLTM)

Das linear-logistische Testmodell (LLTM) ist ein Spezialfall des Rasch-Modells. Es entstammt wie bereits erwähnt aus den Überlegungen, dass die Itemschwierigkeit eine Funktion der Schwierigkeit der beteiligten kognitiven Operationen ist (Scheiblechner, 1972; Fischer, 1973, 1989, 1995b).

Formell ist das LLTM wie folgt definiert (Fischer, 1995b):

$$P\left(+\middle|\xi_v;\sigma_i=\sum_j^p q_{ij}\eta_j\right)=\frac{e^{\xi_v-\sum_j^p q_{ij}\eta_j}}{1+e^{\xi_v-\sum_j^p q_{ij}\eta_j}}$$

Formel 2: LLTM

Die Itemparameter σ des Rasch-Modells werden also in eine gewichtete Summe von Basisparametern $\eta_1, \eta_2, \ldots, \eta_n$ zerlegt, wobei es nur dann zu einem Spezialfall des Rasch-Modells wird wenn die Anzahl der im LLTM geschätzten Basisparameter kleiner als die Anzahl der im Rasch-Modell geschätzten Itemparameter ist (Kubinger, 1989).
Wie auch beim Rasch-Modell erfolgt die Schätzung der Basisparameter η mit der Conditional Maximum Likelihood-Methode (siehe hierzu z. B. Fischer, 1972; 1995b).

Tendenziell können Schwierigkeitsunterschiede zwischen Items damit erklärt werden, dass sich leichtere Items aus weniger kognitiven Operationen zusammensetzen als schwierigere Items (Fischer, 1995b).

Laut Kubinger (1989) erfolgt die Überprüfung der Geltung des LLTM in zwei voneinander unabhängigen Schritten. In einem ersten Schritt wird die Modellkonformität des Rasch-Modells überprüft. Anschließend folgt im zweiten Schritt ein weiterer LRT, der überprüft ob die Daten durch die im LLTM angenommenen Parameter genauso gut erklärt werden können wie durch das Rasch-Modell. Hält das Verfahren der Modellprüfung stand, so kann Rasch-Modell-Konformität für jedes neue Item, das aus den bereits überprüften Teiloperationen besteht, angenommen werden. Zudem kann auch für jedes dieser neuen Items die Schwierigkeit durch eine additive Funktion bestimmt werden. Kubinger (1989) betont, dass im Falle von a-posteriori Gültigkeit des Rasch-Modells, eine Kreuvalidierung des Verfahrens vorgenommen werden sollte.

Weitere Anwendungsmöglichkeiten des LLTM finden sich beispielsweise bei Kubinger (2008; 2009b).

III. Empirischer Teil

5. Zielsetzung der Studie und Hypothesen

Diese Studie hat das Ziel einen Test zum schlussfolgernden Denken mit Hilfe eines Konstruktionsrationals zu entwickeln, um den vorhandenen Itempool kostengünstig und wenig zeitintensiv beliebig erweitern zu können und den Test in weiterer Folge auch adaptiv anwenden zu können. Hierzu werden Items unter Zuhilfenahme eines Konstruktionsrationals, das in Kapitel 6 vorgestellt werden soll, entwickelt. Der erstellte Itempool wird anschließend dahingehend geprüft, ob dieser durch das dichotom logistische Modell von Rasch erklärt werden kann. Des Weiteren soll überprüft werden, ob die inhaltliche Itemstruktur die Schwierigkeit der Items erklären kann. Falls sich die Likelihood des auf der Strukturmatrix des Konstruktionsrationals basierenden LLTM als nicht signifikant unterschiedlich zu jener des Rasch-Modell erweist, kann angenommen werden, dass die den Items zugrundegelegte Struktur ein passendes Erklärungsmodell darstellt.
Aus dieser Vorgehensweise leiten sich die folgenden wissenschaftlichen Hypothesen ab:

$H_{0\text{-}A}$: Der Test AN-TOP erweist sich als Rasch-Modell-konform.
$H_{1\text{-}A}$: Der Test AN-TOP ist nicht Rasch-Modell-konform.

$H_{0\text{-}B}$: Die Itemschwierigkeiten der 56 Items können durch die 17 Basisparameter ebenso gut erklärt werden, wie durch das Rasch-Modell. $\beta_i^{(LLTM)} = \beta_i^{(Rasch\ Modell)}$
$H_{1\text{-}B}$: Die Itemschwierigkeiten der 56 Items werden durch 56 unabhängige Itemparameter erklärt.

6. Alpha-Numerical TOPologies (AN-TOP)

Die vorliegende Studie verwendet aus den in der Einleitung erläuterten Gründen ein neues und bisher nicht publiziertes Testkonzept zur Erfassung des Faktors Reasoning, das im weiteren Verlauf vorgestellt werden soll.

Generell kann der Testkonstruktionsprozess in drei Schritte unterteilt werden. Der erste Schritt bezieht sich auf den Testentwurf, also die Erstellung des Testkonzeptes sowie die Itemerstellung. Im darauffolgenden zweiten Schritt werden die konstruierten Items empirisch überprüft, das heißt erstmalig einer Stichprobe vorgegeben um die Güte des Tests anhand der in der Psychologischen Diagnostik gängigen Gütekriterien zu überprüfen. Schließlich folgt im dritten und letzten Schritt die Normierung des Tests anhand einer repräsentativen Stichprobe. Die vorliegende Studie befasst sich mit den ersten beiden Schritten, die Normierung der AN-TOP erfolgt bei Gelingen der ersten beiden Schritte in einer separaten Studie. Im Folgenden soll das Vorgehen während des ersten Teilschrittes der Testkonstruktion erläutert werden.

6.1 Testkonstruktion

6.1.1 Ein neues Testkonzept

Der Name des Tests, *A*lpha-*N*umerical *TOP*ologies, oder kurz AN-TOP, gibt einen ersten Einblick in dessen Beschaffenheit. Die AN-TOP bestehen aus *Topologien* und arbeiten hauptsächlich mit Buchstaben (*alpha*) und Zahlen (*numerical*) sowie geometrischen Figuren.

Das Prinzip des Topologischen Raums entstammt der Mathematik und kann als ein Mengensystem definiert werden, das aus mehreren Teilmengen besteht. So besteht auch jede Aufgabe der AN-TOP aus drei Teilmengen bzw. Feldern, die wiederum mehrere Elemente enthalten. Die Inhalte der drei Felder sind jeweils durch regelhafte

Beziehungen miteinander verbunden, wobei jedes Mal im letzten Feld ein Element fehlt und von der Testperson in einem dafür vorgesehen Feld ergänzt werden muss. Der Ort an dem das fehlende Element hinzugefügt werden muss ist mit einem Rechteck gekennzeichnet. Abbildung 2 zeigt das erste Beispielitem des Tests sowie die Erklärung der Lösung.

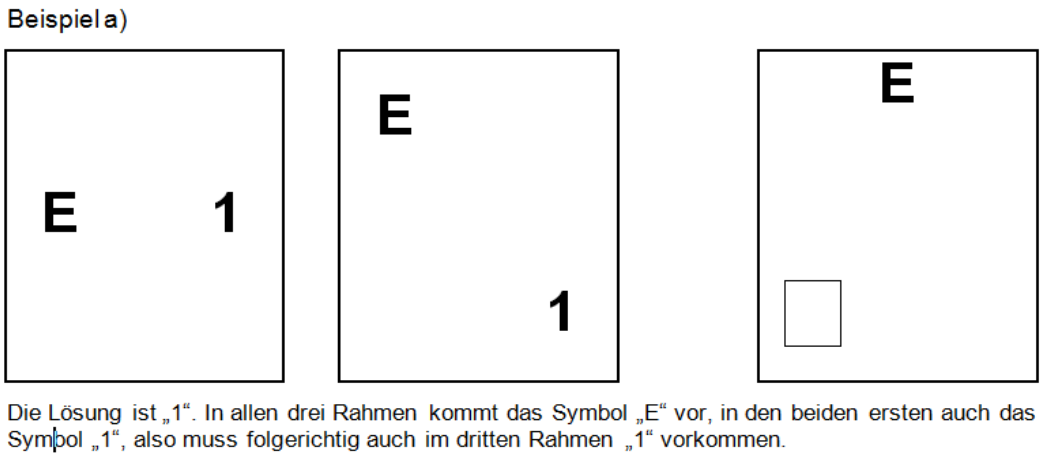

Abbildung 2: Beispielitem A der AN-TOP

6.1.2 Testformat

Leistungstests können hinsichtlich ihrer Konzipierung als Speed bzw. Power-Tests unterschieden werden. Während reine Speed-Tests bei niedrigem Schwierigkeitsgrad der Aufgabenstellung eine enge Zeitbegrenzung aufweisen und die Bearbeitungsgeschwindigkeit als Leistung bewerten, haben Power-Tests keine oder eine sehr großzügig bemessene Zeitbegrenzung, die Richtigkeit der Antwort ist alleiniges Fähigkeitsmerkmal (Kubinger, 2009a). Test, die sowohl eine Speed als auch eine Power Komponente beinhalten werden Speed-and-Power-Tests genannt (Kubinger, 2009a). Da für die Bearbeitung solcher Aufgaben aber vermutlich zwei Fähigkeitsdimensionen (Bearbeitungsgeschwindigkeit und Leistung) benötigt werden und diese Tests das Gütekriterium Skalierung oft nicht erfüllen (Kubinger, 2009a), wurde für die AN-TOP ein reines Power-Testformat mit sehr großzügiger Zeitbegrenzung gewählt.

Der Test wird als Gruppentest mit schriftlicher Vorgabe konstruiert um auf möglichst ökonomische Weise eine große Anzahl an Kindern testen zu können.

6.1.3 Zielgruppe

Die Aufgaben der AN-TOP sind für deutschsprachige Kinder und Jugendliche im Alter von 6 Jahren (6;0) bis 15 Jahren und 11 Monaten (15;11) konzipiert. Obwohl die Aufgaben wie erwähnt neben geometrischen Figuren und Zahlen auch Buchstaben verwenden, wird angenommen, dass der Test den Faktor Reasoning unabhängig von Lese- und Schreibfähigkeit, also sprachfrei misst. Dennoch wird ein grundlegendes Verständnis der deutschen Sprache zum Verständnis der Arbeitsanweisung vorausgesetzt.

6.1.4 Testziel

Das Ziel der Alpha-Numerical Topologies ist es die Fähigkeit zum schlussfolgernden Denken zu erfassen. Somit sind sie Teil der Gruppe von Tests, die individuelle Ausprägungen in einer Fähigkeits- oder Eigenschaftsdimension messen sollen und ist von jener Gruppe von Tests abzugrenzen, die Wissen messen sollen beziehungsweise Testpersonen in Gruppen trennen sollen.

6.1.5 Antwortformat

Für den Test AN-TOP wurde ein freies Antwortformat gewählt. Dieses Antwortformat, bei dem eine Testperson ein Item eigenständig und nach eigenem Ermessen beantwortet ohne aus einem Antwortkatalog auszuwählen, birgt einige Vorteile gegenüber dem gebundenen Antwortformat, bei dem die Testperson unter mehreren vorgegebenen Antwortalternativen wählt. So ist es beispielsweise relativ frei von Zufallseinflüssen wie etwa Rateeffekten, wie in der Literatur mehrfach gezeigt werden konnte (siehe z. B.: Hohensinn & Kubinger, 2009; Vock, 2009).

6.2 Konstruktionsprozess

Nach der Erstellung des Testkonzeptes und der Definition der schwierigkeitsgenerierenden Operationen wurde ein Itempool von 64 Items regelgeleitet erstellt, wobei dieser Itempool in 2 Parallelversionen zu je 32 Items unterteilt werden kann.

Die Aufgaben wurden anschließend anderen Diplomandinnen und Diplomanden vorgegeben und mittels der Methode des lauten Denkens während der Testdurchführung analysiert um die Lösungsprozesse mittels Regelanwendung zu überprüfen.

Anschließend wurden die Aufgaben nach der vermuteten Schwierigkeit gereiht und in vier Parallelversionen (V1, V2, V3 und V4) und zwei Altersgruppen aufgeteilt. Die Versionen wurden jeweils durch linking items (siehe z. B. Kubinger, 2008) miteinander verbunden. Das Vorgabedesign findet sich in Abbildung 3.

Testbogen	an91	an92	an93	an94	an01	an02	an03	an04	an05	an06	an07	an08	an09	an10	an11	an12	an13	an14	an15	an16	an17	an18	an19	an20	an21	an22	an23	an24	an25	an26	an27	an28
Bogen bis 10 Jahre A	A	A	A	A	A	A	A	A	A	A		A	A	A	A	A	A	A	A		A		A	A	A	A	A	A		A	A	A
Bogen bis 10 Jahre B	A	B	A	A	A	A	A	A	A	B		A	A	A	A	B	A	A	A		A		A	A	A	B	A	B		A	A	A
Bogen bis 10 Jahre C	B	A	B	B	B	B	B	B	B	A		B	B	B	B	A	B	B	B		B		B	B	B	A	B	A		B	B	B
Bogen bis 10 Jahre D	B	B	B	B	B	B	B	B	B	B		B	B	B	B	B	B	B	B		B		B	B	B	B	B	B		B	B	B
Bogen ab 10 Jahre A					A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Bogen ab 10 Jahre B					A	A	A	A	A	B	A	A	A	A	A	B	A	A	A	B	A	A	A	A	A	B	A	B	A	A	A	A
Bogen ab 10 Jahre C					B	B	B	B	B	A	B	B	B	B	B	A	B	B	B	A	B	B	B	B	B	A	B	A	B	B	B	B
Bogen ab 10 Jahre D					B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B

Abbildung 3: Aufteilung der Items der AN-TOP auf die acht Testversionen

Vermerk: Die Items sind beim Betreuer dieser Diplomarbeit einzusehen.

6.3 Schwierigkeitsgenerierende Operationen bei AN-TOP

Für die AN-TOP werden, in Übereinstimmung mit der in den Kapiteln 2 und 3 diskutierten Literatur, die folgenden schwierigkeitsgenerierenden Aspekte angenommen: das den Items zugrundeliegende Regelsystem, die auftretende Informationsmenge sowie die perzeptive Organisation der Items. Im Folgenden soll auf diese Gesichtspunkte eingegangen werden.

6.3.1 Das verwendete Konstruktionsrational

Das Konstruktionsrational, das den Alpha-Numerical Topologies zugrunde liegt, besteht aus Regeln, die kognitiven Operationen entsprechen. Dies bedeutet, dass ein Item umso schwieriger ist, je mehr Regeln es beinhaltet. Diese Regeln können in vier Teilbereiche aufgeteilt werden: Basic Operations, Mathematical Operations, Space-related Operations und Perceptual Complexity. Im Folgenden soll auf die einzelnen Regeln näher eingegangen werden. Im Anschluss gibt Tabelle 1 eine Übersicht über das Vorkommen der einzelnen Regeln innerhalb der Items der AN-TOP.

Basic Operations:

1. Continuation:

   Das gesuchte Zeichen kommt in den beiden ersten Feldern in der richtigen Form vor und muss ohne Abwandlung fortgesetzt werden.

2. Number:

   In den beiden ersten Kästchen kommt jeweils ein Zeichen gleich oft vor. Dabei kommt es nur auf die Anzahl der Zeichen an, die Ausprägung ist irrelevant.

Mathematical Operations

3. Counting:

   Dieser Regelkomplex bezeichnet das Vor- oder Zurückgehen in einer festen Reihenfolge (Alphabet, natürliche Zahlen und Anzahl von Symbolen) und wird auf Zahlen, Buchstaben und Symbole angewandt. Dies wird durch die folgenden Regeln ausgedrückt:

a. Sequence across fields forwards

   Hier wird über die Felder hinweg im jeweiligen Bezugssystem vorwärts gezählt.

b. Sequence across fields backwards

   Hier wird über die Felder hinweg im jeweiligen Bezugssystem rückwärts gezählt.

c. Sequence within fields forwards

   Bei Items, die dieser Regel folgen, wird jeweils innerhalb eines Feldes im jeweiligen Bezugssystem vorwärts gezählt.

d. Sequence within fields backwards

   Bei Items, die dieser Regel folgen, wird jeweils innerhalb eines Feldes im jeweiligen Bezugssystem rückwärts gezählt.

e. Counting numbers

   Das Bezugssystem stellen natürliche Zahlen dar.

f. Counting letters

   Das Bezugssystem ist das Alphabet.

4. Addition

   Die Lösung ergibt sich als Addition mehrerer Elemente pro Feld.

5. Multiplication

   Um zur Lösung zu kommen muss ein Element entweder mit sich selbst oder mit einem anderen im selben Feld vorkommenden Element multipliziert werden.

Space-related Operations

6. Rotation

   Das gesuchte Element muss rotiert werden um zur Lösung zu gelangen.

7. Absolute position

   Diese Regel bezeichnet die absolute Position der mit der Regel verbundenen Zeichen innerhalb eines Feldes. Diese ist konstant.

8. Relative position

   Hiermit ist die Position der einzelnen Elemente innerhalb eines Feldes relativ zueinander gemeint.

Perceptual Complexity

9. Distractors

   Der Regelkomplex der Distraktoren gliedert sich in mehrere Teilregeln auf:

   a. Systematic distractors

      Systematische Distraktoren sollen die Testperson insofern auf Abwege führen als dass sie eine implizite Regel suggerieren, die bei näherer Betrachtung und genauer Analyse nicht vorhanden ist. Sie sind daher nicht beliebig (je Feld) austauschbar.

   b. Unsystematic distractors

      Unsystematische Distraktoren sind im Gegensatz zu systematischen Distraktoren je Feld beliebig austauschbar, sofern sie die anzuwendende Regel nicht verändern oder das Item zwei- oder mehrdeutig machen.

   c. Fillers

      Füllzeichen sind jene Zeichen, die in jedem Feld in gleicher Art und Weise vorkommen und zur Lösung des Items nicht verwendet werden müssen.

10. Amount of elements

    Bezeichnet die (im Durchschnitt über alle drei Felder gerechnete) Anzahl der Zeichen im Feld. Hierfür wurden die Items ins 4 Gruppen geteilt: Aufgaben mit durchschnittlich bis zu 3 Zeichen, Aufgaben mit 4 bis 5 Zeichen, Aufgaben mit 6 bis 7 und Aufgaben mit 8 Zeichen und mehr.

11. Demonstration of rule

    Die Anwendung der Regel wird in allen drei Feldern vorgeführt, insbesondere wird die Regel im dritten Feld vorgezeigt.

12. Rule restriction to one field

    Im Gegensatz zum eben beschriebenen Grundsatz, werden die weiteren im Item verwendeten Regeln nur in einem einzigen der beiden ersten Felder vorgeführt.

Tabelle 1: Verteilung der schwierigkeitsgenerierenden Operationen auf die Items der AN-TOP

Regel	1	2	3a	3b	3c	3d	3e	3f	4	5	6	7	8	9a	9b	9c	10	11	12
Item	Basic		Mathematical								Space-related			Perceptual Complexity					
an91A	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
an91B	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
an92A	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
an92B	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
an93A	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	2	1	0
an93B	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0
an94A	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
an94B	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0
an01A	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
an01B	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
an02A	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
an02B	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
an03A	1	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	1	0	0
an03B	1	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	1	0	0
an04A	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0
an04B	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0
an05A	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
an05B	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
an06A	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0
an06B	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0
an07A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
an07B	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0
an08A	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	1	0
an08B	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	2	1	0
an09A	0	0	1	0	0	0	1	0	0	0	0	1	0	3	0	0	2	0	0
an09B	0	0	1	0	0	0	1	0	0	0	0	1	0	3	0	0	2	0	0
an10A	1	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	1	0	0
an10B	1	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	1	0	0
an11A	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	1	0	0
an11B	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	1	0	0
an12A	1	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	2	0	0
an12B	1	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	2	0	0
an13A	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	2	1	0
an13B	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	2	1	0
an14A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	2	0	0
an14B	1	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	2	0	0
an15A	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	2	0	0
an15B	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	2	0	0
an16A	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0
an16B	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0
an17A	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	2	0	0
an17B	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	2	0	0
an18A	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0
an18B	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0
an19A	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	2	1	0
an19B	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	2	1	0
an20A	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
an20B	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
an21A	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	2	0	0
an21B	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	2	0	0
an22A	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
an22B	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
an23A	0	0	0	1	0	0	1	0	0	0	0	0	0	0	4	0	2	0	0
an23B	0	0	0	1	0	0	1	0	0	0	0	0	1	0	4	0	2	0	0
an24A	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	2	0	0
an24B	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	2	0	0
an25A	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0
an25B	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0
an26A	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	2	1	0
an26B	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	2	1	0
an27A	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	1
an27B	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	1
an28A	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	3	1	0
an28B	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3	1	0

7. Datenerhebung

Um die teilnehmenden Kinder auf einem vergleichbaren Wissens- und Entwicklungsstand zu testen, wurde versucht die Testungen in einem möglichst kurzen zeitlichen Rahmen sowie einem eingeschränkten geografischen Gebiet durchzuführen. Die Testungen fanden im Februar 2010 in Oberösterreich an insgesamt 6 Schulen statt.

Die Schulen wurden nach Erhalt der Genehmigung des Landesschulrates für Oberösterreich per Email eingeladen an der Studie teilzunehmen. Nach Einverständnis der Direktion und bei den Volksschulen der betreffenden Klassenlehrerinnen wurden die Eltern durch Elternbriefe über die Studie und deren Ziel in Kenntnis gesetzt. (Schul- und Elternbrief finden sich im Anhang). Von zusätzlichen Einverständniserklärungen der Eltern wurde auf Wunsch der Direktionen abgesehen, wodurch hauptsächlich vollzählige Schulklassen getestet werden konnten.

Um die Datenerhebung zu erleichtern, wurde davon abgesehen, alle Klassenstufen mit in die Erhebung einzubeziehen. Daher wurden nur die Klassenstufen 3, 5 und 7 berücksichtigt. Es wurden nur Schüler der dritten Klassenstufe und höher für diese Studie berücksichtigt, da auch Schreibfähigkeit und -geschwindigkeit sowie Lesefähigkeit und Textverständnis durch die Vorgabe als Paper-Pencil-Verfahren in die Erhebung mit eingehen, wie die Leistung in der eigentlich zu erfassenden Dimension. Dies folgt der Annahme, dass jüngere Kinder durch die schriftliche Vorgabe des Testmaterials überfordert werden könnten.

7.1 Beschreibung der Stichprobe

Die Stichprobe bestand aus insgesamt 423 Kindern und Jugendlichen aus Oberösterreich im Alter zwischen 7 und 14 Jahren. Die Daten wurden hinsichtlich der Verteilung in Bezug auf die Variablen *Geschlecht, Alter, Muttersprache* und *Schulform* untersucht.

7.1.1 Schulform

Abbildung 3 veranschaulicht grafisch die Aufteilung der Stichprobe auf die getesteten Schulformen und Tabelle 1 gibt sowohl die Häufigkeiten als auch den Prozentsatz an. Wie den Daten zu entnehmen ist, konnte eine ungefähre Gleichverteilung der Testpersonen auf die drei getesteten Schulformen erreicht werden.

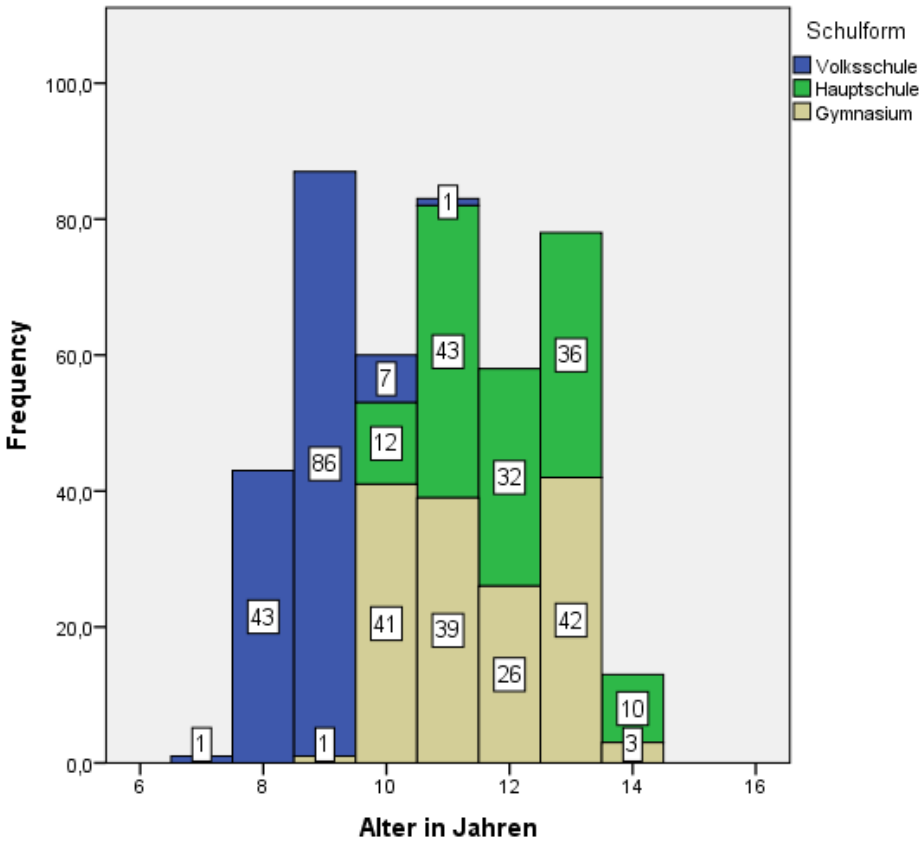

Abbildung 4: Zusammensetzung der Stichprobe

Tabelle 2: Deskriptive Statistik der Variable Schulform

Schulform	Häufigkeit	Anteil in Prozent
Volksschule	138	32,6
Hauptschule	133	31,4
Gymnasium	152	35,9
Gesamt	**423**	**100**

7.1.2 Alter und Geschlecht

Tabelle 2 zeigt die Verteilung der Testpersonen in Bezug auf die Variablen Alter und Geschlecht. Während die Variable Alter eher einer Normalverteilung gleicht, ist die Variable Geschlecht annähernd gleichverteilt. Es nahmen 214 Schüler und 209 Schülerinnen an der Untersuchung teil. Abbildung 4 zeigt die Verteilung der Variable Alter, wobei die Altersbereiche jeweils nach dem Geschlecht geteilt sind.

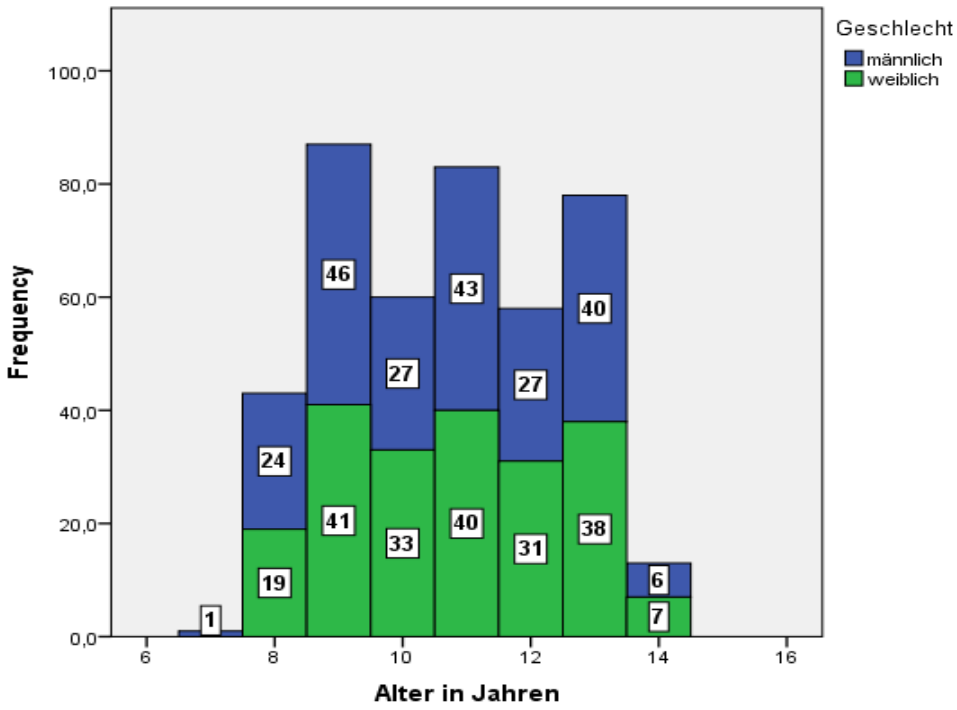

Abbildung 5: Alters- und Geschlechtsverteilung der Stichprobe

Tabelle 3: Deskriptive Statistik der Variablen Alter und Geschlecht

Alter in Jahren	Geschlecht		Gesamt
	männlich	weiblich	
7	1	0	1
8	24	19	43
9	46	41	87
10	27	33	60
11	43	40	83
12	27	31	58
13	40	38	78
14	6	7	13
Gesamt	**214**	**209**	**423**

7.1.3 Muttersprache

113 der 423 Kinder und Jugendlichen gaben an eine andere Muttersprache als Deutsch zu haben, das entspricht 26,7% der Stichprobe. Die neben Deutsch am meisten vertretenen Sprachen waren albanisch, türkisch und bosnisch/kroatisch/serbisch. Eine Aufstellung der Daten findet sich im Anhang..

7.2 Beschreibung der Testdurchführung

Die Gesamtdurchführungszeit mitsamt Instruktion betrug pro Klasse 90 Minuten, wobei die reine Bearbeitungszeit für die vier vorgegebenen Tests bei 75 Minuten lag. Die Testungen fanden vormittags als Gruppentestungen in den gewohnten Klassenräumen der Schülerinnen und Schüler während der Unterrichtszeit statt.

Den Kindern wurden Aufgabenhefte, ein demografischer Fragebogen und ein Lösungsbögen zur Bearbeitung der vier Untertests ausgeteilt (siehe Anhang und weitere Erläuterung zur verwendeten Testbatterie auf den folgenden Seiten).

Zuerst wurde eine allgemeine Testinstruktion gegeben, die die Schüler auf die anschließenden Aufgaben vorbereiten sollte. Anschließend wurde der demografische Fragebogen gemeinsam mit den Schülern ausgefüllt und allfällige Fragen gemeinsam bearbeitet. Erst nachdem Klarheit über den Ablauf und den Aufbau der Testung bestand, wurde zur Instruktion des ersten Untertests (AN-TOP) übergegangen. Zu jedem Untertest wurden mehrere Übungsaufgaben mit der Klasse besprochen, sodass vom Instruktionsverständnis ausgegangen werden konnte. Die Testpersonen wurden bei jedem Untertest instruiert, exakt aber zügig bis zum im Aufgabenheft abgedruckten Stoppschild zu arbeiten. Wenn eine Testperson das Stoppschild vor Ende der Bearbeitungsdauer erreichen würde, hatte sie Zeit, die soeben bearbeiteten Aufgaben des Untertests zu kontrollieren. Gemeinsam wurde dann zum nächsten Untertest übergegangen, die Instruktion erläutert und die Übungsaufgaben gemeinsam besprochen. Erst dann startete die Bearbeitungszeit des jeweiligen Untertests. Bei allen Untertests wurde in gleicher Weise verfahren. Es konnte

beobachtet werden, dass die tatsächlichen Bearbeitungszeiten innerhalb der Klassen und Schultypen sehr unterschiedlich waren. Aufgrund der Rückmeldung der Testpersonen während und nach der Testung kann jedoch davon ausgegangen werden, dass die Bearbeitungszeit ausreichend, wenn nicht sogar zu großzügig bemessen wurde. Da jedoch unterschiedliche Schulformen die gleichen Aufgaben bearbeiteten, sollte sichergestellt werden, dass auch Schüler niedrigeren Bildungsniveaus ausreichend Zeit hatten um die Aufgaben zu bearbeiten. Fehlende Daten, also Items, die von einzelnen Personen nicht bearbeitet wurden, wurden als „nicht gelöst“ verrechnet. Dies folgt der Annahme, dass die Aufgaben aufgrund zu hoher Schwierigkeit, bzw. zu geringer Fähigkeit der Testperson nicht bearbeitet wurden.

Die Testbatterie setzte sich insgesamt aus vier Untertests zusammen. Der erste Untertest bestand aus den bereits beschriebenen Aufgaben der AN-TOP, der zweite Untertest bestand aus 15 Aufgaben des Untertests „Funktionen abstrahieren“ des AID 3 (Kubinger & Holocher-Ertl, 2014). Als dritter Untertest folgte der CFT-20R von Weiß (2006). Schließlich wurden 15 Aufgaben des Family Relations Reasoning Tests (FRRT; Kubinger, Poinstingl, Weidinger & Skoda, in Vorb.) als vierter Untertest vorgegeben. Die vorliegende Studie bezieht sich nur auf die testtheoretische Überprüfung der AN-TOP. Dementsprechend werden die weiteren Untertests nicht näher vorgestellt.

8. Darstellung der Ergebnisse

Das folgende Kapitel zeigt die Ergebnisse der vorliegenden Studie auf. Um die Hypothesen zu überprüfen werden die Items der AN-TOP auf Ihre Rasch-Modell-Konformität überprüft. Die Berechnungen wurden mit Hilfe des Programms R (Version 3.1.1) und der Erweiterung eRm (extended Rasch modelling) von Hatzinger & Mair (2009) durchgeführt.

Zuerst werden die Ergebnisse der Überprüfung der Rasch-Modellkonformität mithilfe von grafischen Modelltests sowie mit Likelihood-Ratio-Tests nach Andersen (LRT's) dargestellt. Im weiteren Verlauf folgen die Ergebnisse zum Linear Logistischen Testmodell. Diese werden ebenfalls durch grafische Modelltests und Likelihood-Ratio-Tests beschrieben.

8.1 Überprüfung der Rasch-Modell Konformität der AN-TOP

8.1.1 Kriterienbildung

Zur Überprüfung des Raschmodells wurden insgesamt vier Kriterien herangezogen, ein internes Kriterium und drei externe Kriterien. Das interne Kriterium bildet der Rohscore, die drei externen Kriterien sind Geschlecht, Alter und Muttersprache.

Score

Als internes Teilungskriterium diente die Anzahl der gelösten Aufgaben (Rohscore). Die Testpersonen wurden hier anhand des Medians des Rohscores in eine Gruppe mit hohem Rohscore und in eine Gruppe mit niedrigem Rohscore aufgeteilt und miteinander verglichen. Das bedeutet, es wurden jene 229 Personen die weniger Items

als der Median gelöst hatten, mit jenen 194 Personen verglichen, die mehr Items als de Median gelöst hatten.

Geschlecht

Als eines der externen Kriterien diente das Geschlecht, das aufzeigen sollte, ob es Unterschiede zwischen Mädchen und Jungen bei der Bearbeitung der Items der AN-TOP gibt. Die beiden Gruppen bestanden aus 214 Jungen und 209 Mädchen.

Alter

Das Alter der Testpersonen diente ebenfalls als externes Teilungskriterium. Die Variable Alter wurde anhand des Medians in zwei Gruppen aufgeteilt. Es wurden «bis 11 Jährige» und «ab 12 Jährige» miteinander verglichen.
Folglich bestand das Teilungskriterium Alter aus einer Gruppe von 192 Kindern, die angaben nicht älter als 11 Jahre zu sein und einer Gruppe mit 231 Kindern, die angaben 12 Jahre oder älter zu sein.

Muttersprache

Auch die Muttersprache der Testpersonen wurde als externes Teilungskriterium herangezogen. Hierbei wurden Kinder mit deutscher Muttersprache mit Kindern mit nicht-deutscher Muttersprache verglichen. Insgesamt nahmen 113 Kinder mit nicht-deutscher Muttersprache an der Studie teil.

Da der Itempool, der für die weitere Analyse der AN-TOP mit dem Linear Logistischen Testmodell verwendet wird, nur Items enthalten soll, deren Itemparameter keinen stichprobenabhängigen Verzerrungen unterliegen, wurden die Daten mittels grafischen Modelltests sowie inferenzstatistisch mit Likelihood-Ratio-Tests (LRT) untersucht. Hierfür wurden die Daten anhand des internen Teilungskriteriums Rohscore (niedriger versus hoher Rohscore, geteilt am Median) und der externen Kriterien Alter (geteilt am Median), Geschlecht (weiblich versus männlich) und Muttersprache (deutsch versus nicht deutsch) geteilt.

Bei einem signifikanten Ergebnis der LRTs *(α* = .01*)* werden die Daten mit Hilfe des Wald-Tests und grafischer Modelltests analysiert um nicht Modell-konforme Items identifizieren und sukzessive ausscheiden zu können. Dieses Ausschlussverfahren wird so lange wiederholt bis keiner der LRTs mehr auffällig ist. Dies würde bedeuten, dass sich die Items der AN-TOP a-posteriori als Rasch-Modell-konform herausstellen.

8.1.2 A-Priori Ergebnisse

Teilungskriterium Rohscore

Aufgrund des ungünstigen Antwortmusters in den einzelnen Teilgruppen konnten 5 Items nicht in die Analysen einbezogen werden. Da diese Items in anderen Teilungskriterien geschätzt werden konnten, mussten sie nicht vom Gesamt-Itempool ausgeschlossen werden.

Der LRT fiel in Bezug auf das Teilungskriterium im ersten Berechnungsdurchgang signifikant aus. Die Hypothese $H_{0\text{-}A}$ – „Die Items der AN-TOP entsprechen dem Rasch-Modell“ muss daher zunächst verworfen werden.

Tabelle 4 zeigt bezüglich des Teilungskriteriums *Rohscore* die geschätzte asymptotisch χ^2-verteilten Testgröße des LRT, die Freiheitsgrade (*df*), die Wahrscheinlichkeit, dass die H_0 gilt (*p*-Wert) sowie den kritischen Wert der χ^2-Verteilung (bei $\alpha = .01$) an.

Tabelle 4: LRT Teilungskriterium Rohscore, erster Berechnungsdurchgang

Teilungskriterium Rohscore			
Andersen χ^2	*df*	p-Wert	Kritischer χ^2-*Wert*
92.25	58	0.003	85.95

Abbildung 6 zeigt den Grafischen Modelltest für das Teilungskriterium Rohscore für alle in die Schätzung mit einbezogenen Items. Bei Betrachtung des Grafischen Modelltests fällt auf, dass vor allem im mittleren Leistungsbereich viele Items nahe der 45°-Geraden liegen. Die Items im höheren und niedrigeren Leistungsbereich liegen

zwar weiter von der Geraden weg, weisen aber größere Konfidenzintervalle auf und sind daher dennoch als Modell-konform zu bewerten. Ein Item ist dann als nicht Rasch-Modell-konform zu bewerten, wenn die Konfidenzellipse die 45°-Gerade nicht schneidet, Abbildung 7 zeigt daher im Ausschnitt jene beiden Items (20A und 27B) inklusiver ihrer Konfidenzellipsen, die im Grafischen Modelltest auffällig sind.

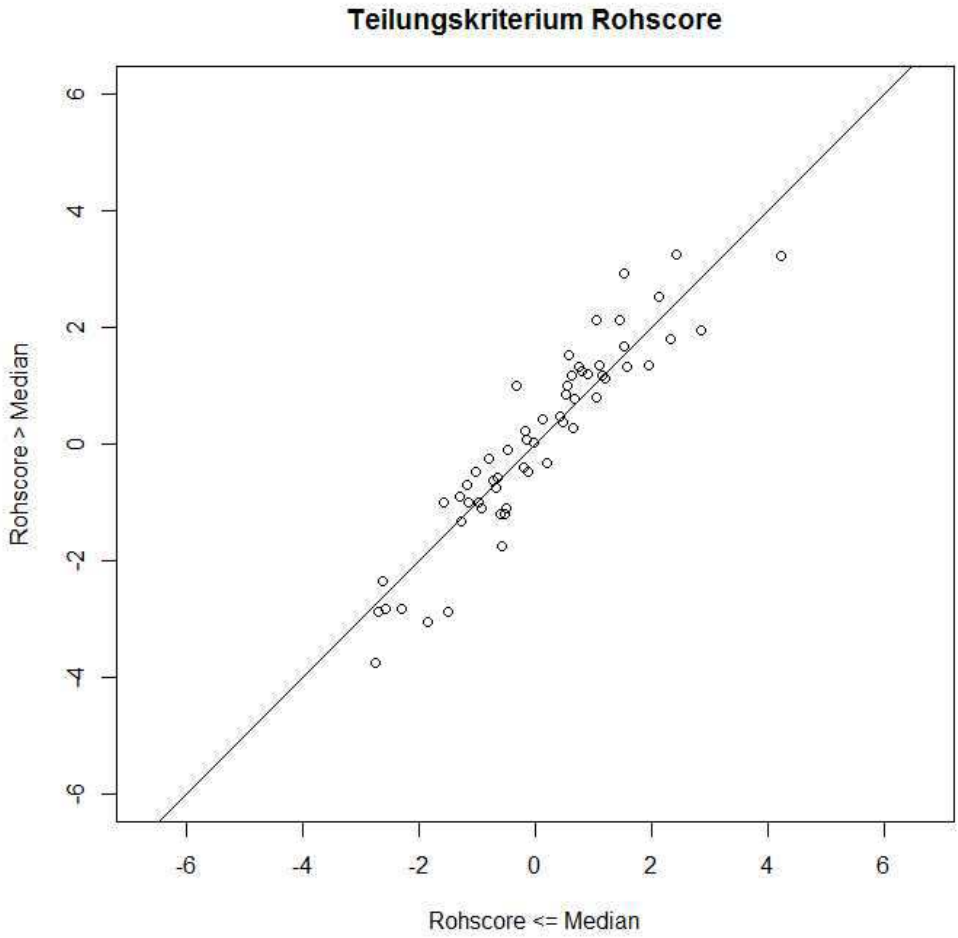

Abbildung 6: Grafischer Modelltest Teilungskriterium Rohscore

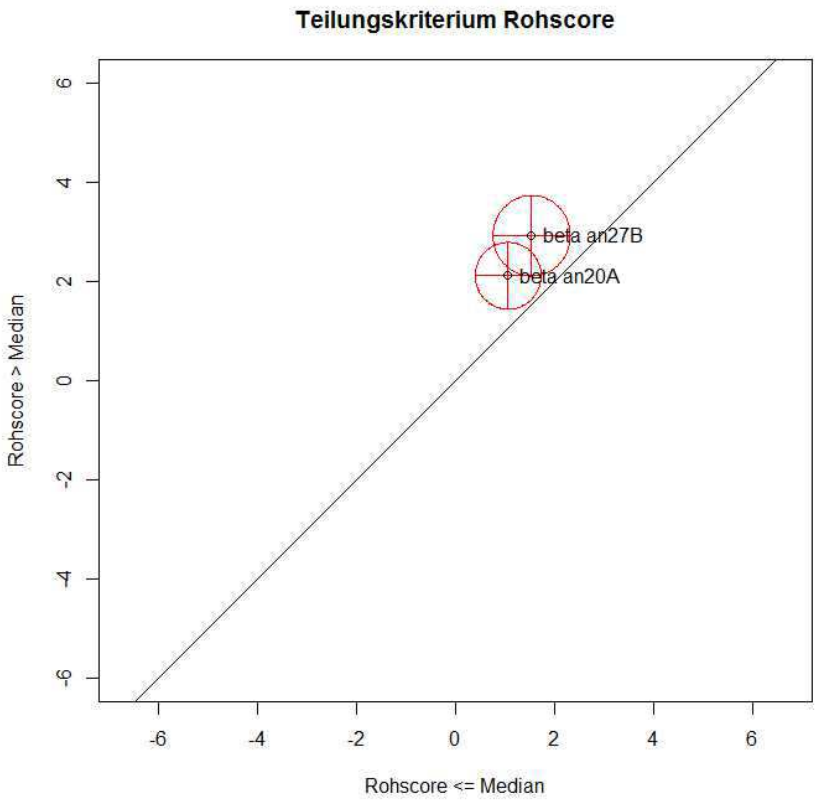

Abbildung 7: Grafischer Modelltest - Teilungskriterium Rohscore - Auffällige Items

Im nächsten Schritt wurde ein Wald-Test durchgeführt, der die Ergebnisse der Grafischen Modelltests stützte, auch hier waren die beiden Items 20A und 27B auffällig. Aus Gründen der Übersichtlichkeit finden sich die genauen Ergebnisse der Wald-Tests in allen Berechnungsschritten im Anhang.

Teilungskriterium Alter

Der LRT für das Teilungskriterium Alter fiel nicht signifikant aus und auch im Grafischen Modelltest konnten keine auffälligen Items identifiziert werden. Neun Items konnten aufgrund des ungünstigen Antwortverhaltens in den einzelnen Subgruppen nicht geschätzt werden.

Tabelle 5 zeigt bezüglich des Teilungskriteriums *Alter* die geschätzte asymptotisch χ^2-verteilten Testgröße des LRT, die Freiheitsgrade (*df*), die Wahrscheinlichkeit, dass die H_0 gilt (*p*-Wert) sowie den kritischen Wert der χ^2-Verteilung (bei $\alpha = .01$) an.

Tabelle 5: LRT Teilungskriterium „Alter“, erster Berechnungsdurchgang

Teilungskriterium Alter			
Andersen χ^2	*df*	p-Wert	Kritischer χ^2-*Wert*
57.37	54	0.351	81.07

Abbildung 8 zeigt den Grafischen Modelltest für alle in Bezug auf das Teilungskriterium *Alter* geschätzten Items.

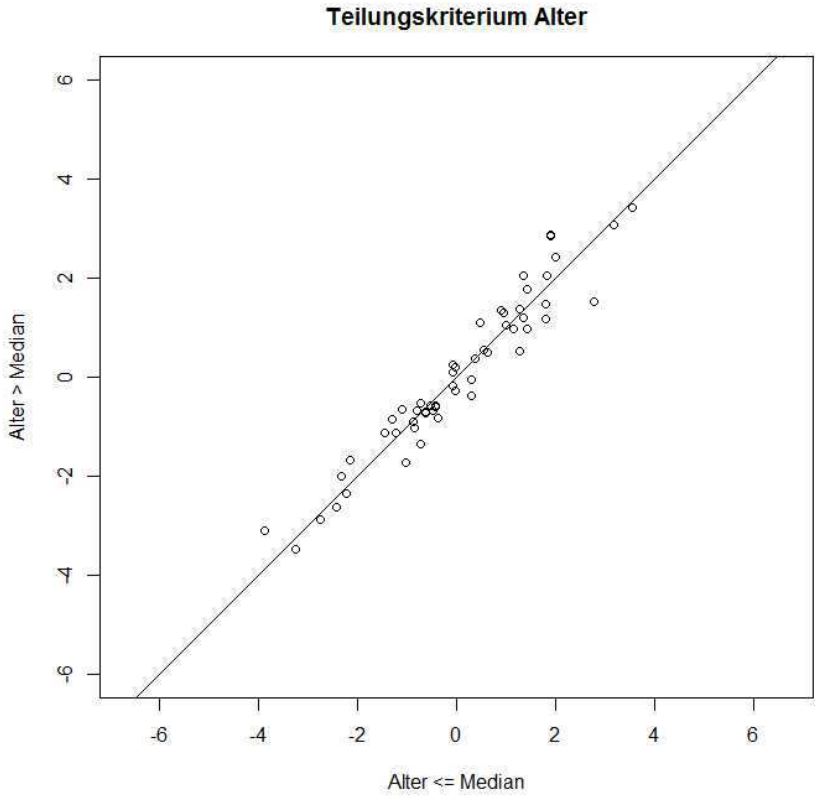

Abbildung 8: Grafischer Modelltest Teilungskriterium "Alter"

Teilungskriterium Geschlecht

Der LRT für das Teilungskriterium *Geschlecht* fiel ebenfalls nicht signifikant aus und auch im Grafischen Modelltest konnten keine auffälligen Items identifiziert werden. Ein Item konnte aufgrund des ungünstigen Antwortverhaltens in den einzelnen Subgruppen nicht geschätzt werden.

Tabelle 6 zeigt bezüglich des Teilungskriteriums *Geschlecht* die geschätzte asymptotisch χ^2-verteilten Testgröße des LRT, die Freiheitsgrade (*df*), die

Wahrscheinlichkeit, dass die H_0 gilt (*p*-Wert) sowie den kritischen Wert der χ^2-Verteilung bei ($\alpha = .01$) an.

Tabelle 6: LRT Teilungskriterium „Geschlecht“, erster Berechnungsdurchgang

Teilungskriterium Geschlecht			
Andersen χ^2	*df*	*p*-Wert	Kritischer χ^2-*Wert*
51.05	62	0.838	90.8

Abbildung 9 zeigt den Grafischen Modelltest für alle in Bezug auf das Teilungskriterium *Geschlecht* geschätzten Items.

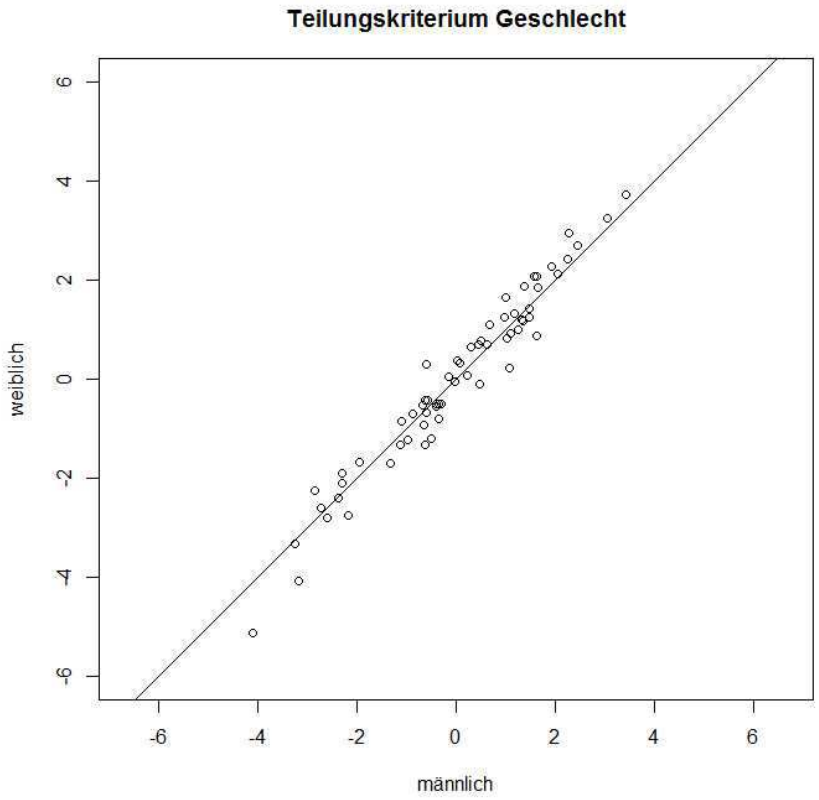

Abbildung 9: Grafischer Modelltest Geschlecht

Teilungskriterium Muttersprache

Der LRT für das Teilungskriterium *Muttersprache* fiel ebenfalls nicht signifikant aus und auch im Grafischen Modelltest konnten keine auffälligen Items identifiziert werden. Ein Item konnte aufgrund des ungünstigen Antwortverhaltens in den einzelnen Subgruppen nicht geschätzt werden.

Tabelle 7 zeigt bezüglich des Teilungskriteriums *Muttersprache* die geschätzte asymptotisch χ^2-verteilten Testgröße des LRT, die Freiheitsgrade (*df*), die Wahrscheinlichkeit, dass die H_0 gilt (*p*-Wert) sowie den kritischen Wert der χ^2-Verteilung bei ($\alpha = .01$) an.

Tabelle 7: LRT Teilungskriterium „Muttersprache“, erster Berechnungsdurchgang

Teilungskriterium Muttersprache			
Andersen χ^2	*df*	*p*-Wert	Kritischer χ^2-Wert
78.77	62	0.704	90.8

Abbildung 9 zeigt den Grafischen Modelltest für alle in Bezug auf das Teilungskriterium *Muttersprache* geschätzten Items.

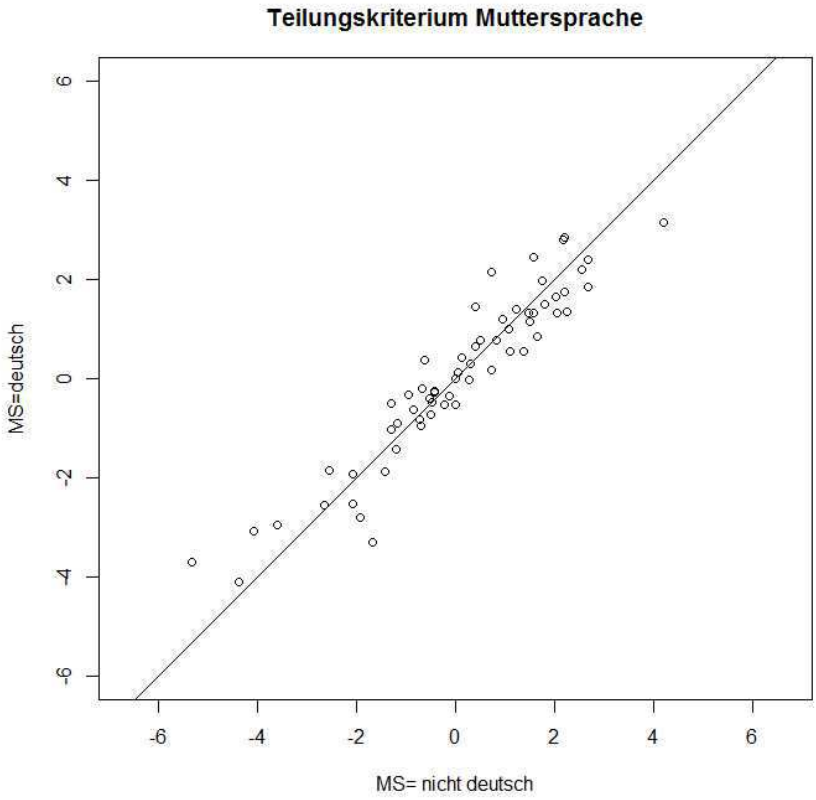

Abbildung 10: Grafischer Modelltest Muttersprache

8.1.3 Ausschluss nicht Rasch-Modell-konformer Items

Bei der Überprüfung des Ausgangsdatensatzes erwiesen sich die beiden Items 20A und 27B als nicht Rasch-Modell-konform. Diese beiden Items und ihre jeweiligen Parallelversionen 20B und 27A waren schon vor der eigentlichen Testvorgabe in den Diskussionen mit anderen Diplomandinnen und Diplomanden bezüglich des Konstruktionsrationals (siehe Kapitel 6.2) aufgefallen. Beide Items folgen derselben inhaltlichen Konstruktionsregel, die besagt, dass die zum Auffinden der Lösung benötigten Informationen in nur einem der beiden ersten Felder präsentiert werden. Die zu verwendende Regel wird also nur einmal präsentiert.

Daraus wurde geschlossen, dass Items, die dieser Regel folgen nicht für alle Testpersonen logisch sind. Daher wurden im ersten Schritt die beiden Items 27A und 27 B ausgeschlossen, obwohl 27A in den Modelltests nicht auffällig war.

Nach Ausschluss der beiden Items ergab der LRT zum Teilungskriterium *Rohscore* erneut ein signifikantes Ergebnis (siehe Anhang). Item 20A war im daraufhin durchgeführten Wald-Test ebenfalls signifikant. Im nächsten Schritt wurden aus den oben für die Items 27A und 27B genannten Gründen die Items 20A und 20B ausgeschlossen. Nach Ausschluss dieser vier Items ergaben die LRTs kein signifikantes Ergebnis mehr. Jedoch war das Item 11A im Grafischen Modelltest zum Teilungskriterium Muttersprache auffällig und wurde daher ebenfalls von den weiteren Berechnungen ausgeschlossen. Die genannten Teilergebnisse sowie die grafischen Modelltests nach Ausschluss der auffälligen Items finden sich im Anhang.

8.1.4 A-Posteriori Ergebnisse

Nach Ausschluss der Items 11A, 20A, 20B, 27A und 27B konnte a-posteriori Modellgültigkeit erzielt werden. Die Nullhypothese kann daher beibehalten werden. Tabelle 8 gibt einen Überblick über die Ergebnisse der LRTs nach Ausschluss der auffälligen Items.

Tabelle 8: Ergebnisse der LRT nach Ausschluss auffällger Items

Teilungskriterium Rohscore			
Andersen χ^2	*df*	*p*-Wert	Kritischer χ^2-Wert
62.31	51	0.133	77.39
Teilungskriterium Alter			
Andersen χ^2	*df*	*p*-Wert	Kritischer χ^2-Wert
45.19	49	0.629	74.92
Teilungskriterium Geschlecht			
Andersen χ^2	*df*	*p*-Wert	Kritischer χ^2-Wert
47.86	57	0.8	84.73
Teilungskriterium Muttersprache			
Andersen χ^2	*df*	*p*-Wert	Kritischer χ^2-Wert
60.21	57	0.361	84.73

Eine Übersicht über die geschätzten Itemschwierigkeitsparameter findet sich in Tabelle 9 auf der folgenden Seite.

Tabelle 9: Itemschwierigkeitsparameter für die Rasch-Modell-konformen Items der AN-TOP

Itemschwierigkeitsparameter mit 0.95 Konfindezintervall							
Item	Schätzung	Unteres KI	Oberes KI	Item	Schätzung	Unteres KI	Oberes KI
an91B	-4.160	-5.604	-2.716	an12B	0,724	0,416	1.032
an92A	-2.244	-2.960	-1.527	an13A	0,437	0,116	0,758
an92B	-2.361	-3.105	-1.616	an13B	0,92	0,598	1.242
an93A	1.245	0,603	1.887	an14A	1.381	1.045	1.717
an93B	0,887	0,322	1.451	an14B	1.629	1.286	1.973
an94A	0,782	0,199	1.366	an15A	-0,566	-0,911	-0,22
an94B	1.506	0,868	2.144	an15B	-0,598	-0,949	-0,247
an01A	-2.488	-3.038	-1.938	an16A	1.904	1.506	2.301
an01B	-1.991	-2.470	-1.513	an16B	2.400	1.964	2.836
an02A	-3.264	-3.976	-2.553	an17A	3.477	2.924	4.031
an02B	-3.096	-3.780	-2.413	an17B	3.881	3.248	4.515
an03A	0,437	0,116	0,758	an18A	0,391	0,009	0,774
an03B	0,186	-0,138	0,509	an18B	0,124	-0,273	0,521
an04A	-2.488	-3.038	-1.938	an19A	-0,27	-0,604	0,064
an04B	-2.235	-2.749	-1.721	an19B	-0,214	-0,548	0,12
an05A	-1.630	-2.058	-1.202	an21A	-0,68	-1.031	-0,328
an05B	-1.882	-2.346	-1.418	an21B	-0,24	-0,575	0,095
an06A	-0,767	-1.113	-0,422	an22A	1.523	1.193	1.853
an06B	-0,542	-0,879	-0,206	an22B	1.511	1.183	1.839
an07A	-1.266	-1.790	-0,742	an23A	-0,296	-0,631	0,039
an07B	-0,361	-0,787	0,064	an23B	0,21	-0,113	0,533
an08A	-0,86	-1.222	-0,498	an24A	-0,19	-0,51	0,13
an08B	-0,162	-0,494	0,17	an24B	-0,322	-0,649	0,004
an09A	1.153	0,825	1.481	an25A	1.634	1.244	2.023
an09B	1.521	1.182	1.859	an25B	2.107	1.686	2.528
an10A	1.178	0,849	1.507	an26A	-0,987	-1.358	-0,616
an10B	1.713	1.365	2.061	an26B	-0,401	-0,742	-0,06
an11B	0,448	0,128	0,768	an28A	2.634	2.209	3.060
an12A	0,893	0,582	1.204	an28B	2.012	1.646	2.377

8.2 Analysen der AN-TOP nach dem LLTM

Im weiteren Verlauf soll das Konstruktionsrational der Aufgaben der AN-TOP mittels LLTM (siehe Kapitel 4) untersucht werden. Die Voraussetzung, dass das Rasch-Modell gilt, ist in Folge der Kapitel 8.1 beschriebenen Analysen erfüllt. Daher soll nun geprüft werden, ob die durch das LLTM geschätzten Basisparameter die Itemschwierigkeiten genauso gut erklären können wie das Rasch-Modell.

8.2.1 Überprüfung des Konstruktionsrationals

Das Konstruktionsrational, das den Aufgaben der AN-TOP zugrundeliegt, wurde bereits in Kapitel 6.3 ausführlich beschrieben. Auch die für die Schätzung des LLTM verwendete Gewichtszahlenmatrix der Basisparameter entspricht der in Kapitel 6.3 angeführten Tabelle 1, es wurden lediglich die nicht Rasch-Modell-konformen Items aus der Matrix entfernt.

Zusätzlich zu den im Rasch-Modell auffälligen Items wurden die Items 22A und B sowie das Item 11B von den weiteren Berechnungen ausgeschlossen. Items 22A und B folgen als einzige Items der Regel „Addition“. Sie wurden ausgeschlossen, da ihr Belassen in den Berechnungen zu einem artifiziellen Ergebnis bezüglich der Schwierigkeit der Regel Addition führen würde.
Item 11B wurde ebenfalls ausgeschlossen, da es isomorph zu Item 11A ist, das im Rasch Modell auffällig war.

Tabelle 10 zeigt das Ergebnis des Vergleichs von RM und LLTM. Das Ergebnis fällt nicht signifikant aus, daher kann die Nullhypothese – „Die Itemschwierigkeiten der 59 Items können durch die 17 Basisparameter ebenso gut erklärt werden, wie durch das Rasch-Modell.“ – beibehalten werden.

Tabelle 10: Ergebnisse der Schätzung des LLTM

$\mathbf{logL}_{LLTM}$	$\mathbf{logL}_{RM}$	$\boldsymbol{\chi 2}$	$\boldsymbol{\chi 2}_{krit\ \alpha = .01}$	*df*	Ergebnis
-3716.68	-3742.93	52.5	61.162	38	nicht signifikant

Abbildung 11 zeigt eine grafische Gegenüberstellung der im RM und im LLTM geschätzten Itemschwierigkeiten.

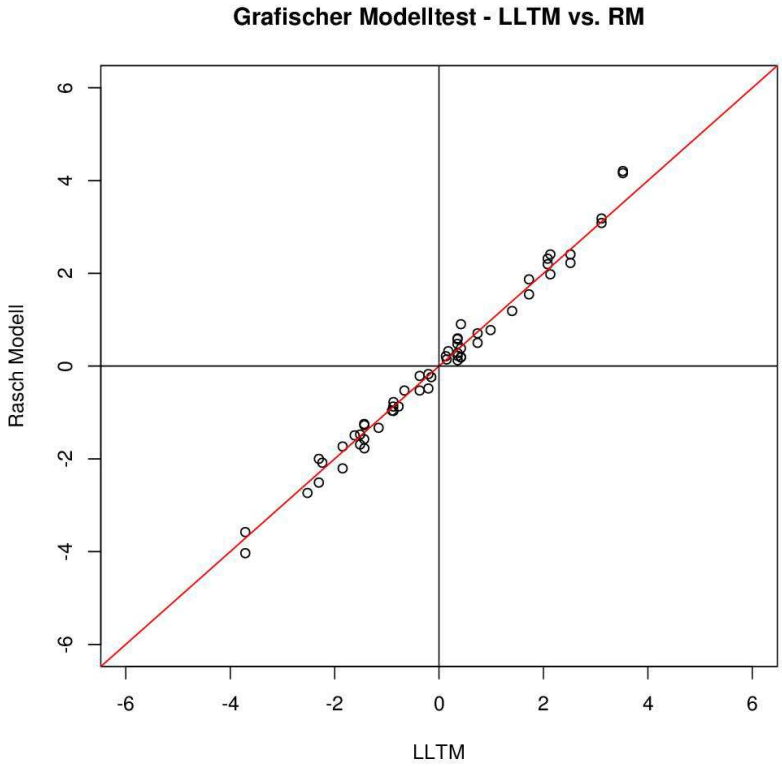

Abbildung 11: Grafischer Modelltest LLTM versus Rasch-Modell

9. Diskussion und Ausblick

Die vorliegende Studie setzt sich mit der Konstruktion eines schriftlichen Gruppentests zum schlussfolgernden Denken bei Kindern und Jugendlichen zwischen 6;0 und 15;11 Jahren auseinander und ist ein Versuch der regelgeleiteten Testentwicklung in diesem Bereich. Die Entwicklung der Items erfolgte formalisiert anhand eines Konstruktionsrationals um die Validität des Tests auf Itemebene zu gewährleisten und auf ökonomische Weise einen möglichst großen Itempool entwickeln zu können.

Die Ergebnisse der Studie zeigen, dass das grundlegende Ziel der Arbeit, nämlich die Entwicklung eines kognitiven Leistungstests, der nachweislich eindimensional misst, erfüllt wurde. Bei dem Verfahren AN-TOP konnte nach dem Ausschluss von 5 Items post-hoc Rasch-Modellgültigkeit angenommen werden. Von den ursprünglichen 64 Items mussten die Items 11A, 20A, 20B, 27A sowie 27B aufgrund mangelnder Modellgültigkeit von den weiteren Analysen ausgeschlossen werden. Für die beiden jeweiligen Parallelitems 20A und B sowie 27A und B konnte eine inhaltliche Erklärung gefunden werden. Beide Itemstämme 20 und 27 folgen demselben Konstruktionsprinzip, das besagt, dass die anzuwendende inhaltliche Regel in nur einem der beiden ersten Felder, die der Regeldemonstration dienen, vorkommt. Dies scheint nicht für alle Testpersonen logisch schlüssig zu sein. Bei Item 11A sind bei genauerer Betrachtung zwei Lösungen denkbar und es wurde aufgrund seiner Auffälligkeit im grafischen Modelltest zum Teilungskriterium Muttersprache ebenfalls ausgeschlossen.
Demzufolge mussten 5 der insgesamt 64 Items ausgeschlossen werden. Das entspricht etwa 8% des Itempools und liegt unter der von Kubinger und Draxler (2007) postulierten Marke von 10%.

Die itemgenerierenden Regeln, anhand derer die Items der AN-TOP konstruiert wurden, wurden im nächsten Schritt mit Hilfe des LLTM überprüft. Neben deren

großen Nützlichkeit bei der Konstruktion eines Itempools konnte über die Gültigkeit des LLTM gezeigt werden, dass diese auch die zugrundeliegenden kognitiven Operationen bei der Bearbeitung der Items wiederspiegeln. Diese Konstruktionsprinzipien, die teilweise der Literatur entstammen (siehe Primi, 2001), lieferten auch eine Erklärung für etwaige mangelnde Rasch-Modell-Gültigkeit der Items, die ausgeschlossen werden mussten. Allerdings ist hier kritisch anzumerken, dass nicht alle Rasch-Modell-konformen Items in die Überprüfung der Gültigkeit des Regelsystems eingingen. Es wurde die Entscheidung getroffen, neben den oben erwähnten fünf Items, drei zusätzliche Items von den weiteren Berechnungen auszuschließen. Für die Items 22A und B geschah dies, da sie als einzige der Regel Addition folgten und ihre Beibehaltung zu einem artifiziellen Ergebnis für die Schwierigkeit der Operation geführt hätte. Für künftige Studien wird empfohlen zusätzliche Items, die mehr Regelkombinationen realisieren, anhand des Konstruktionsrationals zu generieren.

Zwei weitere Kritikpunkt beziehen sich auf die Art der Vorgabe des Tests. Erstens scheint der Test für eine Vorgabe in Gruppenform bei Kindern unter 8 Jahren nicht geeignet, da grundlegende, in der Schule vermittelte Fähigkeiten, wie etwa Schreib- und Lesefähigkeit, für die Bearbeitung des Tests in Gruppenform notwendig sind. Dieses Problem könnte durch die Vorgabe als Einzelverfahren vermieden werden. Zusätzlich erfordern einige der Aufgaben grundlegende Rechenfähigkeiten. Um Schulanfänger bei der Testvorgabe nicht zu frustrieren, sollte für diese Population eine eigene Testform erstellt werden, die keine Items beinhaltet, die altersunangemessene mathematische Operationen beinhaltet.
Zweitens wurde der Test aus Kostengründen mit einem Aufgabenheft und einem Lösungsbogen vorgegeben. Einige, vor allem jüngere Kinder waren durch diese Art der Vorgabe verwirrt. Es wird empfohlen, Testpersonen die Lösungen bei zukünftigen Testungen direkt am Angabebogen ausfüllen zu lassen.

Diese Arbeit stellt somit ein eindimensional messendes psychologisch-diagnostisches Verfahren zur Verfügung. Um das Verfahren in der Praxis verwenden zu können, sollte nun eine Eichung an einer repräsentativen Stichprobe erfolgen.

Für die Weiterentwicklung des Verfahrens sollten anhand des Konstruktionsrationals wie erwähnt weitere Items mit einer höheren Diversität der verwendeten Regelkombinationen konstruiert werden.

Aufgrund der Gültigkeit des Konstruktionsrationals kann laut Embretson (1983) für den Test Validität auf Itemebene angenommen werden. Allerdings stellt dies nur einen Zugang zur Validitätsüberprüfung, nämlich über die Konstruktrepräsentation dar. Wie zuvor erwähnt, sollte für die weitere testtheoretische Überprüfung des Verfahrens noch ein weiterer Zugang über die nomothetische Spanne, etwa durch die Anwendung des MTMM-Verfahrens oder über eine Kreuzvalidierung gewählt werden.

10. Zusammenfassung

Die vorliegende Arbeit verfolgte das Ziel anhand eines Konstruktionsrationals einen Test zum schlussfolgernden Denken zu entwickeln, der dem Rasch-Modell entspricht. Dieses regelgeleitete Vorgehen wurde für die Konstruktion des Verfahrens einerseits gewählt um die Validität bereits auf Itemebene zu gewährleisten, andererseits um Fehler zu vermeiden, die durch intuitive Testkonstruktion entstehen können (vgl. Bühner, 2011; Hornke & Habon, 1986).

Im ersten Schritt wurde daher ein Pool von 64 Items konstruiert, die verschiedene Kombinationen der postulierten schwierigkeitsgenerierenden Elemente realisierten. Aus den 64 resultierenden Items wurden 8 Testversionen zu je 28 Items erstellt, jeweils 4 Parallelversionen für 2 Altersgruppen. Die Altersgruppen unterschieden sich in Bezug auf die angenommene Schwierigkeit der Items.

Im nächsten Schritt wurde der Test einer Stichprobe von 423 Schülerinnen und Schülern aus Oberösterreich im Alter von 7 bis 14 Jahren vorgegeben. Die Variable Geschlecht war annähernd gleichverteilt, die Variable Alter ähnelte eher eine Normalverteilung. Etwa ein Viertel der Teilnehmerinnen und Teilnehmer hatte nicht Deutsch als Muttersprache.

Die Aufgaben der AN-TOP wurden daraufhin auf Rasch-Modell-Konformität geprüft. Nach sukzessiven Ausscheiden von 5 Items, das entspricht etwa 8% des Gesamtitempools, konnte a-posteriori Rasch-Modell-Konformität für die verbleibenden Items erreicht werden. Für die 5 ausgeschiedenen Items konnte eine inhaltliche Erklärung der mangelnden Modellgeltung gefunden werden. Die geschätzten Itemschwierigkeitsparameter zeigen eine breite Streuung der Itemschwierigkeiten, was eine gleichmäßige Abdeckung eines breiten Fähigkeitsbereiches bedeutet.

Im nächsten Schritt wurde das Konstruktionsrational, auf dem die Items der AN-TOP basieren überprüft. Hierfür wurden die im Rasch Modell geschätzten Itemparameter mit jenen verglichen, die sich durch eine additive Operation der im LLTM geschätzten schwierigkeitsgenerierenden Operationen ergaben. Das Konstruktionsrational hielt dem Vergleich stand. Daher kann angenommen werden, dass die Schwierigkeiten der Items der AN-TOP hinreichend genau durch zugrundeliegende kognitive Operationen erklärt werden können. Die Gültigkeit des Modells liefert zudem einen Hinweis auf Vorliegen von Validität auf Itemebene. Dies entspricht allerdings nur einem Validitätsbefund hinsichtlich der Konstruktrepräsentation, offen bleibt die Frage nach der nomothetischen Spanne des Verfahrens. Es wird daher angeraten eine weitere Validierung des Verfahrens durch Ermittlung der konvergenten Validität durchzuführen.
Für den Einsatz des Tests in der Praxis wird empfohlen Eichtabellen an einer repräsentativen Eichstichprobe zu erstellen.

Zusammenfassend kann gesagt werden, dass der Versuch, einen Test zum schlussfolgernden Denken für Kinder zwischen 6;0 und 15;11 Jahren geglückt ist.

Tabellenverzeichnis

Formelverzeichnis

Abbildungsverzeichnis

Literaturverzeichnis

Ackerman, P.L., Beier, M.E. & Boyle, M.O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131,* 30 – 60.

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38 (1),* 123 – 140.

Arendasy, M. E. & Sommer, M. (2005). The effect of different types of perceptual manipulations on the dimensionality of automatically generated figural matrices. *Intelligence, 33,* 307 – 324.

Arendasy, M. E. & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and Individual Differences, 22,* 112 – 117.

Beckmann, J. F. & Guthke, J. (1999). Psychodiagnostik des Schlussfolgernden Denkens. Göttingen: Hogrefe.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S.H. Irvine & P.C. Kyllonen (Eds.), *Item generation for test development* (S. 199 - 218). Hillsdale, NJ: Erlbaum.

Bühner, M., Krumm, S. & Pick, M. (2005). Reasoning is working memory and not attention. *Intelligence, 33,* 251–272.

Carpenter, P. A., Just, M. A. & Shell, P. (1990). What One Intelligence Test Measures: A Theoretical Account of the Processing in the Raven Progressive Matrices Test. *Psychological Review, 97(3),* 404 – 431.

Carroll, J. B. (1993). Human Cognitive Abilities. A survey of factor-analytic studies. New York, Cambridge University Press.

Carroll, J. B. (2005). The three-stratum theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), Contemporary intellectual assessment (2nd ed., S. 67-76). New York: The Guilford Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1 – 22.

Cattell, R. B. & Horn, J. L. (1978). A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement, 15 (3)*, 139 – 164.

Cattell, R. B. (1941). The measurement of adult intelligence. *Psychological Bulletin, 40,* 143 – 193.

Chuderski, A. (2013). When are fluid intelligence and working memory isomorphic and when are they not? *Intelligence, 41*, 244 – 262.

Dimitrov, D. M. & Raykov, T. (2003). Validation of Cognitive Structures: A Structural Equation Modeling Approach. *Multivariate Behavioral Research*, *38 (1)*, 1-23.

Embretson, S. E. (1983). Construct Validity: Construct Representation versus Nomothetic Span. *Psychological Bulletin, 93*, 179 – 197.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175 – 186.

Embretson, S. E. (1995). A measurement model for linking individual learning to process and knowledge: application to mathematical reasoning. *Journal of Educational Measurement, 32*, 277 – 294.

Embretson, S. E. (1998). A Cognitive Design System Approach to Generating Valid Tests: Application to Abstract Reasoning. *Psychological Methods, 3,* 380 – 396.

Embretson, S. E. (1999). Generating Items During Testing: Psychometric Issues And Models. *Psychometrika, 64,* 407 - 433.

Embretson, S. E. & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly, 50 (3)*, 328 – 344.

Embretson, S. E. & Gorin, J. (2001). Improving Construct Validity with Cognitive Psychology Principles. *Journal of Educational Measurement, 38,* 343 – 368.

Fischer, G. H. (1972). Conditional maximum-likelihood estimations of item parameters for a linear logistic test model (Research Bulletin 9). Vienna: University of Vienna, Psychological Institute.

Fischer, G. H. (1973). The Linear Logistic Test Model as an Instrument in Educational Research. *Acta Psychologica, 37,* 359 – 374.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

Fischer, G. H. (1995a). Derivations of the Rasch Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (S. 15 – 38). New York: Springer.

Fischer, G. H. (1995b). The linear logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (S. 131 – 155). New York: Springer.

Fischer, G. H. & Scheiblechner H. H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch. *Psychologische Beiträge, 12,* 23-51.

Forman, A. K. (1973). *Die Konstruktion eines neuen Marizentests und die Untersuchung des Lösungsverhaltens mit Hilfe des Linear Logistischen Testmodells*. Unveröffentlichte Dissertation: Universität Wien.

Forman, A. K. & Piswanger, K. (1979). *Wiener Matrizen-Test (WMT)*. Weinheim: Beltz.

Freund, P. A., Hofer, S. & Holling, H. (2008). Explaining and Controlling for the Psychometric Properties of Computer-Generated Figural Matrix Items. *Applied Psychological Measurement, 32,* 195 – 210.

Fröhlich, W. D. (2010). *Wörterbuch Psychologie*. München: DTV.

Geerlings, H., Glas, C. A. W. & van der Linden, W. (2011). Modeling Rule-Based Item Generation. *Psychometrika, 76,* 337 – 359.

Glas, C. A. W. & Verhelst, N. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (S. 69 – 95). New York: Springer.

Gorin, J. S. & Embretson, S. E. (2006). Item Diffficulty Modeling of Paragraph Comprehension Items. *Applied Psychological Measurement, 30,* 394 – 411.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage.

Heller, K. A., Kratzmeier, H. & Lengfelder, A. (1998a). *Matrizen-Test-Manual, Band 1*. Göttingen: Beltz-Test.

Heller, K. A., Kratzmeier, H. & Lengfelder, A. (1998b). *Matrizen-Test-Manual, Band 2*. Göttingen: Beltz-Test.

Holling, H., Bertling, J. P. & Zeuch, N. (2009). Automatic item generation of probability word problems. *Studies in Educational Evaluation, 35,* 71-76.

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.

Horn, J. L. & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26*, 107 – 129.

Hornke, L. F. (1999). Benefits from Computerized Adaptive Testing as seen in Simulation Studies. *European Journal of Psychological Assessment, 15,* 91 – 98.

Hornke, L. F., Etzel, S. & Rettig, K. (1997). *Adaptiver Matrizentest (AMT)* [Software und Manual]. Mödling: Schuhfried.

Hornke, L. F. & Habon, M. W. (1986). Rule-Based Item Bank Construction and Evaluation within the Linear Logistic Framework. *Applied Psychological Measurement, 10,* 369 – 380.

Irvine, S. H. (2002). Item generation for test development: An introduction. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (S. xv-xxv). Mahwah, NJ: Erlbaum.

Irvine, S. H., Dann, P. L. & Anderson, J. D. (1990). Towards a theory of algorithm-determined cognitive test construction. *British Journal of Psychology, 81,* 173 – 195.

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F. & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods, 14,* 54 – 75.

Kubinger, K. D. (1989). *Moderne Testtheorie*. 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. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly, 50,* 311 – 327.

Kubinger, K. D. (2009a). *Psychologische Diagnostik*. Göttingen: Hogrefe.

Kubinger, K. D. (2009b). Applications of the Linear Logistic Test Model in Psychometric Research. *Educational and Psychological Measurement, 69,* 232 – 244.

Kubinger, K. D. & Draxler, C. (2007). Probleme bei der Testkonstruktion nach dem Rasch-Modell. *Diagnostica, 53 (3),* 131 – 143.

Kubinger, K. D. & Holocher-Ertl, S. (2014). *AID 3: Adaptives Intelligenz Diagnostikum 3*. Göttingen: Beltz-Test.

Kubinger, K. D., Poinstingl, H., Weidinger, S. & Skoda, S. (in Vorb.). *FRRT: Family Relations Reasoning Test.*

Kubose, T. T., Holyoak, K. J. & Hummel, J. E. (2002). The role of textual coherence in incremental analogical mapping. *Journal of Memory and Language, 47 (3),* 407 – 435.

Kunda, M., McGreggor, K. & Goel, A. K. (2013). A computational model for solving problems from the Raven's Progressive Matrices intelligence test using iconic visual representations. *Cognitive Systems Research, 22–23,* 47 – 66.

Kyllonen, P. C. (2002). Item generation for repeated testing of human performance. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (S. 251 – 276). Mahwah, NJ: Erlbaum.

Mackey, A. P., Hill, S. S., Stone, S. I. & Bunge, S. A. (2011). Differential effects of reasoning and speed training in children. *Developmental Science, 14*, 582 – 590.

Marshalek, B., Lohman, D. F. & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7*, 107 – 127.

Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist, 50*, 741 – 749.

Molenaar, I. W. (1995). Some Background for Item Response Theory and The Rasch Model. In G. H. Fischer & I.W. Molenaar (Eds.), *Rasch Models – Foundations, Recent Developments, and Applications* (S. 3 – 14). New York: Springer.

Mulholland, T. M., Pellegrino, J. W. & Glaser, R. (1980). Components of Geometric Analogy Solution. *Cognitive Psychology, 12,* 252 – 284.

Poinstingl, H. (2009). The LLTM as the basis for item generating rules of a new reasoning test: Family Relations Test. *Psychology Science Quarterly, 51,* 123 – 134.

Poinstingl, H., Mair, P. & Hatzinger, R. (2007). Manual zum Softwarepackage eRm (extended Rasch modeling). Anwendung des Rasch-Modells (1-PL Modell) – Deutsche Version. Lengerich: Pabst.

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence, 30*, 41 – 70.

Rasch, G. (1960). *Probabilistic Models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52 (2),* 127 – 141.

Reif, M. (2008). *Zur Konstruktion eines adaptiven Persönlichkeitsfragebogens auf Basis eines Konstruktionsrationals*. Unveröffentlichte Diplomarbeit: Universität Wien.

Rost, J. (2004). *Testtheorie – Testkonstruktion.* (2., überarb. und erw. Aufl.). Bern: Verlag Hans Huber.

Salthouse, T.A., Schroeder, D. H. & Ferrer, E. (2004). Estimating Retest Effects in Longitudinal Assessments of Cognitive Functioning in Adults Between 18 and 60 Years of Age. *Developmental Psychology, 40 (5),* 813 – 822.

Schaie, K. W. (2005). *Developmental influences on adult intelligence*. New York: Oxford University Press.

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. *Zeitschrift für Angewandte und Experimentelle Psychologie, 19,* 476 – 505.

Sonnleitner, P. (2008). Using the LLTM to evaluate an item generating system for reading comprehension. *Psychology Science Quarterly, 50,* 345 – 362.

Stemmler, G., Hagemann, D., Amelang, M. & Bartussek, D. (2011). *Differentielle Psychologie und Persönlichkeitsforschung* (7. vollst. überarb. Aufl.). Stuttgart: Kohlhammer.

Sternberg, R. J. (1977). A component process in analogical reasoning. *Psychological Review, 84 (4),* 353 – 378.

Sternberg, R. J. & Williams, W. M. (1998). *Intelligence, Instruction, and Assessment: Theory Into Practice*. Mahwah, NJ: Lawrence Erlbaum Associates.

Suen, H. K. & French, J. L. (2003). A history of the development of psychological and educational testing. In Reynolds, C. R. & Kamphaus, R. W. (Eds.), *Handbook of psychological and educational assessment of children: Intelligence, aptitude, and achievement,* (S. 3 – 23). New York: The Guilford Press.

Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54,* 426 – 482.

Formann, A. K., Waldherr, K. & Piswanger, K. (2011). *Wiener Matrizen-Test 2*. Göttingen: Beltz Test GmbH.

Weiss, D. J. & Davison, M. L. (1981). Test theory and methods. *Annual Review of Psychology, 32,* 629 – 658.

Weiß, R.H. (2006). *Grundintelligenztest Skala 2 – Revision CFT 20-R*. Göttingen: Hogrefe.

Wilhelm, O. (2005). Measuring Reasoning Ability. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of Understanding and Measuring Intelligence* (S. 373 – 392). London: Sage.

IV. Anhang

11. Ergebnistabellen und Grafische Modelltests der weiteren durchgeführten Berechnungen und Modellschätzungen

11.1 Zusätzliche Ergebnisse der a-priori Modellschätzung

Tabelle 11: Ergebnis des z-Tests nach Fischer und Scheiblechner für das Teilungskriterium Score, erster Berechnungsdurchgang

Item	z-statistic	p-value	Item	z-statistic	p-value	Item	z-statistic	p-value
an92A	0,314	0,753	an09B	1.694	0,09	an19B	-1.504	0,133
an92B	-0,169	0,866	an10A	0,268	0,788	an20A	2.859	0,004
an93A	0,689	0,491	an10B	0,763	0,446	an20B	1.736	0,083
an93B	0,095	0,924	an11A	1.691	0,091	an21A	0,352	0,725
an94A	2.283	0,022	an11B	1.155	0,248	an21B	0,255	0,799
an94B	1.498	0,134	an12A	-0,338	0,735	an22A	0,922	0,356
an01A	-0,974	0,33	an12B	0,901	0,368	an22B	1.420	0,156
an01B	-0,681	0,496	an13A	-1.599	0,11	an23A	-1.727	0,084
an03A	0,148	0,882	an13B	-1.113	0,266	an23B	-0,975	0,329
an03B	-0,563	0,573	an14A	-0,734	0,463	an24A	0,219	0,827
an04A	-0,974	0,33	an14B	0,047	0,962	an24B	-1.559	0,119
an04B	-0,33	0,741	an15A	-0,065	0,948	an25A	-0,221	0,825
an05A	-1.589	0,112	an15B	1.244	0,214	an25B	0,328	0,743
an06A	1.053	0,292	an16A	-0,587	0,557	an26A	1.443	0,149
an06B	-0,424	0,671	an16B	-1.008	0,314	an26B	1.545	0,122
an07A	-1.334	0,182	an17A	1.490	0,136	an27A	0,847	0,397
an07B	-1.820	0,069	an17B	-0,946	0,344	an27B	3.169	0,002
an08A	-0,131	0,895	an18A	0,525	0,6	an28A	-1.611	0,107
an08B	1.559	0,119	an18B	0,877	0,381	an28B	-1.387	0,165
an09A	1.008	0,313	an19A	-0,207	0,836			

11.2 Ergebnisse der Modellschätzung nach Ausschluss der Items 27A und 27B

Tabelle 12: Ergebnisse der LRTs nach Ausschluss der Items 27A und 27B

Teilungskriterium Rohscore			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
85.24	56	0.007	83.51
Teilungskriterium Alter			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
48.77	52	0.602	78.61
Teilungskriterium Geschlecht			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
49.77	60	0.824	88.37
Teilungskriterium Muttersprache			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
75.745	60	0.083	88.37

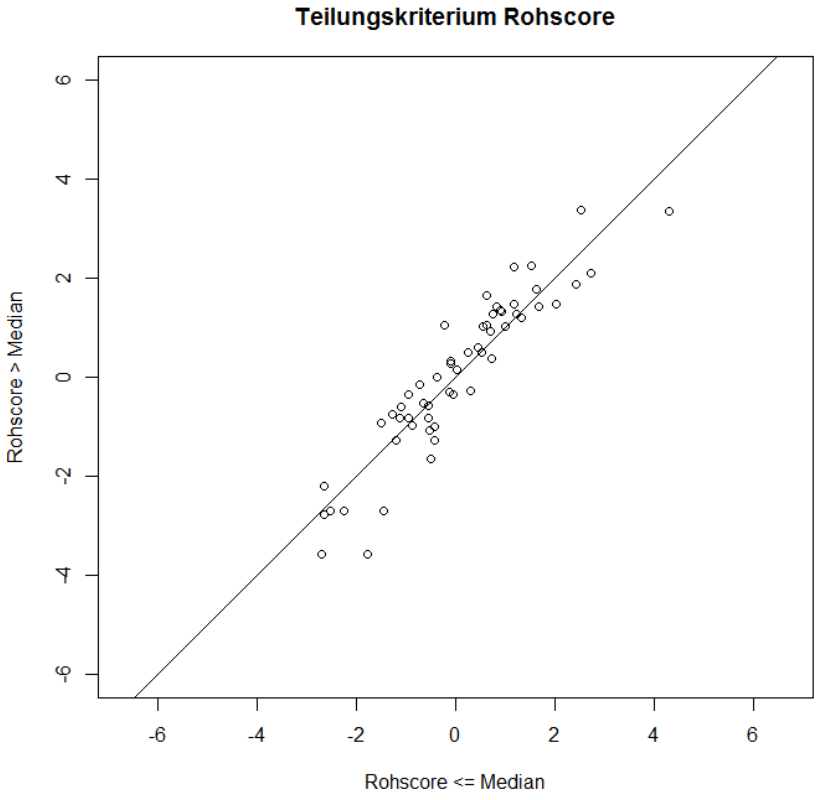

Abbildung 12: Grafischer Modelltest Teilungskriterium Score nach Ausschluss von 27A und 27B

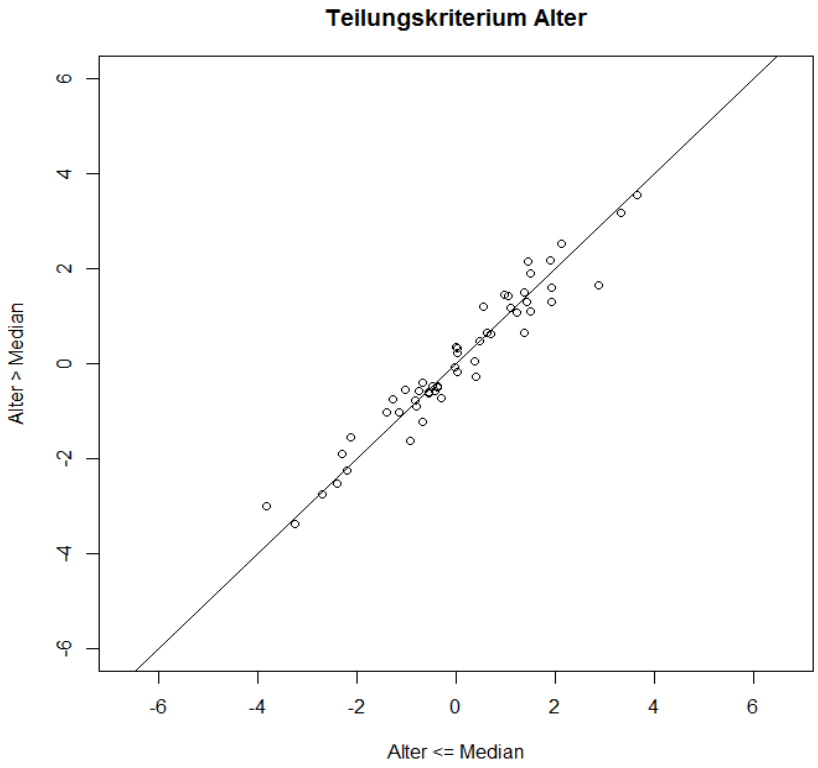

Abbildung 13: Grafischer Modelltest Teilungskriterium Alter nach Ausschluss von 27A und 27B

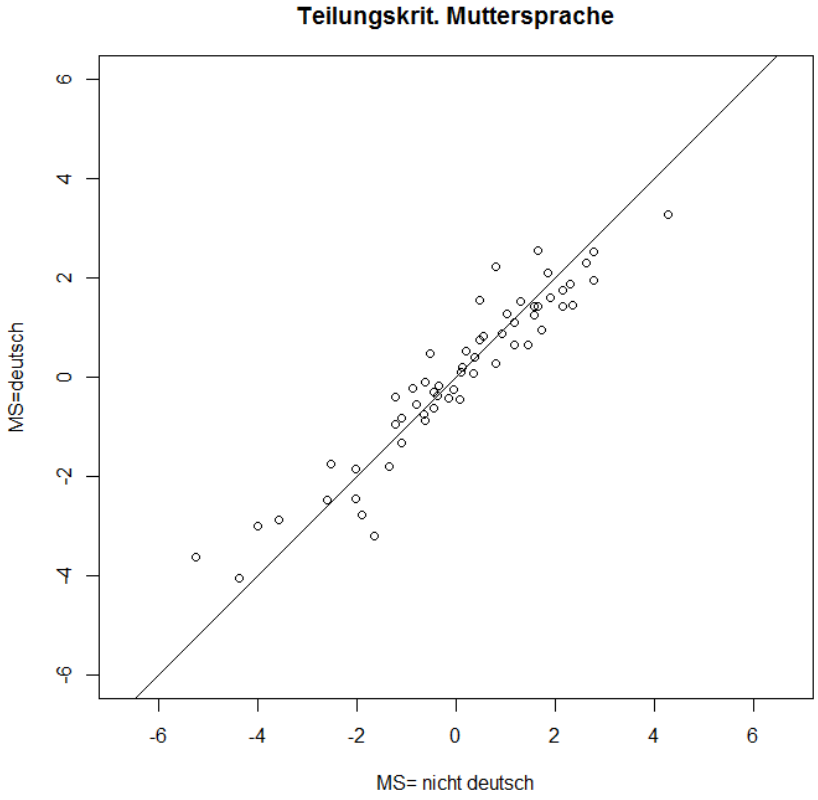

Abbildung 14: Grafischer Modelltest Teilungskriterium Muttersprache nach Ausschluss von 27A und 27B

Tabelle 13: Ergebnisse des z-Tests nach Fischer und Scheiblechner für das Teilungskriterium Rohscore nach Ausschluss der Items 27A und 27B

Item	z-statistic	p-value	Item	z-statistic	p-value	Item	z-statistic	p-value
an92A	0,528	0,598	an09A	1.432	0,152	an18B	0,898	0,369
an92B	-0,12	0,905	an09B	1.759	0,079	an19A	-0,775	0,439
an93A	0,633	0,527	an10A	0,698	0,485	an19B	-1.417	0,157
an93B	0,214	0,83	an10B	0,824	0,41	an20A	2.811	0,005
an94A	2.203	0,028	an11A	1.566	0,117	an20B	1.792	0,073
an94B	1.602	0,109	an11B	1.242	0,214	an21A	0,727	0,467
an01A	-0,853	0,394	an12A	-0,117	0,907	an21B	0,353	0,724
an01B	-0,607	0,544	an12B	0,822	0,411	an22A	1.228	0,219
an03A	0,411	0,681	an13A	-1.655	0,098	an22B	1.386	0,166
an03B	-0,475	0,635	an13B	-1.040	0,298	an23A	-2.002	0,045
an04A	-0,853	0,394	an14A	0,051	0,959	an23B	-0,889	0,374
an04B	-0,254	0,8	an14B	0,108	0,914	an24A	-0,065	0,948
an05A	-1.782	0,075	an15A	0,317	0,751	an24B	-1.414	0,157
an06A	1.402	0,161	an15B	1.352	0,176	an25A	-0,307	0,759
an06B	-0,276	0,783	an16A	-0,584	0,559	an25B	0,347	0,728
an07A	-1.199	0,231	an16B	-1.039	0,299	an26A	1.409	0,159
an07B	-1.804	0,071	an17A	1.506	0,132	an26B	1.652	0,099
an08A	-0,182	0,856	an17B	-0,92	0,358	an28A	-1.208	0,227
an08B	1.661	0,097	an18A	0,948	0,343	an28B	-1.337	0,181

11.3 Ergebnisse der Modellschätzung nach Ausschluss der Items 20A und 20B

Tabelle 14: Ergebnisse der LRTs nach Ausscheiden der Items 20A, 20B, 27A und 27B

Teilungskriterium Rohscore			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
70.761	54	0.063	81.06
Teilungskriterium Alter			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
45.029	50	0.673	76.15
Teilungskriterium Geschlecht			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
48.11	58	0.82	85.95
Teilungskriterium Muttersprache			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
69.042	58	0.152	85.95

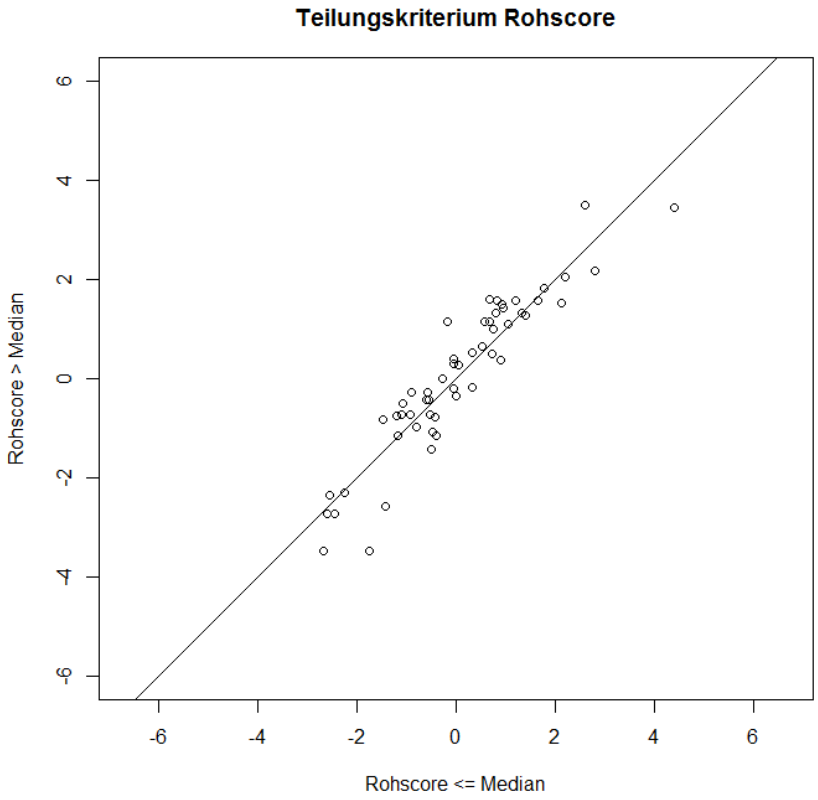

Abbildung 15: Grafischer Modelltest Teilungskriterium Score nach Ausschluss von 20A, 20B, 27A und 27B

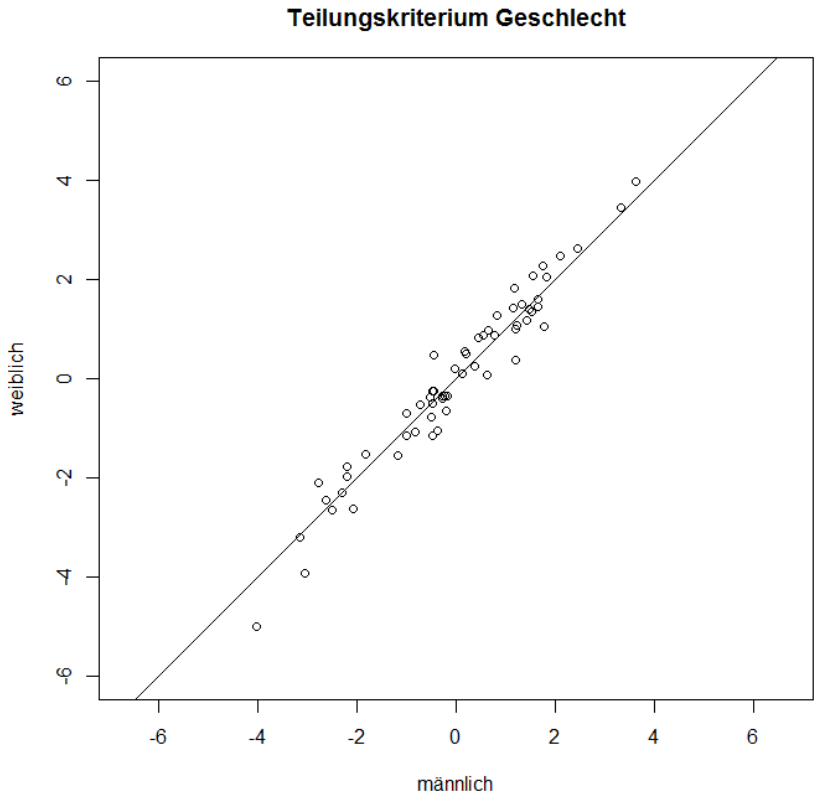

Abbildung 16: Grafischer Modelltest Teilungskriterium Geschlecht nach Ausschluss von 20A, 20B, 27A und 27B

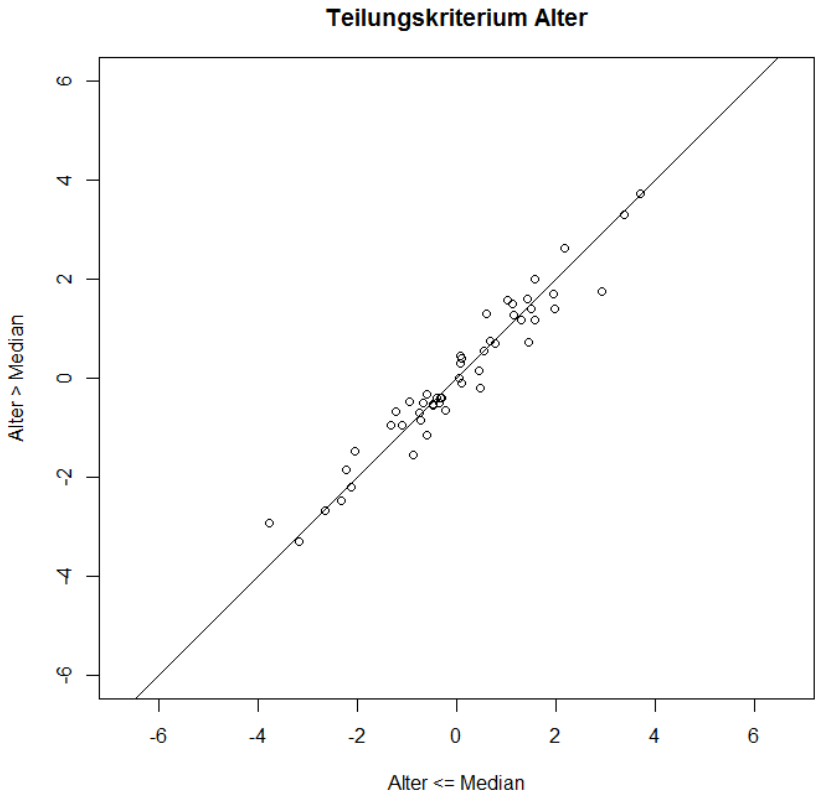

Abbildung 17: Grafischer Modelltest Teilungskriterium Alter nach Ausschluss von 20A, 20B, 27A und 27B

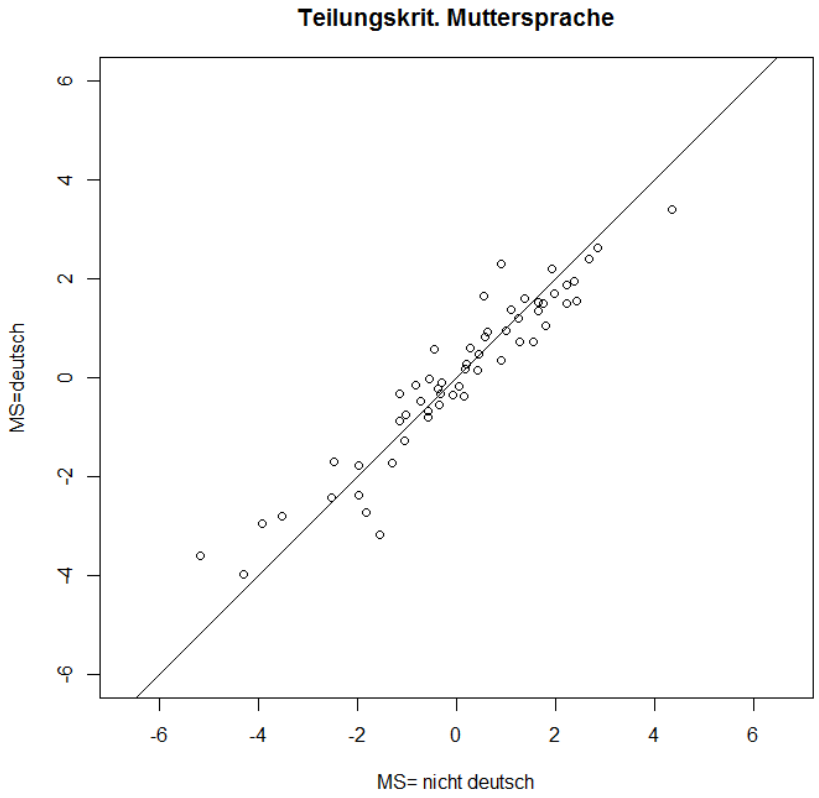

Abbildung 18: Grafischer Modelltest Teilungskriterium Muttersprache nach Ausschluss von 20A, 20B, 27A und 27B

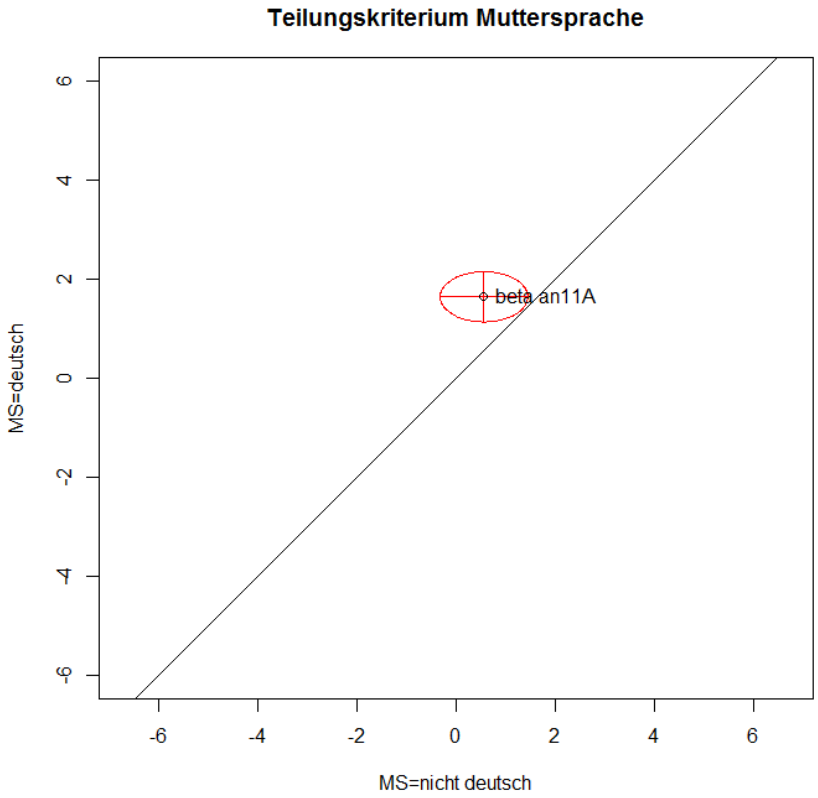

Abbildung 19: Grafischer Modelltest für das Item 11A, Teilungskriterium Muttersprache

11.4 Ergebnisse der Rasch-Modell-Schätzungen nach Ausschluss des Items 11A

Tabelle 15: Ergebnisse des LRTs nach Ausschluss des Items 11A

Teilungskriterium Rohscore			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
62.312	51	0.133	77.39
Teilungskriterium Alter			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
45.185	49	0.629	74.92
Teilungskriterium Geschlecht			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
47.863	57	0.8	84.73
Teilungskriterium Muttersprache			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
60.206	57	0.361	84.73

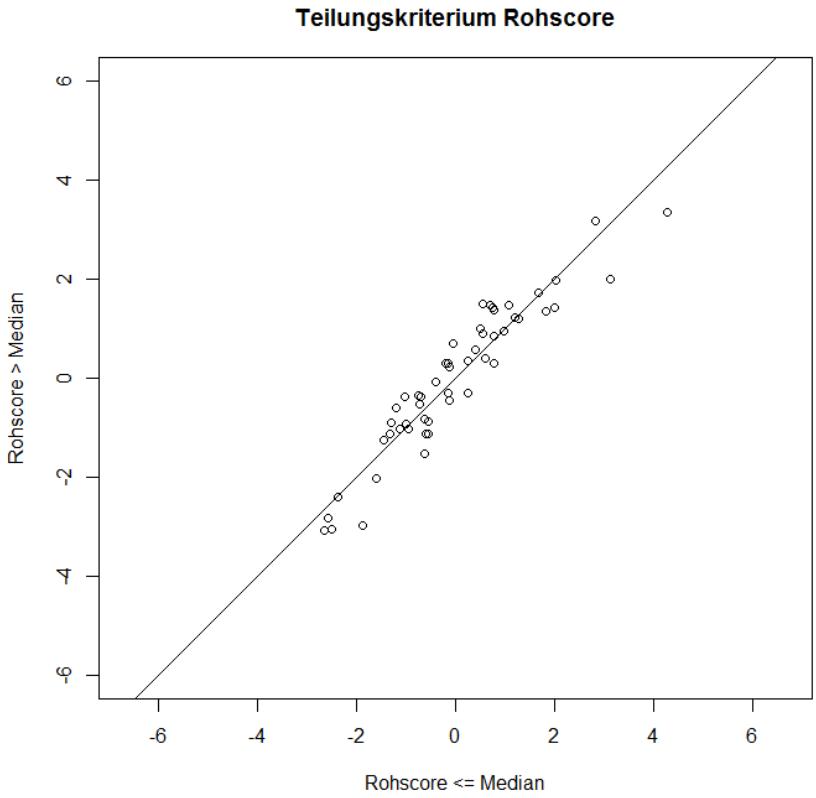

Abbildung 20: Grafischer Modelltest der AN-TOP nach Ausschluss der nicht Modell-konformen Items, TK: Score

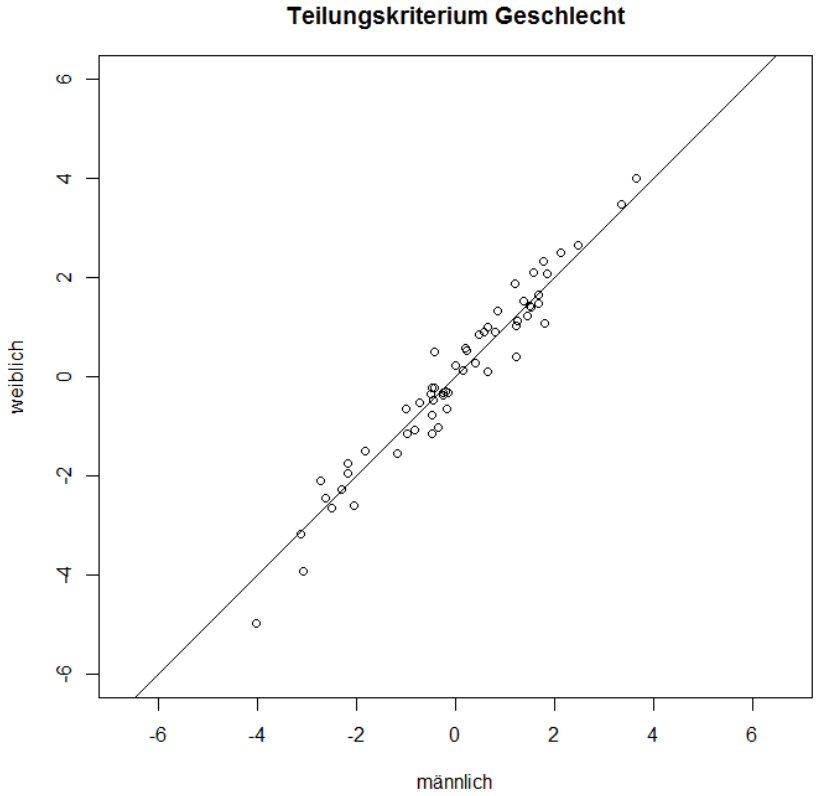

Abbildung 21: Grafischer Modelltest der AN-TOP nach Ausschluss der nicht Modell-konformen Items, TK: Geschlecht

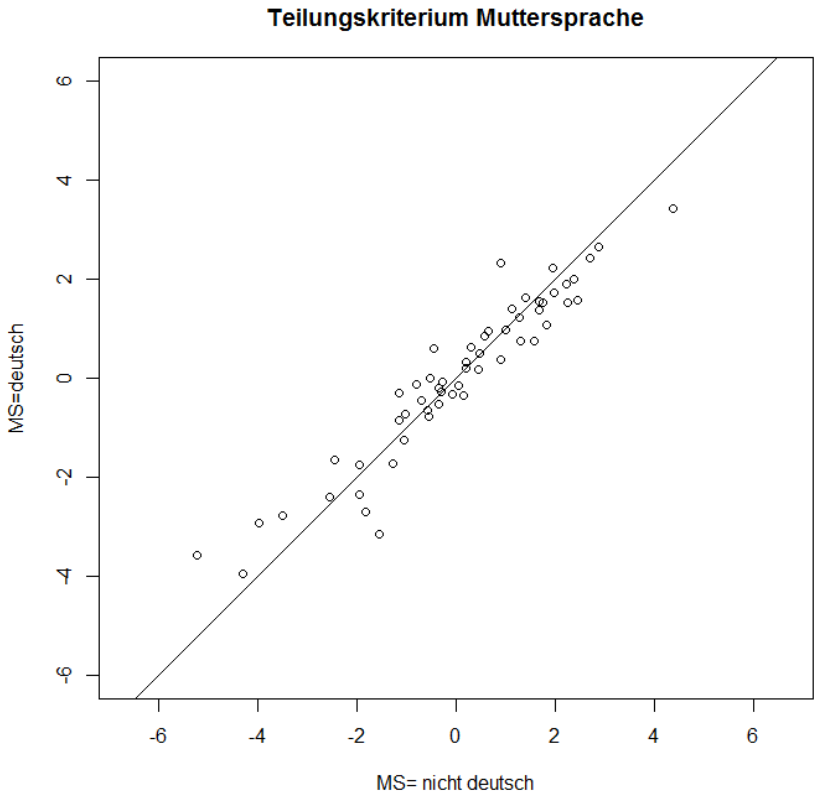

Abbildung 22: Grafischer Modelltest der AN-TOP nach Ausschluss der nicht Modell-konformen Items, TK: Muttersprache

11.5 Ergebnisse der Rasch Modell Schätzung des für den Vergleich mit dem LLTM herangezogenen Itempools

Tabelle 16: Ergebnisse der LRTs der für den Vergleich mit dem LLTM herangezogenen Itempools

Teilungskriterium Rohscore			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
47.687	49	0.526	74.92
Teilungskriterium Alter			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
40.732	46	0.692	71.2
Teilungskriterium Geschlecht			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
45.309	54	0.794	81.06
Teilungskriterium Muttersprache			
Andersen χ^2	*df*	*p* -Wert	Kritischer χ^2- Wert
61.392	54	0.228	81.06

Elternbrief

Nina Heuberger
Fakultät für Psychologie
Inst. für Entwicklungspsychologie & psychologische Diagnostik
Liebiggasse 5
A- 1010 Wien
E-Mail: heuberger.nina@gmail.com

Liebe Eltern und Erziehungsberechtigte,

hiermit möchte ich Sie über eine Studie informieren, die ich im Rahmen meiner Diplomarbeit im Fach Psychologie an der VS1 in Enns am 1. und 3. März 2010 durchführe. In dieser Studie soll die Qualität eines neuen Untertests für die Intelligenztestbatterie AID 2 überprüft werden. Dieser neue Untertest misst schlussfolgerndes Denken durch sprachfreies Material, dadurch sollen Kinder mit Sprachproblemen nicht benachteiligt werden.

Insgesamt werden ungefähr 450 Kinder aus den Schulstufen 3, 5 und 7 in verschiedenen oberösterreichischen Schulen getestet. Die Untersuchungen finden während der Schulzeit statt und dauern ca. 2 Schulstunden. Durchgeführt werden die Testungen im Klassenverband, wobei jedes Kind für sich die Aufgaben bearbeitet. Möchte Ihr Kind die Untersuchung aus irgendeinem Grund frühzeitig abbrechen, ist das natürlich jederzeit möglich.

Die gewonnenen Daten werden im Sinne des Datenschutzes ausschließlich für wissenschaftliche Zwecke genutzt. Die Ergebnisse werden anonym behandelt, weswegen leider keine individuelle Rückmeldung der Ergebnisse möglich ist.

Bei individuellen Rückfragen können Sie mich gerne unter: heuberger.nina@gmail.com kontaktieren.

Mit freundlichen Grüßen,
Nina Heuberger

Lebenslauf

NINA HEUBERGER,

BERUFSERFAHRUNG

11/2015 – jetzt	**Airbus, Toulouse, FRANKREICH** *A330 Completion and Delivery (C&DC) Center Project - Tianjin, China und Toulouse, Frankreich* **HR Business Partner und Recruitment / Expatriation Work Package Leader** In 2015 entschloss sich Airbus gemeinsam mit chinesischen Partnern ein Completion und Delivery Center für die A330 Familie neben der bereits existierenden Endmontagelinie für die A320 Familie in Tianjin, China zu eröffnen. Das C&DC wird ab Oktober 2017 die gesamten Flugzeugfinitionsarbeiten abdecken: Empfang des in Toulouse endmontierten „grünen" Flugzeugs, Kabineninstallation, Lackierung, Motoren- und Flugtests sowie Auslieferung und Kundenabnahme. ▪ HR Business Partner für die europäische Projektorganisation: o Steuerung von personalwirtschaftlichen Prozessen wie Personalbeschaffung und -verwaltung, Arbeitsrecht etc. o Qualifizierung, Coaching und Weiterbildung von Führungskräften und MitarbeiterInnen ▪ Verantwortlich für die Definition der Recruitment und Employment Marketing Strategie ▪ Definition des Recruitment-Prozesses und der Auswahlmethoden ▪ Leitung und Durchführung des operativen Recruitments in Europa und China ▪ Erstellung einer Ressource-Planning Datenbank für die Projekt- und die operationelle Phase für alle am Projekt beteiligten Workstreams unter Berücksichtigung der Trainingspfade der benötigten Ressourcen ▪ Impat / Expat Management: o Koordination der Unterkünfte sowie Sicherstellen der Integration und der lokalen Unterstützung für chinesische Impats während deren Schulung in Europa o Unterstützung der reibungslosen Übergabe europäischer Expats nach China ▪ Zentraler Ansprechpartner auf HR-Seite für die Projektkommunikation in Europa und China ▪ Leiten der zweiwöchentlichen Austauschrunden mit dem Business

01/2013 – 10/2015	**Airbus Group, Toulouse, FRANKREICH** *Airbus Group Employment Operations – Recruitment Center – Global Sourcing Team* **Sourcing Expert für Airbus Defence & Space Frankreich und Deutschland** ▪ **Recruitment (Beschreibung der Aktivitäten siehe Erfahrung von 03/12 bis 12/12)** **Zusätzliche Aufgaben:** ▪ **Lean-Koordinatorin** für das Global Sourcing Team: o Koordination eines **5S-Projektes**: Implementierung und Durchführung von Audits o Implementierung und kontinuierliche Verbesserung eines Visual Management Systems (SQCDP) ▪ Leitung eines Projektes zur **kontinuierlichen Verbesserung**: o Leitung von monatlichen Innovations-Workshops zur Harmonisierung und Verbesserung von der Recruitment-Tools, -Methoden und -Prozesse o Teilnahme an der Entwicklung eines Qualitätsmanagementsystems **(ISO 9001)** ▪ **Kommunikation** zu den Resultaten und Schulung des Teams bezüglich der Neuerungen ▪ Teilnahme an einem Talentförderungsprogramm im Bereich HR: o Leitung eines Projektes zur Verbesserung der Kundenzufriedenheit in einem Service-Kontext o Präsentation der Resultate vor den HR-Führungskräften der Airbus Group ▪ Zentrale Kontaktperson für FEMTEC (deutsche Organisation zur Förderung von Frauen in MINT-Studienfächern und –berufen) ▪ Training von Neuzugängen im Team und Führung von PraktikantInnen

03/2012 – 12/2012	**EXPECTRA on behalf of Airbus Group, Toulouse, FRANKREICH** *Airbus Group Employment Operations – Recruitment Center – Global Sourcing Team*
	Sourcing Expert für Airbus Deutschland ▪ **Recruitment von MitarbeiterInnen in technischen und Support-Funtionen** o Führen von Briefinggesprächen mit einstellenden Führungskräften zur Bestimmung der benötigten Kompetenzen und Qualifikationen o Definition der Sourcing-Strategie für die jeweilige Vakanz o Ausschreiben der Stellen auf geeigneten Job-Boards, internen und externen Karriereseiten und anderen Professional Media (Xing, LinkedIn, etc.) o Suche nach potentiellen KandidatInnen in CV-Theken, auf Professional Media sowie in internen Datenbanken o Community Management o Vorauswahl von geeigneten KandidatInnen und Führen von Telefoninterviews o Auswahl der am besten geeigneten Bewerbung mit der einstellenden Führungskraft ▪ Teilnahme an Karrieremessen und an weiteren Employment Marketing Initiativen ▪ Förderung der internen Mobilität von MitarbeiterInnen und Betreuung von Redeployment-Fällen

09/2011 – 02/2012	**Airbus Group, Toulouse, FRANKREICH** *Airbus Group Employment Operations – Recruitment Center – Global Sourcing Team (Praktikum)*
	▪ Suche und Vorauswahl von KandidatInnen für VIE-Stellen bei Airbus Deutschland ▪ Community Management ▪ Vorauswahl von Bewerbungen für das Talententwicklungsprogramm PROGRESS

12/2010	**AKH Linz, ÖSTERREICH** *Klinische Psychologie (Praktikum)*
	▪ Vorgabe von psychologisch-diagnostischen Verfahren ▪ Animation von Entspannungstrainings für Gruppen ▪ Führen von Anamnesegesprächen

AUSBILDUNG

2005 –	**Universität Wien, ÖSTERREICH** *Diplomstudium Psychologie*
01 - 05/2011	**University of Illinois, Urbana-Champaign, USA** *Auslandssemester*
1997 – 2005	**Akademisches Gymnasium, Linz, ÖSTERREICH** *Matura*

SPRACHEN- UND IT-KENNTNISSE

Sprachen	**Deutsch**	Muttersprache
	Englisch	fließend
	Französisch	fließend

IT Kenntnisse	**MS Office**	Gute Kenntnisse
	SPSS und R	Erweiterte Kenntnisse
	SAP R/3	Erweiterte Nutzerkenntnisse
	Weitere	▪ Tägliche professionelle Nutzung von Social und Professional Media sowie Jobboards im deutschsprachigen und internationalen Raum (LinkedIn, Xing, Monster, Stepstone, Viadeo, Apec, etc....) ▪ Erweiterte Kenntnisse in der Online Recherche inklusive wissenschaftlicher Datenbanken