universität wien

Diplomarbeit

Titel der Arbeit

Konstruktion des sprachlichen Untertests „Antonyme finden“ für die Intelligenztestbatterie AID 3

Verfasser

Benjamin Weber

Angestrebter akademischer Grad

Magister der Naturwissenschaften (Mag. rer. nat.)

Wien, im Februar 2011

Studienkennzahl:	A 298
Studienrichtung:	Psychologie
Betreuerin:	Mag. Dr. Stefana Holocher-Ertl

Abstract

Die Zielsetzung dieser Arbeit besteht in der Konstruktion eines sprachlichen Untertests (*Antonyme finden*) für die Intelligenztestbatterie AID 3. Der AID 3 ist die dritte Generation der erstmals 1985 erschienenen Testbatterie AID (Adaptives Intelligenz Diagnostikum, Kubinger & Wurst, 1985) zur Erfassung komplexer und basaler Kognitionen (Intelligenz) bei Kindern und Jugendlichen. Um den diagnostischen Informationswert in Bezug auf das *elementare Sprachverständnis* eines Kindes bzw. Jugendlichen zu erhöhen, ist im AID 3 neben dem Untertest *Synonyme finden* die Vorgabe des Untertests *Antonyme finden* vorgesehen. Im Zuge der Itemkonstruktion wurden 67 Aufgaben entwickelt, die einer Stichprobe von 711 Schüler(innen) vorgegeben wurde. In einer anschließenden Datenanalyse wurde der Untertest *Antonyme finden* auf die Geltung des Rasch-Modells überprüft, um Aussagen über die Gütekriterien Skalierung und Fairness treffen zu können. Nach Ausschluss von 8 Items konnte a posteriori die Gültigkeit des Rasch-Modells für die restlichen Items angenommen werden. Die Summe aller gelösten Items ist im Sinne des Gütekriteriums Skalierung somit ein faires Maß für die erbrachte Testleistung. Positiv zu werten ist, dass die Items des Untertests *Antonyme finden* einen sehr breiten Fähigkeitsbereich gleichmäßig abdecken. Ferner benachteiligt der Untertest keine Personen aufgrund ihrer Geschlechtszugehörigkeit. Um festzustellen, ob der Subtest *Antonyme finden* auch das misst, was er zu messen beansprucht, wurde eine konvergente Validität mit dem Untertest *Synonyme finden* berechnet, der ebenfalls das elementare Sprachverständnis prüft. Es stellte sich ein hoher Zusammenhang der Testleistung in beiden Untertests heraus. Bislang ungeklärt bleibt die Frage, ob durch die Testwerte des Untertest *Antonyme finden* Kinder mit nicht deutscher Muttersprache benachteiligt werden. Es scheint daher wünschenswert, für den AID 3 eine türkische sowie eine bosnisch/kroatisch/serbische Version zu erstellen, um Kinder mit dementsprechender Muttersprache optimal fair diagnostizieren zu können.

Abstract – English

The aim of this study was to develop a language subtest (*finding antonyms*) as part of the AID 3 test battery of intelligence. The AID 3 is the third generation of the AID test battery (Adaptives Intelligenz Diagnostikum, Kubinger & Wurst, 1985) and was first published in 1985. Its aim is to assess complex and basic cognition (intelligence) of children and adolescents. To improve the diagnostic value of information in the area of *elementary language understanding* of children and adolescents, the AID 3 suggests not only using the subtest *finding synonyms* but also the subtest *finding antonyms*. The process of item construction succeeded in developing 67 items that were then tested among 711 school students. The subsequent analysis of data for the subtest *finding antonyms* was tested using the Rasch model to enable statements concerning the quality criteria for scaling and fairness. After having eliminated 8 items, a validity of the Rasch model could be assumed for the remaining items. The sum of all items solved is consequently in relation to the criteria of scaling a fair measure for the test performance. A positive aspect of the subtest *finding antonyms* is the fact that its items cover a wide range of ability levels. Furthermore it can be said that the subtest does not discriminate tested people because of their *gender*. To prove whether the subtest *finding antonyms* measures what it claims to assess, convergent validity was analysed in relation to another subtest, *finding synonyms*, which also assesses *elementary understanding of language*. A high correlation of the performance in these two subtests could be proved. What remains unexplained at this point is whether the results of the subtest *finding antonyms* discriminates children who do not have German as their mother tongue. It therefore seems useful to develop AID 3 versions in Turkish, and Bosnian / Croatian / Serbian to ensure a fair assessment of children with these languages as their mother tongue.

Danksagung

An erster Stelle möchte ich meinen Eltern danken, die mich während meiner ganzen Studienzeit sowohl emotional als auch in finanzieller Hinsicht immer unterstützt haben und mir dadurch mein Studium überhaupt erst ermöglicht haben.

Großer Dank gilt natürlich meiner Betreuerin Frau Dr. Stefana Holocher-Ertl für ihre wertschätzende und fachlich hochwertige Betreuung während der gesamten Diplomarbeitsphase.

Weiters gilt ein großer Dank meinen vielen Helferleins, die mich in verschiedenen Phasen meiner Diplomarbeit unterstützt haben. Hierbei sei Nina Heuberger besonders für ihre große Hilfe bei den Rasch-Modell-Analysen gedankt.

Auch Jan Steinfeld sei an dieser Stelle für seine schnelle Hilfe bei diversen statistischen Problemen Dank ausgesprochen.

Ebenso danke ich besonders Lisa Janschek und Lara Pivodic für das Korrekturlesen meiner Diplomarbeit.

Dank gebührt auch meiner Cousine Veronika Bukovec und meiner lieben Freundin Angelika Längle für ihre Hilfestellungen im Endspurt der Diplomarbeit.

Ein großes Dankeschön gilt auch meinen Kolleginnen des AID 3 – Diplomand(innen)-Teams für die gute Zusammenarbeit.

Auch Mag. Doris Fleck danke ich für die enorme Unterstützung, ohne die die Testungen am BRG 9, Glasergasse nicht möglich gewesen wären.

Zu guter Letzt danke ich besonders meinen Freunden und meiner Familie, die während meiner gesamten Studienzeit immer für mich da waren und sind.

Inhaltsverzeichnis

I Einleitung

Die intellektuellen Fähigkeiten eines Kindes müssen immer im zeitlichen und gesellschaftlichen Kontext betrachtet werden. So waren zur Messung der intellektuellen Fähigkeiten eines Kindes bei der Veröffentlichung der Intelligenztestbatterie AID (Adaptives Intelligenz Diagnostikum, Kubinger & Wurst) im Jahr 1985 andere Wissensinhalte relevant als bei der Revision im Jahr 2000 (AID 2). Aufgaben in Intelligenztests müssen somit von Zeit zu Zeit aktualisiert werden, damit sie auch wirklich jene intelligenzbezogenen Fähigkeiten und Wissen messen, welche in der heutigen Zeit relevant sind. So müssen beispielsweise sprachliche Untertests dem heutigen Sprachgebrauch angepasst werden. Wörter, die in den 80er-Jahren verwendet wurden, sind heutzutage teilweise nicht mehr im täglichen Sprachgebrauch zu finden. Ebenso müssen die geografischen, politischen oder wissenschaftlichen Veränderungen der letzten Jahrzehnte bei Aufgabenbereichen angepasst werden, die alltägliches Wissen oder das Verstehen von gesellschaftlichen Zusammenhängen zu messen beanspruchen.

Neben der Aktualisierung der Aufgabeninhalte muss auch die Kritik von Psycholog(innen) aus der Praxis ernst genommen und entsprechend bei der Konzeption einer weiteren Revision miteinbezogen werden. So kam beispielsweise die Rückmeldung, dass einige Untertests in einigen Fähigkeits- oder Altersbereichen zu wenig differenzieren, wodurch die Messgenauigkeit beeinträchtigt wird.

Obwohl der AID 2 in einer 2. Version 2009 neu normiert wurde (Kubinger, 2009a), sind seit 2000 nur minimale inhaltliche Veränderungen vorgenommen worden. Aus diesem Grund wurde im Sommer 2009 das Projekt AID 3 gestartet. Neben Aktualisierungen der Untertests wurden auch einige neue Untertests konzipiert, die bisher nicht beachtete Dimensionen messen oder ergänzen sollten. Ein Beispiel ist der Untertest *Antonyme finden* zur Erfassung von Sprachlogik und Wortschatz, welcher Inhalt und Zielsetzung dieser Diplomarbeit ist.

Die Arbeit ist in einen theoretischen und empirischen Teil gegliedert. Der **theoretische Teil** befasst sich zunächst mit der Frage, wie sprachliche Intelligenz im Kinder- und Jugendalter erfasst werden kann. Weiters erfolgt eine detaillierte Beschreibung der Testbatterie AID 2 sowie eine Auseinandersetzung mit dem Begriff der *Antonymie* aus sprachwissenschaftlicher Sicht. Schließlich werden Methoden und Modelle der *Item-Response-Theorie* vorgestellt, die sowohl für die Testkonstruktion als auch für die Datenanalyse von hoher Relevanz sind. Im **empirischen Teil** wird zunächst der Testkonstruktionsprozess samt theoretischem

Hintergrund beschrieben. Anschließend erfolgt eine Diskussion zu den Gütekriterien des Untertests *Antonyme finden.* Im Kapitel *Methoden* wird der Untersuchungsplan einschließlich der Hypothesen vorgestellt. Darauf folgen eine Beschreibung der Stichprobe sowie die Ergebnisse der Datenanalyse. In der *Diskussion* werden schließlich die positiven Aspekte und Mängel der empirischen Arbeit beleuchtet.

II Theoretischer Teil

1 Adaptives Intelligenz Diagnostikum 2 (AID 2)

Der Untertest *Antonyme finden* ist als sprachlicher Untertest der Intelligenztestbatterie AID 3 vorgesehen. Im folgenden Kapitel wird die aktuelle Version des Verfahrens (AID 2.2) genau beschrieben.

Das Adaptive Intelligenz Diagnostikum 2 (Kubinger und Wurst, 2000) ist eine Intelligenztestbatterie für Kinder und Jugendliche im Alter von 6 bis 15 Jahren zur Erfassung komplexer und basaler Kognitionen (*Intelligenz*) (Kubinger, 2009a, S.2). Erstmals ist die Testbatterie im Jahre 1985 unter dem Namen **AID** erschienen, 2000 kam eine inhaltlich überarbeitete sowie neu normierte zweite Version als **AID 2** auf den Markt. Da die DIN 33430 (Norm zur berufsbezogene Eignungsbeurteilung, siehe dazu Westhoff et. al, 2004) die Forderung stellt, einschlägige Verfahren spätestens alle 8 Jahre einer neuen Eichung zu unterziehen, wurde eine 2. neu geeichte Auflage (AID 2, Version 2.2) kürzlich publiziert (Kubinger, 2009a).

Der AID 2 (Version 2.2) besteht aus 11 Untertests und 3 Zusatztests, die verschiedene Aspekte intellektueller Fähigkeiten zu erfassen versuchen. Intelligenz wird im AID 2 als „die Gesamtheit aller kognitiven Voraussetzungen, die notwendig sind, um Wissen zu erwerben und Handlungskompetenzen zu entwickeln“ definiert (Kubinger, 2009a, S.23). Inhaltlich ist der AID am Testkonzept von David Wechsler orientiert. Die Untertests sind denen der Intelligenztestbatterie HAWIK (Hamburg-Wechsler Intelligenztest für Kinder, aktuellste Version HAWIK-IV, Petermann & Petermann, 2007) thematisch ähnlich, unterscheiden sich aber hinsichtlich ihrer Testkonzeption deutlich. Der AID 2 ist nach Methoden der *Item-Response-Theorie* (siehe Kapitel 4) konstruiert, die eine *adaptive*[1] Testvorgabe ermöglicht. Anders als bei anderen Intelligenztests werden einem Kind im AID 2 nur diejenigen Aufgaben vorgegeben, die dem individuellen Leistungsniveau entsprechen. Dies ermöglicht eine hohe testökonomische Vorgehensweise, da die Messgenauigkeit trotz geringerer Aufgabenanzahl im Vergleich zu Verfahren mit konventioneller Testvorgabe gleich hoch bleibt. Zusätzlich kann die Motivation des Kindes aufrechterhalten werden, da es durch die adaptive Testvorgabe keine Aufgaben bearbeiten muss, die ihm zu leicht oder zu schwer fallen (Kubinger, 2009a, 2009b).

Der AID 2 ist ein Individualverfahren, das Kind wird daher alleine und nicht in der Gruppe getestet. Die Untertests konnten mit wenigen Ausnahmen als reine *power-Tests* konzipiert werden. Die Einzelvorgabe ermöglicht weiters die Verwendung des *freien Antwortformats*[2].

Die meisten Intelligenztests im Kinder- als auch im Erwachsenenbereich sehen die Berechnung eines Intelligenzquotienten (*IQ*) vor, definiert als globales Maß für die intellektuelle Leistungsfähigkeit einer Person (Häcker & Stapf , 2004). Die Autoren des AID propagieren hingegen einen förderungsorientierten Ansatz. Anstatt einen Gesamtwert zu interpretieren, der Aufschluss über die globalen Fähigkeiten eines Kindes geben soll, ist im AID eine detaillierte *Profilinterpretation* in Bezug auf die einzelnen Testwerte je Untertest vorgesehen. Dadurch können Leistungsstärken und relative Schwächen eines Kindes identifiziert werden. Fakultativ (da von vielen Eltern oft gewünscht) kann als globales Leistungsmaß die sog. *Intelligenzquantität* sowie der *Range* der Intelligenz berechnet werden. Die Intelligenzquantität, zu interpretieren als *kognitive Mindestfähigkeit*, ergibt sich aus der niedrigsten Untertestleistung. Der *Range* beschreibt die Streuung der Testleistungen als *Grad der Differenziertheit* der Fähigkeit eines Kindes. Eine genaue Profilinterpretation ist der alleinigen Berechnung der Intelligenzquantität und des Ranges allerdings eindeutig vorzuziehen.
Die Vorgabe des AID 2 ermöglicht ferner ein Screening zur Erfassung bestimmter Teilleistungsschwächen wie bspw. visumotorische Störungen. Der interessierte Leser sei auf Leiss (2003) verwiesen.

Zur Beurteilung des *Arbeitshaltungen* der Testperson ist im Protokollbogen des AID 2 ein Beiblatt enthalten, das dem/der Testleiter(in) helfen soll, das *Arbeits- und Kontaktverhalten* des Kindes in einer Leistungssituation zu beurteilen (Kubinger, 2009b).

Die Anwendungsmöglichkeiten des AID 2 sind vielfältig. So kommt der Test neben der Anwendung in der Entwicklungsdiagnostik auch in der neuropsychologischen Diagnostik und Berufs- und Bildungsberatung zum Einsatz. Für die Schulpsychologie ist das Verfahren unter anderem interessant, da mit dem AID 2–Türkisch Kinder mit Türkisch als Muttersprache *fair* getestet werden können (Kubinger, 2009a, 2009b). Der AID 2 ist auch bei der Abklärung von Hochbegabung einsetzbar. Speziell für die förderungsorientierte Diagnostik von

---

[1] Die Form des *adaptiven* Testens wird im Kapitel 6.6 genauer erklärt.
[2] Für eine umfangreiche Erklärung der *Speed-Power-Problematik* und des *freien Antwortformats* sei auf Abschnitt 6.1 verwiesen.

Hochbegabung in Anlehnung an das *Wiener Diagnosemodell zum Hochleistungspotential* ist der AID 2 ein sehr geeignetes Verfahren (Holocher-Ertl, Kubinger & Hohensinn, 2008).

1.1 Die Subtests des AID 2

Die 11 Skalen (Untertests) und 3 Zusatztests des AID 2 lassen sich in Aufgabengruppen einteilen, die sowohl „*manuell-visuelle*" als auch „*verbal-akustische*" Fähigkeiten erfassen. Während Aufgaben der erstgenannten Gruppe *visuelles Erfassen und manuelles Agieren* erfordern, muss das Kind bei Aufgaben der zweiten Gruppe Information *akustisch erfassen* und damit *verbal agieren*.

Alle Untertests des AID 2 beruhen auf *operationalen Definitionen*[3], welche die gemessenen Fähigkeiten jeweils festlegen. Im folgenden Abschnitt wird jeder Untertest des AID 2 einzeln beschrieben[4]. Untertests, die *manuell-visuelle* Fähigkeiten erfassen, werden mit (M) versehen, jene, die *verbal-akustische* Fähigkeiten messen, mit (V).

1) **Alltagswissen (V):** *Der Untertest Alltagswissen soll die Fähigkeit prüfen, sich Sachkenntnisse über Inhalte anzueignen, die in der heutigen Gesellschaft alltäglich sind.*
   Der Testperson werden Wissensfragen gestellt, die sie mündlich beantworten muss.

2) **Realitätssicherheit (M):** *Der Untertest Realitätssicherheit soll prüfen, inwieweit die Wirklichkeit um Dinge des Alltags verstanden wird, bzw. kontrolliert werden kann.*
   Die Testperson soll auf Bildkarten ein fehlendes Detail entdecken.

3) **Angewandtes Rechnen (V):** *Der Untertest Angewandtes Rechnen soll weitgehend unabhängig von schulischen Rechenfertigkeiten prüfen, inwieweit die Testperson bei der Problemlösung alltäglicher Aufgabenstellungen durch entsprechende Schlussfolgerungen die passenden Rechenoperationen anzuwenden imstande ist.*
   Die Testperson bekommt Textrechenaufgaben vorgegeben, die sie lösen muss.

---

[3] Eine „operationale Definition" beschreibt eine Variable lediglich dadurch, dass sie die Operation festlegt, mit Hilfe derer man diese Variable messen kann." (Rost, 2004, S.22)

[4] Die Definitionen des Untertests sind dem Testmanual des AID 2 (Version 2.2) (Kubinger, 2009a, S. 9-13) entnommen.

4) **Soziale und Sachliche Folgerichtigkeit (M):** *Mit dem Untertest Soziale und Sachliche Folgerichtigkeit soll die Fähigkeit erfasst werden, die Abfolge sozialen Geschehens bzw. alltäglicher Sachgegebenheiten zu verstehen und zu kontrollieren.*
   Die Testperson soll ungeordnete Bildfolgen verschiedener Geschichten in eine logische Reihung bringen.

5) **Unmittelbares Reproduzieren** –numerisch **(V):** *Der Untertest Unmittelbares Reproduzieren –numerisch soll die Kapazität der seriellen Informationsverarbeitung (im verbal-akustischen Bereich) messen.*
   Der Testperson werden Zahlenreihen vorgesagt, welche sie zunächst „vorwärts" und anschließend „rückwärts" wiedergeben soll.

6) **Synonyme finden (V):** *Der Untertest Synonyme finden soll das elementare Sprachverständnis prüfen, nämlich inwieweit die Testperson die Bedeutung sprachgebundener Begriffe erfasst bzw. über einen Wortschatz verfügt, der solche Begriffe alternativ ausdrücken lässt*.
   Der Testperson werden mündlich Wörter vorgegeben, für die sie jeweils ein anderes Wort mit derselben Bedeutung finden muss.

7) **Kodieren und Assoziieren (M):** *Mit dem Untertest Kodieren und Assoziieren sollen zwei voneinander partiell unabhängige Fähigkeiten erfasst werden: Die Informationsverarbeitungsschnelligkeit und die Fähigkeit zum inzidentellen Lernen.*
   Die Testperson muss zu Objekten auf einem Arbeitsblatt die passenden Symbole aus einer Vorlage abzeichnen und sie in einem zweiten Schritt ohne Zuhilfenahme der Vorlage wiedergeben.

8) **Antizipieren und Kombinieren** -figural **(M):** *Der Untertest Antizipieren und Kombinieren –figural soll schlussfolgerndes Denken in der Hinsicht prüfen, Teile eines (konkreten) Ganzen erkennen und dieses Ganze gestalten zu können.*
   Die Testperson muss die Teile einer Figur zusammensetzen.

9) **Funktionen abstrahieren (V):** *Mit dem Untertest Funktionen abstrahieren soll die Fähigkeit erfasst werden, durch Abstraktion zu einer Begriffsbildung zu gelangen*.
   Die Testperson soll aus zwei Begriffen die gemeinsame Funktion erschließen.

10) **Analysieren und Synthetisieren** –abstrakt **(M):** *Der Untertest Analysieren und Synthetisieren –abstrakt soll die Fähigkeit prüfen, komplexe (abstrakte) Gestalten durch eine geeignete Strukturierung reproduzieren zu können.*

Die Testperson soll mithilfe von Würfeln, die unterschiedliche Seiten aufweisen, ein geometrisches Muster nachbauen.

11) **Soziales Erfassen und Sachliches Reflektieren (V):** *Mit dem Untertest Soziales Erfassen und Sachliches Reflektieren soll geprüft werden, inwieweit die Testperson Sachzusammenhänge der „gesellschaftlichen" Umwelt begreift bzw. inwieweit sie sozialisiert in dem Sinne ist, dass sie über sozial angepasste Verhaltensweisen und gesellschaftliche Bedingungen Bescheid weiß.*

Der Testperson werden Fragen zu den eben beschriebenen Inhalten gestellt.

1.1.1 Optionale Zusatztests

Die Zusatztests können bei spezifischen Fragestellungen vorgegeben werden. Vor allem im Zusammenhang mit dem Screening von Teilleistungsstörungen sind sie besonders relevant (Preusche & Leiss, 2003).

5b) **Unmittelbares Reproduzieren** –figural/abstrakt: *Der Zusatztest Unmittelbares Reproduzieren–figural/abstrakt soll die Kapazität der seriellen Informationsverarbeitung (im visumotorischen Bereich) messen.*

Der/die Testleiter(in) tippt Bilder einer Bildertafel in einer bestimmten Reihenfolge an. Die Testperson soll dies in derselben Reihenfolge nachmachen.

5b) **Merken und Einprägen:** *Mit dem Zusatztest Merken und Einprägen soll die Behaltenskapazität erfasst werden, wie sie durch eine einmalige Wiederholung der Reizdarbietung erreichbar ist.*

Die Testperson soll Wortlisten mit sinnfreien Silben nachsprechen.

10a) **Strukturieren** –visumotorisch: *Der Zusatztest Strukturieren –visumotorisch soll die Fähigkeit erfassen, komplexe (abstrakte) Gestalten in elementare Teilkomponenten zerlegen zu können.*

Die Testperson soll geometrische Muster durch das Zeichnen von Linien in die verschiedenen Seiten eines Würfels einteilen.

2 Die Erfassung sprachlicher Intelligenz

Die *sprachliche* oder *verbale Intelligenz* ist ein zentraler Bestandteil vieler Intelligenzmodelle. Sie kann durch verschiedenste Intelligenztests erfasst werden. Sprachliche Intelligenz wird je nach Anlehnung an verschiedene Intelligenzmodelle in Intelligenztests unterschiedlich operationalisiert[5] (eine genaue Darstellung verschiedener Intelligenzmodelle liefern Amelang, Bartussek, Stemmler & Hagemann, 2006). Dieses Kapitel befasst sich mit der Frage, wie sprachliche Fähigkeiten oder verbale Intelligenz im *Kinder- und Jugendalter* gemessen oder erfasst werden können. Dabei werden jene Verfahren genauer beschrieben, die im deutschsprachigen Raum im Rahmen psychologischen Diagnostizierens am meisten Anwendung finden (Kastner-Koller, pers. Mitteilung, 17.01.2011).[6] Die Darstellung beschränkt sich auf Tests, die für den Altersbereich gelten, für den auch der AID 2 konzipiert ist (6-15 Jahre), da die Verfahren sonst schlecht miteinander verglichen werden können.

Der AID 2 wurde bereits im Kapitel 1 genau dargestellt und wird daher erst in Abschnitt 2.5 mit den anderen Verfahren in Bezug auf Gemeinsamkeiten und Unterschiede bei der Erfassung sprachlicher Fähigkeiten verglichen.

2.1 HAWIK-IV (Hamburg-Wechsel-Intelligenztest für Kinder – IV)

Der HAWIK-IV (Petermann & Petermann, 2007) ist eine Intelligenztestbatterie zur Erfassung allgemeiner und spezifischer intellektueller Fähigkeiten bei Kindern von 6 bis 16 Jahren. Der HAWIK-IV ist ein Individual-Verfahren und besteht aus 10 Untertests und 5 Zusatztests, die den 4 Indizes *Sprachverständnis*, *wahrnehmungsgebundenes logisches Denken*, *Arbeitsgedächtnis* und *Verarbeitungsgeschwindigkeit* zugeordnet werden können. Die Vorgabe der einzelnen Untertests beruht auf einer *konventionellen Strategie*. Jeder Testperson werden somit prinzipiell alle Aufgaben eines Untertests vorgegeben (im Gegensatz zur *adaptiven* Testvorgabe des AID 2), bis ein definiertes Abbruchkriterium erreicht ist. Der HAWIK-IV besteht aus *Power* sowie *Power-Speed*-Tests und verwendet ein *freies*

[5] Operationalisierung bedeutet, dass eine nicht direkt beobachtbare Variable (sprachliche Intelligenz) für die Beobachtung bzw. für die experimentelle Manipulation zugänglich gemacht werden kann. Es geht somit darum, wie man ein theoretisches Konstrukt *messbar* machen kann (frei nach Häcker & Stapf, 2004).

[6] Als Expertin wurde Ass.-Prof. Dr. Ursula Kastner-Koller, Leiterin des Arbeitskreises „Erziehungsberatung" der Test- und Beratungsstelle der Universität Wien herangezogen, die jene diagnostischen Verfahren zur Erfassung sprachlicher Fähigkeiten nannte, die im deutschsprachigen Raum zur Abklärung der sprachlichen Intelligenz am häufigsten zum Einsatz kommen.

Antwortformat[7]. Bezüglich der Gütekriterien werden dem HAWIK-IV eine zufriedenstellend hohe Messgenauigkeit sowie eine annähernde Konstruktvalidierung attestiert. Die Normtabellen sind für den deutschsprachigen Raum weitgehend repräsentativ (Kubinger, 2009b). Der HAWIK-IV sieht die Berechnung eines Gesamt-IQ vor. Da sich das Kapitel mit der Erfassung der sprachlichen Intelligenz beschäftigt, wird der Fokus auf den Index *Sprachverständnis* gelegt.

Der Index *Sprachverständnis* misst die *sprachliche Begriffsbildung*, das *sprachliche Schlussfolgern* und das *erworbene Wissen*. Er besteht aus den Kernuntertests „Gemeinsamkeiten finden", „Wortschatz-Test" und „Allgemeines Verständnis" sowie den Optionalen Zusatztests „Allgemeines Wissen" und „Begriffe erkennen" angegeben (Kastner-Koller & Deimann, 2008; Petermann & Petermann, 2007; Preusche & Leiss, 2003). In Tabelle 1 werden die Kern-Untertests sowie die optionalen Zusatztests (kursiv) beschrieben und der jeweils gemessene Fähigkeitsbereich angegeben.

Tabelle 1: *Beschreibung der Untertests des Index Sprachverständnis (HAWIK-IV)*

Untertest	**Beschreibung der Aufgabe**	**Gemessene Fähigkeit**
Gemeinsamkeiten finden	Die Testperson soll die Gemeinsamkeit von einem Begriffspaar nennen.	Verbales Schlussfolgern; sprachliche Konzeptbildung
Wortschatztest	Das Kind soll eine verbale Definition zu einem vorgelesenen Wort oder gezeigten Bild geben.	Umfang des Wortschatzes, Stand der Sprachentwicklung;
Allgemeines Verständnis	Die Testperson muss Fragen zu alltäglichen Problemen und sozialen Situationen oder Regeln beantworten.	Praktisches Urteilsvermögen, Kenntnis sozialer Regeln und ihrer Bedeutung;
Allgemeines Wissen	Das Kind muss Wissensfragen beantworten.	Breite des erworbenen Wissens („kristalline Intelligenz");
Begriffe erkennen	Dem Kind werden Hinweissätze vorgelesen, aus denen ein Begriff erschlossen werden soll.	Allgemeines Schlussfolgern, verbale Abstraktion, Bereichswissen;

[7] Genauere Erklärungen zum Testkonzept eines Tests sowie zu den Begriffen *freies Antwortformat* sowie *Power-Speed-Tests* werden in Abschnitt 6 gegeben.

2.2 K-ABC (Kaufman – Assessment Battery for Children)

Die K-ABC (Melchers & Preuß, 2009) ist ein Individual-Verfahren zur Messung von Intelligenz und spezifischen Fertigkeiten für Kinder im Alter von 2;6 bis 12;5 Jahren. Er kann somit auch zur Beurteilung intellektueller Fähigkeiten von Vorschulkindern herangezogen werden. Intelligenz wird im K-ABC als *Fähigkeit, wie ein Individuum Probleme löst und Informationen verarbeitet*, definiert (Melcher & Preuß, 2009). Ziel war eine klare Unterscheidung zwischen angeeignetem Wissen und intellektuellen Fähigkeiten. Der K-ABC besteht aus 16 Untertests, von denen je nach Alter maximal 13 durchgeführt werden. Die Untertests werden 4 übergeordneten Skalen zugeordnet. Die *Skala einzelheitlichen Denkens* sowie die *Skala ganzheitlichen Denkens* werden zur „*Skala intellektueller Fertigkeiten"* zusammengefasst und bilden das Maß für die Gesamtintelligenz. Die zwei weiteren Skalen bestehen aus der *Fertigkeitenskala* und der *sprachfreien Skala*. Die Untertests des K-ABC werden nach der *konventionellen* Strategie vorgegeben und sehen ein *freies Antwortformat* vor. Der Test besteht wie der HAWIK-IV aus *Power* sowie *Power-Speed*-Tests. Bei Betrachtung der Gütekriterien ergeben sich für die einzelnen Untertest mittlere bis hohe Reliabilitätsmaße, weiters scheinen die einzelnen Skalen faktorenanlytisch weitgehend konstruktvalidiert zu sein. Negativ hervorzuheben sind die veralteten Eichtabellen (Kubinger, 2009b; Preusche & Leiss, 2003; Testzentrale, 2009).

Im Zusammenhang mit der Erfassung sprachlicher Intelligenz ist die *Fertigkeitenskala* bedeutsam. Sie erfasst das gelernte Wissen sowie schulische Fertigkeiten der Kinder, setzt allerdings sprachliches Verständnis und Ausdrucksvermögen voraus. Die Skala umfasst die Untertests „Wortschatz", „Gesichter und Orte", „Rechnen", „Rätsel", „Lesen/Buchstabieren" und „Lesen/Verstehen", wovon die für die sprachliche Intelligenz relevanten Subtests in Tabelle 2 beschrieben werden.

Tabelle 2: *Beschreibung der sprachlichen Untertests der Fertigkeitenskala (K-ABC)*

Untertest	Beschreibung der Aufgabe	Gemessene Fähigkeit
Wortschatz	Das Kind soll das richtige Wort für Gegenstände auf Fotos nennen.	Erinnern sprachlicher Beziehungen;
Gesichter und Orte	Dem Kind werden Bilder von berühmten und fiktiven Persönlichkeiten sowie von Orten gezeigt, welche es benennen muss.	Umfang des allgemeinen Faktenwissens;

2.3 KFT 4-12+ R (Kognitiver Fähigkeitstest für 4. – 12. Klassen, Revision)

Der KFT 4-12+ R (Heller & Perleth, 2000) ist ein differentieller Intelligenztest zur Ermittlung der kognitiven Ausstattung von Schülern der 4. bis 12. Klasse. Er ist als Gruppen- und Einzeltest anwendbar. Der Test besteht aus 9 Untertests, die sich auf die Bereiche *verbale Fähigkeiten*, *quantitative (numerische) Fähigkeiten* sowie *figural-räumliche Fähigkeiten* verteilen. Es liegen für alle Untertests zeitliche Beschränkungen vor, weswegen der KFT 4-12+ R als *Power-Speed-Test* bezeichnet werden kann. Die Reliabilitätsmaße der einzelnen Untertests reichen von geringen bis hohe Werte. Bezüglich der Validität des KFT 4-12+ R kann eine faktorenanalytisch begründete Konstruktvalidität weitgehend angenommen werden. Im Manual sind schulstufen- sowie schultypenspezifische Eichtabellen angegeben, die für den deutschsprachigen Raum repräsentativ sind (Heller & Perleth, 2000; Kubinger, 2009b; Testzentrale, 2009).

Der *Verbal-Teil* des KFT umfasst die Untertests „Wortschatz“, „Wortklassifikationen“ und „Wortanalogien“. Die Untertests sind im Multiple-Choice-Format gestaltet. Aus 5 Antwortmöglichkeiten ist jeweils eine richtige zu wählen. Tabelle 3 gibt Auskunft über die Beschreibung der Untertests sowie deren Messintention.

Tabelle 3: *Beschreibung der Untertests des Verbal-Teils des KFT 4-12+ R*

Untertest	Beschreibung der Aufgabe	Gemessene Fähigkeit
Wortschatz	Zu einem Wort muss ein Oberbegriff oder Synonym gefunden werden.	Sprachverständnis
Wortklassifikationen	Das Kind muss zu drei Wörtern einen gemeinsamen Oberbegriff finden.	Sprachgebundenes logisches Denken
Wortanalogien	Ein Wortpaar steht zueinander in einer bestimmten Relation. Zu einem dritten Begriff ist dasjenige Wort zu finden, das mit dem dritten Begriff in gleicher Relation steht wie die beiden ersten zueinander.	Sprachgebundenes logisches Denken

2.4 BUEGA (Basisdiagnostik Umschriebener Entwicklungsstörungen im Grundschulalter)

Die BUEGA (Esser, Wyschkon & Ballaschk, 2008) ist ein Verfahren zur Erfassung von Entwicklungsstörungen nach dem Klassifikationssystem psychischer Störungen ICD-10 (Dilling, Mombour & Schmidt, 2010) im Grundschulalter und wird im Einzelsetting durchgeführt. Ziel der BUEGA ist die Erfassung relevanter *Teilleistungsstörungen*. Der Test besteht aus den sieben Skalen *verbale Intelligenz*, *nonverbale Intelligenz*, *expressive Sprache*, *Lesen*, *Rechtschreibung*, *Rechnen* und *Aufmerksamkeit.* Die Skala *verbale Intelligenz* besteht aus dem Untertest „Analogien“. Die Beschreibung des Untertests sowie der gemessene Fähigkeitsbereich sind in Tabelle 4 dargestellt.

Als Reliabilitätsmaß wurden innere Konsistenzen berechnet, die als ausreichend bis sehr gut einzuschätzen sind. Während die inhaltliche Validität gesichert scheint, ist die Kriteriumsvalidität als fragwürdig zu beurteilen. Auch die Repräsentativität der Eichtabellen für den gesamten deutschsprachigen Raum scheint zweifelhaft, da die Normierung ausschließlich in einem deutschen Bundesland durchgeführt wurde (Renner, 2009; Testzentrale, 2010).

Tabelle 4: *Beschreibung der Skala Verbale Intelligenz der BUEGA*

Untertest	Beschreibung der Aufgabe	Gemessene Fähigkeit
Analogien	Die Testperson muss einen verbal dargebotenen Satz ergänzen.	Sprachlich-schlussfolgerndes Denken;

2.5 Gemeinsamkeiten und Unterschiede

Die Frage, welcher Test nun am besten dafür geeignet ist, sprachliche Intelligenz zu erfassen, lässt sich nicht eindeutig beantworten. *Je nach Fragestellung* ist das eine oder andere Verfahren besser oder weniger gut geeignet. Die Intelligenztests unterscheiden sich zunächst hinsichtlich ihres Anwendungsbereichs. Während der HAWIK-IV, AID 2 sowie der K-ABC Individualverfahren zur Abklärung der kognitiven Fähigkeiten eines Kindes sind, ist der KFT 12+ R eher als Gruppentestung konzipiert und für die Schullaufbahnberatung sowie die Evaluation von Schulversuchen und Förderprogrammen geeignet (Testzentrale, 2010). Die BUEGA ist hingegen kein Intelligenztest im klassischen Sinne, sondern dient eher der Erfassung von Entwicklungs- und Teilleistungsstörungen.

Neben dem Anwendungsbereich unterscheiden sich die Verfahren auch hinsichtlich der *Anzahl der Untertests*, die zur Beurteilung der sprachlichen Kompetenzen herangezogen werden, sowie in Bezug auf die Operationalisierung des Konstrukts *verbale Intelligenz*. Es können insgesamt 4 Bereiche unterschieden werden, die zur Beurteilung der sprachlichen Intelligenz verwendet werden.

- *Wortschatz als Indikator für Sprachverständnis/Sprachentwicklung*
- *Sprachlich-schlussfolgerndes Denken*
- *Erworbenes Wissen*
- *Kenntnis sozialer Regeln und deren Bedeutung*

Sprachliche Intelligenz scheint somit nicht als eine Dimension gesehen zu werden, sondern als ein Konstrukt, das sich aus mehreren Fähigkeiten zusammensetzt. Neben dem Wortschatz als Indikator für den Stand der Sprachentwicklung kommt auch eine logisch-schlussfolgernde Komponente hinzu, nämlich inwiefern ein Kind durch sprachliche Abstraktion zu einer

Lösung (meist einem Wort) kommt. Weiters werden das erworbene (Fakten-)Wissen sowie die Kenntnis sozial angepassten Verhaltens und sozialer Regeln bei einigen Tests zur sprachlichen Intelligenz gezählt.

Alle vorgestellten Intelligenztests bis auf die BUEGA beinhalten zumindest einen *Wortschatztest.* Ebenso verfügen alle Verfahren bis auf den K-ABC über einen Untertest zum *sprachlich-schlussfolgernden Denken.* Ansonsten unterscheiden sich die Verfahren hinsichtlich der Anzahl und Art der Untertests. Die größte Anzahl an Untertests, die zur Beurteilung der verbalen Intelligenz herangezogen werden, weist der **HAWIK-IV** auf. Er verfügt über einen eigens definierten Index *Sprachverständnis,* der die sprachliche Begriffsbildung, das sprachliche Schlussfolgern sowie erworbenes Wissen erfasst. Bei genauerer Betrachtung besteht der Index Sprachverständnis aus 5 Untertests (3 Kernuntertests sowie 2 optionale Zusatztests) , die jeweils einem der 4 oben genannten Bereiche zugeordnet werden können, sowie einem weiteren Subtest zum verbalen Schlussfolgern.

Der **AID 2** gibt im Manual keine eigene Skala oder eigenen Index an, welche explizit sprachliche Intelligenz erfassen. Vielmehr beruhen die Untertests auf *operationalen Definitionen*, welche die gemessene Fähigkeit genau festlegen. Man muss somit selbst entscheiden, welche Untertests man zur Beurteilung der sprachlichen Fähigkeiten heranzieht. Dieser auf den ersten Blick mühselig wirkende Umstand hat allerdings einige Vorteile. Dadurch, dass die Untertests zu keiner übergeordneten Skala zusammengefasst werden, kann eine Interpretation zu jedem Untertest, der jeweils eine Fähigkeit misst, im Einzelnen erfolgen. Beim HAWIK-IV kann ein Kind, das hinsichtlich seines Wortschatzes durchschnittlich begabt ist, durch schlechte Werte in den sprachlich-schlussfolgernden Untertests insgesamt zu einem unterdurchschnittlichen Ergebnis bezüglich des Gesamtindex Sprachverständnis kommen. Die Sinnhaftigkeit einer derartigen Verrechnung, wie sie auch der IQ vornimmt, ist zweifelhaft. Zwar korrelieren die einzelnen Untertests des Index Sprachverständnis miteinander und die faktorenanalytischen Untersuchungen ergeben einen Faktor, auf dem alle Untertests teilweise hoch laden, allerdings ist die Verrechnung der Untertestleistungen zu einem Indexwert inhaltlich problematisch. Die schlechte Leistung beim Index Sprachverständnis kann alleine auf eine Schwäche beim logisch-schlussfolgernden Denken zurückzuführen sein. Die Gültigkeit des Gütekriteriums der Skalierung ist für den Index Sprachverständnis im HAWIK-IV fragwürdig, da schon per Definition nicht nur eine Fähigkeit in die Beurteilung miteinfließt. Viel sinnvoller wäre es, die unterschiedlichen Aspekte sprachlicher Fähigkeiten auf Untertestebene einzeln zu

interpretieren. Dies ist im AID 2 gelungen. Es gibt jeweils einen Untertest, der die 4 Bereiche *Wortschatz*, *Sprachlich-schlussfolgerndes Denken*, *Erworbenes Wissen* sowie die *Kenntnis sozialer Regeln und deren Bedeutung* misst, auch wenn die Definitionen im Manual des AID 2 etwas abweichen. Allerdings muss kritisch angemerkt werden, dass für einen ungeübten Testleiter die Beurteilung der unterschiedlichen Aspekte sprachlicher Fähigkeiten schwierig ist, da auch die operationalen Definitionen sprachlich sehr komplex sind.

Der **KFT 4-12+ R** beinhaltet eine eigene *Skala verbaler Fähigkeiten,* die aus einem Untertest besteht, der den Wortschatz misst, sowie 2 Untertests, die sprachgebundenes logisches Denken erfassen. Die übergeordnete Skala misst neben Wortschatz somit zu einem großen Anteil logisch-schlussfolgerndes Denken.

Der **K-ABC** ist aufgrund seiner Konzeption insofern anders, da er auch im Vorschulbereich angewendet wird. Die Gestaltung der Untertests ist somit meist eher visuell gestaltet und im Vergleich zu anderen Verfahren weit weniger sprachlastig. Zwei sprachliche Untertests sind in die *Fertigkeitenskala* miteinbezogen, die das *erlernte Wissen,* sowie *schulische Fertigkeiten der Kinder* erfasst, allerdings sprachliches Verständnis und Ausdrucksvermögen voraussetzt. Die Untertests messen ein wortschatzähnliches Konstrukt („das Erinnern sprachlicher Beziehungen“) sowie das Faktenwissen.

Die **BUEGA** beansprucht durch die Skala *verbale Intelligenz* eben jene zu messen. Dies scheint durch die alleinige Abdeckung durch *einen* Untertest, der sprachlich-schlussfolgerndes Denken misst, doch eher fragwürdig.

Auch in Bezug auf die *testtheoretische Konzeption* unterscheiden sich die Intelligenzverfahren voneinander. Während der HAWIK-IV auf Modellen der klassischen Testtheorie beruht und eine konventionelle Testvorgabe verfolgt, ist der AID 2 nach Modellen der Item-Response-Theorie konstruiert und verfolgt eine adaptive Strategie[8]. Der AID 2 erfüllt für die meisten Untertests das Gütekriterium Skalierung, dessen Berechnung aufgrund der Affinität zur klassischen Testtheorie beim HAWIK-IV nicht möglich ist. Der HAWIK-IV ist hingegen ein Verfahren mit langjähriger Tradition, das auf den neuesten Erkenntnissen kognitionspsychologischer wie auch klinischer Forschung basiert. HAWIK-IV und AID 2 sind somit beide gut zur Beurteilung der sprachlichen Fähigkeiten eines Kindes bzw. Jugendlichen geeignet.

[8] Zur Gegenüberstellung der adaptiven und konventionellen Testvorgabe sei auf Kapitel 6.6 verwiesen.

Im Manual des KFT 12+ R ist nachzulesen, dass sich zwar in den meisten Fällen signifikante Abweichungen vom Rasch-Modell aufgrund einzelner Items ergaben, die meisten Items jedes Subtests jedoch als „Rasch-homogen“ [sic] angesehen werden können. Auf weitere Befunde könne aber nicht eingegangen werden (Heller & Perleth, 2000, S.19). Das Rasch-Modell sowie das Gütekriterium Skalierung gilt demnach nicht. Weiters ist die Konzeption als *Speed-and-Powertest* mit Multiple-Choice-Antwort-Format kritisch. Durch die Zeitbegrenzung kann nicht gesagt werden, ob die Fähigkeit oder die Geschwindigkeit der Testperson gemessen wird. Das Multiple-Choice-Format ermöglicht zudem, durch Raten zu einer Lösung zu kommen. Der KFT ist somit zur Beurteilung der verbalen Fähigkeiten *eines* Kindes eher weniger geeignet.

Ebenso unterscheiden sich die Verfahren hinsichtlich ihrer Konzeption als Individual- oder Gruppentestverfahren. Während der AID 2, HAWIK-IV, K-ABC sowie die BUEGA Individual-Verfahren sind, ist der KFT 4-12+ R aufgrund seines Testkonzepts eher als Gruppenverfahren konzipiert. Individual-Verfahren haben den Vorteil, besondere Testmaterialien (Würfel, Puzzles) verwenden zu können, die im Gruppensetting nicht administrierbar sind. Weiters liefert ein Einzelsetting die Möglichkeit, zusätzliche diagnostische Information über eine Verhaltensbeobachtung der Testperson zu erhalten. Ein großer Nachteil von Individual-Verfahren betrifft die wenig ökonomische Vorgehensweise. Während bei einer Gruppentestung bspw. in einer Stunde eine ganze Schulklasse getestet werden kann, erhält man bei gleichen zeitlichen Ressourcen im Einzelsetting definitionsgemäß nur Testergebnisse *einer* Testperson. Neben diesem wirtschaftlichen Aspekt haben Gruppenverfahren allerdings auch den Vorteil, testleiterunabhängig zu sein. Es kommt zu weit weniger persönlicher Interaktion zwischen dem/der Testleiter(in) und der Testperson als bei einer Einzeltestung. Nachteile von Gruppenverfahren betreffen hauptsächlich die Notwendigkeit von Zeitbegrenzungen für jeden Untertest. Dies hat einer Vermischung der Speed und Power-Komponente zur Folge (siehe Abschnitt 6.1). Weiters besteht bei Gruppenverfahren immer die Gefahr des Abschreibens (Kubinger, 2009b).

Die Frage, welches Verfahren zur Beurteilung der verbalen Intelligenz herangezogen wird, muss somit immer im Kontext der Fragestellung beantwortet werden. Je nachdem, ob es sich um eine Einzel- oder Gruppentestung vorgesehen ist, ob Teilleistungsschwächen identifiziert werden sollen oder ob man eine IQ- Diagnostik oder einen förderungsorientierten Ansatz verfolgt, ist eines der beschriebenen Verfahren auszuwählen.

3 Antonymie

Das Ziel der vorliegenden Arbeit ist, einen sprachlichen Untertest zu entwickeln, der das jeweilige Antonym eines Wortes erfragt. Demzufolge muss der Begriff der *Antonymie* genauer definiert werden.

Der Begriff *Antonymie* ist abgeleitet aus dem griechischen *anti/ant* = „gegen“ und, *o´nyma* = „Name“ (Bußmann, 2008). Die Antonymie ist der Oberbegriff für semantische[9] Gegenrelationen. Der Begriff des Gegenwortes ist sehr weit gefasst und enthält alle Arten von Bedeutungsbeziehungen, die im gesellschaftlichen Sprachbewusstsein als Ausdruck eines aufeinander bezogenen *Kontrastes* gelten (Agricola & Agricola, 1992). Die Antonymie wird auch als Spezialfall der *Synonymie* (Bedeutungsähnlichkeit bzw. –gleichheit) angesehen, da sich zwei Bedeutungen bis auf ein semantisches Merkmal, das bei beiden entgegengesetzt ist, gleichen. Der Übergang von Bedeutungsgleichheit zu Bedeutungsähnlichkeit und schließlich zum Bedeutungsgegensatz ist somit fließend. Antonyme müssen daher trotz ihrer Bedeutungsverschiedenheit gemeinsame Bedeutungsmerkmale aufweisen, um überhaupt miteinander in Beziehung gesetzt werden zu können (Agricola, 1992, Häcker & Stapf, 2004). Die Sprachwissenschaft ist bestrebt, die *Gegenwortpaare* in Gruppen einzuteilen, die sich im Grad der Genauigkeit der Gegensatzrelation unterscheiden. Die Klassifikation wird von verschiedenen Sprachwissenschaftlern unterschiedlich vorgenommen, es lassen sich allerdings drei Kategorien der Antonymie unterscheiden (nach Agricola & Agricola, 1992; Geckeler, 1979; Bußmann, 2008; Lutzeier, 1995).

1) **Komplementarität**

   Zwei Elemente, für die die Komplementaritäts-Relation gilt, stehen zueinander in einer *Entweder-oder* – Beziehung. Der Gegensatz ist *nicht graduierbar* und es gibt *keinen Zwischenbereich*.

   Bsp.: tot – lebendig; Inland – Ausland;

[9] Die Semantik ist die *Lehre der Wortbedeutung* (nach Häcker & Stapf, 2004).

2) **Antonymie (im eigentlichen Sinne)**

   Die Antonymie-Relation (auch *konträre Antonymie* genannt) von zwei Elementen unterscheidet sich von der Komplementaritäts-Relation dadurch, dass der Gegensatz *graduierbar* ist. Es sind häufig Zwischenstufen oder eine *neutrale* Bedeutungseinheit vorhanden.

   Bsp.: lieben – (gleichgültig sein) – hassen; heiß – (warm) – kalt;

3) **Konversität**

   Die Konversität beschreibt die Beziehung zwischen zwei Elementen, von denen die eine die *semantische Umkehrung* der anderen darstellt. Es handelt sich dabei um zwei unterschiedliche Perspektiven desselben Sachverhalts.

   Bsp.: kaufen – verkaufen; mieten – vermieten;

Agricola & Agricola (1992) beschreiben noch eine weitere Kategorie von Antonymen – die *fakultativen Gegenwortpaare*. Dabei handelt es sich um keine Bedeutungsgegensätze im eigentlichen Sinne, sondern um Elemente, deren Bedeutungsabstand groß genug ist, um eine gemeinsame übergeordnete Bedeutung erkennen zu können. So lassen sich bspw. in Bezug auf die verschiedenen Bedeutungen des Wortes *Sonne* fakultative Gegenwörter finden:

Sonne – Mond; Sonne – Erde; Sonne – Schatten; Sonne – Regen;

Für die Konstruktion des Untertests *Antonyme finden* werden Gegenwortpaare herangezogen, die einer der drei Kategorien *Komplemetarität*, *Antonymie* & *Konversität* zugeordnet werden können. *Fakultative Gegenwortpaare* werden nach Möglichkeit vermieden, da sie erstens streng genommen keine Bedeutungsgegensätze widerspiegeln und zweitens bei der Kodierung der Antwort (richtig oder falsch) zu Problemen führen (siehe Abschnitt 6.7).

4 Item Response Theory

Die Konstruktion eines Tests sowie die Analyse der erhobenen Daten muss immer auf testtheoretischen Modellen basieren. Generell kann man zwei Herangehensweisen unterscheiden. Der Großteil der publizierten psychologischen Tests ist nach Modellen der *klassischen Testtheorie* konstruiert, während in letzter Zeit immer mehr Verfahren mithilfe von Modellen der *probalistischen Testtheorie* entwickelt werden. International hat sich für die probabilistische Testtheorie die Bezeichnung *Item-Response-Theory* (IRT) durchgesetzt.

Die *klassische Testtheorie* beurteilt die Qualität eines Tests anhand festgelegter Gütekriterien, vor allem anhand der Hauptgütekriterien Validität (Gültigkeit), Reliabilität und Objektivität. Die *Item-Response-Theory* befasst sich zuvor mit einem grundlegenden, in der klassischen Testtheorie weitgehend unbeachteten Kriterium – dem Gütekriterium *Skalierung*. „Ein Test erfüllt das Gütekriterium *Skalierung*, wenn die laut Verrechnungsvorschriften resultierenden Testwerte die empirischen Verhaltensrelationen adäquat abbilden“ (Kubinger, 2009b, S.82). Es geht folglich darum, ob der in einem Test gewählte Verrechnungsmodus der Testleistungen zu Testwerten empirisch gerechtfertigt ist. Während die Methoden der *klassischen Testtheorie* für die Beantwortung dieser Frage ungeeignet sind, ermöglichen die Modelle der *Item-Response-Theorie* eine Überprüfung des Gütekriteriums Skalierung. Wenn ein Verfahren das Gütekriterium Skalierung nicht erfüllt, sind Überlegungen hinsichtlich der Hauptgütekriterien im Grunde müßig (Kubinger & Proyer, 2004b). Für eine genauere Darstellung der Probleme der Methoden der klassischen Testtheorie sei auf Kubinger (2009b) verwiesen.

Allein schon die im AID 3 realisierte Form des adaptiven Testens macht es unumgänglich, den Untertest *Antonyme finden* nach Methoden der *Item-Response-Theorie* zu konstruieren. Aber auch um gewährleisten zu können, dass durch die Skala *Antonyme finden* tatsächlich nur eine Fähigkeitsdimension erfasst wird und der Verrechnungsmodus im Sinne des Gütekriteriums Skalierung *fair* ist, sind Analysen nach Modellen der *Item-Response-Theorie* nötig.

Modelle der IRT treffen Annahmen darüber, wie eine Antwort auf ein *Item* (Synonym für *Aufgabe* in der testtheoretischen Terminologie) zustande kommt. Das Antwortverhalten (die *item responses*) ist somit von Interesse. Genauer formuliert treffen die Modelle der IRT

Annahmen darüber, von welchen Parametern die Lösungswahrscheinlichkeit eines Items abhängt (Bühner, 2011; Rost, 2004). Ferner ist allen Modellen der *Item-Response-Theorie* die Annahme gemein, dass den beobachtbaren Reaktionen einer Testperson in einem Test, *eine* nicht beobachtbare, latente Eigenschaft (trait) zugrunde liegt (Kubinger, 2003). Die zentrale Idee der IRT ist, dass die Wahrscheinlichkeit, ein Item zu lösen, nur von der Fähigkeit der Person, sowie von einem oder mehreren Parametern, die das Item charakterisieren, abhängt (Molenaar, 1995). Man unterscheidet einige Modelle der IRT anhand der Anzahl der im Modell enthaltenen Itemparameter. Das Rasch-Modell beruht auf Analysen mit *einem* Parameter und wird daher auch *1-parametrisches logistisches Testmodell* (*1-PL-Modell*) genannt. Beim *2-Parameter-Modell* (*2-PL-Modell*) wird neben dem Itemschwierigkeitsparameter ein weiterer Parameter verwendet, mit dem Items unterschiedlich gewichtet werden können. Beim 3-PL-Modell wird zusätzlich noch ein Rateparameter geschätzt (Kubinger, 1989; Rost, 2004). Die Datenanalyse des Untertests *Antonyme finden* erfolgt mithilfe des 1-PL-Modells.

Die IRT verfügt über einen breiten Anwendungsbereich. Sie findet in einem Testentwicklungsprozess im Zuge des Designs, der Testvorgabe, der Erstellung eines Itempools und der Eichung eines Tests Anwendung. Sie ist ferner aber auch dazu geeignet, die Qualität eines Tests zu verbessern, indem Items ausfindig gemacht werden können, die bestimmte Gruppen von Testpersonen benachteiligen.

4.1 Rasch-Modell (1-PL-Modell)

Das Rasch-Modell wurde vom dänischen Statistiker Georg Rasch entwickelt und ist inzwischen eines der meist verwendeten Modelle der IRT.

Ein großer Vorteil des Rasch-Modells besteht in der Möglichkeit, einen Test auf das Gütekriterium *Skalierung* zu überprüfen. Die einfachste Möglichkeit, um bei einem Test zu einem Testwert zu kommen, besteht darin, die Summe der gelösten Aufgaben zu bilden. Dabei wird zwischen zwei Reaktionskategorien unterschieden: Die Aufgabe wird gelöst (+) oder nicht gelöst (-). Damit wird postuliert, dass der Rohwert, als Anzahl der richtigen Antworten einer Person, eine „erschöpfende Statistik“ für ihren Personenparameter darstellt (Fischer, 1989, 1995). Es ist daher für den Testwert irrelevant, welche Items die Testperson gelöst hat und welche nicht. Dies führt zu einer wichtigen Annahme des Modells – der **lokalen stochastischen Unabhängigkeit**. Ob eine Testperson eine Aufgabe löst oder nicht,

hängt nur von ihrer Fähigkeit und von der Schwierigkeit des Items ab, nicht aber davon, welche anderen Aufgaben sie schon gelöst hat oder noch lösen wird (Kubinger, 1989, 2003, 2009b). Die Beantwortung der Items muss somit unabhängig voneinander erfolgen. Wenn die Anzahl der gelösten Aufgaben in einem Test ein faires Maß (im Sinne des Gütekriterium Skalierung) für die Fähigkeit einer Person sein soll, muss das Rasch-Modell gelten. Einen Beweis dazu liefert Fischer (1995).

Im Rahmen des Rasch-Modells wird angenommen, dass die Wahrscheinlichkeit, ein bestimmtes Item zu lösen, abgesehen vom Zufall, nur von der Fähigkeitsausprägung einer Person und der Schwierigkeit des Items abhängt. Da es somit nur um eine einzige Eigenschaftsdimension geht, kann die Fähigkeit einer Person durch eine einzige Zahl, den sog. *Personenparameter* ξ repräsentiert werden. Die Schwierigkeit des Items wird durch den *Item(schwierigkeits)parameter* σ dargestellt.

Genauer spezifiziert beschreibt das Rasch-Modell die Wahrscheinlichkeit, dass eine Testperson v ein Item i löst, in Abhängigkeit vom Personenparameter ξ_v und dem Schwierigkeitsparameter des Items σ_i. Diese Annahme wird durch eine logistische Wahrscheinlichkeitsfunktion verdeutlicht:

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

Das Rasch-Modell wie auch jedes andere Modell der IRT beruht somit auf einem *wahrscheinlichkeitstheoretischen* Ansatz. Man kann bei einer gegebenen Fähigkeit ξ einer Person v nicht deterministisch vorhersagen, ob die Person eine Aufgabe lösen wird oder nicht, sondern nur, wie wahrscheinlich sie zu einer Lösung kommen wird. Je größer die Fähigkeit ξ bei konstanter Schwierigkeit des Items, desto höher ist die Wahrscheinlichkeit, das Item zu lösen. Abbildung 1 zeigt für drei Items des Untertests *Antonyme finden* die sog. *Item Characteristik Curve* (ICC). Die ICC eines Items stellt die Lösungswahrscheinlichkeit als Funktion der latenten Fähigkeitsdimension ξ grafisch dar (Molenaar, 1995).

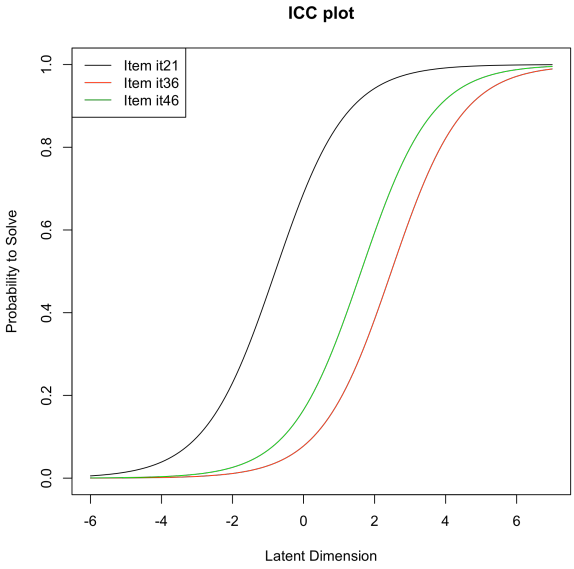

Abbildung 1: ICC-Kurven von drei Items des Untertests *Antonyme finden*

Die x-Achse gibt den Wertebereich der latenten Fähigkeitsdimension ξ an, die y-Achse stellt die Lösungswahrscheinlichkeit dar. Theoretisch liegt der Personenparameter ξ zwischen - ∞ und + ∞, praktisch liegt der Wertebereich aber zwischen -5 und +5 (Kubinger, 1989). Der Itemschwierigkeitsparameter σ für ein Item i definiert die Position auf der ICC-Kurve, bei der die Wahrscheinlichkeit einer richtigen Antwort einer Person mit der Fähigkeit ξ_v 50 % ist. σ_i gibt folglich die Schwierigkeit des Items an. Ist σ_i positiv, das Item daher tendenziell schwieriger, liegt die ICC-Kurve weiter rechts. Ist σ_i negativ, das Item ist daher tendenziell leichter, befindet sich die Kurve weiter links (Rost, 2004). Abbildung 1 zufolge ist demnach Item 21 das leichteste, gefolgt von Item 36 und 46. Je schwieriger das Item ist (je höher σ_i), desto fähiger muss eine Person sein, um mit einer 50%-Wahrscheinlichkeit das Item zu lösen (Hambleton, Swaminathan & Rogers, 1991). Die Wahrscheinlichkeit ein Item i zu lösen, ist im Grunde von der Differenz zwischen dem Personenparameter und dem Itemschwierigkeitsparameter abhängig.

Eine besondere Eigenschaft des Modells besteht in der Möglichkeit der **spezifisch objektiven Vergleiche** (Kubinger, 1989). Der Unterschied in der Fähigkeitsausprägung ξ_v und ξ_w zweier Personen, kann unabhängig davon bestimmt werden, welche Items des Tests sie bearbeitet haben. Fast wichtiger ist der Umstand, dass der Vergleich zweier Items i und j bezüglich σ_i und σ_j unabhängig davon ist, welche Stichprobe dafür herangezogen wurde. Das bedeutet,

dass die Schätzungen der Itemparameter **stichprobenunabhängig** sind, da die Wahl der Stichprobe aus einer bestimmten Population für die statistische Inferenz dieser Parameter keine Rolle spielt. Dies hat eine wichtige Implikation zur Folge. Durch das Postulat der *Stichprobenunabhängigkeit* wird das Rasch-Modell im Gegensatz zu Modellen der klassischen Testtheorie *prüfbar*. Im Sinne des Rasch-Modells müssten die Itemparameterschätzungen in unterschiedlichen Stichproben (bspw. Österreich vs. Schweiz) statistisch gleich sein. Wenn sich allerdings empirisch ergibt, dass die Itemparameterschätzungen zumindest für ein Item nicht gleich sind, gilt das Rasch-Modell nicht (Kubinger, 2003).

Wenn das Rasch-Modell für einen Test gilt, können folgende Schlussfolgerungen gezogen werden. Der Test misst *eindimensional* und die Verrechnung der Testleistung zu Testwerten ist *fair* – das Gütekriterium Skalierung ist erfüllt. Je nachdem, welche *Teilungskriterien* zur Überprüfung der spezifischen Objektivität herangezogen werden, können auch Aussagen zum Gütekriterium *Fairness* getroffen werden (siehe Abschnitt 4.2 und 7.5).

Um einen Test auf Geltung des Rasch-Modells zu prüfen, stehen mehrere Möglichkeiten in Form von Modelltests zur Verfügung.

4.2 Prüfung der Gültigkeit des Rasch-Modells

Ob ein Test oder gegebener Itempool tatsächlich dem Rasch-Modell entspricht, kann mithilfe unterschiedlicher Modelltests geprüft werde. Die Modelltests bedienen sich dem Postulat der *Stichprobenunabhängigkeit*. Die Schätzungen des Itemparameters müssen in unterschiedlichen Teilstichproben gleich sein. Die einfachste Form einer Überprüfung ist der **grafische Modelltest,** bei dem die Itemparameter, die in zwei unterschiedlichen Stichproben geschätzt wurden, pro Item in einem rechtwinkeligen Koordinatensystem gegenübergestellt werden (siehe dazu Abbildung 2).

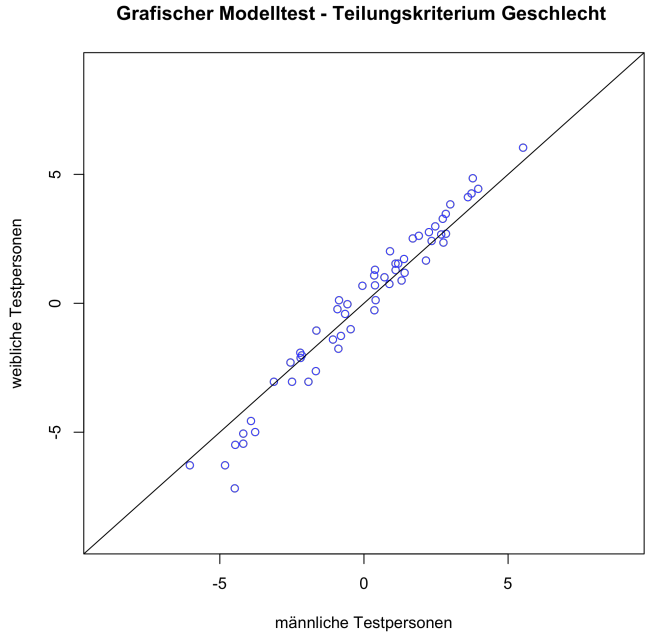

Abbildung 2: Grafischer Modelltest, Teilungskriterium Geschlecht

Bei Geltung des Rasch-Modells müssten alle Punkte nahe der 45°- Gerade liegen, da die Parameterschätzungen für jedes Item identisch wären. Weichen einzelne Punkte von der Geraden deutlich ab, ist die Stichprobenunabhängigkeit nicht erfüllt und der Test misst nicht *fair* (Kubinger & Proyer, 2004b; Kubinger, 1989, 2009b).

Inferenzstatistisch kann die Modellgültigkeit mittels **Likelihood-Ratio-Test** (LR-Test) nach Andersen überprüft werden. Er überprüft, ob die empirischen Daten durch die Itemparameterschätzungen in verschiedenen Teilstichproben besser beschrieben werden können als durch die Parameterschätzung der Items anhand der Gesamtstichprobe. Eine Stichprobe kann anhand unterschiedlicher Kriterien geteilt werden. Neben dem *externen Teilungskriterium* wie etwa Geschlecht, Alter oder Muttersprache kann man Stichproben für den LR-Test auch nach *internen* Kriterien wie dem Rohscore teilen (Glas & Verhelst, 1995; Kubinger, 1989). Mithilfe des LR-Tests kann man somit einen Test auch auf das Gütekriterium *Fairness* hin überprüfen, indem man feststellt, ob die Itemparameterschätzungen für interessierende Personengruppen (bspw. männliche vs. weibliche Testpersonen oder Gymnasiumschüler(innen) vs. Hauptschüler(innen)) gleich sind.

Der LR-Test ermöglicht allerdings nur eine globale Modellprüfung über *alle* Items eines Tests. Eine Überprüfung der Güte jedes *einzelnen* Items ermöglicht ein anderer Modelltest,

der sog. **Wald-Test**. Er ist dem LR-Test sehr ähnlich, da ebenfalls aufgrund von Teilungskriterien verschiedene Teilstichproben miteinander verglichen werden (Glas, Verhelst, 1995). Wenn die unterschiedlichen Parameterschätzungen eines Items in den zwei Teilstichproben stark voneinander abweichen, wird es signifikant und passt somit nicht zum Modell. Der Wald-Test ermöglicht somit, ungeeignete Items eines Tests zu identifizieren, die als Folge aus dem Itempool ausgeschlossen werden können. Teilweise kann dadurch, zumindest a posteriori, ein Test doch noch Rasch-Modell-Konformität erreichen (Kubinger, 1989).

Auf die Frage, welche Modelltests für die Überprüfung auf Rasch-Modell-Konformität herangezogen werden sollen, meinen Glas & Verhelst (1995, S. 94): „So the scientific way, may, after all, be to choose „statistics all“ and to give the alternative hypothesis that the RM [Rasch-Modell] does not hold as much chance as possible.“

III Empirischer Teil

5 Hintergrund und Ziel der Untersuchung

Die Idee zur Konzeption des Untertests *Antonyme finden* entstand aus Problemen des Untertests *Synonyme finden* des AID 2. Bei diesem Untertest wird dem Kind ein Wort vorgelesen, wonach es ein anderes Wort finden soll, das dasselbe bedeutet. Damit soll das elementare Sprachverständnis eines Kindes erfasst werden, nämlich inwieweit es in der Lage ist, die Bedeutung eines Begriffes zu verstehen und ihn alternativ ausdrücken zu können. Sowohl theoretisch als auch praktisch ergaben sich Probleme bezüglich dieses Untertests. Zwei Begriffe zu finden, die wirklich *dieselbe* Bedeutung haben, ist sehr schwierig. Oft sind Begriffe, die als synonym angesehen werden, nur *assoziativ* miteinander verknüpft und im Grunde keine Synonyme im Sinne der Bedeutungs*gleichheit*. Wenn man nun Assoziationen doch mitberücksichtigt, ergeben sich bei entsprechenden Items oft sehr viele Antwortmöglichkeiten, die als richtig zu werten sind. Wenn man beispielsweise nach einem synonymen Begriff zum Wort „schnell“ fragt, wären die Antworten „rasch“, „flink“, „geschwind“, „flott“, „in Windeseile“, „unverzüglich“, „hurtig“, „rapide“, „zügig“, „eilig“ als assoziativ richtig zu werten. Als Folge wird der Antwortkatalog endlos lang, was zu großen Problemen bei der Itemkonstruktion sowie bei der Kodierung der Aufgaben führt.
Bei der Itemkonstruktion wiederum ist es schwierig, alle Assoziationen zu einem Begriff zu finden und zu beurteilen. Die Entscheidung, ob ein Begriff nun synonym, assoziativ oder eben nur ein ähnliches Wort ist, ist oft schwierig, da es beispielsweise sein kann, dass Begriffe im Sprachgebrauch synonym verwendet werden, in entsprechenden Lexika aber eine unterschiedliche Bedeutung haben.
Auch bei der Kodierung der Items kommt es zu Schwierigkeiten. Einige Psycholog(inn)en kodieren nur jene Antworten des Kindes als richtig, die auch im Antwortkatalog stehen, andere hingegen kodieren auch assoziative Wörter, die sinnvoll und richtig erscheinen ebenfalls als richtig. Als Konsequenz leidet die *Auswertungsobjektivität*, da nicht jede/r Psychologe(in) aufgrund unterschiedlicher Kodierung bei gleichen Items zum gleichen Ergebnis kommt.
Generell lässt sich feststellen, dass assoziative Begriffe keine alternativen und damit synonymen Ausdrucksweisen eines ursprünglichen Begriffes sind.

Auch im Hinblick auf die Internationalisierung des AID treten Probleme bezüglich des Untertests *Synonyme finden* auf. Es existiert bereits eine publizierte Version des AID 2 in

einer anderen Sprache (AID 2 – Türkisch). Versionen zu AID 2 – Ungarisch, AID 2 - Niederländisch und AID 2 – Englisch liegen vor, sind aber noch nicht publiziert. Wie bereits erläutert, treten bereits bei der deutschen Version des AID Probleme mit alternativen Begriffen von Wörtern und der dazugehörigen Assoziationsproblemantik auf. Wenn man die Items des Untertests samt Lösungen nun in andere Sprachen einfach übersetzt, wird die Problematik noch verschärft. Wenn schon in der deutschen Version teils Unklarheit darüber herrscht, ob ein Begriff synonym oder assoziativ ist, birgt eine Übersetzung in eine andere Sprache zusätzlich Fehlerquellen. Womöglich gibt es synonyme Begriffe, für die es kein deutsches Äquivalent gibt und die somit im Antwortkatalog fehlen. Weiters können gerade bei langen Antwortkatalogen einige übersetzte Lösungen eine unterschiedliche Bedeutung haben. Damit nun die sprachlichen Fähigkeiten eines Kindes zu beurteilen erscheint fragwürdig.

Überlegungen gaben Anlass dazu, für den AID 3 einen weiteren Untertest zu konzipieren, der das elementare Sprachverständnis prüft und die angesprochenen Probleme des Untertests *Synonyme finden* zu lösen versucht – der Untertest *Antonyme finden.* Die Idee dahinter ist grundsätzlich simpel. Anstatt das Synonym eines Begriffes zu erfragen, ist nun das Antonym, somit das Gegenteil eines Begriffes von Interesse. Daraus ergeben sich im Vergleich zum Untertest *Synonyme finden* einige Vorteile. Die Idee, das Gegenteil eines Begriffes zu erfragen, führt grundsätzlich zu eindeutigeren Ergebnissen, da es für viele Wörter Gegenwörter gibt. Das Konstrukt der Bedeutungsgleichheit des Untertests *Synonyme finden* ist hingegen umstritten. Ebenso wird das Problem der assoziativen Begriffe im Untertest *Antonyme finden* nahezu gelöst, da das Problem der Assoziation eben nur auf synonyme Wörter beschränkt ist. Auch die Übersetzung in andere Sprachen ist weniger problematisch als beim Untertest *Synonyme finden,* da man davon ausgehen kann, dass Gegensätze in Fremdsprachen noch eher den Gegensatzrelationen der deutschen Sprache entsprechen und somit eher übersetzbar sind als synonyme Relationen.

Guthke (1996) weist darauf hin, dass sich der/die Testkonstrukteur(in) fragen sollte, welchen Beitrag der konstruierte Test in Verbindung mit anderen Informationsquellen leisten kann. Der Untertest *Antonyme finden* soll demnach im AID 3 neben dem Untertest *Synonyme finden* und *Funktionen abstrahieren* zusätzliche Informationen zur sprachlichen Fähigkeit eines Kindes liefern.

6 Testkonstruktion

Bühner (2011) unterteilt den Prozess der Testkonstruktion in drei Teilabschnitte: (1) die Erstellung des Testentwurfs, (2) die empirische Überprüfung sowie (3) die Normierung.
Die vorliegende Arbeit befasst sich mit den Schritten (1) und (2), die Normierung des Untertests *Antonyme finden* erfolgt im Zuge der Normierung des AID 3.

Ein intuitives Vorgehen bei der Erstellung des Testentwurfs birgt viele Fehler, die zu einem späteren Zeitpunkt nicht mehr korrigierbar sind. Für die Konstruktion des Untertests *Antonyme finden* wurde ein anschauliches Modell von Bühner (2011) gewählt, das den Teilschritt „Erstellung des Testentwurfs" in mehrere Einzelschritte aufgliedert. Die Beschreibung des Testkonzepts des Untertests *Antonyme finden* orientiert sich an jenem Modell.

6.1 Testart und Festlegen der Art der Indikatoren

Cattell (1965, zitiert nach Bühner, 2011) unterscheidet Daten hinsichtlich der Art, wie sie erhoben werden. T-Daten (Test data) sind Daten aus Tests, deren Aufgaben eindeutig mit richtig oder falsch bewertet werden können. Demgegenüber werden Q-Daten (Questionnaire data) aus Fragebögen gewonnen und L-Daten (Life data) aus Verhaltensbeurteilungen eines Beobachters. Der Untertest *Antonyme finden* ist ein sprachlicher Leistungs*test* mit dichotomem Antwortformat und kann dadurch eindeutig den T-Daten zugeordnet werden. Leistungstests bestehen meist aus objektiven Indikatoren, während Persönlichkeitsfragebogen aus subjektiven Indikatoren bestehen (Bühner, 2011).
Die objektiven Indikatoren des Untertests *Antonyme finden* sind Items, die das Konstrukt *Wortschatz unter Beachtung sprachlogischer Regeln* zu messen beanspruchen. „Ein Item ist die kleinste Beobachtungseinheit in einem Test, sozusagen der elementare Baustein, aus dem ein Test aufgebaut ist." (Rost, 2004. S.55)
Ein Vorteil, der allen Leistungstests gemein ist, ist der Umstand, dass die Testperson das Ergebnis nur in eine Richtung *verfälschen* kann. Man kann sich bei Leistungstests weniger fähig darstellen als man ist, versuchen die Antwort zu erraten oder keine Motivation zeigen, es ist allerdings nicht möglich sich intelligenter darzustellen, als man ist (Rost, 2004). Somit

hat jeder Leistungstest gegenüber jedem Persönlichkeitsverfahren den Vorteil, weitaus weniger verfälschbar zu sein.
Innerhalb der Gruppe der Leistungstests lassen sich zudem *Speed-* und *Power-Tests* unterscheiden. Bei *Speed-Tests* wird die Bearbeitungsgeschwindigkeit als Leistung bewertet, während der Schwierigkeitsgrad der Aufgaben sehr niedrig ist. *Powertests* sind Test der Leistungshöhe, die keine oder eine großzügig bemessene Zeitbegrenzung aufweisen (Häcker & Stapf, 2004). Daneben gibt es noch *Speed-and-Power-Tests*, die sowohl eine *Speed-* als auch eine *Power*-Komponente beinhalten (Kubinger, 2009b). Da es aber zur Bearbeitung einer entsprechenden Testaufgabe zwei Fähigkeitsdimensionen benötigt (Leistung *und* Bearbeitungsgeschwindigkeit), erfüllen *Speed-and-Power-Tests* oft nicht das Gütekriterium Skalierung. Bei einer schwachen Leistung ist bspw. nicht klar ersichtlich, ob das schlechte Abschneiden auf eine mangelnde Fähigkeit oder langsame Bearbeitung zurückzuführen ist.
Der Untertest *Antonyme finden* ist ein reiner Powertest, da es keine Zeitbegrenzung für die Bearbeitung gibt und allein die Leistungshöhe von Interesse ist.

6.2 Festlegen der Zielgruppe

Der Untertest *Antonyme finden* ist für deutschsprachige Kinder und Jugendliche im Alter von 6 Jahren (6;0) bis 15 Jahre und 11 Monate (15;11) konzipiert.
Dabei ist für die Itemkonstruktion zu beachten, dass nur solche Begriffe generiert werden, die nicht bestimmte Personengruppen aufgrund von Herkunft, Geschlechtszugehörigkeit oder soziokulturellem Status systematisch benachteiligen (Kubinger, 2009b). So muss beispielsweise beachtet werden, dass der AID 3 für den gesamten deutschsprachigen Raum (Deutschland, Österreich, Schweiz) konzipiert ist und somit keine Begriffe im Test enthalten sein dürfen, die für ein bestimmtes deutschsprachiges Land eine höhere Itemschwierigkeit aufweisen als für ein anderes. So wäre das Item des Untertests Synonyme finden: *Nenne mir ein anderes Wort für „Sessel“* ungeeignet, da das Wort in der Schweiz und Deutschland eine andere Bedeutung hat als in Österreich.
Ferner sollte bei der Itemkonstruktion beachtet werden, dass sich die Itemschwierigkeiten der einzelnen Aufgaben für Gruppen von Personen mit unterschiedlicher Geschlechtszugehörigkeit oder Herkunft nicht unterscheiden; ein grundlegendes Verständnis der deutschen Sprache muss bei der Bearbeitung der Aufgaben natürlich vorausgesetzt werden.

6.3 Testziel

Der Untertest *Antonyme finden* hat das Ziel, das *elementare Sprachverständnis* zu erfassen. Damit gehört er zur Gruppe der Tests, die die *Bestimmung einer Eigenschafts- oder Fähigkeitsausprägung* zum Ziel haben. Er ist somit von Tests abzugrenzen, deren Testziel die *Gruppentrennung* oder die *Erfassung von Wissen* ist.

Bei Tests, deren Ziel es ist, Eigenschaften oder Fähigkeiten zu messen, ist es vor allem relevant, inhaltsvalide Items zu konstruieren. Dabei ist es wichtig, dass die Aufgaben nur eine zugrunde liegende Dimension erfassen und miteinander korrelieren (Bühner, 2011).

Somit ist es für die Konstruktion zu beachten, dass ausschließlich Items generiert werden, die nur das zu messen beanspruchte Konstrukt erfassen, nicht aber andere Fähigkeitsdimensionen wie beispielsweise *allgemeines Wissen*.

6.4 Erstellen einer Definition des Messgegenstandes

Der Untertest *Antonyme finden* ist ein Test zum elementaren Sprachverständnis und misst den Wortschatz unter Beachtung sprachlogischer Regeln. Die operationale Definition lautet: *Beim Untertest Antonyme finden wird die Fähigkeit gemessen*, *inwieweit die Testperson imstande ist, die Bedeutung eines Begriffes zu erfassen und die gegensätzliche Bedeutung dieses Begriffes wiedergeben zu können.*

6.5 Wahl des Antwortformats

Der Tradition der Untertests des AID und AID 2 folgend wurde für den Untertest *Antonyme finden* ein *freies Antwortformat* gewählt. Die freie Aufgabenbeantwortung ist dadurch gekennzeichnet, dass die Testperson die gestellte Aufgabe verbal oder nonverbal (bspw. bei der Bearbeitung von Testmaterial) nach eigenem Ermessen selbst beantworten soll. (Lienert & Raatz, 1998).

Beispielitem*: Nenne mir das Gegenteil von „warm“*.

Ein freies Aufgabenformat hat gegenüber gebundenenen Antwortformaten wie etwa dem Multiple-Choice-Antwortformat einige Vorteile. Ein sehr bedeutsamer Vorteil des freien Antwortformats ist, dass es quasi frei von Zufallseinflüssen ist. (Bühner, 2011; Lienert & Raatz, 1998). Durch Raten zu einer Lösung zu kommen, wie es bei Multiple-Choice-Items möglich ist, ist beim freien Antwortformat nicht sinnvoll. Es kommt daher zu keiner

Verfälschung des Testergebnisses aufgrund von Rateeffekten. Das freie Antwortformat scheint auch diagnostisch aufschlussreicher zu sein als das Multipe-Choice-Format, da die Testperson selbst eine Antwort generieren muss, anstatt aus einer Vorgabe an Antwortmöglichkeiten zu wählen (Kubinger, 2009b). Demzufolge ist beim freien Antwortformat auch kein bloßes Wiedererkennen der Lösung möglich. Allerdings ergeben sich auch Nachteile gegenüber gebundenen Antwortformaten. So ist der Zeitaufwand für die Bearbeitung von Aufgaben mit freiem Antwortformat größer (Bühner, 2011; Lienert & Raatz, 1998). Ein gravierendes Problem ist die teilweise eingeschränkte Auswertungsobjektivität von Aufgaben mit freiem Antwortformat. Das Multiple-Choice-Format ist verrechnungssicher (im Sinne der Auswertungsobjektivität), während sich beim freien Antwortformat eine mangelnde Auswertungsobjektivität ergeben kann, wenn es für ein Item beispielsweise mehrere Lösungen gibt oder die Richtigkeit einer Antwort nicht eindeutig ist (Bühner, 2011; Lienert & Raatz, 1998; Kubinger, 2009b). So kann die Antwort einer Testperson für den/die Testleiter(in) richtig oder originell erscheinen, obwohl sie nicht im Antwortkatalog aufgelistet ist. Einige Psycholog(innen) kodieren jene kritischen Antworten als richtig, andere halten sich streng an den Antwortkatalog und kodieren die Aufgaben als falsch. Als Folge leidet die Auswertungsobjektivität, da bei gleicher Aufgabe verschiedene Testleiter(innen) nicht zum gleichen Ergebnis kommen. Kubinger (2009b) ist allerdings der Ansicht, dass auch Aufgaben, die nach dem freien Antwortformat konstruiert sind, durchaus verrechungssicher sein können. Wenn Probleme hinsichtlich der Objektivität auftreten, liegt das oft an Mängeln der Testkonstruktion und nur teilweise am Testkonzept als solchem.

6.6 Testvorgabe

Bei der Testvorgabe wird grundsätzlich zwischen *adaptiver* und *konventioneller* Testvorgabe unterschieden. Bei der *konventionellen* Vorgabe werden jeder Testperson dieselben Aufgaben in derselben Reihenfolge vorgegeben, während bei *adaptiver* Vorgabe jede Testperson nur jene Aufgaben bearbeiten muss, die ihrem Leistungsniveau entsprechen (Kubinger, 2009a). Während die *konventionelle* Testvorgabe bei Verfahren eingesetzt wird, die auf der klassischen Testtheorie beruhen, ist die *adaptive* Vorgehensweise zwingend mit der *Item-Response-Theorie* verbunden. (Kubinger & Proyer, 2004b) Die konventionelle Vorgabe führt zu Problemen hinsichtlich des Gütekriteriums *Ökonomie*. Da jeder Testperson dieselben Aufgaben vorgegeben werden, kann es dazu führen, dass leistungsschwachen Personen viele

Aufgaben zu schwer und leistungsfähigen Personen viele Aufgaben zu leicht fallen. Die Vorgabe zu leichter bzw. zu schwieriger Items führt kaum zu einem Informationsgewinn und ist folglich unökonomisch. Ferner besteht die Gefahr von Motivationsverlust oder Frustration. Die *adaptive* Testvorgabe ist im Vergleich dazu viel effizienter. Man versucht der Testperson in Abhängigkeit davon, welche Aufgaben sie schon gelöst hat (adaptiv), nur solche Aufgaben vorzugeben, die möglichst informativ sind. Informativ ist eine Aufgabe dann, wenn die Wahrscheinlichkeit, dass die Testperson die Aufgabe löst, 50% beträgt.
Da nur informative Items vorgegeben werden, benötigt man im Vergleich zur *konventionellen* Vorgabe weniger Aufgaben und erzielt dennoch die gleiche Messgenauigkeit.
Man unterscheidet bei der adaptiven Testvorgabe zwischen zwei Strategien: dem *tailored-testing* sowie dem *branched-testing*. Das *tailored-testing* ist dem *branched-testing* insofern überlegen, dass nach jeder Aufgabenbearbeitung der Testperson auch wirklich jenes Item als nächstes vorgegeben wird, dass entsprechend der Fähigkeit der Testperson am informativsten ist. Die Berechnung ist allerdings an den Computer gebunden. Beim *branched-testing* werden die verschiedenen Items in Aufgabengruppen zusammengefasst. Nach der Bearbeitung einer dieser Aufgabengruppen (und eben nicht nach jeder Aufgabe), wird der Testperson in Abhängigkeit davon, wie viele Aufgaben dieser Gruppe sie gelöst hat, eine leichtere, gleich schwierige oder schwierigere Aufgabengruppe vorgegeben. Ein Vorteil des *brached-testing* liegt darin, dass es auch für Papier-Bleistift-Verfahren eingesetzt werden kann.

Beim Untertest *Antonyme finden* wurde, in Anlehnung an nahezu alle Untertests des AID 2, eine adaptive Testvorgabe nach der *branched-testing*-Strategie als Ziel gesetzt. Für die adaptive Vorgabe müssen allerdings die Schwierigkeitsparameter jedes Items bekannt sein. Da die Items bislang noch keiner Stichprobe vorgegeben wurden, muss die erste Testvorgabe nach der konventionellen Strategie erfolgen. Anschließend können mit Modellen und Analysemethoden der *Item-Response-Theorie* die Schwierigkeitsparameter jedes Items berechnet werden, um eine adaptive Vorgehensweise nach der *branched-testing*-Strategie zu ermöglichen.
Die Datenerhebung des Untertests *Antonyme finden* erfolgt somit nach der *konventionellen* Testvorgabe, ohne aber für die anschließende Analyse auf Methoden der klassischen Testtheorie zurückzugreifen.

6.7 Regeln zur Itemkonstruktion

Um zu gewährleisten, dass die Itemkonstruktion des Untertests *Antonyme finden* nicht intuitiv – fehlerhaft, sondern regelgeleitet geschieht, wurden vor Beginn der Aufgabengenerierung Itemkonstruktionsregeln, insbesondere spezielle *„Ausschlusskriterien"* für Items definiert. Ein Item wird nur in ein vorläufiges Itemuniversum aufgenommen, wenn es nicht in eine der folgenden *Ausschlusskategorien* fällt:

- *Mehrere synonyme Lösungen*
- *Assoziative Lösungen*
- *Homonyme*
- *Begriffe, die Alltagswissen messen*
- *Fremdwörter*
- *Fachbegriffe*
- *Begriffe, bei denen die Vorsilbe un- eine Lösung ist*

Werden diese Regeln bei der Konstruktion der Aufgaben nicht beachtet, kommt es bei der Testdurchführung und in weiterer Folge bei der statistischen Auswertung zu Problemen. Bei Items mit vielen *synonymen Lösungen* ist die Wahrscheinlichkeit erhöht, dass von den Testpersonen Antworten gegeben werden, die nicht im Antwortkatalog stehen. Das führt zur unangenehmen Situation, dass der/die Testleiter(in) nach eigenem Ermessen entscheiden muss, ob das Item als richtig oder falsch bewertet werden muss. Als Folge leidet die Verrechnungssicherheit (Auswertungsobjektivität). Wenn man diesem Umstand entgegenwirken will, muss man möglichst alle Lösungen angeben, was einen endlos langen Antwortkatalog zur Folge hat.

Beispiel: *Nenne das Gegenteil des Wortes „interessant"*.

Die richtige Antwort auf dieses Item wäre *„langweilig"*. Aber auch die Antworten *„eintönig", „öde", „uninteressant", „einfallslos", „fad"* oder *„einschläfernd"* sind nicht falsch. Items, die mehrere synonyme Lösungen besitzen, sind somit zu vermeiden. Außerdem würde man die Probleme des Untertests *Synonyme finden* einfach übernehmen.

Ein weiteres Beispiel für ein Ausschlusskriterium, das aufgrund der Probleme des Untertests *Synonyme finden* entwickelt wurde, sind *Assoziationen*.

Beispiel: *Nenne das Gegenteil des Wortes „Staatsanwalt"*.

Die naheliegende Antwort *(Straf-)Verteidiger* ist eine Assoziation des Begriffes Staatsanwalt, nicht aber das Gegenteil. Ein Staatsanwalt ist ein assoziativ ähnlicher Beruf, aber es gibt in diesem Fall kein Gegenteil. In Anlehnung an die unterschiedlichen Kategorien der *Antonymie* (Kapitel 3) handelt es sich beim vorliegenden Beispiel um ein *fakultatives Gegenwortpaar*. Diese Antonymie-Relation soll vermieden werden, da sie streng genommen keine Bedeutungsgegensätze widerspiegelt.

Eine weitere Kategorie, die zum Ausschluss eines Items führt, ist die Kategorie *Homonyme* (mehrdeutige Begriffe).

Beispiel: *Nenne das Gegenteil des Wortes „vormachen“.*

Die Lösung „*nachmachen“* scheint trivial, jedoch ist es sehr schwierig, ein Antonym zu finden, wenn man *„vormachen“* im Sinne von „*jemandem etwas vormachen*“ versteht. Jene Items erweisen sich bei der Datenanalyse oft als nicht Rasch-Modell-konform, da womöglich leistungsschwache Kinder die einfache Lösung nennen können, wo hingegen leistungsstarke Kinder zu keiner Lösung kommen, da sie an die schwierigere Bedeutung des Begriffes denken. Ebenso ist das Item: *Nenne das Gegenteil des Wortes: „binden“* kritisch, da neben der eigentlichen Bedeutung (wie bspw. Schuhbänder binden) auch die Bedeutung „*ein Buch binden*“ oder „*eine Sauce binden*“ existiert. Ferner muss bei jedem Item beachtet werden, dass die Testperson den Begriff vorgelesen bekommt, demnach nicht klar einschätzen kann, ob es sich um ein Nomen, Adjektiv oder Verb handelt. Im konkreten Fall wäre es für die Testperson wahrlich schwierig einzuschätzen, ob es sich beim Wort „binden“ um ein Nomen (im Sinne der Mehrzahl von Damenbinde) oder ein Verb handelt.

Items, die in die Ausschlusskategorien „*Begriffe, die Alltagswissen messen*“, sowie „*Fachbegriffe*“ fallen, laufen ebenso Gefahr, nicht Rasch-Modell-konform zu sein.

So misst das Item: *Nenne das Gegenteil des Wortes „Säure“* (Lösung: „*Base*“) eher Alltagswissen als sprachliche Fähigkeit, womit die Voraussetzung verletzt wird, dass für die Beantwortung der Aufgaben nur eine zugrunde liegende Fähigkeit benötigt wird. Auch die Lösung der Aufgabe: *Nenne mir das Gegenteil des Wortes „absorbieren“* (Lösung: „*ausscheiden*“), scheint durch spezielles Fachwissen wahrscheinlicher als durch sprachliche Fähigkeiten.

Begriffe, die der Kategorie „*Fremdwörter“* zuzuordnen sind, müssen ebenso ausgeschlossen werden, da sie gegen das Gütekriterium Skalierung verstoßen.

Bsp.: *Was ist das Gegenteil von „impulsiv*“?

Kinder aus Familien mit niedrigem sozioökonomischen Status werden durch derartige Aufgaben vermutlich systematisch benachteiligt, da sie meist keinen Zugang zu fremdsprachlichem Wortschatz haben. Dieses Item misst demzufolge eher Bildung als sprachliche Fähigkeiten.

Zuletzt sind auch Items, bei denen die Vorsilbe *un-* zu einer Lösung führt, auszuscheiden.

Bsp.: *Nenne mir das Gegenteil des Wortes „vorteilhaft“.*

Die Lösung „*unvorteilhaft*“ ist derart trivial, dass das Items wohl keinerlei Informationswert besitzt. Weiters erscheint das Item *Nenne mir das Gegenteil des Wortes „ohne Gewähr*“ zunächst durchaus anspruchsvoll (Lösung/en: „garantiert“, „unter Garantie“, „sicher“), bei genauer Betrachtungsweise erweist sich aber auch die Antwort „*mit Gewähr*“ als richtig, wonach das Item ebenfalls auszuschließen ist.

6.8 Konstruktionsprozess

Im Juli 2009 wurde mit der Erstellung eines Itempools begonnen. Für eine adaptive Testvorgabe nach dem *branched-testing* werden mehr Items benötigt als bei der konventionellen Testvorgabe. Das Verzweigungsschema im AID 2 sieht 60 verschiedene Items vor, die einen sehr breiten Schwierigkeitsbereich abdecken, auf den die Aufgaben gleichmäßig verteilt sind (Kubinger, 2009a). Da man davon ausgehen muss, dass einige Items nicht dem Rasch-Modell entsprechen und ausgeschieden werden müssen, sollten zwischen 65 und 70 Items konstruiert werden.

Zunächst wurde versucht aus unterschiedlichen Themengebieten wie bspw. Natur, Schulbereich, Gefühle, Freizeit, Eigenschaften, Sport, etc. Items zu konstruieren, die anschließend einer genauen Recherche unterzogen wurden. Zunächst wurde die genaue Bedeutung bzw. Definition des Begriffes recherchiert und darauf aufbauend versucht, das entsprechende Antonym zu finden. Als Quellen für die Recherche wurden spezielle Wörterbücher (Agricola & Agricola, 1992; Bulitta & Bulitta, 2003) sowie Internetlexika (Synonym.com, 2007; Wictionary, 2009 & Woxikon, 2009) verwendet. Die angegebenen Quellen dienten einerseits als Ideenhilfe zur Iteminstruktion, auf der anderen Seite lieferten sie Lösungsvorschläge zu Antonymen. Da auch Gegenwörter ähnliche oder synonyme Wörter haben können, wurden zu jedem Antonym entsprechende Synonyme gesucht, um auch alle

möglichen Lösungen des Ursprungsbegriffes zu erfassen. Die Antwortmöglichkeiten, die Internetlexika sowie Wörterbücher zu Synonymen anbieten, fallen allerdings oft nicht in die Definition von Antonymen, wie sie für die Konstruktion des Tests verwendet wurden. Viele Vorschläge sind zu assoziativ oder umgangssprachlich, sodass die verschiedenen Antwortmöglichkeiten bewertet werden mussten. Dazu diente ein Online-Forum, in dem zu jedem Untertest ein eigenes Diskussionsforum eingerichtet wurde. Es diente dazu, vorgeschlagene Items zu diskutieren und zu bewerten. Am Diskussionsprozess beteiligten sich 5 Diplomand(innen), die Projektleiter(innen) Univ.-Prof. Dr. Kubinger und Dr. Holocher-Ertl sowie Mitarbeiter des Arbeitsbereichs Psychologische Diagnostik.

Zunächst wurden mithilfe der Ausschlusskategorien etliche Items samt Lösungsvorschlägen konstruiert und ins Forum gestellt. Ferner wurde jedes Item hinsichtlich seiner Schwierigkeit einer der drei Kategorien, *leicht*, *mittel* oder *schwierig* zugeordnet. Die einzelnen Aufgaben wurden von den Forumbeteiligten ausführlich diskutiert, worauf sie entweder ausgeschieden, einer weiteren Recherche unterzogen oder in den endgültigen Itempool aufgenommen wurden. So mussten einige Items ausgeschlossen werden, da sie dennoch in eine der Ausschlusskategorien fielen oder bspw. in einem anderen Untertest schon vorkamen. So kam es auch nicht selten vor, dass ein Item sich als ungünstig für den Untertest *Antonyme finden* herausstellte, für den Untertest *Synonyme finden* oder *Alltagswissen* aber gut geeignet war. Teilweise ergab die Diskussion, dass ein Item zwar generell brauchbar, aber noch unpräzise war, da weitere Antwortalternativen zur Diskussion standen. Jene Items wurden dann erneut auf ihre exakte Bedeutung und Gegensatz-Relation recherchiert und anschließend nochmals zur Diskussion gestellt. Auch die Schwierigkeitseinschätzung der Items wurde diskutiert und teilweise entsprechend verändert. Der gesamte Diskussionsprozess gestaltete sich sehr langwierig, da auf jede Kritik eingegangen wurde und ein Item erst dann in den Itempool aufgenommen wurde, wenn alle an der Diskussion beteiligten ihre Zustimmung gaben. Mitte August 2009 konnte die Testkonstruktion abgeschlossen werden, da der Itempool mit 67 Items groß genug war. Anschließend wurden die drei Schwierigkeitskategorien auf 7 Kategorien erweitert und jedes Item subjektiv einer der Kategorien zugeteilt. Die Schwierigkeitskategorien lauteten nun „sehr leicht“, „leicht“, „leicht bis mittel“, „mittel“, „mittel bis schwierig“, „schwierig“ und „sehr schwierig". Diese Differenzierung war notwendig, um die Items zumindest subjektiv der Schwierigkeit nach genauer reihen zu können, da die empirischen Itemschwierigkeitsparameter ja noch nicht bekannt waren. Die Verteilung der Items pro Schwierigkeitskategorie ist Tabelle 5 zu entnehmen.

Tabelle 5: *Verteilung der Items bezüglich ihrer Schwierigkeit*

Schwierigkeitskategorie							
Itemanzahl	sehr leicht	leicht	leicht-mittel	mittel	mittel-schwierig	schwierig	Sehr schwierig
	9	11	5	17	5	12	8

Anschließend wurden 6 Testhefte erstellt, jeweils zwei Parallelversionen für 3 Alterskategorien. Somit ergaben sich zwei Testhefte für die Altersgruppe *6-8 Jahre*, zwei Testhefte für die Alterskategorie *9-11 Jahre* und zwei Testhefte für die Altersgruppe *12-15 Jahre*. Jedes Testheft enthielt 20 Items, wobei die Altersgruppe der 6-8-jährigen großteils Items der Kategorie „sehr-leicht“ und „leicht“ erhielten, die Altersgruppe *9-11 Jahre* vermehrt Aufgaben der Kategorien „leicht-mittel“, „mittel“ und „ mittel-schwierig“ und der Kategorie der 12-15-jährigen hauptsächlich Items der Kategorien „mittel-schwierig“, „schwierig“ und „sehr schwierig“ zugeteilt wurden. Da aufgrund der großen Aufgabenanzahl nicht jedes Item jedem Kind vorgegeben werden kann, kam ein *balanciertes Block-Design* zur Anwendung, das ein Verzweigungsschema für die Zuteilung der Items zu jedem Testheft vorsieht (Kubinger & Rasch, 2006 zitiert nach Kubinger, 2009a). Dieser Vorgang ist notwendig, um bei der Datenanalyse die Itemschwierigkeitsparameter berechnen zu können. Da nicht jedes Kind jedes Item bearbeitet, sind einige sogenannte *linking-items* nötig, um alle Itemparameter schätzen zu können. Nach der Fertigstellung der Testformen konnte der Testkonstruktionsprozess abgeschlossen werden.

7 Gütekriterien des Untertests *Antonyme finden*

7.1 Objektivität

„Unter *Objektivität* eines Tests verstehen wir den Grad, in dem die Ergebnisse eines Tests unabhängig vom Untersucher sind“ (Lienert & Raatz, 1998, S. 7).

Lienert & Raatz (1998) unterscheiden drei Aspekte der Objektivität: die *Durchführungsobjektivität*, die *Auswertungsobjektivität* und die *Interpretationsobjektivität*, sprich inwieweit die Durchführung, Auswertung und Interpretation eines Tests unabhängig vom Untersucher dieselben Ergebnisse liefern.

Durch eine standardisierte schriftliche Instruktion bei der Vorgabe des Tests sollten Mängel hinsichtlich der *Durchführungsobjektivität* vermieden werden. Die Forderung, die Untersuchungssituation zu standardisieren, war im Zuge der Testungen in verschiedenen Schulen allerdings nicht zu erfüllen. Genaue statistische Untersuchungen zur *Testleiterunabhängigkeit* wie im AID 2 waren im Rahmen dieser Diplomarbeit nicht möglich.

Hinsichtlich der *Auswertungsobjektivität* (Kubinger, 2009b, spricht in diesem Zusammenhang von *Verrechnungssicherheit*) ist aufgrund des *freien Antwortformats* mit Problemen zu rechnen. Im Falle des Untertests *Antonyme finden* kann weitgehend Auswertungsobjektivität postuliert werden, da die Instruktion gegeben wurde, auch wirklich nur jene Antworten als richtig zu kodieren, die im Antwortkatalog enthalten sind. Etwaige kritische Items, die womöglich doch je nach Testleiter(in) unterschiedlich kodiert wurden, sollten durch die Rasch-Modell-Analysen identifizierbar sein.

Die Interpretationsobjektivität kann als gegeben betrachtet werden, da im Anschluss an die Testungen Analysen vorgesehen sind, die für jede Testperson einen Prozentrang ergeben. Somit kann das Testergebnis unabhängig vom Untersucher interpretiert werden.

7.2 Reliabilität

„Unter Reliabilität oder Zuverlässigkeit eines Tests versteht man den Grad der Genauigkeit, mit der er ein bestimmtes Persönlichkeits- oder Verhaltensmerkmal mißt [sic], gleichgültig, ob er dieses Merkmal auch zu messen beansprucht (...)“ (Lienert & Raatz, 1998, S. 9).

Sollten sich die Items des Tests *Antonyme finden* als Rasch-Modell-konform herausstellen kann die *innere Konsistenz* als gegeben betrachtet werden, da alle Items dasselbe messen.

7.3 Validität

„Die Validität oder Gültigkeit eines Tests gibt den Grad der Genauigkeit an, mit dem dieser Test dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, das (die) er messen oder vorhersagen soll, tatsächlich mißt [sic] oder vorhersagt“ (Lienert & Raatz, 1998, S. 10).

Es können drei Arten der Validität unterschieden werden: *Inhaltliche Validität*, *Konstruktvalidität* sowie *Kriteriumsvalidität* (Lienert & Raatz, 1998; Kubinger, 2009b).
Inhaltliche Validität eines Tests wird erreicht, wenn der Test selbst das optimale Kriterium für das zu erfassende Merkmal darstellt (Lienert & Raatz, 1998). Dieses Validitätskonzept kann bspw. über *Experten-Ratings* hergestellt werden (Kubinger, 2009b), was aber vor allem die ökonomischen Ressourcen dieser Diplomarbeit sprengen würde.
Konstruktvalidität eines Tests ist dann gegeben, wenn er theoriegeleitete Annahmen in Bezug auf ein bestimmtes Konstrukt erfüllt (Kubinger, 2009b). Die Konstruktvalidierung eines Tests kann bspw. mithilfe der Faktorenanalyse überprüft werden. Dies wird in einer Diplomarbeit von Karmann (in Vorbereitung) im Arbeitsbereich Psychologische Diagnostik der Universität Wien realisiert. Darin wird untersucht, inwieweit die sprachlichen Untertests des AID 3 (*Synonyme finden, Antonyme finden* sowie *Funktionen abstrahieren*) das Konstrukt Sprachkompetenz abdecken. Somit sei in Bezug auf die Konstruktvalidität des Untertests *Antonyme finden* auf die Ergebnisse von Karmann (in Arbeit) verwiesen.
Um die *Kriteriumsvalidität* eines Tests zu überprüfen werden die Testergebnisse mit einem sog. *Außenkriterium* korreliert, welches dasselbe Merkmal zu messen beansprucht (Lienert, 1998; Kubinger, 2009b). Der Vorteil gegenüber den vorher genannten Validierungsarten ist die Möglichkeit der Berechnung einer statistischen Maßzahl. Im Falle des Untertests *Antonyme finden* ergibt sich die Möglichkeit, eine *konvergente Validität*[10] mit einem anderen Untertest des AID 3 zu berechnen, der dasselbe Konstrukt erfasst – der Untertest *Synonyme finden*. Die konvergente Validität kann mithilfe des Statistiksoftware PASW (SPSS) überprüft werden.

[10] Der Begriff „konvergente Validität“ zielt darauf ab, dass ein Test mit einem anderen Test, der ein ähnliches Konstrukt erfasst, hoch korrelieren sollte (Rost, 2004).

7.4 Skalierung

„Ein Test erfüllt das Gütekriterium *Skalierung*, wenn die laut Verrechnungsvorschriften resultierenden Testwerte die empirischen Verhaltensrelationen adäquat abbilden“ (Kubinger, 2009b, S. 82).

Der resultierende Testwert des Untertests *Antonyme finden* ist die Summe aller gelösten Items. Dieser kann aber nur ein faires Maß für die erbrachte Testleistung sein, wenn das Rasch-Modell gilt (Kubinger, 2009a, 2009b).
Diese *Verrechnungsfairness* im Sinne des Gütekriteriums Skalierung wird innerhalb dieser Diplomarbeit mithilfe des Rasch-Modells überprüft.

7.5 Fairness

„Ein Test erfüllt das Gütekriterium *Fairness,* wenn die resultierenden Testwerte zu keiner systematischen Diskriminierung bestimmter Testpersonen zum Beispiel aufgrund ihrer ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppenzugehörigkeit führen“ (Kubinger, 2009b, S. 123).

Laut Schober (2003) bezieht sich *Fairness* auf Aspekte, die unmittelbar mit den Inhalten des Tests und seiner Durchführung verknüpft sind. Bezüglich der Testinhalte kann es zu Benachteiligungen aufgrund eines *Item-Bias* kommen. So ist es denkbar, dass Personen aufgrund ihrer Geschlechtszugehörigkeit durch einzelne Items systematisch benachteiligt werden, da jene Items entweder für männliche oder weibliche Testpersonen leichter zu lösen sind. Um diesem Problem entgegenzuwirken, werden bei den Rasch-Modell-Analysen jene Aufgaben entfernt, die für ein Geschlecht höhere Itemschwierigkeitsparameter aufweisen.

Die Durchführungsfairness kann beeinträchtigt sein, wenn beispielsweise Kinder aufgrund sprachlicher Schwierigkeiten die Testinstruktion nicht verstehen. Der Untertest *Antonyme finden* würde in Bezug auf die Durchführung unfair messen, wenn ein schlechtes Testergebnis bei Kindern, deren Muttersprache nicht Deutsch ist, dadurch zustande kommt, dass sie die Testinstruktion nicht verstanden haben. Die Möglichkeit einer sprachfreien Instruktion ist bei einem sprachlichen Untertest definitionsgemäß nicht gegeben. Um den angesprochenen Problemen hinsichtlich der Durchführungsfairness entgegenzuwirken, muss vom (von der) Testleiter(in) am besten vor oder während der Testung entschieden werden, ob das Kind die

Testinstruktion verstanden hat. Ist dies nicht der Fall, dürfen die Testergebnisse des Kindes nicht interpretiert werden.

7.6 Weitere Gütekriterien

Im nachfolgenden Absatz wird kurz auf weitere Gütekriterien eingegangen, ohne genauere Definitionen anzuführen. Für genauere Beschreibungen sei auf Guthke (1996), Kubinger (2003, 2009b), Kubinger & Proyer (2004a) sowie Lienert & Raatz (1998) verwiesen.

Obwohl die Testvorgabe des Untertests *Antonyme finden* innerhalb dieser Diplomarbeit nach der konventionellen Strategie durchgeführt wurde, um Itemschwierigkeitsparameter berechnen zu können, hat der Test eine adaptive Vorgabe als Ziel. Dies beansprucht im Sinne des Gütekriteriums **Ökonomie** relativ wenig Ressourcen, da trotz wenig vorgegebener Aufgaben relativ genau gemessen werden kann. Da der Test im Rahmen des AID 3 einzeln vorgegeben wird, ist der Testvorgabeaufwand im Vergleich zu Gruppenverfahren definitionsgemäß natürlich höher.

Da der Untertest *Antonyme finden* ein Leistungstest ist, erfüllt er weitgehend das Gütekriterium der **Unverfälschbarkeit.**

Um die Testergebnisse verschiedener Personen miteinander vergleichen zu können, benötigt man einen Maßstab (Guthke, 1996). Die Vorgabe des Untertests *Antonyme finden* an einer großen, für den deutschsprachigen Raum repräsentativen Stichprobe wird im Zuge der Normierung des AID 3 erfolgen. Die resultierenden Eichtabellen werden aktuell sein, womit das Gütekriterium **Eichung** erfüllt sein wird.

Die innerhalb dieser Diplomarbeit realisierte konventionelle Testvorgabe kann zu motivationalen Problemen führen, da die Aufgaben in einer aufsteigenden Schwierigkeitsfolge gereiht sind. Ein leistungsschwaches Kind muss somit alle Aufgaben einer Testform bearbeiten, auch wenn es die leichteren nicht beantworten kann. Dies kann zu motivationalen Einbrüchen und Frustration führen und ist im Sinne des Gütekriteriums **Zumutbarkeit** kritisch zu sehen. Wie schon mehrmals erwähnt, ist die konventionelle Vorgabe unumgänglich, wenn ein adaptives Testkonzept geplant ist. Mit der *adaptiven Vorgabe* wird das Kind künftig bei der Bearbeitung des Untertest *Antonyme finden* im Zuge der Testung mit dem AID 3 sowohl in körperlicher, psychischer (insbesondere motivationaler und emotionaler) und zeitlicher Hinsicht *geschont* werden (Kubinger, 2009b).

Da der Untertest *Antonyme finden* zusätzliche Informationen zu sprachlichen Fähigkeiten im AID 3 liefern soll ist ihm generell **Nützlichkeit** zu attestieren.

8 Methode

8.1 Untersuchungsplan

Da die Überarbeitung und Aktualisierung einer Intelligenztestbatterie einen großen Arbeitsaufwand bedeutet, wurden 5 Diplomand(innen) des Arbeitskreises Psychologische Diagnostik mit deren Durchführung betraut. Zeitgleich wurden die Konstruktions- und Durchführungsschritte von Dr. Stefana Holocher-Ertl und Univ.-Prof. Mag. Dr. Klaus Kubinger sowie einigen Mitarbeiter(innen) des Arbeitskreises supervidiert und inhaltlich begleitet.

Obwohl an sich jeder/jede Diplomand(in) mit der Konstruktion oder Überarbeitung eines einzelnen Untertests betraut war, sollte bei der Datenerhebung von jeder/m jeweils die gesamte Rohfassung des AID 3 vorgegeben werden. Das hatte den Vorteil, dass der Stichprobenumfang weitaus größer war, als es durch eine alleinige Vorgabe möglich gewesen wäre. Die Testungen sollten von jedem/jeder Diplomand(in) im gleichen Zeitraum (Jänner bis März 2010) durchgeführt werden, um anschließend eine Auswertung mit dem gesamten Datenmaterial durchführen zu können. Anschließend sollten gegen Ende des Schuljahres (Ende Juni 2010) schriftliche Ergebnisberichte der Leistungen der getesteten Kinder und Jugendliche an die Eltern verschickt werden.

Das verwendete Testmaterial war bei allen Testungen identisch, sodass die Durchführung unter den gleichen Bedingungen stattfinden konnte. Um Fehler bei der Kodierung sowie der Vorgabe der Untertests zu vermeiden und somit die Verrechnungssicherheit zu gewährleisten, wurde allen Diplomand(innen) vor den Testungen kostenlos ein AID 2 – Zertifizierungskurs angeboten. Ferner wurde zu Beginn der Testungen von Dr. Stefana Holocher-Ertl ein Workshop durchgeführt, wo Fragen gestellt und Unklarheiten beseitigt werden konnten.

8.2 Hypothesen

Nicht allein die im AID 3 realisierte *adaptive Testvorgabe* macht es notwendig, den Untertest *Antonyme finden* dahingehend zu überprüfen, ob er dem logistischen Testmodell von Rasch entspricht. Auch um feststellen zu können, ob der Test *eindimensional* misst und die

Verrechnung der Testleistung zu Testwerten im Sinne des Gütekriteriums Skalierung *fair* ist, muss der Test auf Rasch-Modell-Konformität überprüft werden.

Auch in Bezug auf das Gütekriterium *Fairness* kann man im Zuge der Modellprüfung feststellen, ob der Untertest *Antonyme finden* Personen in Bezug auf ihre Geschlechtszugehörigkeit und Muttersprache systematisch benachteiligt.

Daraus ergeben sich eine Haupthypothese sowie zwei Nebenhypothesen:

H_0-1: Die Items des Untertests *Antonyme finden* entsprechen dem Rasch-Modell.

H_1-1: Die Items des Untertests *Antonyme finden* sind nicht Rasch-Modell-konform.

H_0-2: Es kommt durch die resultierenden Testwerte des Untertest *Antonyme finden* zu keiner Benachteiligung von Personen in Bezug auf ihre Geschlechtszugehörigkeit.

H_1-2: Der Untertest *Antonyme finden* benachteiligt ein Geschlecht.

H_0-3: Es kommt durch die resultierenden Testwerte des Untertest *Antonyme finden* zu keiner Benachteiligung von Personen in Bezug auf ihre Muttersprache.

H_1-3: Der Untertest *Antonyme finden* benachteiligt Personen hinsichtlich ihrer Muttersprache.

8.3 Erhebungsinstrument

Als Erhebungsinstrument diente die überarbeitete und aktualisierte Form des AID 2 – die Rohform des AID 3. Die Untertests *„Alltagswissen“*, *„Realitätssicherheit“, „Angewandtes Rechnen“, „Synonyme finden“, „Funktionen abstrahieren“ und „Soziales Erfassen und Sachliches Reflektieren“*, die auch im AID 2 enthalten sind, wurden in der Konstruktionsphase inhaltlich überarbeitet und in der aktualisierten Form im AID 3 vorgegeben. Der AID 3 enthält zusätzlich drei neu konstruierte Untertests: *„Visuelle Merkfähigkeit“, „Antonyme finden“* und *„Formale Folgerichtigkeit“*.

Der Untertest *„visuelle Merkfähigkeit“* erfasst die kurzfristige Merkfähigkeit bei visuellem Stimulusmaterial. Der Untertest *„Formale Folgerichtigkeit“* dient der Erfassung von Reasoning bei figuralem Aufgabenmaterial (Hagenmüller, in Vorbereitung).

Die 6 aktualisierten Untertests des AID 2 gemeinsam mit den drei neu konstruierten Tests ergaben 9 Subtests, die jeder Testperson vorgegeben werden. Die weiteren Untertests des AID

2 *„Soziale und Sachliche Folgerichtigkeit“, „Unmittelbares Reproduzieren – numerisch“, „Kodieren und Assoziieren“, „Antizipieren und Kombinieren – figural“,* sowie *„Analysieren und Synthetisieren – abstrakt“,* wurden einer von 4 Testzusammenstellungen als Zusatztests zugeordnet. Die Untertests wurden ebenfalls in der Vorphase überarbeitet und mit neuen Items versehen. Der Test sollte bei der gesamten Datenerhebung in allen 4 Testzusammenstellungen etwa gleich oft vorgegeben werden, um für die Zusatztests eine vergleichbar große Datenmenge zu erhalten.
Die optionalen Zusatztests des AID 2 („*Unmittelbares Reproduzieren – figural/abstrakt“, „Merken und Einprägen“* sowie *„Strukturieren – visumotorisch*“) wurden im Rahmen des AID 3 nicht vorgegeben.

8.3.1 Vorgabe des Untertests Antonyme finden

Wie bereits in Abschnitt 6.8 beschrieben wurde, wurden für 3 unterschiedliche Alterskategorien jeweils 2 Testhefte erstellt. Somit beinhaltete das Testmanual 6 unterschiedliche Testformen, wobei eine Testform aus jeweils 20 Items bestand, die konventionell vorgegeben wurden. Die Testperson musste somit alle Items der ihr vorgegebenen Testform bearbeiten. Aus urheberrechtichen Gründen können die Items des Untertests *Antonyme finden* innerhalb dieser Diplomarbeit nicht angeführt werden. Nachfolgend werden aber zwei Beispielitems genannt.

Beispielitem 1: *Sag´ mir das Gegenteil von „warm“.*

Beispielitem 2*: Sag´ mir das Gegenteil von „nass“.*

Die Lösung des ersten Beispielitems wäre „kalt“, die des zweiten Beispielitems „trocken“. Die standardisierte Instruktion des Untertests *Antonyme finden,* die jeder Testperson zu Beginn der Durchführung des Untertests verbal vorgegeben wurde, ist im Anhang zu finden.

8.4 Stichprobe

8.4.1 Aquirierung der Stichprobe

Nachdem die Arbeiten zur Konstruktion bzw. Überarbeitung und Aktualisierung abgeschlossen waren, wurden im Herbst 2009 für die Stichprobenaquirierung Schulen gesucht, die sich bereit erklärten bei der Datenerhebung mitzuwirken. Dazu wurde an etliche Schulen im Raum Wien und Niederösterreich ein Lehrer(innen)brief verschickt, teilweise wurde die Untersuchung vor Ort an den Schulen vorgestellt. Der Lehrer(innen)brief ist im Anhang zu finden. Acht Schulen aus Wien und zwei Schulen aus Niederösterreich erklärten sich bereit, an der Untersuchung teilzunehmen. Für die offizielle Bewilligung wurde dem Stadtschulrat Wien im Oktober 2009 eine Beschreibung der Untersuchung zugesendet. Diese kann ebenfalls im Anhang nachgelesen werden. Im November 2009 wurden die geplanten Erhebungen vom Stadtschulrat bewilligt, sodass eine genauere Planung mit den Schulen beginnen konnte.

Die Schule, an der Diplomand die Testungen durchführte, war das BRG 9 Erich-Fried-Realgymnasium in Wien, welches von ihm in seiner Schulzeit selbst 8 Jahre lang besucht wurde.

Die ursprüngliche Vorgabe war, von der 1. – 5. Schulstufe mindestens 50-70 Schüler(innen) zu testen. Die Anzahl der Testungen sowie das Geschlechterverhältnis sollten pro Schulstufe in etwa gleich verteilt sein. Pro Schulstufe wurden zwei Klassen gewählt, deren Klassenvorständen ich den Lehrerinnenbrief sowie die Elternbriefe samt den Einverständniserklärungen zur Testung zukommen ließ. Im Elternbrief wurde neben einer kurzen Projektbeschreibung den Eltern in Aussicht gestellt, ihnen einen schriftlichen Ergebnisbericht über die intellektuellen Stärken und relativen Schwächen ihres Kindes zuzuschicken, falls sie ihr Kind teilnehmen ließen. Der Elternbrief sowie ein Muster des schriftlichen Ergebnisberichts sind ebenfalls im Anhang nachzulesen. Insgesamt wurden 246 Elternbriefe ausgeteilt, woraus 118 Zusagen resultierten. Das ergibt eine Rücklaufquote von fast 48%, somit zeigte sich die Hälfte der Eltern bereit, ihr Kind an der Untersuchung teilnehmen zu lassen. Die hohe Rücklaufquote hängt möglicherweise mit dem Umstand zusammen, dass viele Lehrer vom Diplomanden persönlich angesprochen wurden und um deren Engagement gebeten wurde. Bei der Analyse der Rücklaufquoten pro Klasse fällt auf, dass ein deutlich höherer Prozentsatz an Elternbriefen unterschrieben wurde, wenn ich den

Klassenvorstand der Klasse zuvor persönlich angesprochen hatte. Eine Übersicht über die Rücklaufquote liefert Tabelle 6.

Tabelle 6: *Rücklaufquote pro Klasse*

Klasse	Anzahl Schüler	Zusagen	Rücklaufquote
1B	25	8	32%
1C	26	12	46%
2A	25	17	68%
2C	25	16	64%
3A	22	18	82%
3B	20	9	45%
4A	23	11	48%
4B	22	4	18%
5A	29	14	48%
5B	29	9	31%

Anmerkung: Die grau unterlegten Felder markieren jene Klassen, mit deren Klassenvorständen vor Ausgabe der Elternbriefe ein persönliches Gespräch stattfand.

Mit Ausnahme der 1C liegen die Rücklaufquoten jener Klassen alle über 50% (durchschnittlich 65%), während aus den anderen Klassen weniger als die Hälfte (durchschnittlich 37%) der Elternbriefe unterschrieben zurückgesendet wurden. Aus der Klasse, deren Klassenvorstand meine Ansprechperson für die Testungen war, erklärten sich gar 82% der Eltern bereit, ihr Kind an der Testung teilnehmen zu lassen.

Der Unterschied in den Rücklaufquoten ist insofern interessant, da die Klassenvorstände die Eltern nicht direkt motivieren konnten, ihr Kind an der Testung teilnehmen zu lassen, da die Information ausschließlich über den Elternbrief übermittelt wurde. Es muss also einen indirekten Effekt auf die elterliche Entscheidung gegeben haben. Es ist naheliegend, anzunehmen, dass jene Lehrer, denen vom Diplomanden persönlich der Sinn und Nutzen der Untersuchung erklärt werden konnte, den Elternbrief mit einer anderen Erklärung ausgeteilt haben als jene Lehrer, mit denen ich nicht persönlich sprechen konnte. Es scheint ihnen

gelungen zu sein, das Interesse der Kinder für die Testung zu wecken. Dass sich allerdings ein derart großer Unterschied in der Rücklaufquote ergibt, war nicht zu erwarten, da sehr viele Faktoren entscheidend sind, ob Eltern ihre Zustimmung zu einer Testung geben oder nicht.
Eine Non-Responder-Analyse[11] ist definitionsgemäß schwer durchführbar, allerdings ließen einige Eltern die Elternbriefe auch dann zurückkommen, wenn sie ihr Kind nicht teilnehmen ließen. Eine qualitative Analyse der Einverständniserklärungen sowie ein Gespräch mit einigen Klassenvorständen ergab, dass viele Eltern einerseits grundsätzlich Angst bzw. Bedenken hätten, ihr Kind *testen* zu lassen, andererseits seien in letzter Zeit an jener Schule in einigen Schulstufen verpflichtend Leistungserhebungen des Bildungsministeriums durchgeführt worden, weswegen Eltern teilweise kritisch reagierten, wenn sie erneut mit einer Testung ihres Kindes konfrontiert waren.

8.4.2 Beschreibung der Teilstichprobe

Im Zeitraum von Jänner bis März 2010 konnten 125 Kinder getestet werden, obwohl nur 118 Kinder eine Einverständniserklärung der Eltern abgegeben hatten. Einige Kinder brachten die Einverständniserklärungen erst während des Zeitraums, in dem die Testungen stattfanden. Andere Kinder wurden auch durch die Berichte der bereits getesteten Klassenkameraden neugierig und nahmen ebenfalls an der Testung teil, vorausgesetzt, sie hatten die Einverständniserklärung unterschreiben lassen. Da die Kinder einzeln und nicht in der Gruppe getestet werden mussten, wurde von der Schule ein Raum zur Verfügung gestellt, in dem die Testungen weitgehend ungestört durchgeführt werden konnten. Dabei wurde sehr darauf geachtet, dass die Kinder nicht während den Hauptgegenständen oder in Fächern, wo sie gefährdet waren, getestet wurden, sondern nur, wenn sowohl der/die Schüler(in) als auch die Lehrkraft einverstanden waren. Die Testdauer erstreckte sich von 55 bis 105 Minuten, wobei tendenziell Testungen in den 1. und 2. Klassen schneller durchgeführt werden konnten als in den Schulstufen 3, 4 & 5. Auf eine deskriptive Analyse der Teilstichprobe wird hier verzichtet, da für die statistische Analyse die Daten aller Schulen verwendet wurden. Die Gesamtstichprobe wird im nächsten Abschnitt genau beschrieben.

[11] Eine „Non-Responder-Analyse“ ist eine Untersuchung jener Personen, die eine Testung verweigern bzw. nicht freiwillig an einer Testung teilnehmen (Kubinger, 2009b).

8.4.3 Beschreibung der Gesamtstichprobe

Die Daten wurden hinsichtlich der Verteilung in Bezug auf die Variablen *Schulform*, *Geschlecht*, *Alter* sowie *Muttersprache* analysiert. Insgesamt wurden 711 Kinder getestet, wovon 16 Kinder wegen fehlender Angaben ausgeschlossen werden mussten. Der endgültige Datensatz, mit dem auch die statistischen Analysen durchgeführt wurden, umfasste somit 695 Schüler(innen).

8.4.3.1 Schulform

Tabelle 7 gibt die Häufigkeit sowie den Prozentsatz der getesteten Schüler(innen) für die Variable *Schulform* an. Die größte Anzahl an Kindern wurde in der Volkssschule sowie im Gymnasium getestet. Die Testpersonen aus Hauptschulen, Kooperativen Mittelschulen und Berufsbildenden höheren Schulen machen zusammen 20 % der Stichprobe aus. Abbildung 3 veranschaulicht grafisch die Verteilung in Bezug auf die unterschiedlichen Schulformen.

Tabelle 7: *Deskriptive Statistik der Variable Schulform*

Schulform	Häufigkeit	Anteil in Prozent (%)
Volkssschule	286	41.2
Gymnasium	269	38.7
Hauptschule	56	8.1
Kooperative Mittelsschule	67	9.6
Berufsbildende höhere Schule	17	2.4
Gesamt	695	100

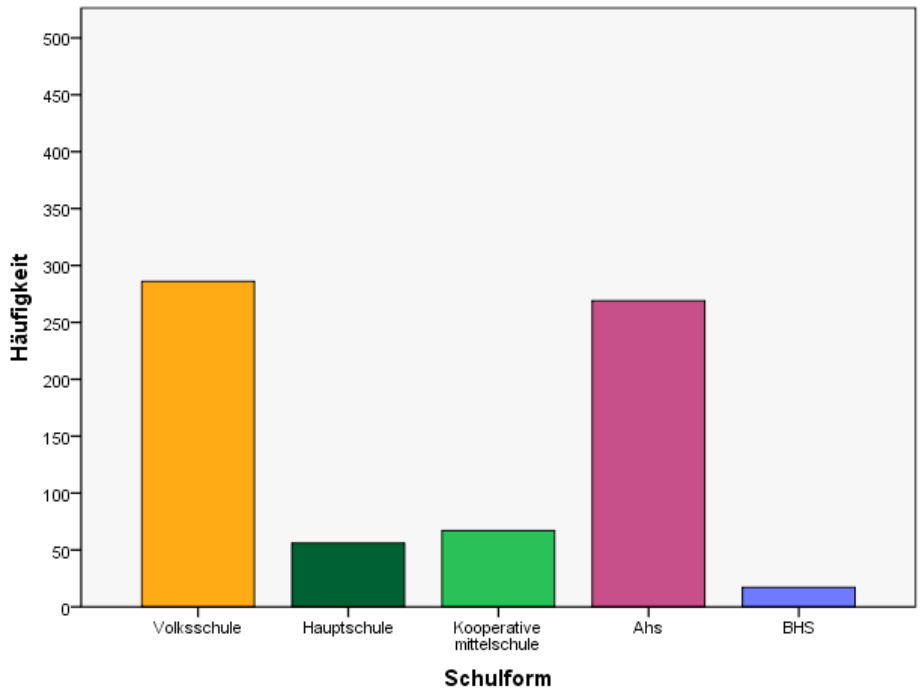

Abbildung 3: Balkendiagramm der Variable *Schulform*

8.4.3.2 Alter & Geschlecht

Tabelle 8 zeigt die Verteilung der Testungen über die Variablen *Geschlecht* und *Alter*. Die Variable *Geschlecht* ist annähernd gleich verteilt. Es wurden insgesamt 323 Schüler (46.5%) und 372 Schülerinnen (53.5% getestet). Die Variable *Alter* folgt eher dem Bild einer Normalverteilung (siehe Abbildung 4). Die meisten der getesteten Kinder befinden sich in der Altersgruppe von 10 – 12 Jahren. Die wenigsten Testungen wurden bei den 6- und 15-jährigen durchgeführt. Abbildung 4 zeigt die Verteilung der Variable *Alter*, wobei jeder Altersbereich nach *Geschlecht* geteilt ist.

Tabelle 8: *Deskriptive Statistik der Variablen Geschlecht & Alter*

Alter in Jahren	**Geschlecht**		**Gesamt**
	männlich	weiblich	
6	13	23	**36**
7	33	29	**62**
8	35	25	**60**
9	34	38	**72**
10	52	42	**94**
11	42	51	**93**
12	39	51	**90**
13	29	35	**64**
14	30	49	**79**
15	15	29	**44**
Gesamt	**323**	**372**	**695**

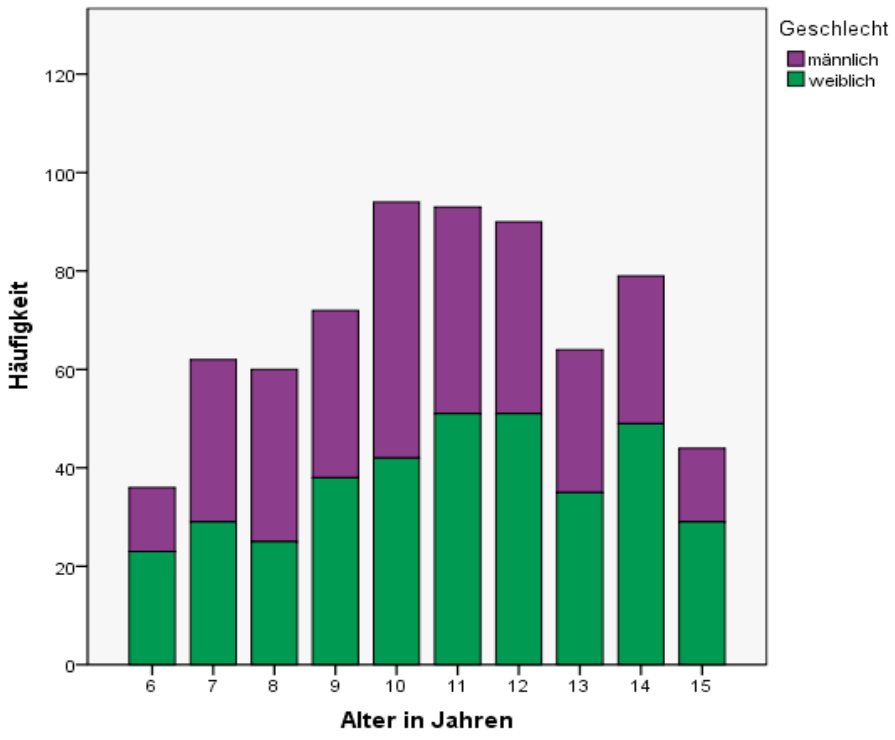

Abbildung 4: Balkendiagramm der Variablen *Geschlecht* & *Alter*

8.4.3.3 Muttersprache

Aus Tabelle 9 wird ersichtlich, dass annähernd 2/3 der getesteten Kinder Deutsch als Muttersprache angegeben haben. 31.6 % haben demnach eine andere Muttersprache. Davon ist BKS[12] die Gruppe jener Muttersprachen mit dem höchsten prozentuellen Anteil an der Gesamtstichprobe (10.6%), gefolgt von Türkisch (6.3 %). 14.8 % der Schüler(innen) sind einer von 6 weiteren Muttersprachen(gruppen) zuzuordnen. Abbildung 5 veranschaulicht grafisch die Verteilung der Variable *Muttersprache*.

Tabelle 9: *Deskriptive Statistik der Variable Muttersprache*

Muttersprache	**Häufigkeit**	**Anzahl in Prozent (%)**
Deutsch	475	68.3
Türkisch	44	6.3
BKS	73	10.5
Andere Muttersprache	103	14.8
Gesamt	695	100

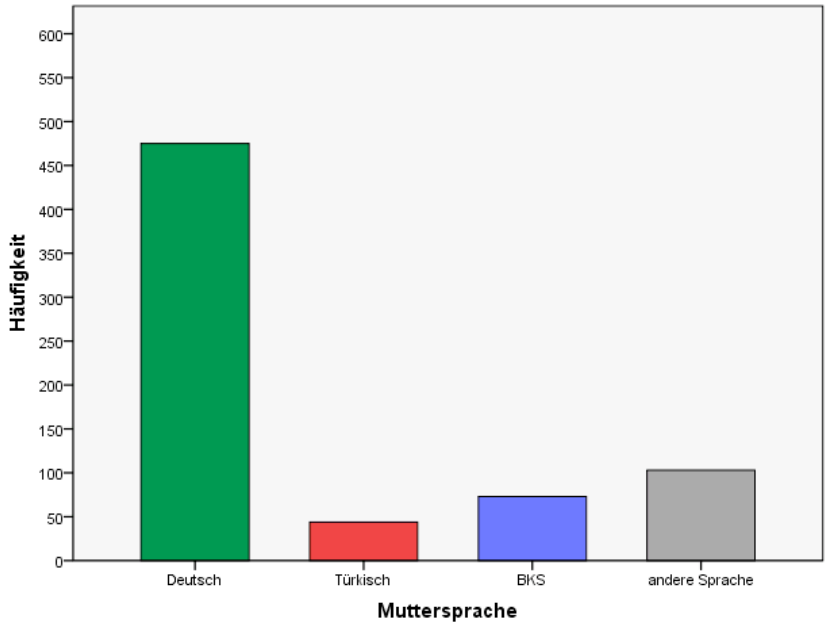

Abbildung 5: Balkendiagramm der Variable *Muttersprache*

[12] BKS gilt als Abkürzung für die Sprachen Bosnisch & Kroatisch, Serbisch; zusätzlich wurde auch die Sprache Slowenisch zu dieser Gruppe hinzugefügt.

9 Ergebnisse

Zur Prüfung der Hypothesen werden die Items des Untertests *Antonyme finden* auf ihre Rasch-Modell-Konformität überprüft. Die Daten wurden mithilfe des Statistiksoftware PASW 18 (Predictive Analysis SoftWare) archiviert. Für die Rasch-Modell-Berechnungen wurde das Programm R Version 2.12.0 gemeinsam mit dem Paket eRm (extended Rasch modelling) von Mair & Hatzinger (2009) verwendet. Die Übereinstimmungsvalidität wurde ebenfalls mit dem Paket PASW 18 berechnet.

Der ursprüngliche Datensatz beinhaltete die Testwerte von 711 Kindern. 13 Kinder mussten vom Datensatz ausgeschlossen werden, da ihnen der Untertest *Antonyme finden* nicht vorgegeben werden konnte. Den Berichten der anderen Diplomandinnen zufolge verstanden einige Kinder aufgrund von schlechten Deutschkenntnissen die Instruktion nicht, sodass der Untertest richtigerweise nicht durchgeführt wurde. Bei drei weiteren Kindern war die Muttersprache nicht angegeben, worauf sie ebenfalls aus dem Datensatz ausgeschlossen wurden. Folglich resultierte ein Datensatz mit Testwerten von 695 Kindern, mit dem die Analyse durchgeführt wurde.

9.1 Überprüfung des Untertests *Antonyme finden* auf Geltung des Rasch-Modells

Die Daten des Untertests *Antonyme finden* wurden zunächst inferenzstatistisch mithilfe des Likelihood-Ratio-Tests (LR-Test) von Anderson überprüft. Dazu wurde die Stichprobe anhand folgender Kriterien geteilt:

Internes Teilungskriterium

- Rohscore (niedriger vs. hoher Rohscore, geteilt durch den Median)

Externe Teilungskriterien

- Geschlecht (männliche vs. weibliche Testpersonen)
- Alter (<11 Jahre vs. ≥11 Jahre)
- Muttersprache (Deutsch vs. andere Muttersprache)

Fällt der Modelltest hinsichtlich eines Teilungskriteriums (TK) signifikant aus *(α=.01)* werden unter Zuhilfenahme weiterer Modelltests (grafischer Modelltest und Wald-Test) nicht modell-konforme Items identifiziert und sukzessive ausgeschieden. Auch die Rückmeldung anderer Testleiter(innen) aus ihren Testerfahrungen bezüglich inhaltlich kritischer Items wird bei diesem Schritt berücksichtigt. Der Ausschlussprozess wird solange sukzessive fortgeführt, bis sich hinsichtlich der genannten Teilungskriterien keine signifikante Modellabweichung mehr feststellen lässt. Der Test erwiese sich somit *a posteriori* (im Nachhinein) Rasch-Modell-konform. Sollte der LR-Test eines Teilungskriteriums nach dem Itemausschluss noch immer signifikant ausfallen, muss zur Beurteilung der Modellgültigkeit der grafische Modelltest miteinbezogen werden.

9.2 Erste Modellprüfung

9.2.1 Teilungskriterium Rohscore

Der LR-Test erbrachte im Bezug auf das Teilungskriterium *Rohscore* ein signifikantes Ergebnis. Die Hypothese H_0-1: „Die Items des Untertests *Antonyme finden* entsprechen dem Rasch-Modell“ muss demnach zunächst verworfen werden. Tabelle 10 gibt bezüglich des Teilungskriteriums *Rohscore* die asymptotisch χ^2-verteilten Testgrößen des LR-Tests, die Anzahl berücksichtigter Aufgaben (df[13]), die Wahrscheinlichkeit, dass die H_0 gilt (p-Wert) sowie die kritischen Werte der χ^2-Verteilung bei *(α=.01)* an.

Für den ersten Berechnungsdurchgang konnten 10 Items aufgrund ungünstiger Antwortmuster nicht in die Analyse miteinbezogen werden. Da jene Items bei Berechnungen mit anderen Teilungskriterien sehr wohl geschätzt werden konnten, mussten sie nicht vom Itempool ausgeschlossen werden.

Abbildung 6 zeigt den grafischen Modelltest für das Teilungskriterium *Rohscore* über alle in die Analyse miteinbezogenen Items. Ein Item ist als nicht modell-konform zu bewerten, wenn die Konfidenz-Ellipse die 45°-Gerade nicht schneidet. Abbildung 7 stellt diejenigen Aufgaben dar, die dem Modell nicht entsprechen. Bei der Betrachtung der grafischen Modelltests fällt auf, dass viele Items eine gute Passung zeigen, während im mittleren Fähigkeitsbereich 9

[13]Die Freiheitsgrade (df) beschreiben die Beobachtungswerte einer Stichprobe, die voneinander unabhängig sind. Df=53 bedeutet im vorliegenden Fall, dass 54 Aufgaben in die Analyse eingegangen sind. Eine detailierte Beschreibung liefert Bortz (2005).

Items mit dem Modell nicht konform sind. Als dritter Modelltest wurde der Wald-Test durchgeführt, der 8 signifikante Items identifizierte, die dem Rasch-Modell nicht entsprechen. Die Ergebnisse der Wald-Tests für alle Berechnungsschritte werden aus Gründen der Übersichtlichkeit im Anhang dargestellt. Die Ergebnisse des Wald-Tests in Bezug auf das Teilungskriterium *Rohscore* sind unter Tabelle 25 zu finden.

Tabelle 10: *LR-Test für das TK „Rohscore", erster Berechnungsdurchgang*

Teilungskriterium Rohscore			
Andersen χ^2	*df*	*p-Wert*	*Kritischer* χ^2 *–Wert*
143.19	56	<.001	83.51

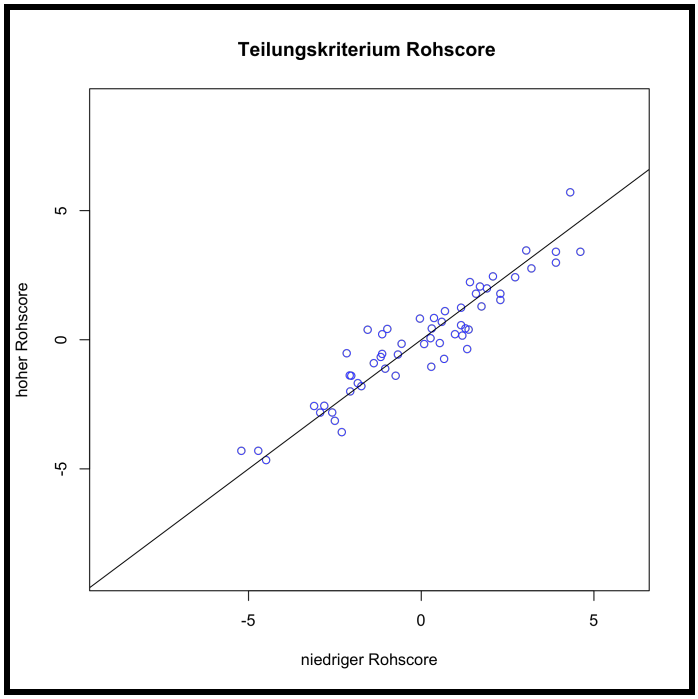

Abbildung 6: Grafischer Modelltest, TK *Rohscore*

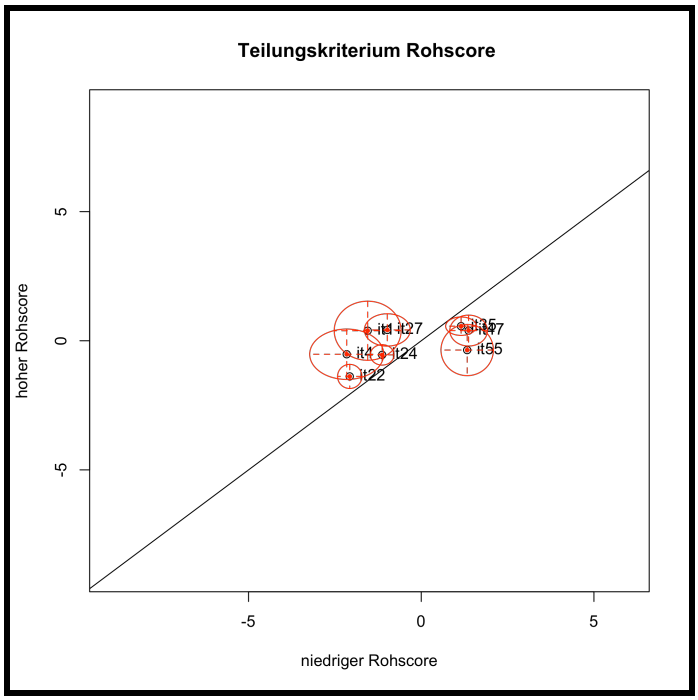

Abbildung 7: Grafischer Modelltest, TK *Rohscore*, nicht modell-konforme Items

9.2.2 Teilungskriterium Geschlecht

In Bezug auf das Teilungskriterium *Geschlecht* wird der LR-Test ebenfalls signifikant (siehe Tabelle 11). Zwei Items mussten aufgrund ungünstiger Antwortmuster von der Analyse ausgeschlossen werden. Abbildung 8 zeigt den grafischen Modelltest, Abbildung 9 die nicht modellkonformen Items mit zugehörigen Konfidenz-Ellipsen. Viele Items liegen nahe der 45°-Geraden, die Itemschätzungen im unteren Fähigkeitsbereich weichen zwar von der Geraden ab, weisen aber große Konfidenzintervalle auf, sodass sie als noch mit dem Modell konform angesehen werden können. Vier Items weisen hingegen keine Modellanpassung auf. Der Wald-Test ergibt, dass ebenfalls vier Items signifikant sind und somit als nicht Rasch-Modell-konform gelten (siehe Tabelle 26 im Anhang).

Tabelle 11: *LR-Test für das TK „Geschlecht", erster Berechnungsdurchgang*

Teilungskriterium Geschlecht			
Andersen χ^2	*df*	*p-Wert*	*Kritischer* χ^2 *–Wert*
167.69	64	<.001	93.22

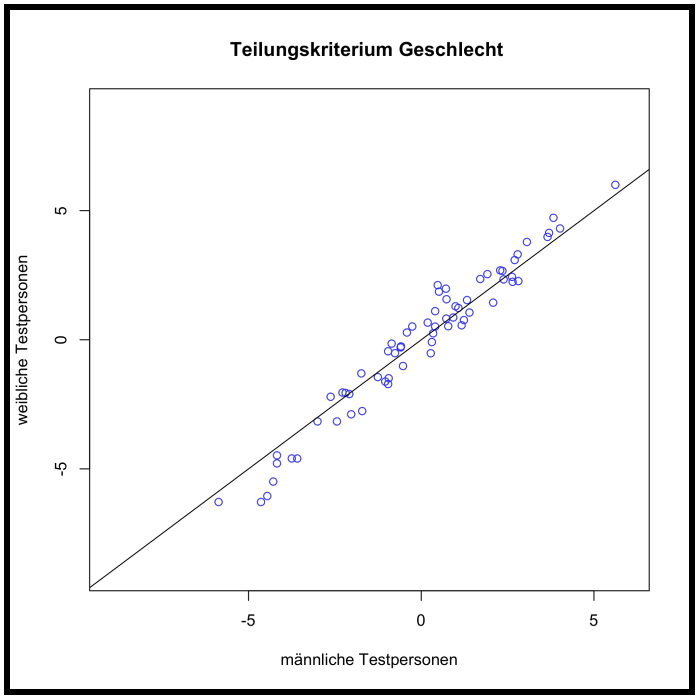

Abbildung 8: Grafischer Modelltest, TK *Geschlecht*

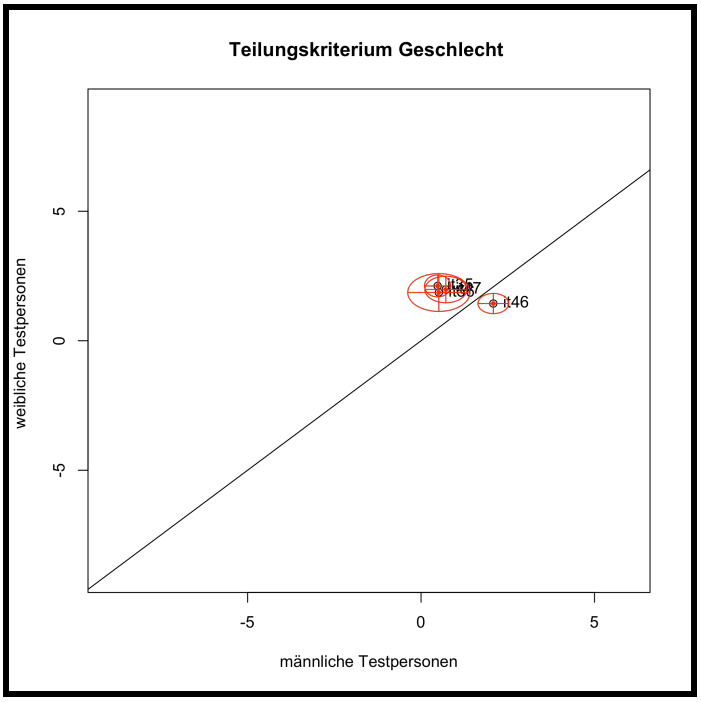

Abbildung 9: Grafischer Modelltest, TK *Geschlecht*, nicht modell-konforme Items

9.2.3 Teilungskriterium Muttersprache

Der LR-Test für das Teilungskriterium Muttersprache erbringt ein signifikantes Ergebnis (siehe Tabelle 12). Die Analyse wurde mit 61 Items durchgeführt, folglich konnten 6 Items nicht in die Analyse miteinbezogen werden. Abbildung 10 zeigt die Grafische Modellkontrolle, Abbildung 11 diejenigen Items, die dem Modell nicht entsprechen. Der grafische Modelltest dieses Teilungskriteriums fällt schlechter aus als die der anderen Teilungskriterien. Die Items streuen mehr um die 45°-Gerade und die Konfidenz-Ellipsen von 10 Items schneiden die Gerade nicht. Der Wald-Test identifiziert 11 signifikante Items (siehe Tabelle 27 im Anhang).

Tabelle 12: *LR-Test für das TK „Muttersprache", erster Berechnungsdurchgang*

Teilungskriterium Muttersprache			
Andersen χ^2	*df*	*p-Wert*	*Kritischer* χ^2 *–Wert*
214.53	60	<.001	88.38

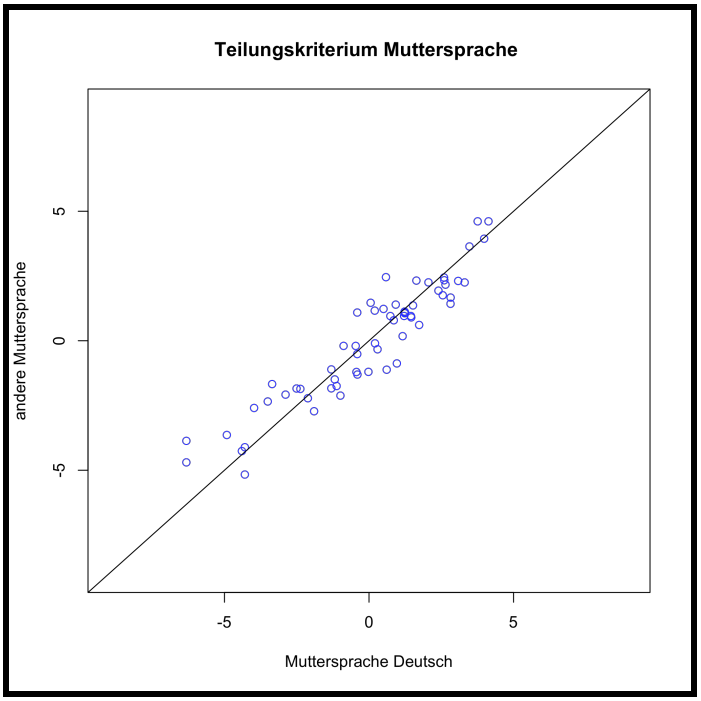

Abbildung 10: Grafischer Modelltest, TK *Muttersprache*

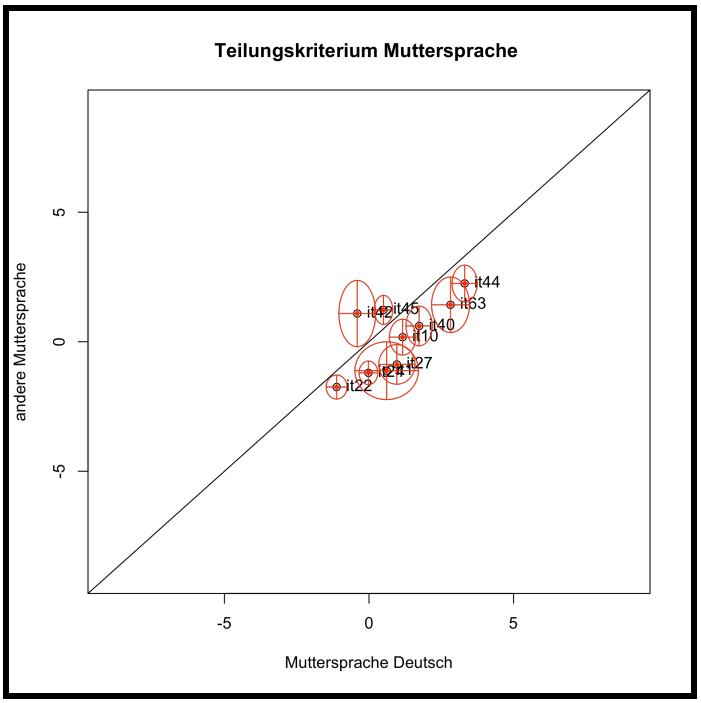

Abbildung 11: Grafischer Modelltest, TK *Muttersprache*, nicht modell-konforme Items

9.2.4 Teilungskriterium Alter

Der LR-Test für das Teilungskriterium Alter fällt zwar ebenfalls signifikant aus, der kritische χ^2 –Wert liegt allerdings nur noch knapp unter dem empirischen χ^2-Wert (siehe Tabelle 13). Es wurden 29 Items in die Analyse miteinbezogen. Die geringe Anzahl an Aufgaben hat den Grund, dass je nach Altersgruppe unterschiedliche Testformen mit verschiedenen Items vorgegeben wurden, sodass für etliche Items eine Parameterschätzung in einer Teilgruppe nicht möglich war. Die Items, die jedoch durch die verzweigte Vorgabe geschätzt werden konnten, wiesen eine gute grafische Modellpassung auf (siehe Abbildung 12) Viele Items werden in beiden Teilstichproben exakt gleich geschätzt und liegen somit auf der 45°-Geraden. Die restlichen Itemschätzungen liegen sehr nahe der Geraden, nur zwei Items erweisen sich in als nicht modell-konform (siehe Abbildung 13). Auch im Wald-Test werden zwei Items signifikant (siehe Tabelle 28 im Anhang).

Tabelle 13: *LR-Test für das TK „Alter“, erster Berechnungsdurchgang*

Teilungskriterium Alter			
Andersen χ^2	*df*	*p-Wert*	*Kritischer* χ^2 *–Wert*
53.39	28	<.001	48.28

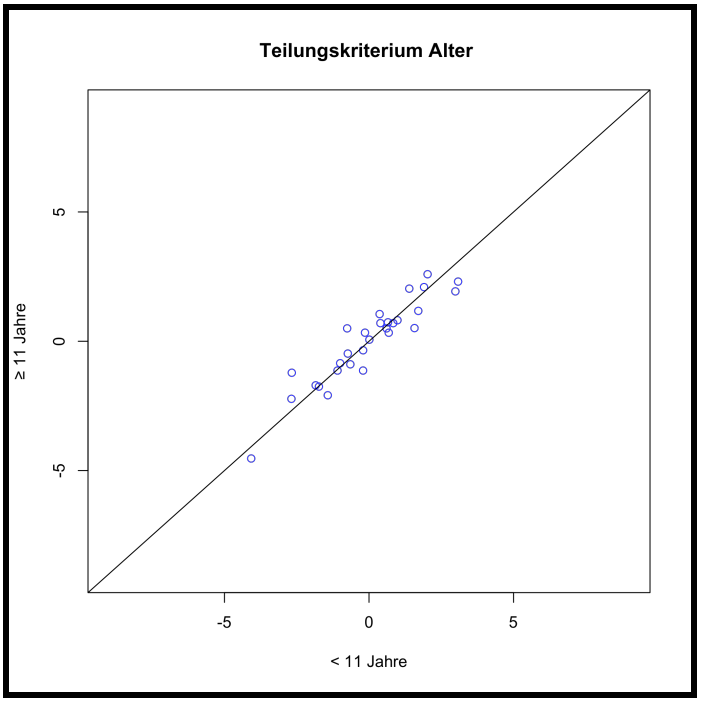

Abbildung 12: Grafischer Modelltest, TK *Alter*

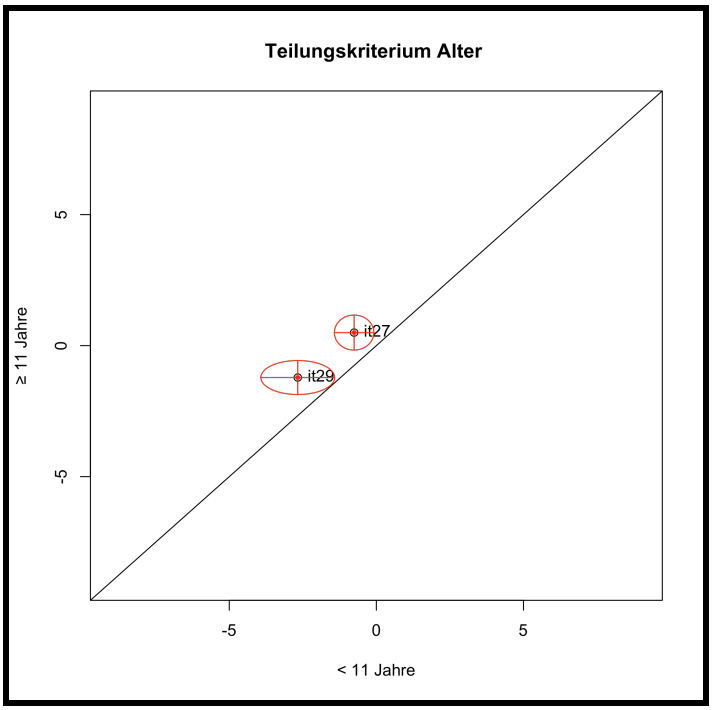

Abbildung 13: Grafischer Modelltest, TK *Alter*, nicht modell-konforme Items

9.3 Ausschluss nicht Rasch-Modell-konformer Items

Die Tabellen 14-18 geben jene Items an, die sich in der ersten Modellprüfung als nicht Rasch-Modell–konform erwiesen haben. Dafür wurden die Ergebnisse aus den grafischen Modellkontrollen und Wald-Tests aller Teilungskriterien zusammengeführt. Die signifikanten Items der Wald-Tests sind in den Tabellen 25-28 im Anhang grau markiert. Bis auf ein Item erwiesen sich alle kritischen Aufgaben in beiden Modelltests als nicht modellkonform. Die signifikanten bzw. nicht modellkonformen Items sind grau unterlegt. Somit lässt sich feststellen, welche Items in Bezug auf mehr als ein Teilungskriterium auffällig sind.

Tabelle 14: *Nicht Rasch-Modell-konforme Items*

Teilungs-kriterien	Items														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Rohscore															
Geschlecht															
Mutterspr.															
Alter															

Tabelle 15: *Nicht Rasch-Modell-konforme Items*

Teilungs-kriterien	Items														
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Rohscore															
Geschlecht															
Mutterspr.															
Alter															

Tabelle 16: *Nicht Rasch-Modell-konforme Items*

Teilungs-kriterien	Items														
	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
Rohscore															
Geschlecht															
Mutterspr.															
Alter															

Tabelle 17: *Nicht Rasch-Modell-konforme Items*

Teilungs-kriterien	Items														
	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Rohscore															
Geschlecht															
Mutterspr.															
Alter															

Tabelle 18: *Nicht Rasch-Modell-konforme Items*

Teilungs-kriterien	Items						
	61	62	63	64	65	66	67
Rohscore							
Geschlecht							
Mutterspr.							
Alter							

Die Analyse der nicht modellkonformen Items ergibt, dass Item 27 in *drei* Teilungskriterien signifikant wird. In jeweils *zwei* Teilungskriterien werden die Items 1, 22, 24, 35 und 47 signifikant. Interessant ist der Umstand, dass die Mehrzahl der genannten Items von den übrigen Diplomandinnen auch *inhaltlich* kritisch bewertet wurde. Viele Items fallen aufgrund von Erfahrungswerten bei der Vorgabe in eines der definierten Ausschlusskriterien. So wurde auf das Item 27: *Nenne mir das Gegenteil von „gestern"* (Lösung: „*morgen*") von vielen Kindern die Antwort „*heute*" gegeben. Das Wort gestern wurde bei der Konstruktion als „der Tag *vor* dem heutigen Tag" definiert, dessen Antonym „der Tag *nach* dem heutigen Tag", demnach „*morgen*" wäre. Umgangssprachlich wird aber auch das Wort „heute" als Antonym zu „gestern" betrachtet. Folglich hat das Item „*gestern*" zwei Lösungen und ist demnach auch aufgrund inhaltlicher Mängel aus dem Itempool auszuscheiden. Ein weiteres Beispiel für ein nicht modellkonformes Item ist Item 35: „Nenne mir das Gegenteil des Wortes *zärtlich*", (Lösung: „*grob*", „*unsanft*"). Das Item fällt auch nach Analyse der Antworten nicht in eine Ausschlusskategorie, aber es verstößt gegen die Eindimensionalität der Messung. Das Item fällt männlichen Testpersonen schwerer als weiblichen Testpersonen. Damit wird neben sprachlicher Fähigkeit auch in gewisser Weise das Geschlecht gemessen, ein Umstand, der in diesem Test nicht vorgesehen ist.

Der Methode des sukzessiven Itemausschlusses zufolge wurde zunächst das Item 27 ausgeschlossen und anschließend wurden erneut Modellprüfungen für alle Teilungskriterien durchgeführt. Da die Ergebnisse aller LR-Tests noch immer signifikant waren, wurden mittels Wald-Test und grafischem Modelltest erneut Items gesucht, die dem Rasch-Modell nicht entsprechen. Dieser Prozess wurde solange fortgeführt, bis der LR-Test für drei Teilungskriterien nicht signifikant wurde. In diesem Zuge wurden die Items 1, 22, 24, 27, 35, 47, 55, 56 aus dem Itempool ausgeschlossen.

9.4 Letzter Berechnungsdurchgang

9.4.1 Teilungskriterium Rohscore

Nach Ausschluss von 8 Items wurde der LR-Test im Bezug auf das Teilungskriterium *Rohscore* nicht signifikant (siehe Tabelle 19). Der empirische χ^2 –Wert liegt folglich unter dem kritischen χ^2 –Wert. Abbildung 14 zeigt den grafischen Modelltest für das Teilungskriterium *Rohscore*. Sieht man von einigen Aufgaben ab, streuen die Parameterschätzungen der Items eng um die 45°-Gerade. Aber auch jene Items, die etwas

weiter entfernt liegen, passen zum Modell, da die jeweiligen Konfidenz-Ellipsen alle die Gerade schneiden (siehe Abbildung 15). Im Wald-Test ergeben sich für alle Items nicht signifikante Ergebnisse (siehe Tabelle 29 im Anhang).

Tabelle 19: *LR-Test für das TK „Rohscore", letzter Berechnungsdurchgang*

Teilungskriterium Rohscore			
Andersen χ^2	*df*	*p-Wert*	*Kritischer* χ^2 *–Wert*
71.57	48	.015	73.68

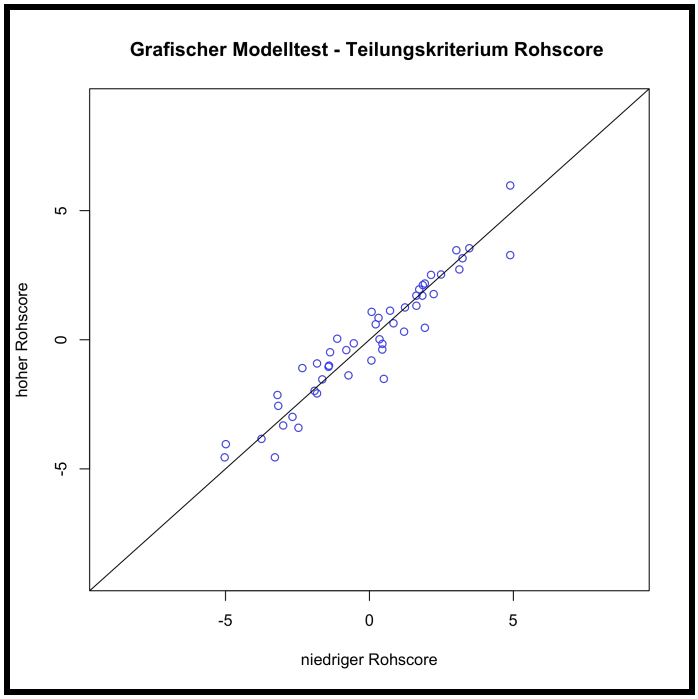

Abbildung 14: Grafischer Modelltest, TK *Rohscore*

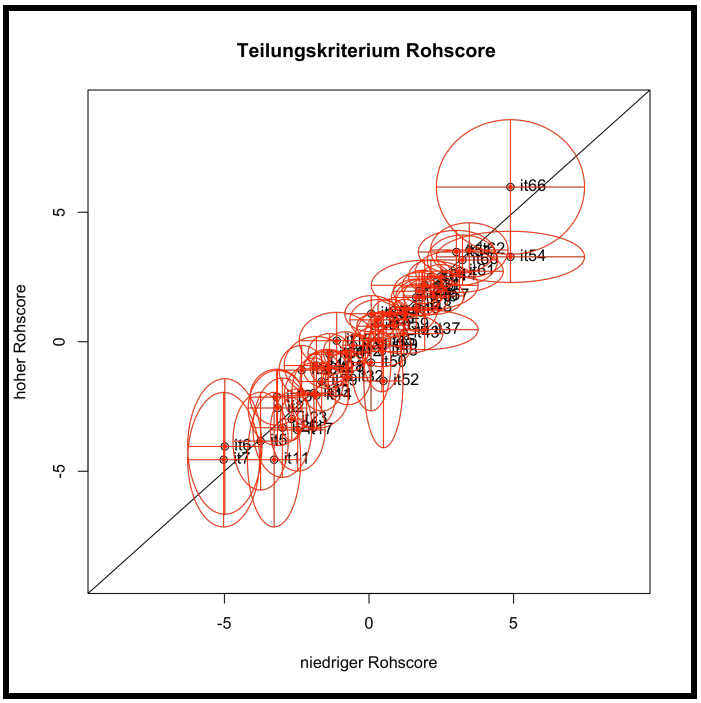

Abbildung 15: Grafischer Modelltest mit Konfidenz-Ellipsen, TK *Rohscore*

9.4.2 Teilungskriterium Geschlecht

Der LR-Test für das Teilungskriterium Geschlecht ergibt ebenfalls ein nicht signifikantes Ergebnis (siehe Tabelle 20). Dass die Daten dem Rasch-Modell entsprechen, wird auch bei Betrachtung des grafischen Modelltests ersichtlich (siehe Abbildung 16). Die Itemparameterschätzungen im unteren Fähigkeitsbereich, die nicht eng an der 45°-Geraden liegen, entsprechen nach Betrachtung der Konfidenz-Ellipsen sehr wohl dem Modell (siehe Abbildung 17). Auch der Wald-Test liefert keine Hinweise auf signifikante Items (siehe Tabelle 30 im Anhang).

Aufgrund der Geltung des Rasch-Modells der Items in Bezug auf das Teilungskriterium Geschlecht lässt sich die Hypothese H0-2 beantworten. Die Nullhypothese gilt, da die Parameterschätzungen in Bezug auf die Variable Geschlecht gleich sind. Es kommt demnach durch die resultierenden Testwerte des Untertests *Antonyme finden* zu keiner systematischen Benachteiligung aufgrund ihrer geschlechtsspezifischen Gruppenzugehörigkeit. Der Untertest misst diesbezüglich *fair*.

Tabelle 20: *LR-Test für das TK „Geschlecht“, letzter Berechnungsdurchgang*

Teilungskriterium Geschlecht			
Andersen χ^2	*df*	*p-Wert*	*Kritischer* χ^2 *–Wert*
83.18	56	.011	83.51

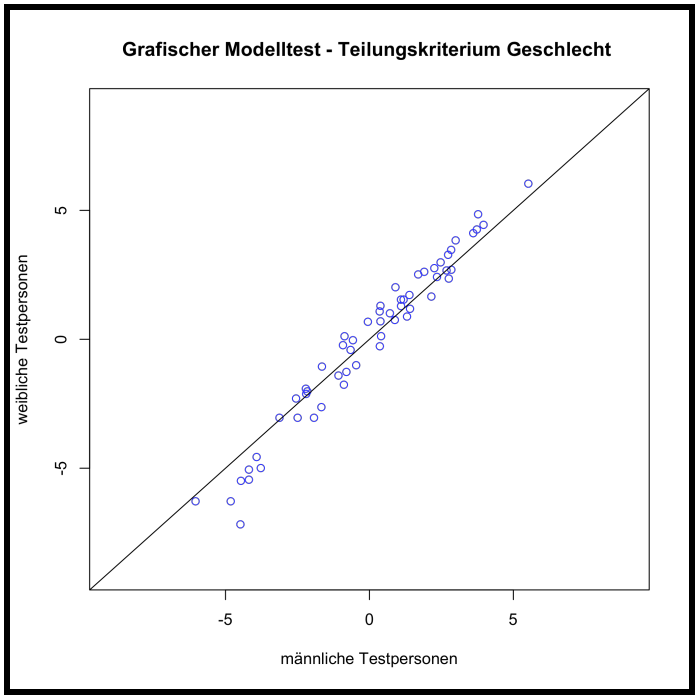

Abbildung 16: Grafischer Modelltest, TK *Geschlecht*

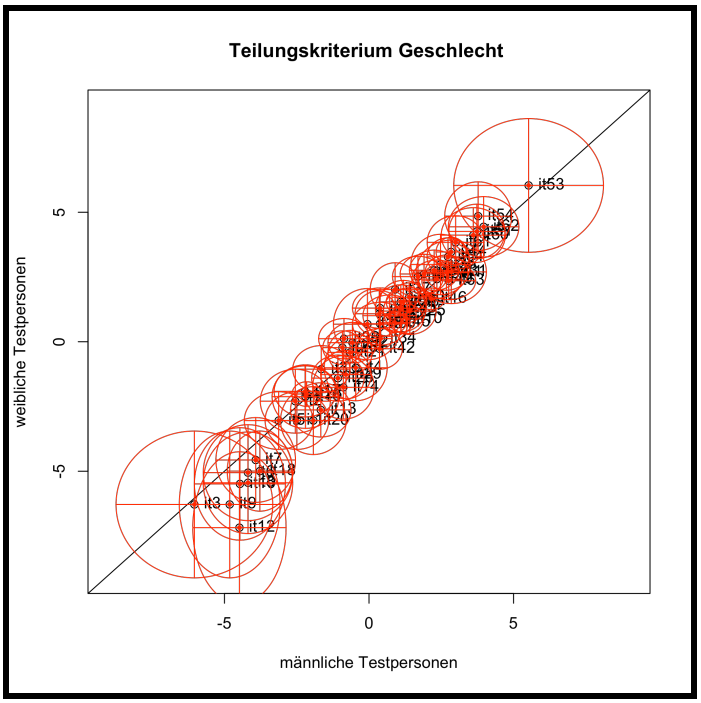

Abbildung 17: Grafischer Modelltest mit Konfidenz-Ellipsen, TK *Geschlecht*

9.4.3 Teilungskriterium Muttersprache

Im Bezug auf das Teilungskriterium Muttersprache wurde der LR-Test auch nach Ausschluss von 8 Items noch immer signifikant (siehe Tabelle 21). Obwohl die Itemschätzungen im grafischen Modelltest (siehe Abbildung 18) relativ eng um die 45°-Gerade streuen, entsprechen 5 Items nicht dem Modell (siehe Abbildung 20). Die Konfidenz-Ellipsen aller Items sind in Abbildung 19 dargestellt. Der Wald-Test identifiziert ebenfalls 5 signifikante Items (siehe Tabelle 31 im Anhang). Eine Diskussion, inwieweit für die Hypothese H0-3[14] die Nullhypothese gilt, wird im Abschnitt 10 (Diskussion und Ausblick) genauer erläutert.

Tabelle 21: *LR-Test für das TK „Muttersprache", letzter Berechnungsdurchgang*

Teilungskriterium Muttersprache			
Andersen χ^2	*df*	*p-Wert*	*Kritischer* χ^2 *–Wert*
138.07	52	<.001	78.62

[14] H0-3: Es kommt durch die resultierenden Testwerte des Untertests *Antonyme finden* zu keiner Benachteiligung von Personen in Bezug auf ihre *Muttersprache*.

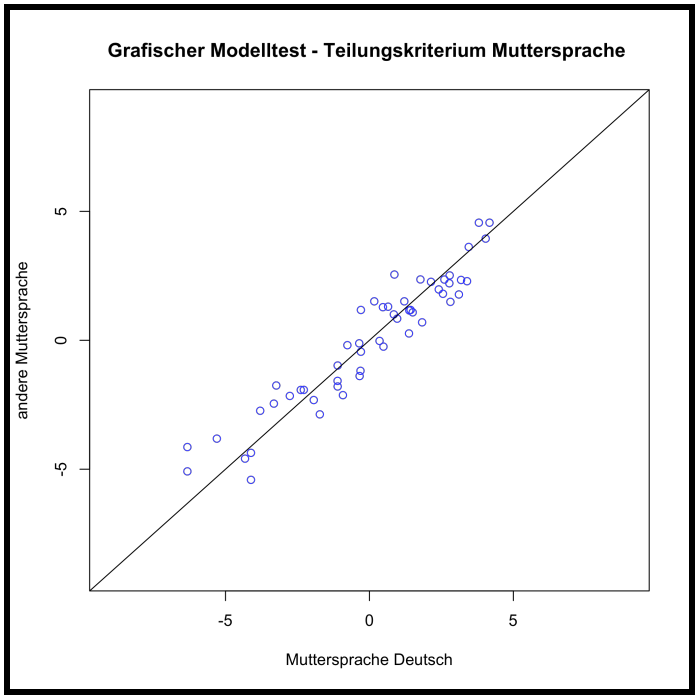

Abbildung 18: Grafischer Modelltest, TK *Muttersprache*

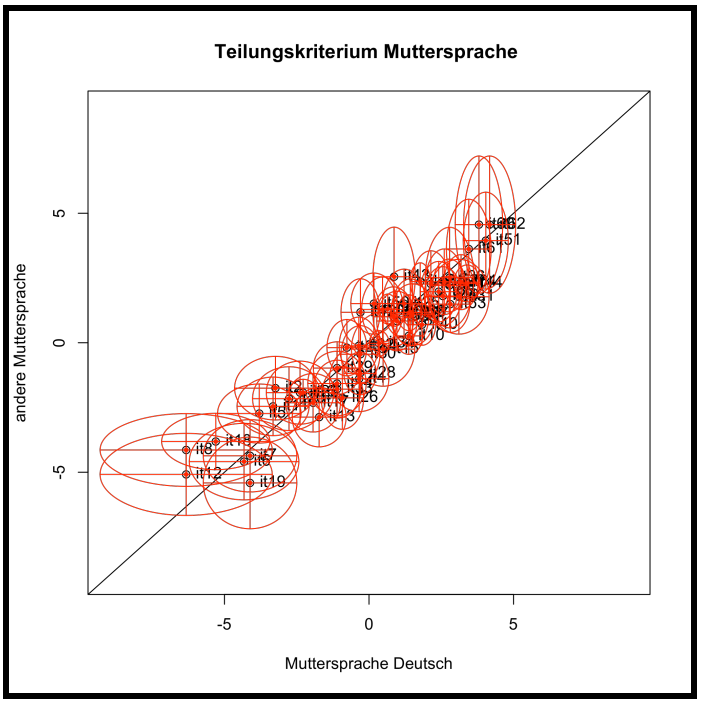

Abbildung 19: Grafischer Modelltest mit Konfidenz-Ellipsen, TK *Muttersprache*

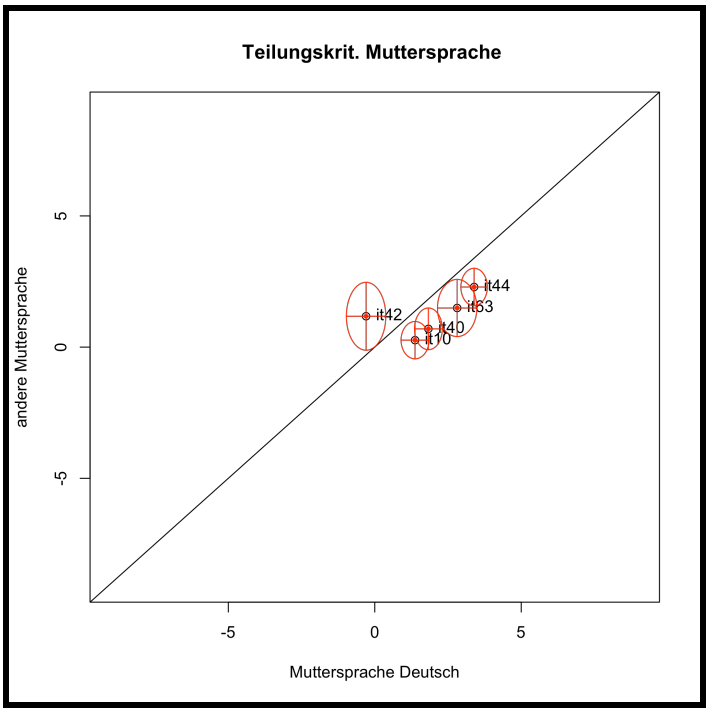

Abbildung 20: Grafischer Modelltest, TK *Muttersprache*, nicht modell-konforme Items

9.4.4 Teilungskriterium Alter

Der LR-Test für das Teilungskriterium Alter war nach der letzten Modellprüfung nicht signifikant (siehe Tabelle 22). Der grafische Modelltest zeigt eine nahezu ideale Passung der Items (siehe Abbildung 21), welche auch im Wald-Test nicht signifikant werden (siehe Tabelle 32 im Anhang). Abbildung 22 veranschaulicht die Konfidenz-Ellipsen aller Items.

Tabelle 22: *LR-Test für das TK „Alter", letzter Berechnungsdurchgang*

Teilungskriterium Alter			
Andersen χ^2	*df*	*p-Wert*	*Kritischer* χ^2 *–Wert*
38.28	25	.043	44.31

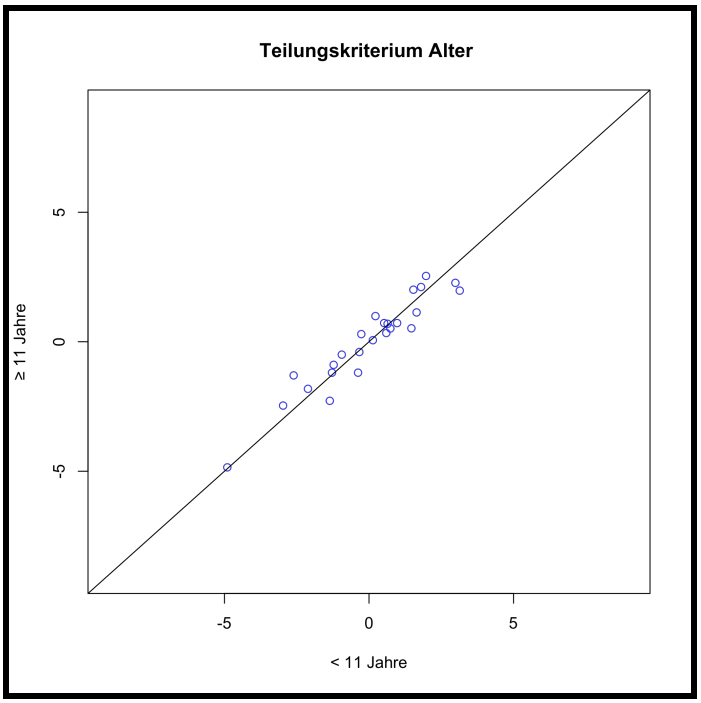

Abbildung 21: Grafischer Modelltest, TK *Alter*

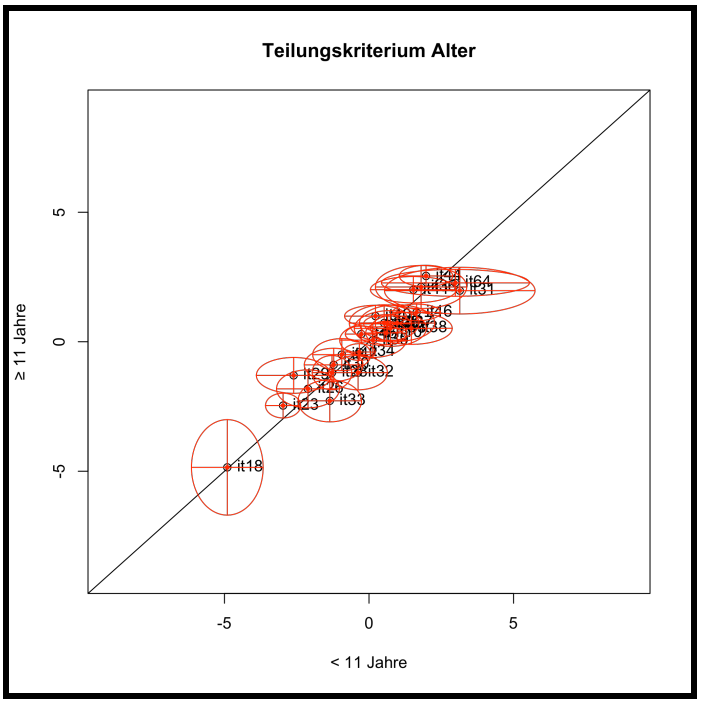

Abbildung 22: Grafischer Modelltest mit Konfidenz-Ellipsen, TK *Alter*

Nach Ausschluss von 8 Items wird der LT-Test für drei der vier Teilungskriterien nicht signifikant. Abgesehen von der Variable *Muttersprache* darf für die restlichen Items des Untertests *Antonyme finden* die Geltung des Rasch-Modells angenommen werden.

9.5 Itemschwierigkeitsparameter des Untertests Antonyme finden

Nach Ausschluss der nicht-modell-konformen Aufgaben, konnten für die verbleibenden Items die Itemschwierigkeitsparameter berechnet werden. Wie schon in Kapitel 4 beschrieben, liegt der Wertebereich von Personen- bzw. Itemschwierigkeitsparametern generell zwischen -∞ und +∞, praktisch allerdings zwischen -5 und +5. Die Itemschwierigkeitsparameter des Untertests *Antonyme finden* reichen von -6 bis +6 und sind gleichmäßig verteilt. Sie decken somit einen sehr breiten Fähigkeitsbereich gleichmäßig ab. Die Forderung für adaptive Verfahren, gerade den mittleren Fähigkeitsbereich mit vielen Aufgaben abzudecken, ist eindeutig erfüllt. Die Hälfte der Items befindet sich im Fähigkeitsbereich von -2 bis +2. Die Itemschwierigkeitsparameter des Untertests *Antonyme finden* sind in Tabelle 23 aufgeführt, wobei die Items der Schwierigkeit nach gereiht sind.

Tabelle 23: *Itemschwierigkeitsparameter des Untertests „Antonyme finden“*

Item	**σ_i**	Lower KI	Upper KI	**Item**	**σ_i**	Lower KI	Upper KI
Item 3	**-6.368**	-7.846	-4.890	Item 45	**0.564**	0.350	0.778
Item 12	**-5.667**	-6.703	-4.630	Item 52	**0.634**	0.178	1.089
Item 9	**-5.604**	-6.718	-4.489	Item 58	**0.677**	0.224	1.131
Item 19	**-5.122**	-6.079	-4.165	Item 10	**0.846**	0.540	1.152
Item 8	**-4.931**	-5.843	-4.019	Item 43	**0.948**	0.468	1.428
Item 6	**-4.769**	-5.660	-3.878	Item 25	**1.052**	0.566	1.539
Item 18	**-4.501**	-5.185	-3.816	Item 59	**1.086**	0.821	1.351
Item 7	**-4.445**	-5.243	-3.657	Item 38	**1.122**	0.681	1.564
Item 5	**-3.332**	-3.943	-2.701	Item 37	**1.160**	0.667	1.653
Item 11	**-3.017**	-3.609	-2.424	Item 40	**1.324**	1.016	1.632
Item 20	**-2.672**	-3.326	-2.017	Item 46	**1.621**	1.391	1.851
Item 2	**-2.636**	-3.282	-1.990	Item 48	**1.916**	1.476	2.355
Item 13	**-2.412**	-2.968	-1.855	Item 65	**2.060**	1.639	2.480
Item 16	**-2.380**	-3.011	-1.750	Item 49	**2.135**	1.711	2.560
Item 23	**-2.317**	-2.580	-2.054	Item 63	**2.252**	1.820	2.684
Item 17	**-2.286**	-2.838	-1.735	Item 57	**2.292**	1.857	2.727
Item 33	**-1.572**	-2.077	-1.066	Item 41	**2.419**	2.040	2.797
Item 14	**-1.505**	-2.117	-0.894	Item 36	**2.473**	1.889	3.056
Item 26	**-1.490**	-1.975	-1.005	Item 31	**2.529**	1.871	3.188
Item 29	**-1.260**	-1.744	-0.776	Item 64	**2.773**	2.382	3.164
Item 4	**-0.938**	-1.566	-0.309	Item 44	**2.910**	2.621	3.200
Item 21	**-0.795**	-1.225	-0.364	Item 61	**3.198**	2.661	3.734
Item 28	**-0.794**	-1.147	-0.441	Item 60	**3.636**	3.055	4.217
Item 30	**-0.566**	-1.000	-0.131	Item 51	**3.751**	3.210	4.292
Item 32	**-0.524**	-0.957	-0.091	Item 62	**3.974**	3.336	4.612
Item 42	**-0.200**	-0.624	0.224	Item 54	**4.071**	3.414	4.728
Item 34	**0.011**	-0.202	0.224	Item 53	**5.558**	4.167	6.948
Item 15	**0.035**	-0.633	0.703	Item 66	**6.029**	4.620	7.439
Item 50	**0.316**	-0.156	0.788	Item 67	**6.260**	4.313	8.207
Item 39	**0.462**	0.002	0.923				

9.6 Weitere Auswertungen

Um eine Maßzahl für die Validität des Untertests *Antonyme finden* zu bestimmen, wird die *konvergente Validität* mit der Skala *Synonyme finden* berechnet. Zur Bestimmung wird mittels PASW Statistics 18 eine Korrelation zwischen den Personenparametern der Untertests *Antonyme finden* und *Synonyme finden* berechnet. Um auch die Signifikanz von Korrelationen interpretieren zu können, müssen die Variablen intervallskaliert und normalverteilt sein (Field, 2009). Die Voraussetzung der Intervallskalierung ist bei beiden Variablen gegeben, allerdings zeigt die statistische Prüfung auf Normalverteilung, dass sowohl die Variable *Antonyme finden* als auch *Synonyme finden* nicht normalverteilt sind. Somit wird statt der Produkt-Moment-Korrelation auf ein nicht parametrisches Verfahren – der Rangkorrelation nach Spearman zurückgegriffen. Der Korrelationskoeffizient, die Stichprobenanzahl sowie die Ergebnisse des Signifikanztests *(α= 0.01*) sind in Tabelle 24 angeführt.

Tabelle 24: *Rangkorrelation der Untertests „Antonyme finden" und „Synonyme finden"*

Spearman´s Rangkorrelation		Personenparameter *Antonyme finden*	Personenparameter *Synonyme finden*
Personenparameter *Antonyme finden*	*Korrelationskoeffizient r_s*	1	.87
	p-Wert		.000
	Stichprobenanzahl	695	689
Personenparameter *Synonyme finden*	*Korrelationskoeffizient r_s*	.87	1
	p-Wert	.000	
	Stichprobenanzahl	689	689

Der Rangkorrelationskoeffizient ist in Bezug auf das Signifikanzniveau von *α= 0.01* signifikant, es besteht somit ein Zusammenhang zwischen den beiden Variablen. Viel bedeutsamer ist allerdings der Korrelationskoeffizient *r_s= 0.871*. Nach Cohen (1988, zitiert nach Field, 2009) spricht man ab einem Korrelationskoeffizient *r= 0.50* von einem großen Effekt. Es besteht somit ein hoher Zusammenhang zwischen den Testleistungen des Untertests *Antonyme finden* und den Testleistungen der Skala *Synonyme finden*. In Abbildung

23 ist der Zusammenhang der beiden Variablen in einem Streudiagramm anschaulich dargestellt.

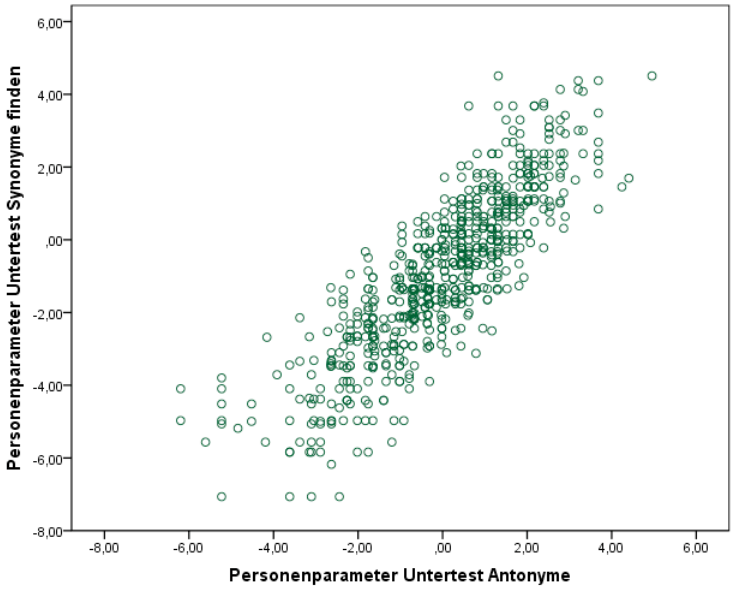

Abbildung 23: Streudiagramm bezüglich der Untertests *Antonyme finden* und *Synonyme finden*

Der lineare Zusammenhang zwischen den Personenparametern der beiden Untertests ist klar zu erkennen. Zusammenfassend kann die *konvergente Validität* des Untertests *Antonyme finden* in Bezug auf die Skala *Synonyme finden* als hoch angesehen werden. Der Untertest *Antonyme finden* misst demnach ein ähnliches Konstrukt wie die Skala *Synonyme finden*.

10 Diskussion und Ausblick

Das Ziel der vorliegenden Arbeit bestand darin, einen sprachlichen Untertest für den AID 3 zu entwickeln, der neben dem teilweise problematischen Untertest *Synonyme finden* diagnostische Information zum elementaren Sprachverständnis eines Kindes bzw. Jugendlichen liefern kann. Im Zuge der Testkonstruktion wurde ein Itempool geschaffen, der 67 Items umfasste. Nach der Datenerhebung wurden die Items auf Konformität mit dem dichotom logistischen Modell von Rasch überprüft. Nach Ausschluss von 8 Items konnte für die verbleibenden Aufgaben *a posteriori* Rasch-Modell-Konformität angenommen werden. Guthke (1996) zufolge muss der Testkonstrukteur damit rechnen, nach der Überprüfung der Testform ein Drittel der Aufgaben eliminieren zu müssen. In der vorliegenden Untersuchung mussten hingegen nur 12 % der Items ausgeschieden werden. Bei Betrachtung der Itemschwierigkeitsparameter lässt sich erkennen, dass die 59 verbliebenen Items des Untertests *Antonyme finden* einen breiten Fähigkeitsbereich von -6 bis +6 gleichmäßig abdecken. Die Hälfte der Items befindet sich im Fähigkeitsbereich von -2 bis +2, wonach die Forderung für adaptive Verfahren, den mittleren Fähigkeitsbereich mit vielen Aufgaben abzudecken, eindeutig erfüllt ist.

Der Likelihood-Ratio-Test wurde für die Teilungskriterien *Rohscore*, *Alter* sowie *Geschlecht* nach dem Ausschluss der 8 Items nicht mehr signifikant. Einzig das Teilungskriterium *Muttersprache* erwies sich nach der letzten Modellschätzung noch immer als signifikant. Der grafische Modelltest zeigt hingegen eine gute Modellgeltung, die Itemparameterschätzungen streuen relativ eng um die 45°- Gerade. Ebenso muss der Umstand in Betracht gezogen werden, dass der Likelihood-Ratio-Test bei großen Stichproben eher signifikant ausfällt (Kubinger, 2009a). Trotzdem soll nun auf mögliche Gründe eingegangen werden, warum der LR-Test nur in Bezug auf das Teilungskriterium Muttersprache signifikant wurde, nicht aber hinsichtlich anderer Teilungskriterien.

Ein möglicher Grund besteht darin, dass bei der Testung von Kindern mit schlechten Deutschkenntnissen, gerade im Volksschulalter, teilweise nicht beurteilt werden konnte, ob das Kind überhaupt die Instruktion des Untertests verstanden hat. Während einige Testleiter(innen) sich richtigerweise entschieden, den Untertest sicherheitshalber nicht vorzugeben, wurde in einigen Fällen der Untertest leider sehr wohl gewertet. Ein weiterer Grund, warum die Parameterschätzungen der Items zwischen den Stichproben *Deutsch als*

Muttersprache sowie *andere Muttersprache* unterschiedlich waren, liegt möglicherweise im unterschiedlichen Spracherwerb. So scheint es möglich, dass Kinder mit anderer Muttersprache als Deutsch über einen qualitativ anderen Wortschatz verfügen. Das Rasch-Modell gilt nur, wenn die Reihung der Items hinsichtlich ihrer Schwierigkeit in beiden Teilstichproben gleich ist. Bei zwei Items i und j könnte für Kinder mit deutscher Muttersprache das Item i leichter sein als das Item j, während Kindern mit anderer Muttersprache das Item j leichter fällt. Die Items würden somit neben sprachlicher Fähigkeit quasi die Muttersprache messen, woraufhin das Rasch-Modell nicht gilt. Es muss somit in Erwägung gezogen werden, dass der Untertest *Antonyme finden* hinsichtlich des Gütekriteriums *Fairness* Kinder mit nicht deutscher Muttersprache benachteiligt.

Der Untertest *Antonyme* lässt sich aber auch in Bezug auf andere Gütekriterien beurteilen. Als *Validität*smaß wurde eine konvergente Validität mit dem Untertest *Synonyme finden* berechnet, der ebenfalls das elementare Sprachverständnis misst. Statistische Analysen ergaben eine hohe Korrelation der beiden Untertests. Die konvergente Validität als Maß für die Kriteriumsvalidität ist somit in Bezug auf den Untertest *Synonyme finden* als hoch zu werten.

Die innere Konsistenz kann aufgrund der Geltung des Rasch-Modells als gegeben betrachtet werden. Das Gütekriterium *Reliabilität* (Messgenauigkeit) ist damit erfüllt, da die Items nur eine Fähigkeit messen. Der Test erfüllt aufgrund der Geltung des Rasch-Modells ebenfalls das Gütekriterium *Skalierung*. Die Summe aller gelösten Items ist somit ein faires Maß für die erbrachte Testleistung.

Das Gütekriterium *Objektivität* muss hingegen differenziert betrachtet werden. Obwohl die Instruktion standardisiert wurde, können keine Aussagen zur Testleiterunabhängigkeit getroffen werden, da diesbezügliche statistische Untersuchungen aufgrund der Zusammensetzung der Stichprobe sowie der Anzahl der Testleiter(innen) nicht vorgenommen werden konnten. Die Auswertungsobjektivität ist schon allein durch die Verwendung des freien Antwortformats kritisch zu betrachten. Obwohl alle Testleiter(innen) die Instruktion hatten, nur Antworten als richtig zu kodieren, die im Antwortkatalog stehen, wurden, wie aus Erfahrungsberichten der Testleiter(innen) bekannt wurde, teilweise gleiche Antworten von verschiedenen Testleiter(innen) unterschiedlich kodiert. Obwohl dies eher die Ausnahme als die Regel war, kann die *Auswertungsobjektivität* daher nicht als gegeben betrachtet werden. Da für jede Testperson ein Fähigkeitsparameter sowie ein Prozentrang berechnet wurde, ist die *Interpretationsobjektivität* erfüllt.

Hinsichtlich des Gütekriteriums *Fairness* ergibt die Parameterschätzung in Bezug auf die Variable *Geschlecht* keine signifikanten Unterschiede. Der Untertest *Antonyme finden* misst diesbezüglich *fair*. Inwiefern der Test das Gütekriterium Fairness bezüglich der Variable *Muttersprache* erfüllt, wurde bereits diskutiert. Die Gütekriterien *Ökonomie, Unverfälschbarkeit, Nützlichkeit* und *Eichung* können als erfüllt betrachtet werden (siehe Abschnitt 7.6).

Bei der Durchführung der Testungen stellte sich aufgrund von Erfahrungsberichten der Testleiter(innen) heraus, dass die *Akzeptanz* des Untertests *Antonyme finden* höher war als die des Subtests *Synonyme finden*. Es fiel den Kindern sichtlich leichter, das Gegenteil eines Wortes zu nennen als ein Wort, das dasselbe bedeutet. Oftmals sagten Kinder bei der Vorgabe des Subtests *Synonyme finden*, ob sie nicht einfach das Gegenteil nennen dürfen.

Der Untertest *Antonyme finden* wird im Zuge der Normierung des AID 3 erneut einer großen Stichprobe unterzogen werden. Dabei sollte der Untertest erneut daraufhin untersucht werden, ob sich die Parameterschätzungen der Items bezüglich der Variable *Muttersprache* als unterschiedlich erweisen. Wenn möglich sollten alle Items, die sich im grafischen Modelltest sowie im Wald-Test als nicht modell-konform ergeben, aus dem Itempool ausgeschlossen werden. Diese Möglichkeit bestand auch innerhalb dieser Untersuchung, allerdings hätten somit 5 weitere Items aus dem Itempool entfernt werden müssen. In jedem Fall scheint für die Vorgabe des Subtests *Antonyme finden* im AID 3 jene Strategie sinnvoll, die auch beim AID 2 –Türkisch angewendet wird. Um optimal fair zu diagnostizieren, sollte der Untertest *Antonyme finden* in derjenigen Sprache vorgegeben werden, die das Kind besser beherrscht. Demnach ist es für die Version des AID 3 wünschenswert, ebenfalls eine türkische Version zu erstellen. Vielmehr oder ebenso angebracht scheint die Entwicklung einer Testversion für Kinder, deren Muttersprache BKS[15] ist, da jene Gruppe den größten prozentuellen Anteil an Kindern mit nicht-deutscher Muttersprache in der untersuchten Stichprobe hatte. Wenn trotz mangelnder Deutschkenntnisse ein Kind mit der deutschsprachigen Version getestet wird, muss sichergestellt werden, ob das Kind zumindest die Instruktion verstanden hat. Etliche Kinder verstanden schlichtweg das Wort *Gegenteil* nicht. Sollte dies nicht der Fall sein, darf der Untertest nicht vorgegeben oder das Ergebnis nicht interpretiert werden.

Ein weiterer Aspekt, der im AID 3 Beachtung finden sollte, ist die Art und Weise, wie sehr sich der/die Testleiter(in) bei der Kodierung der Aufgaben an den Antwortkatalog halten soll.

[15] Bosnisch/Kroatisch/Serbisch

Den Erfahrungen dieser Untersuchung zufolge herrschte teilweise Unklarheit darüber, wie streng man sich an die Lösungen im Antwortkatalog zu halten hat. Während einige Testleiter(innen) auch Antworten als richtig kodierten, die kreativ waren und durchaus einen umfangreichen Wortschatz widerspiegelten, werteten andere die entsprechende Antwort als falsch. Als Folge leidet die *Verrechnungssicherheit.* Wenn man sich für die Strategie entscheiden sollte, dem/der Testleiter(in) die Entscheidung zu überlassen, ob nun eine Antwort als richtig oder falsch zu kodieren ist, muss man davon ausgehen, dass alle Testleiter(innen) ihrerseits dasselbe Ausmaß an sprachlicher Intelligenz aufweisen. So wäre beispielsweise ein sprachlich hochleistender Jugendlicher mit kreativen Antworten benachteiligt, wenn der/die Testleiter(in) eine Antwort nur wegen eigener sprachlicher Unsicherheit als falsch kodiert.

Insgesamt lässt sich feststellen, dass die Konstruktion eines sprachlichen Untertests für die Intelligenztestbatterie AID 3 gelungen ist. Die Items des Untertests *Antonyme finden* decken gleichmäßig einen breiten Fähigkeitsbereich ab und weisen eine hohe testtheoretische Güte in Bezug auf verschiedene Gütekriterien auf, die zur Beurteilung eines diagnostischen Verfahrens herangezogen werden. Im Hinblick auf die Veröffentlichung der dritten Version der Intelligenztestbatterie AID (AID 3), müssen allerdings noch Analysen erfolgen, ob Kinder mit nicht-deutscher Muttersprache durch den Untertest *Antonyme finden* benachteiligt werden.

11 Zusammenfassung

Die Zielsetzung dieser Arbeit bestand in der Konstruktion eines sprachlichen Untertests für die 3. Version der Intelligenztestbatterie AID (AID 3). Die Idee zur Konzeption des Untertests *Antonyme finden* entstand aus Problemen des Untertests *Synonyme finden* des AID 2. Durch die Vorgabe des Subtests *Antonyme finden* soll in Bezug auf das *elementare Sprachverständnis* validere Information gesammelt werden als durch die alleinige Vorgabe des Untertests *Synonyme finden*.

In einem ersten Schritt wurde unter Beachtung spezieller *Ausschlusskriterien* ein hinreichend großer Itempool konstruiert. Aus den 67 resultierenden Items wurden 6 Testhefte erstellt, jeweils zwei Parallelversionen für 3 Altersgruppen. Die Testhefte unterschieden sich je nach Altersgruppe hinsichtlich der Schwierigkeit der Items. Um eine hinreichend große Stichprobe aquirieren zu können, wurde in einem Team von 5 Diplomand(innen) die gesamte Rohversion des AID 3 vorgegeben. Dadurch konnten auch die Daten der anderen Diplomand(innen) in die Analysen miteinbezogen werden. Die Stichprobe umfasste 711 Kinder und Jugendliche im Alter von 6 bis 15 Jahren. Die Variable *Geschlecht* war gleich verteilt, während die Variable *Alter* eher einer Normalverteilung ähnelte. Etwa 2/3 der Kinder hatten Deutsch als Muttersprache. Die zweitgrößte Sprachengruppe umfasste Kinder mit BKS als Muttersprache, gefolgt von muttersprachlich türkischen Schüler(innen).

Der Untertest *Antonyme finden* wurde auf die Geltung des Rasch-Modells überprüft, um Aussagen über die Gütekriterien *Skalierung* und *Fairness* treffen zu können. Weiters ist Rasch-Modell-Konformität der Items notwendig, um den Untertest *Antonyme finden* im AID 3 *adaptiv* nach dem *branched-testing-design* vorgeben zu können. Nach Ausschluss von 8 Items konnte *a posteriori* die Gültigkeit des Rasch-Modells für die restlichen Items angenommen werden. Das Gütekriterium *Skalierung* ist somit erfüllt. Die resultierenden Itemschwierigkeitsparameter zeigen, dass die Items des Untertests *Antonyme finden* einen breiten Fähigkeitsbereich gleichmäßig abdecken. Ebenso ist der Test im Sinne der *Reliabilität* „messgenau“, da aufgrund der Geltung des Rasch-Modells alle Items dasselbe Konstrukt messen. Als *Validität*sbefund wurde eine konvergente Validität mit dem Untertest *Synonyme finden* berechnet. Es resultierte ein hoher linearer Zusammenhang. Der Untertest *Antonyme finden* misst *fair* in Bezug auf die Variable *Geschlecht*. Ungeklärt bleibt die Frage, ob durch die Testwerte des Subtests *Antonyme finden* Kinder mit nicht deutscher Muttersprache

benachteiligt werden. Für den AID 3 scheint sowohl eine türkische als auch wie bosnisch/serbisch/kroatische Version wünschenswert, um jene Kinder optimal fair diagnostizieren zu können.

Tabellenverzeichnis

Abbildungsverzeichnis

Literaturverzeichnis

Amelang, M., Bartussek, D., Stemmler, G. & Hagemann, D. (2006). *Differentielle Psychologie und Persönlichkeitsforschung*. (6., überarb. Aufl.). Stuttgart: Kohlhammer.

Agricola, C. & Agricola E. (1992). *Duden – Wörter und Gegenwörter*. (2., durchges. Aufl.). Mannheim: Dudenverlag.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. (6., vollst. überarb. und erw. Aufl.). Heidelberg: Springer.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion*. (3. akt. und erw. Aufl.). München: Pearson Studium.

Bulitta, E. & Bulitta, H. (2003). *Wörterbuch der Synonyme und Antonyme*. Frankfurt am Main: Fischer-Taschenbuch-Verlag.

Bußmann, H. (Hrsg.) (2008). *Lexikon der Sprachwissenschaft.* (4., durchges. und bibliogr. erg. Aufl.). Stuttgart: Kröner.

Dilling, H., Mombour, W. & Schmidt, M.H. (2010*). Internationale Klassifikation psychischer Störungen – ICD-10 Kapitel V (F)*. (7., überarb. Aufl.). Bern: Huber.

Field, A. (2009). *Discovering Statistics Using SPSS*. (3. Ed.). Los Angeles: Sage.

Fischer, G.H. (1989). Spezifische Objektivität: Eine wissenschaftstheoretische Grundlage des Rasch-Modells. In K.D. Kubinger (Hrsg.). *Moderne Testtheorie*. (S. 87-111). Weinheim: Beltz.

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

Geckeler, H. (1979). Antonymie und Wortart. In E. Bülow & P. Schmitter (Hrsg.). *Integrale Linguistik.* Amsterdam: Benjamins.

Glas, C.A.W., Verhelst, N.D. (1995). Testing the Rasch Model. In G.H. Fischer & I.W. Molenaar (Eds.). *Rasch Models – Foundations, Recent Developments, and Application*s (p. 69-95). New York: Springer.

Guthke, J. (1996). *Intelligenz im Test – Wege der psychologischen Intelligenzdiagnostik*. Göttingen: Vandenhoeck.

Häcker, H. & Stapf, K.H. (Hrsg.). (2004). *Dorsch Psychologisches Wörterbuch*. (14., vollst. überarb. und erw. Aufl.). Bern: Verlag Hans Huber.

Hagenmüller, B. (in Vorbereitung). *Entwicklung des Untertests „Formale Folgerichtigkeit" zur Erfassung von Reasoning in der Intelligenz-Testbatterie AID 3*. Unveröff. Dipl.Arbeit, Universität, Wien.

Hambleton, R.K., Swaminathan H. & Rogers, J.H. (1991). *Fundamentals of Item Response Theory*. *Volume 2*. Newbury Park: Sage.

Heller, K. & Perleth, C. (2000). *Kognitiver Fähigkeitstest KFT 4-12+ R* (für 4. bis 12. Klassen, Revision). Göttingen: Beltz.

Holocher-Ertl, S., Kubinger, K. D. & Hohensinn, C. (2008). Hochbegabungsdiagnostik: HAWIK-IV oder AID 2. *Kindheit und Entwicklung, 17*, (2), 99-106.

Karmann, A. (in Vorbereitung). *Wie gut decken die sprachbezogenen Untertests des AID 3 Sprachkompetenz ab?* Unveröff. Dipl.Arbeit, Universität, Wien.

Kastner-Koller, U. & Deimann, P. (2008). Testbesprechung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 40*, (3), 161-165.

Kubinger, K.D. (2009a). *Adaptives Intelligenz Diagnostikum 2 (Version 2.2).* (2., neu geeichte und überarb. Aufl.). Göttingen: Beltz.

Kubinger, K. D. (2009b). *Psychologische Diagnostik – Theorie und Praxis psychologischen Diagnostizierens*. (2., überarb. und erw. Aufl.). Göttingen: Hogrefe.

Kubinger, K.D. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie. In K.D. Kubinger (Hrsg.). *Moderne Testtheorie* (S.19-83). Weinheim: Beltz.

Kubinger, K. D. & Wurst, E. (1985). *Adaptives Intelligenz Diagnostikum (AID).* Weinheim: Beltz.

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

Kubinger, K.D. (2003). Gütekriterien. In K.D. Kubinger & R.S. Jäger (Hrsg.). *Schlüsselbegriffe der psychologischen Diagnostik* (S. 195-204). Weinheim: Beltz.

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

Kubinger, K.D. & Proyer R. (2004a). Gütekriterien. In K. Westhoff, L.J. Hellfritsch, L.F. Hornke, K.D. Kubinger, F. Lang, H. Moosbrugger, A. Püschel, G. Reimann (Hrsg.). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (S. 186-194). Lengerich: Pabst.

Kubinger, K.D. & Proyer R. (2004b). Testtheorien. In K. Westhoff, L.J. Hellfritsch, L.F. Hornke, K.D. Kubinger, F. Lang, H. Moosbrugger, A. Püschel, G. Reimann (Hrsg.). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (S. 173-186). Lengerich: Pabst.

Leiss, U. (2003). *Erstellung und Erprobung einer optimalen Strategie zur Diagnostik von Teilleistungsschwächen*. Unveröff. Diss., Universität, Wien.

Lienert, G.A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. (6. Aufl.). Weinheim: Beltz.

Lutzeier, P. R. (1995). *Lexikologie.* Tübingen: Stauffenburg.

Mair, P. & Hatzinger, R. (2009). *Extended Rasch Modeling: The R Package eRm*. PDF-Dateianhang zum Programmpaket eRm.

Melchers, P. & Preuß, U. (2009). *Kaufman – Assessment Battery for Children – deutschsprachige Fassung*. (8., unveränd. Aufl.). Frankfurt am Main: Pearson.

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

Petermann, F. & Petermann, U. (2007). HAWIK-IV. *Hamburg-Wechsler-Intelligenztest für Kinder – IV*. Bern: Huber.

Preusche, I. & Leiss, U. (2003). Intelligenztests für Kinder. HAWIK-III, AID 2 und K-ABC im Vergleich. *Report Psychologie*, *28,* (1), 12-26.

Renner, G. (2009). Testbesprechung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 41*, (1), 45-48.

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

Schober B. (2003). Fairness. In K.D. Kubinger & R.S. Jäger (Hrsg.). *Schlüsselbegriffe der psychologischen Diagnostik* (S.136-137). Weinheim: Beltz.

Synonym.com. (2007). [Online im Internet]. URL: http://www.synonym.com/antonym [Juni - September 2009].

Testzentrale. (2010). *Basisdiagnostik umschriebener Entwicklungsstörungen im Grundschulalter (BUEGA)*. [Online im Internet]. URL: http://www.testzentrale.de/programm/basisdiagnostik-umschriebener-entwicklungsstorungen-im-grundschulalter.html [24.01.2011].

Testzentrale. (2009). *Kaufman-Assessment Battery for Children (K-ABC) – deutsche Version.* [Online im Internet]. URL: http://www.testzentrale.ch/de/tests/testabkuerzungen-a-z/alphabet/K/flexShow/testDetail/testUid/437/ [24.01.2011].

Testzentrale. (2010). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4-12+ R)*. [Online im Internet]. URL: *http://www.testzentrale.de/programm/kognitiver-fahigkeitstest-fur-4-bis-12-klassen-revision.html* [24.01.2011].

Westhoff K., Hellfritsch L.J., Hornke, L.F., Kubinger K.D., Lang F., Moosbrugger H., Püschel A., Reimann G. (Hrsg.). (2004). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430*. Lengerich: Pabst.

Wictionary. (2009). [Online im Internet]. URL: http://de.wiktionary.org/wiki/Main_Page [Juni-September 2009].

Woxikon. (2009). [Online im Internet]. URL: http://synonyme.woxikon.de/ [Juni-September 2009].

Anhang

A) Instruktion des Untertests *Antonyme finden*

B) Lehrer(innen)brief

C) Elternbrief

D) Schriftlicher Ergebnisbericht (Muster)

E) Ergebnisse der Wald-Tests

A) Instruktion des Untertests *Antonyme finden*

„Ich nenne dir ein paar Wörter und wir wollen sehen, wie viele du davon kennst. Für jedes Wort, das ich dir sage, sollst du das Gegenteil finden. Wenn ich zum Beispiel sage: *warm*; dann sagst du: *kalt*. Verstehst du? Das Gegenteil von warm ist kalt. Probieren wir noch ein anderes Beispiel: Sag´ mir das gegenteil von *nass*."

- Der/die *Tl* hilft, wenn die *Tp* alleine nicht die richtige Antwort findet. –

„So, nun versuchen wir es mit anderen Wörtern. Sag´mir das Gegenteil von..."[16]

[16] Die Instruktion wurde dem vorläufigen Testmanual des AID 3 entnommen, welches zu jedem Untertest eine Testinstruktion samt Aufgabenkatalog beinhaltet.

B) Lehrer(innen)brief

Wien, Dezember 2009

Sehr geehrte Lehrerinnen und Lehrer!

Viele Kinder werden im Laufe ihrer schulischen Karriere aus den unterschiedlichsten Gründen mit psychologischen Tests untersucht. Bei schulpsychologischen Fragestellungen wie z.B. schulische Unter- bzw. Überforderung, Verhaltensprobleme im schulischen Kontext, Aufmerksamkeitsproblemen, Abklärung einer möglichen Legasthenie/Dyskalkulie, Schullaufbahnberatungen etc. kommt dabei meist ein Intelligenztest zum Einsatz, um die intellektuellen Stärken und Schwächen des Kindes abschätzen zu können.

Das *Adaptive Intelligenz Diagnostikum AID – aktuelle Version 2.2 (AID 2.2, Kubinger, 2009)* - ist ein im deutschen Sprachraum sehr etabliertes Verfahren zur Erfassung der intellektuellen Fähigkeiten von Kindern und Jugendlichen zwischen 6 und 15 Jahren. Diese Intelligenz-Testbatterie wurde nun um neue Wissensgebiete und historische und geografische sowie sprachliche Entwicklungen aktualisiert.

Im Zuge eines Forschungsprojekts der Universität Wien (Leitung: Univ. Prof. Dr. Mag. Klaus D. Kubinger) findet nun eine Schüler(innen)-Testung mit der aktualisierten Form AID 3 statt.

Wir wenden uns daher mit der Bitte an die Eltern Ihrer Schüler(innen), diese an der Untersuchung teilnehmen zu lassen, vorausgesetzt natürlich, dass das Kind damit einverstanden ist. Die Untersuchung findet während der Schulzeit einzeln statt und dauert ca. eine Stunde. Durchgeführt werden die Testungen von speziell dafür ausgebildeten Testleitern(innen). Erfahrungsgemäß macht den Kindern die Mitarbeit an den Aufgaben viel Spaß. Möchte das Kind jedoch einmal eine Pause einlegen oder die Untersuchung aus irgendeinem Grund frühzeitig abbrechen, ist das natürlich jederzeit möglich.

Auf Wunsch werden wir den Eltern als kleines Dankeschön einen kurzen schriftlichen Ergebnisbericht über die intellektuellen Stärken und Schwächen des Kindes zuschicken.

Die gewonnen Daten werden im Sinne des Datenschutzes ausschließlich für wissenschaftliche Zwecke genutzt. Sämtliche Ergebnisse der Schüler(innen) werden (noch während der Testung) von den Testleitern(innen) anonymisiert. Es können keine Ergebnisse an Sie oder die Schuldirektion weitergegeben werden.

Wir möchten Sie bitten, die Elternbriefe an Ihre Schüler(innen) zu übergeben und die Rückmeldungen der Eltern gesammelt in der Direktion Ihrer Schule abzugeben. Wenn sich Eltern Ihrer Schüler(innen)

mit der Teilnahme an der Untersuchung einverstanden erklären, werden wir uns bei der Terminvereinbarung sehr bemühen, den Ablauf Ihres Unterrichts so wenig wie möglich zu stören.

Für eventuelle Rückfragen stehen als Ansprechpersonen Frau Dr. Stefana Holocher-Ertl, Projektkoordinatorin (Tel: +43-1-4277 47851, email: stefana.holocher-ertl@univie.ac.at), und Frau Nicole Görner, Projektassistentin (email: nicole.goerner@gmx.at), jederzeit gerne zur Verfügung.

Mit der Bitte um Ihre Unterstützung, freundlichen Grüßen und herzlichen Dank im Voraus!

Dr. Stefana Holocher-Ertl (im Auftrag der Projektleitung)

C) Elternbrief

Wien, Dezember 2009

Liebe Eltern!

Viele Kinder werden im Laufe ihrer schulischen Karriere aus den unterschiedlichsten Gründen mit psychologischen Tests untersucht. Bei schulpsychologischen Fragestellungen wie z.B. schulische Unter- bzw. Überforderung, Schullaufbahnberatungen, Verhaltensprobleme im schulischen Kontext, Aufmerksamkeitsprobleme, Abklärung einer möglichen Legasthenie/Dyskalkulie etc. kommt dabei meist ein Intelligenztest zum Einsatz, um die intellektuellen Stärken und Schwächen des Kindes abschätzen zu können.

Das *Adaptive Intelligenz Diagnostikum AID – aktuelle Version 2.2 (AID 2.2, Kubinger, 2009)* - ist ein im deutschen Sprachraum sehr etabliertes Verfahren zur Erfassung der intellektuellen Fähigkeiten von Kindern und Jugendlichen zwischen 6 und 15 Jahren. Diese Intelligenz-Testbatterie wurde nun um neue Wissensgebiete erweitert und eine gesellschaftliche und sprachliche Aktualisierung vorgenommen.

Im Zuge eines Forschungsprojekts der Universität Wien (Leitung: Univ. Prof. Dr. Mag. Klaus D. Kubinger) findet nun eine Schüler(innen)-Testung mit der aktualisierten Form AID 3 statt.

Wir würden es sehr begrüßen, in dieses Forschungsprojekt auch Ihr Kind einbeziehen zu können.

Wir wenden uns daher mit der Bitte an Sie, Ihr Kind an dieser Untersuchung teilnehmen zu lassen, vorausgesetzt natürlich, Ihr Kind ist einverstanden. Die Untersuchung findet während der Schulzeit einzeln statt und dauert ca. eine Stunde. Durchgeführt werden die Testungen von speziell dafür ausgebildeten Testleitern(innen). Erfahrungsgemäß macht den Kindern die Mitarbeit an den Aufgaben viel Spaß (natürlich kann Ihr Kind dabei eine kleine Pause einlegen).

Auf Wunsch ist es auch möglich, Ihnen als kleines Dankeschön einen kurzen schriftlichen Ergebnisbericht über die intellektuellen Stärken und relativen Schwächen Ihres Kindes zuzuschicken.

Für eventuelle Rückfragen stehen als Ansprechpersonen Frau Dr. Stefana Holocher-Ertl, Projektkoordinatorin (Tel: +43-1-4277 47851, email: stefana.holocher-ertl@univie.ac.at), und Frau Nicole Görner, Projektassistentin (email: nicole.goerner@gmx.at), jederzeit gerne zur Verfügung.

Die gewonnen Daten werden im Sinne des Datenschutzes ausschließlich für wissenschaftliche Zwecke genutzt. Sämtliche Ergebnisse der Schüler(innen) werden (noch während der Testung) von den Testleitern(innen) anonymisiert. Direktion bzw. Lehrer(innen) der Schule werden selbstverständlich nicht über die Ergebnisse informiert.

Wir bitten Sie, mit Ihrer Unterschrift auf dem beiliegenden Formular, Ihr Einverständnis zur Teilnahme Ihres Kindes an der oben beschriebenen Untersuchung zu erteilen.

Mit freundlichen Grüßen und herzlichen Dank im Voraus!

Dr. Stefana Holocher-Ertl (im Auftrag der Projektleitung)

Ich erkläre mich mit der Teilnahme meiner Tochter/meines Sohnes

_____, geboren am _____,

Name des Kindes

an der *Schüler(innen)-Erhebung zum AID 3* einverstanden.

☐ Ich bitte um die Zusendung eines kurzen schriftlichen Ergebnisberichts an die Adresse:

_____

_____

_____

☐ Ich wünsche keinen Ergebnisbericht.

_____ _____

Datum Unterschrift des/der Erziehungsberechtigten

D) Schriftlicher Ergebnisbericht (Muster)

Wien, im Juni 2010

Liebe Eltern!

Vielen Dank für Ihr Einverständnis zur Teilnahme Ihres Kindes an der Schüler(innen)erhebung zur Intelligenz-Testbatterie AID 3 im Rahmen eines Forschungsprojekts der Universität Wien (Leitung: Univ. Prof. Dr. Mag. Klaus D. Kubinger). Das Adaptive Intelligenz Diagnostikum AID – aktuelle Version 2.2 (AID 2.2, Kubinger, 2009) - ist ein im deutschen Sprachraum sehr etabliertes Verfahren zur Erfassung der intellektuellen Fähigkeiten von Kindern und Jugendlichen zwischen 6 und 15 Jahren. Diese Intelligenz-Testbatterie wurde in der Version AID 3 um neue Wissensgebiete erweitert und aktualisiert und nun erstmals an Schüler/innen in Wien und Niederösterreich erprobt.

Wir wollen Ihnen nun über die Testergebnisse Ihres Kindes berichten. Für eine anonymisierte Verarbeitung und Speicherung der Daten haben wir Ihrem Kind folgenden Probandencode zugeteilt: ***bh1***

Die Testung Ihres Kindes mit dem AID 3 fand innerhalb der Schulzeit statt und umfasste die Dauer von ungefähr einer Stunde. Durchgeführt wurde diese von einer/einem speziell dafür ausgebildeten Testleiter/in.

Testergebnisse:

Die Leistungen in den einzelnen Untertests wurden jeweils mit einer altersspezifischen Stichprobe aus Wien und Niederösterreich verglichen. Die Testergebnisse werden in Prozenträngen (PR) angegeben, wobei ein Prozentrangwert (PR) von 25 bis 75 als durchschnittlich (alterstypisch) gilt. Der PR gibt an, wie viel Prozent der Gleichaltrigen in der Vergleichsstichprobe eine gleich gute oder niedrigere Leistung erbringen.

Untertest	**Interpretation**
Alltagswissen PR = 76	Es wird die Fähigkeit gemessen, sich Sachkenntnisse über Inhalte anzueignen, die in der heutigen Gesellschaft alltäglich sind (Wissen zu den Themen: Geschichte, Erdkunde, Sport, Kunst, Biologie). Die Leistungen des/der Schülers/in liegen hier über dem Altersdurchschnitt.
Antonyme PR = 66	Gemessen wird die Fähigkeit, die Gegensätzlichkeit von Begriffen zu erkennen und die Größe des Wortschatzes, der solche Gegensätze auszudrücken vermag. Der/die Schüler/in erbrachte hier eine durchschnittlich gute Leistung.
Realitätssicherheit PR = 79	Es wird die Fähigkeit gemessen, wesentliche Merkmale von Dingen des Alltags zu erkennen, wenn diese auf Bildern fehlen. Weiters zeigt es auch die Ausprägung der visuellen Differenzierungsfähigkeit. Der/die Schüler/in erreicht hier ein überdurchschnittliches Ergebnis.
Angewandtes Rechnen PR = 82	Dieser Untertest zeigt die rechnerische Fähigkeit, unabhängig von schulischen Rechenfertigkeiten. Es zeigt die Fähigkeit, Problemstellungen des Alltags durch Anwendung passender Rechenoperationen lösen zu können. Die Leistung des/der Schülers/in liegen hier über dem Durchschnitt.
Synonyme finden PR = 70	Es wird die Fähigkeit gemessen, die Bedeutung sprachgebundener Begriffe zu erkennen, und die Größe des Wortschatzes, der solche Begriffe durch andere Worte auszudrücken vermag. Der/die Schüler/in erbrachte hier eine durchschnittliche Leistung.
Formale Folgerichtigkeit PR = 40	Es zeigt sich die Fähigkeit, des Erkennens und Zuordnens der Zugehörigkeit von Figuren zu einer vorgegebenen Figurenreihe. Es handelt sich hierbei um die Fähigkeit zum logisch-schlussfolgernden Denken bei visuellem Aufgabenmaterial. Die Leistungen des/der Schülers/in liegen dabei im durchschnittlichen Bereich.

Funktionen abstrahieren PR = 45	Geprüft wird die Fähigkeit, durch schlussfolgerndes Denken im sprachlichen Bereich Funktionen zu abstrahieren und diese sprachlich ausdrücken zu können. Der/die Schüler/in erreichte ein durchschnittliches Ergebnis.
Soziales Erfassen und sachliches Reflektieren PR = 76	Gemessen wird das Verständnis, über Sachzusammenhänge der gesellschaftlichen Umwelt und über soziale angepasste Verhaltensweisen und gesellschaftliche Bedingungen bescheid zu wissen. Die Leistung des/der Schülers/in liegt dabei über dem Durchschnitt.

Ergänzende Bemerkungen:

Die Leistungen Ihres Kindes wurden im Rahmen eines Forschungsprojektes gewonnen und sind daher nur eingeschränkt aussagekräftig. Sollten Sie genauere Informationen zu der Leistungsfähigkeit Ihres Kindes wünschen, so raten wir Ihnen zu einer nochmaligen Testung mit ausführlicher Beratung bei einem/einer niedergelassenen Kinderpsychologen/in (www.psychologie.at).

Nochmals vielen Dank für Ihre Teilnahme und alles Gute für Sie und für die Zukunft Ihres Kindes,

Dr. Stefana Holocher-Ertl

E) Ergebnisse der Wald-Tests

Tabelle 25: *Wald-Test für Teilungskriterium „Rohscore" – Erster Berechnungsdurchgang*

Item	*z-Wert*	*p-Wert*	Item	*z-Wert*	*p-Wert*
Item 1	3.33	.001	Item 38	1.09	.276
Item 2	0.16	.875	Item 39	-0.53	.598
Item 4	2.91	.004	Item 40	1.44	.149
Item 6	0.38	.701	Item 41	0.23	.819
Item 10	0.42	.672	Item 42	0.25	.802
Item 13	0.43	.671	Item 43	-1.28	.200
Item 14	0.26	.793	Item 44	-0.89	.372
Item 15	2.22	.026	Item 45	-1.01	.311
Item 16	-1.53	.126	Item 46	0.40	.691
Item 17	-0.38	.707	Item 47	-3.11	.002
Item 18	-0.16	.871	Item 48	-1.05	.295
Item 19	0.81	.418	Item 49	-1.56	.120
Item 20	0.86	.391	Item 50	-2.05	.040
Item 21	-0.17	.863	Item 51	0.70	.485
Item 22	3.03	.002	Item 52	-2.39	.017
Item 23	-1.78	.076	Item 53	0.99	.321
Item 24	2.88	.004	Item 54	1.13	.258
Item 25	1.86	.063	Item 55	3.49	.000
Item 26	1.30	.194	Item 56	-2.18	.029
Item 27	3.96	.000	Item 57	-1.05	.295
Item 28	1.37	.170	Item 58	-1.35	.177
Item 29	-0.12	.902	Item 59	0.40	.692
Item 30	1.22	.222	Item 60	-1.16	.246
Item 31	0.45	.651	Item 61	-0.64	.520
Item 32	-1.34	.181	Item 62	-0.61	.539
Item 33	0.10	.919	Item 63	2.00	.045
Item 34	2.08	.038	Item 64	0.97	.331
Item 35	-2.65	.008	Item 65	0.48	.630
Item 37	-1.28	.201			

Tabelle 26: *Wald-Test für Teilungskriterium „Geschlecht“ – Erster Berechnungsdurchgang*

Item	*z-Wert*	*p-Wert*	Item	*z-Wert*	*p-Wert*
Item 1	1.10	.272	Item 34	-1.94	.052
Item 2	0.66	.504	Item 35	7.42	.000
Item 3	-0.28	.782	Item 36	0.53	.597
Item 4	-0.80	.423	Item 37	1.66	.097
Item 5	-0.28	.783	Item 38	0.38	.721
Item 6	-0.37	.714	Item 39	1.06	.290
Item 7	-1.09	.278	Item 40	0.69	.488
Item 8	-0.72	.472	Item 41	-0.50	.614
Item 9	-1.30	.195	Item 42	-1.93	.054
Item 10	-2.04	.041	Item 43	-0.14	.892
Item 11	-1.29	.198	Item 44	1.78	.075
Item 12	-1.54	.123	Item 45	-1.25	.213
Item 13	-1.98	.048	Item 46	-2.82	.005
Item 14	-1.29	.199	Item 47	4.12	.000
Item 15	1.17	.244	Item 48	1.46	.144
Item 16	0.39	.698	Item 49	-0.11	.913
Item 17	0.25	.804	Item 50	0.23	.816
Item 18	-1.57	.117	Item 51	0.80	.425
Item 19	-1.20	.229	Item 52	0.21	.836
Item 20	-1.38	.168	Item 53	0.27	.789
Item 21	0.57	.566	Item 54	1.33	.184
Item 22	-0.85	.397	Item 55	-0.85	.396
Item 23	-0.08	.934	Item 56	3.01	.003
Item 24	1.59	.113	Item 57	0.91	.361
Item 25	-1.00	.319	Item 58	1.55	.121
Item 26	-1.24	.214	Item 59	1.17	.241
Item 27	0.21	.802	Item 60	0.56	.576
Item 28	1.50	.135	Item 61	1.33	.330
Item 29	-1.19	.234	Item 62	0.45	.653
Item 30	1.66	.097	Item 63	-1.20	.231
Item 31	-0.61	.541	Item 64	0.97	.331
Item 32	0.71	.478	Item 65	1.47	.142
Item 33	0.90	.366			

Tabelle 27: *Wald-Test für Teilungskriterium „Muttersprache“ – Erster Berechnungsdurchgang*

Item	*z-Wert*	*p-Wert*	Item	*z-Wert*	*p-Wert*
Item 1	-2.85	.004	Item 34	-1.39	.165
Item 2	2.41	.016	Item 35	-2.30	.021
Item 4	-1.50	.134	Item 36	-0.19	.850
Item 5	2.05	.041	Item 37	-0.48	.634
Item 6	0.16	.874	Item 38	-0.29	.775
Item 7	0.23	.816	Item 39	1.75	.080
Item 8	2.00	.046	Item 40	-3.25	.001
Item 10	-3.06	.002	Item 41	-1.07	.284
Item 11	1.89	.058	Item 42	2.70	.007
Item 12	1.29	.196	Item 43	2.37	.018
Item 13	-1.56	.119	Item 44	-3.28	.001
Item 14	-0.52	.602	Item 45	2.90	.004
Item 15	-0.95	.341	Item 46	2.22	.026
Item 16	1.06	.289	Item 47	-0.45	.650
Item 17	-0.20	.841	Item 48	0.36	.719
Item 18	1.75	.081	Item 49	-1.59	.111
Item 19	-0.97	.334	Item 50	2.63	.008
Item 20	1.24	.216	Item 51	-0.05	.957
Item 21	1.43	.153	Item 52	0.43	.669
Item 22	-2.75	.006	Item 55	-0.23	.818
Item 23	2.02	.043	Item 56	-1.01	.315
Item 24	-5.44	.000	Item 57	-0.49	.626
Item 25	0.81	.420	Item 58	-0.13	.894
Item 26	-2.17	.030	Item 59	-0.56	.575
Item 27	-4.79	.000	Item 60	0.79	.428
Item 28	-2.13	.033	Item 61	0.21	.836
Item 29	0.41	.680	Item 62	0.44	.660
Item 30	-0.23	.822	Item 63	-2.86	.004
Item 31	-1.68	.092	Item 64	-1.66	.096
Item 32	0.54	.591	Item 65	-0.92	.357
Item 33	-1.10	.273			

Tabelle 28: *Wald-Test für Teilungskriterium „Alter“ – Erster Berechnungsdurchgang*

Item	*z-Wert*	*p-Wert*
Item 10	-1.11	.266
Item 18	-0.59	.554
Item 22	-0.09	.391
Item 23	1.80	.072
Item 24	-1.09	.276
Item 25	0.58	.562
Item 26	0.29	.770
Item 27	3.41	.001
Item 28	-0.13	.901
Item 29	2.65	.008
Item 30	0.33	.739
Item 31	-0.99	.321
Item 32	-2.06	.039
Item 33	-1.33	.183
Item 34	-0.64	.524
Item 35	-0.62	.532
Item 36	0.28	.781
Item 37	-0.24	.809
Item 38	-1.81	.071
Item 39	0.10	.921
Item 40	1.59	.112
Item 41	1.07	.284
Item 42	0.59	.557
Item 43	-0.20	.842
Item 44	1.48	.140
Item 45	2.06	.040
Item 46	-1.59	.112
Item 59	0.22	.823
Item 64	-0.77	.445

Tabelle 29: *Wald-Test für Teilungskriterium „Rohscore“ – Letzter Berechnungsdurchgang*

Item	*z-Wert*	*p-Wert*	Item	*z-Wert*	*p-Wert*
Item 2	0.88	.377	Item 39	-1.21	.226
Item 4	1.47	.141	Item 40	1.33	.185
Item 5	-0.11	.915	Item 41	0.70	.485
Item 6	0.84	.403	Item 42	0.94	.345
Item 7	0.43	.670	Item 43	-1.48	.140
Item 10	1.21	.226	Item 44	0.13	.901
Item 11	-1.19	.235	Item 45	-1.53	.126
Item 13	1.92	.055	Item 46	0.07	.943
Item 14	-0.39	.695	Item 48	-0.69	.493
Item 15	1.76	.078	Item 49	-0.29	.773
Item 17	-1.31	.189	Item 50	-1.13	.258
Item 20	-0.38	.702	Item 51	0.74	.461
Item 21	0.84	.399	Item 52	-1.95	.052
Item 23	-0.88	.379	Item 54	-1.51	.130
Item 25	1.10	.273	Item 57	-0.94	.347
Item 26	2.59	.010	Item 58	-1.25	.210
Item 28	1.12	.264	Item 59	-0.72	.473
Item 29	0.20	.845	Item 60	-0.13	.897
Item 30	2.08	.037	Item 61	-0.60	.549
Item 31	0.31	.755	Item 62	0.11	.911
Item 32	-1.28	.200	Item 63	0.50	.620
Item 33	-0.12	.905	Item 64	0.95	.342
Item 34	1.93	.054	Item 65	0.18	.856
Item 37	-1.88	.060	Item 66	0.76	.445
Item 38	2.25	.024			

Tabelle 30: *Wald-Test für Teilungskriterium „Geschlecht“ – Letzter Berechnungsdurchgang*

Item	*z-Wert*	*p-Wert*	Item	*z-Wert*	*p-Wert*
Item 2	0.38	.706	Item 34	-1.31	.192
Item 3	-0.16	.874	Item 36	0.85	.397
Item 4	-0.84	.401	Item 37	2.14	.033
Item 5	0.13	.897	Item 38	0.78	.437
Item 6	-0.93	.354	Item 39	1.53	.127
Item 7	-0.78	.436	Item 40	1.03	.301
Item 8	-1.30	.194	Item 41	-0.03	.978
Item 9	-1.13	.256	Item 42	-1.46	.145
Item 10	-1.36	.175	Item 43	0.39	.700
Item 11	-0.90	.366	Item 44	2.11	.035
Item 12	-2.08	.038	Item 45	-0.61	.542
Item 13	-1.68	.094	Item 46	-2.06	.039
Item 14	-1.39	.165	Item 48	1.79	.073
Item 15	1.04	.297	Item 49	0.16	.874
Item 16	0.12	.904	Item 50	0.64	.523
Item 17	0.52	.606	Item 51	0.95	.344
Item 18	-1.64	.101	Item 52	0.63	.530
Item 19	-0.99	.325	Item 53	0.36	.717
Item 20	-1.66	.097	Item 54	1.57	.116
Item 21	0.55	.585	Item 57	1.12	.261
Item 23	0.51	.610	Item 58	1.93	.056
Item 25	-0.46	.646	Item 59	1.63	.104
Item 26	-0.67	.504	Item 60	0.85	.394
Item 28	1.90	.058	Item 61	1.52	.128
Item 29	-0.94	.348	Item 62	0.72	.469
Item 30	2.16	.031	Item 63	-0.88	.381
Item 31	-0.22	.825	Item 64	1.35	.176
Item 32	1.21	.228	Item 65	1.64	.102
Item 33	1.13	.257			

Tabelle 31: *Wald-Test für Teilungskriterium „Muttersprache“ – Letzter Berechnungsdurchgang*

Item	*z-Wert*	*p-Wert*	Item	*z-Wert*	*p-Wert*
Item 2	2.04	.041	Item 34	-1.62	.106
Item 4	-1.65	.099	Item 36	-0.33	.740
Item 5	1.49	.137	Item 37	-0.76	.445
Item 6	-0.29	.774	Item 38	-0.45	.652
Item 7	-0.30	.763	Item 39	1.45	.147
Item 8	1.71	.088	Item 40	-3.16	.002
Item 10	-3.34	.001	Item 41	-1.25	.212
Item 11	1.32	.186	Item 42	2.60	.009
Item 12	0.95	.343	Item 43	2.12	.034
Item 13	-2.00	.045	Item 44	-3.36	.001
Item 14	-0.75	.453	Item 45	2.52	.012
Item 15	-1.05	.295	Item 46	1.88	.061
Item 16	0.70	.483	Item 48	0.23	.818
Item 17	-0.67	.503	Item 49	-1.48	.138
Item 18	1.77	.077	Item 50	2.43	.015
Item 19	-1.39	.163	Item 51	-0.12	.903
Item 20	0.89	.375	Item 52	0.30	.767
Item 21	1.15	.249	Item 57	-0.45	.653
Item 23	1.31	.192	Item 58	-0.24	.810
Item 25	0.50	.614	Item 59	-0.73	.469
Item 26	-2.19	.028	Item 60	0.71	.481
Item 28	-2.26	.024	Item 61	0.21	.833
Item 29	0.25	.805	Item 62	0.38	.721
Item 30	-0.29	.769	Item 63	-2.66	.008
Item 31	-1.93	.054	Item 64	-1.79	.074
Item 32	0.47	.641	Item 65	-0.85	.396
Item 33	-1.28	.201			

Tabelle 32: *Wald-Test für Teilungskriterium „Alter“ – Letzter Berechnungsdurchgang*

Item	z-Wert	p-Wert
Item 10	-0.80	.426
Item 18	0.06	.953
Item 23	1.69	.091
Item 25	0.37	.715
Item 26	0.58	.565
Item 28	0.24	.814
Item 29	2.30	.021
Item 30	0.71	.477
Item 31	-1.09	.276
Item 32	-1.73	.083
Item 33	-1.74	.081
Item 34	-0.25	.801
Item 36	0.45	.652
Item 37	-0.44	.662
Item 38	-1.58	.114
Item 39	-0.14	.887
Item 40	1.72	.085
Item 41	0.78	.436
Item 42	0.97	.334
Item 43	-0.41	.684
Item 44	1.45	.148
Item 45	2.28	.023
Item 46	-1.49	.136
Item 59	0.14	.888
Item 64	-0.70	.481

Lebenslauf

Persönliche Daten

Name:	Benjamin Weber
Geburtsdatum:	20.09.1986
Geburtsort:	Wien
Staatsbürgerschaft:	Österreich & Schweiz

Ausbildung

03/2009 – 01/2010	**Ausbildung zum Student Mentor** im Rahmen eines universitären Mentoring-Projekts (*Cascaded Blended Mentoring*, CBM) der Fakultät für Psychologie, Universität Wien
Seit 10/2004	**Studium der Psychologie an der Universität Wien** Schwerpunkt: Angewandte Kinder- und Jugendpsychologie & Klinische Psychologie
1996-2004	**Erich-Fried-Realgymnasium**, Wien
11/2000-01/2001	**Ausbildung zum diplomierten Babysitter** im Eltern-Kind-Zentrum Gilgegasse, Wien
1992-1996	**Volksschule Gilgegasse**, Wien

Berufserfahrung

11/2009 - 01/2010	**Praktikum an der Test- und Beratungsstelle der Universität Wien**
Seit 10/2008	**Schachtrainertätigkeit und Kinderbetreuung** im Kinderhort Vorgartenstraße, Wien
10/2005 - 9/2006	**Absolvierung des Zivildienstes beim Verein Wiener Jugendzentren**, Jugendzentrum Rennbahnweg, Wien

Seit 2004	**Betreuertätigkeit bei Kinder- und Jugendferienlagern** jeweils 2 Wochen/Sommer (Schachimedes-Feriencamp), Steiermark
Seit 2001	**Arbeit als Babysitter, Kinderbetreuer & Kinderanimateur** im Eltern-Kind-Zentrum Gilgegasse, Wien

Weitere Qualifikationen

Sprachkenntnisse:	**Deutsch**: Muttersprache **Schweizerdeutsch**: 2. Muttersprache **Englisch**: fundierte Kenntnisse in Wort und Schrift **Französisch**: Grundkenntnisse in Wort und Schrift
EDV:	**MS Office**, **SPSS**