universität wien

Diplomarbeit

Titel der Arbeit

Dimensionalitätsprüfung des Untertests 9 „Funktionen Abstrahieren“ der beiden Intelligenz-Testbatterien AID 3 und AID-Gruppe

Verfasserin:

Doris Mayerhofer

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, 2014

Studienkennzahl: 298

Studienrichtung: Psychologie

Betreuer: Univ.-Prof. i. R. Dr. Mag. Klaus Kubinger

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Danksagung

Ein herzliches Dankeschön möchte ich an Herrn Univ.-Prof. i.R. Dr. Mag. Klaus D. Kubinger und Frau Mag. Bettina Hagenmüller für die Betreuung dieser Diplomarbeit richten.

Frau Mag. Hagenmüller hat diese Diplomarbeit subbetreut und stand meiner Kollegin Sandra Weichselbaum und mir stets mit Expertise und bestmöglicher fachlicher Beratung zur Seite.

Ebenso großer Dank gebührt meiner Familie, die mir während der Studienzeit eine bedeutende Stütze in finanzieller und seelischer Hinsicht bot. Ihnen habe ich sowohl meinen schnellen Studienverlauf als auch den guten Studienerfolg zu verdanken, da sie mir auf jede erdenkliche Art und Weise den Rücken stärkten.

Weiteren Dank möchte ich meiner Diplomarbeitskollegin Sandra Weichselbaum widmen, da sich die Zusammenarbeit mit ihr äußerst unkompliziert und kurzweilig gestaltet hat.

Meine beiden Studienkolleginnen und Freundinnen, Katharina Schossleitner und Mirjam Haag, möchte ich in dieser Danksagung ebenfalls erwähnen. Es war mir eine Freude, den Weg durchs Studium mit ihnen zu teilen.

Last but not least bedanke ich mich auch herzlich bei den Schulen, die an der Erhebung teilgenommen haben. Allen voran dem Direktor des Gymnasiums Zwettl, Herrn Mag. Wolfgang Steinbauer, und der Direktorin der Volksschule Kottes/Purk, Frau Elisabeth Apolt, sowie natürlich auch den Lehrerinnen und Lehrern der Schulen, deren Klassen und Unterrichtszeit ich für meine Studie nutzen durfte. Und zu guter Letzt bedanke ich mich bei den Schülerinnen und Schülern, die so engagiert teilgenommen haben.

Abstract

Diese Diplomarbeit behandelt die Dimensionalitätsprüfung des Untertest 9 „Funktionen Abstrahieren“ in den beiden Testbatterien AID 3 (Kubinger & Holocher-Ertl, 2014, in Druck) und AID – Gruppe (Kubinger & Hagenmüller, in Vorb.).

Der Untertest 9 „Funktionen Abstrahieren“ soll dabei sowohl in Einzel- als auch in der Gruppenvorgabe das Fähigkeitskonstrukt „Begriffsbildung durch Abstraktion erfassen“.

In Ober- und Niederösterreich bearbeiteten 219 Schülerinnen und Schüler der 3. bis 9. Schulstufe eine Auswahl von insgesamt 30 Items des Untertests 9 in den beiden Vorgabemodi AID 3 und AID Gruppe. Die erhobenen Daten wurden mittels der Statistiksoftware „R“ (Version 3.0.2) mit dem Package „eRm“ analysiert. Zur Prüfung der Itemhomogenität wurde der Martin-Löf-Test berechnet. Die Ergebnisse der Dimensionalitätsprüfung sprechen dafür, dass der Untertest 9 „Funktionen Abstrahieren“ in beiden Vorgabemodi eindimensional das zu messende Konstrukt erfasst.

Abstract

The aim of this thesis is to test the dimensionality of the subtest 9 “Abstracting” in both test batteries AID 3 (Kubinger & Holocher-Ertl, 2014, in print) and AID – Gruppe (Kubinger & Hagenmüller, in work). We tested the hypotheses if both modes of testing examine the same construct (common functionalism of things)[1] one-dimensionally in spite of different item designs. A total of 219 pupils from Lower and Upper Austria from 3rd up to 9th level of education attended to an item selection of subtest 9 “Abstraction”. To analyse data the statistic software “R” (version 3.0.2) with the package “eRm” was used. Martin-Löf-Test was applied to test homogeneity of items. The output illustrates homogeneity of items in subtest 9 in AID 3 and AID – Gruppe, both modes measure the mental concept of ‘definition’ one-dimensionally.

[1] Kubinger, 2004

Vorwort

Meine Kollegin, Sandra Weichselbaum, behandelt in ihrer Diplomarbeit das Thema „Dimensionalitätsprüfung des Untertests 4 ‚Soziale und Sachliche Folgerichtigkeit' der beiden Intelligenz-Testbatterien AID 3 und AID-Gruppe" (in Arbeit) während diese Diplomarbeit die Dimensionalitätsprüfung für den Untertest 9 „Funktionen Abstrahieren" zum Thema hat.
Für die Datenerhebung unserer beider Diplomarbeiten wurden aus praktischen Gründen sowohl Untertest 4 „Soziale und Sachliche Folgerichtigkeit" sowie 9 „Funktionen Abstrahieren" von beiden Testleiterinnen gemeinsam erhoben. Zur Datenanalyse wurde in beiden Diplomarbeiten derselbe Datensatz herangezogen.

Diese Diplomarbeit behandelt zwar ausschließlich Untertest 9 „Funktionen Abstrahieren", dennoch wird im Zuge der Beschreibung der Vorgabe auch immer wieder auf Untertest 4 eingegangen.

Die Ergebnisdarstellung ist speziell Untertest 9 „Funktionen Abstrahieren" gewidmet.

1) Einleitung

Diese Diplomarbeit behandelt die Dimensionalitätsprüfung des Untertests 9 „Funktionen Abstrahieren“ des AID 3 (Kubinger & Holocher-Ertl, in Vorbereitung) und AID Gruppe (Kubinger & Hagenmüller, in Vorbereitung).
Es wird untersucht, ob und inwieweit es Unterschiede bei der Bearbeitung beider Vorgabemodi, AID 3 und AID-Gruppe, bei Kindern und Jugendlichen in den Schulstufen 3 bis 9 in Schulen aus Ober- und Niederösterreich gibt.

Bei einer Dimensionalitätsprüfung werden die Items der beiden Administrationsformen AID 3 (Kubinger & Holocher-Ertl, 2014, in Druck) und AID-Gruppe (Kubinger & Hagenmüller, in Vorb.) darauf untersucht, ob sie eindimensional messen, das heißt, dasselbe Konstrukt erfassen. Die beiden Vorgabemodi gelten als konstruktäquivalent, wenn die Items homogen sind und ein und dieselbe Fähigkeitsdimension erfassen. Zur Messung der Homogenität der Items in der Einzel- und der Gruppenversion wird der Martin-Löf-Test berechnet (genaueres dazu im empirischen Teil dieser Arbeit).

Da sich die Vorgabemodi des Untertests 9 in AID 3 und AID – Gruppe unterscheiden, war es wichtig zu untersuchen, ob „Funktionen Abstrahieren“ trotz verschiedener Gestaltungsweisen dasselbe Fähigkeitskonzept erfasst.

Zusammengefasst lautet die Fragestellung dieser Diplomarbeit folgendermaßen:
Misst der Untertest 9 „Funktionen Abstrahieren“ in beiden Vorgabemodi, AID 3 und AID-Gruppe, dieselbe Fähigkeit?

Im einleitenden Abschnitt wird näher auf die Grundlagen der Intelligenzdiagnostik und die testtheoretischen Grundlagen der Dimensionalitätsprüfung sowie auf die psychologisch-diagnostischen Verfahren AID 3 und AID-Gruppe eingegangen. Der empirische Teil dieser Arbeit behandelt die Durchführung der Testungen, die Datenanalyse und die Darstellung der Ergebnisse.

II) Theorieteil

2) Grundlagen

2.1) Intelligenzdiagnostik

2.1.1) Intelligenz

Darüber, was Intelligenz ist und was intelligentes Verhalten ausmacht, gibt es eine Vielzahl an Definitionen.

„Die Intelligenzdefinition, die dem jeweiligen Test zugrunde liegt, beeinflusst auch das verwendete Aufgabenmaterial.“ (Schlagheck & Petermann, 2006, S. 94)
Da das AID (Kubinger & Wurst, 1985) am Testmodell David Wechslers orientiert ist, wird in dieser Diplomarbeit dessen Definition thematisiert:

„Intelligenz ist die zusammengesetzte oder globale Fähigkeit des Individuums, zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umgebung wirkungsvoll auseinander zu setzen.“ (Wechsler 1956, S.13).

Determinanten der Intelligenz sind nach Wechsler ebenso „[…] das Gedächtnis und die Fähigkeit zur sozialen Anpassung […]“ (Kubinger, 2009a, S. 59), also auch Faktoren, die auf indirektem Wege intellektuelles Verhalten verursachen beziehungsweise ausmachen.

Wechsler zielte in der Zusammenstellung seiner Intelligenz-Testbatterie darauf ab, durch unterschiedliche Untertestarten die Faktoren verbaler Intelligenz und praktischer (Handlungs-)Intelligenz zu erfassen (vgl. Kubinger, 2009a).

Damit war es möglich, neben einem sogenannten Intelligenzquotienten die Teilaspekte verbale Intelligenz und Handlungsintelligenz zu berechnen (Gerrig & Zimbardo, 2008).

2.1.2) Intelligenztests

„Ein *psychologisch-diagnostisches Verfahren* (vereinfacht oft ‚*Test*' genannt) erhebt unter standardisierten Bedingungen eine Informationsstichprobe über einen (oder mehrere) Menschen, indem systematisch erstellte Fragen/Aufgaben interessierende Verhaltensweisen oder psychische Vorgänge auslösen; Ziel ist es, die fragliche Merkmalsausprägung zu bestimmen." (Kubinger, 2009a, S.10)

Dieser Definition zufolge, sind Intelligenztests psychologisch-diagnostische Verfahren, deren Einsatz darauf abzielt, anhand der Antworten oder Reaktionen einer Testperson auf bestimmte Items, auf ihre Befähigung zur Erbringung geistiger Leistungen zu schließen. Mithilfe spezieller Testkennwerte wird versucht, je Aufgabengruppe die Ausprägungen in der zu messen beabsichtigten Fähigkeit zu quantifizieren.

Die Items von Intelligenztests sind so gestaltet, dass sie bei der Testperson intellektuelles Verhalten provozieren. Diese Verhaltensweisen, die durch die Antwort auf das Item (richtig/falsch) registriert werden, lassen auf die latente Eigenschaft „Intelligenz", beziehungsweise „kognitive Leistungsfähigkeit" schließen (vgl. Kubinger, 2009a).

Intelligenztests sind neben Entwicklungs-, Schul- und allgemeinen Leistungstests (z.B. Erfassung der Konzentration) sowie speziellen Funktions- und Eignungsüberprüfungen eine weitere Form von Leistungstests.

„Leistungstests zeichnen sich dadurch aus, dass von den Personen die Lösung von Aufgaben oder Problemen verlangt wird, die Reproduktion von Wissen, das Unterbeweisstellen von Können, Ausdauer oder Konzentrationsfähigkeit.“ (Rost, 2004, S. 43).

Ein Charakteristikum für diese Art von psychologisch-diagnostischen Verfahren ist demnach, dass die Testperson im Zuge dieser eine „maximum performance“ erbringen soll, also die momentan bestmögliche Leistung (Kubinger, 2003).

Zumeist handelt es sich um Testbatterien, die sich aus verschiedenen Untertests zusammensetzen.

Die Gestaltungsprinzipien für diese Art psychologischer Diagnostik lassen sich folgendermaßen beschreiben:

(1) Power-Test:

Bei einem Power-Test sollen Testpersonen die vorgegebenen Items ohne Zeitbegrenzung bearbeiten. Daran ist zu kritisieren, dass die Vorgabe von Gruppenverfahren eine gewisse Zeitkomponente braucht, um eine wirtschaftliche Durchführung und Standardisierung der Testsituation gewährleisten zu können (vgl. Kubinger, 2003).

(2) Speed-Test:

Bei einem reinen Geschwindigkeitstest werden laut Moosbrugger und Kelava (2008) einfache Testaufgaben vorgegeben, die von den meisten Testpersonen bearbeitet werden können. Somit wird nicht die Fähigkeit der Testpersonen erfasst, sondern die Bearbeitungsgeschwindigkeit. Mit dieser Art psychologisch-diagnostischer Verfahren können zum Beispiel Konzentration oder Aufmerksamkeit gemessen werden.

(3) Speed-and-Power-Tests:

Hier wird Testpersonen für die Bearbeitung der Testaufgaben ein gewisses Zeitbudget eingeräumt – dafür gibt es zwei Möglichkeiten: Einerseits kann die Bearbeitungszeit für ganze Untertests beziehungsweise auch für den gesamten Test vorgegeben sein, andererseits ist es auch denkbar, dass einzelne Items in einer gewissen Zeit zu beantworten sind (Kubinger, 2009a).

An der Vorgabe von Speed-and-Power-Tests ist zu kritisieren, dass es zur Benachteiligung von Testpersonen mit einem langsameren Arbeitsstil kommen kann. Personen, die eher reflexiv vorgehen und sich mehr Zeit für die richtige Beantwortung nehmen, können womöglich in der vorgegebenen Zeit weniger Items bearbeiten als impulsive Testpersonen. „Offensichtlich schneiden Testpersonen nur dann in einem *Speed-and-Power*-Test gut ab, wenn sie sowohl leistungsstark als auch schnell arbeiten.“ (Kubinger, 2009a, S. 144)

2.1.3) Intelligenzquotient

Der Intelligenzquotient stellt in der traditionellen Intelligenzdiagnostik ein Globalmaß für die Leistung in einem Intelligenztest dar. Er dient der Quantifizierung und Vergleichsmöglichkeit geistiger Leistungsfähigkeit (vgl. Kubinger, 2003).

Der bloße IQ-Wert ist dabei aber ohne weitere Informationen relativ aussageleer. Dieser einzelne Wert ermöglicht keine adäquaten Rückschlüsse auf die kognitiven Stärken oder Schwächen einer Testperson, da er eine „kompensatorische Wirkung“ der additiv verrechneten Leistungen im Intelligenztest voraussetzt, was aus praktischer Sicht zu kritisieren ist. Demnach müssten die kognitiven Schwächen durch kognitive Stärken ausgebessert werden – unabhängig von den betroffenen Unter-

tests beziehungsweise Leistungsdomänen. Somit könnte zum Beispiel eine schwache Leistung in einem Untertest zur Ermittlung des logischen Denkens durch starke Leistungen in einem Untertest zur Ermittlung der Merkfähigkeit ausgebessert werden. Da die Ergebnisse eines Intelligenztests auf alltägliche Situationen der Testperson verallgemeinert werden können sollten, deshalb erscheint es aus praktischer Perspektive wenig plausibel, dass intellektuelle Teilkompetenzen, unabhängig von deren Bereich, auch im Alltag einander einfach kompensieren (vgl. Kubinger, 2009a).

Die Konstruktion des AID weicht davon ab, einen einzigen IQ-Wert für eine Testperson zu ermitteln. Vielmehr geht es darum, ihre individuellen Stärken und Schwächen in den verschiedenen Domänen geistiger Leistungsfähigkeit aufzuzeigen (vgl. Kubinger, 2009b). Es wird eine sogenannte Profil-Interpretation angestrebt. „Eine Profilanalyse besteht darin, die gewonnenen Testwerte, zumeist für eine einzige Person, in ihrer Relation zueinander aufzuschlüsseln.“ (Kubinger, 2004, S. 337). Hierbei werden die Testleistungen in den einzelnen Untertests miteinander in Beziehung gesetzt (vgl. Kubinger, 2003).

2.2) Einzel – und Gruppenverfahren

2.2.1) Individual- vs. Gruppentests

Einzeltestungen ermöglichen eine bessere Interaktion zwischen Testperson und Testleiterin bzw. Testleiter – man kann gut auf Fragen eingehen, es fällt eher auf, wenn die Testperson die Instruktion nicht richtig verstanden hat, das Verhalten in der Testsituation kann gut beobachtet werden, es kommt zu keinen Störungen durch an-

dere anwesende Testpersonen und das Testmaterial kann speziell gestaltet sein (z.B. in Form von Puzzles oder Bausteinen) (Kubinger, 2003) oder adaptives Testen kann realisiert werden. Es ist auch möglich, Testpersonen ohne Lese- oder Schreibkompetenzen zu testen.

Gruppentestungen hingegen sind an eine Verschriftlichung der Items gebunden, eine besondere Gestaltung des Testmaterials ist ebenso kaum möglich.

2.2.2) AID 3

Der AID 3 (Kubinger & Holocher-Ertl, 2014, in Druck) unterscheidet sich in seiner Gestaltung des Testmaterials sowie der besonderen Möglichkeiten zur Testung und Beobachtung deutlich von anderen, zum Teil gruppentauglichen Intelligenztests. Hier ist die Testperson dazu aufgefordert, in einer interaktiven Situation auf das durch die Testleiterin bzw. den Testleiter vorgegebene Testmaterial zu reagieren.

2.3) Antwortformate

Die Gestaltung der Items spielt eine wichtige Rolle in der Psychologischen Diagnostik, da sie einen wesentlichen Anteil an der Standardisierung der Testsituation haben.

2.3.1) Freies Antwortformat

Hier muss die Testperson selbständig die Antwort auf ein Testitem generieren, es sind keine Antwortmöglichkeiten vorgegeben. Im AID 3 zum Beispiel antwortet die Testperson mündlich auf die von der Testleiterin bzw. dem Testleiter gestellten Fragen. Als Beispiel kann man hier die Vorgabe der Items des Untertests 6 „Synonyme finden“ genannt werden. Der Testleiter bzw. die Testleiterin nennt der Testperson einen Begriff, zu welchem das Kind einen gleichbedeutenden Begriff finden soll.
Ein Vorteil des freien Antwortformats in der Leistungsdiagnostik liegt darin, dass durch die Testperson selbst generierte Antworten diagnostisch sehr aufschlussreich sind.
Nachteilig ist hingegen, dass die Testdauer etwas länger ist und sowohl die Bearbeitung als auch die Auswertung der Aufgaben mit einem größeren Aufwand verbunden sind. Ebenso kann es zur Beeinträchtigung der Auswertungsobjektivität kommen, da die Beurteilung teilrichtiger Antworten nach „richtig“ oder „falsch“ im Ermessen des Testleiters liegen kann (frei nach Kubinger, 2009a).

2.3.2) Multiple-Choice-Format

Multiple-Choice-Format bedeutet, dass die Testperson nach der Frage-, Aufgabenstellung aus Antwortalternativen die richtige(n) Antwort(en) finden und markieren soll. Vorteile finden sich hier vor allem in der Wirtschaftlichkeit in Auswertung und Bearbeitung und der höheren Verrechnungssicherheit (Kubinger, 2009a).
Ein Kritikpunkt an diesem Vorgabemodus ist, dass Testpersonen bei Unkenntnis der richtigen Antwort versuchen könnten, die richtige Lösung zu erraten. Falls die Test-

person die Antwort erraten sollte, wird das Item als gelöst verrechnet, obwohl die Testperson nicht die zu messen beabsichtige Fähigkeit besitzt. Durch die Vorgabe solcher Multiple-Choice-Items kann die Itemschwierigkeit reduziert werden.
Dieses Problem könnte jedoch durch eine spezielle Konzeptualisierung des Tests umgangen werden: Kubinger und Gottschall führten 2007 eine Studie zu Ratewahrscheinlichkeiten bei Itempaaren mit identischem Inhalt durch. Laut dieser Studie unterscheidet sich die Schwierigkeit eines Items in den Antwortformaten „freies Antwortformat“ und „x aus 5“, also 0-5 richtige Antwortalternativen pro Item möglich, nicht signifikant von einander (Kubinger & Gottschall, 2007).

2.4) Item-Response-Theorie

Die Grundlagen dieser Studie basieren auf den Annahmen der Item-Response-Theorie, die sich zum Ziel gemacht hat, anhand der Itemantworten (Itemresponses) „Rückschlüsse auf interessierende Einstellungs-, Persönlichkeits- oder Fähigkeitsmerkmale“ zu ziehen (Moosbrugger, 2008, S. 216).

Laut Moosbrugger (2008) wird zwischen zwei Arten von Variablen unterschieden:
- *manifest*: hierbei handelt es sich um beobachtbare Merkmale, die sich zum Beispiel im Antwortverhalten zeigen
- *latent*: sind die Ausprägungen auf dem Personenmerkmal, zum Beispiel ihrer Fähigkeit, auf die anhand der manifesten Variablen, z.B. der Leistungen in einem Intelligenztest, geschlossen werden soll

Wichtig in diesem Zusammenhang ist die Itemhomogenität (genaueres dazu im Empirieteil dieser Arbeit), da diese eine Voraussetzung für den Schluss von Antwortverhalten auf die latente Variable darstellt. Lediglich die Fähigkeit der Testperson soll Einfluss auf die Beantwortung der Items eines Intelligenztests haben.

2.5) Prüfung der Konstruktäquivalenz

2.5.1) Konstrukt und Konstruktvalidität

Laut Tewes und Wildgrube (1999) handelt es sich bei einem Konstrukt um eine nicht direkt beobachtbare Eigenschaft, die mittels Beobachtungen erschließbar ist. „Konstrukte sind also allgemein anerkannte, aber eben nicht direkt beobachtbare ‚Phänomene', wie z.B. Intelligenz, Angst oder Stress." (Kubinger, 2009a, S.57).

Laut Kubinger (2009a) ist Konstruktvalidität gegeben, wenn ein psychologisch-diagnostisches Verfahren theoriegeleitete Vorstellungen bezüglich des zugrundeliegenden Konstrukts erfüllt, also zum Beispiel ein Intelligenztests tatsächlich nur intelligentes Verhalten der Testpersonen erfasst.
Demnach gilt Wechslers Intelligenztheorie als „ […] typisches Beispiel des faktorenanalytischen Ansatzes einer Konstruktvalidierung" (Kubinger, 2009a S. 59). Seine Intelligenz-Testbatterie erfasst nämlich voneinander unabhängige Bereiche der Intelligenz, verbal und praktisch.

Auf die Prüfung der Konstruktäquivalenz wird im Empirieteil dieser Arbeit näher eingegangen. In Bezug darauf sind noch Grundlagen zu Items und zur Prüfung der Homogenität dieser Items zu erläutern.

2.5.2) Prüfung Konstruktäquivalenz

Die Prüfung der Konstruktäquivalenz bezieht sich darauf, die „[…] Gleichwertigkeit zweier Versionen eines psychologisch-diagnostischen Verfahrens." (Kubinger 2003, S. 32) zu erfassen. Das heißt, dass es bei der psychologisch-diagnostischen Untersuchung einer Testperson also keinen Unterschied machen sollte, ob sie beispielsweise die Einzel- oder die Gruppenversion eines psychologisch-diagnostischen Verfahrens bearbeitet.

(1) „Item"

„Item" ist die Bezeichnung für die Aufgaben eines psychologisch-diagnostischen Verfahrens, auf die die Testperson antworten beziehungsweise reagieren soll.
Das Item ist der kleinste Baustein eines psychologisch-diagnostischen Tests. Es besteht aus zwei Komponenten: dem sogenannten Itemstamm und dem Antwortformat:
Der *Itemstamm* lässt sich als „Aufgabe" beschreiben, die die Testperson zu bearbeiten hat. Dieser kann laut Rost (2004) „[…] aus einer Frage, einer Aussage, einem Bild, einer Geschichte, einer Zeichnung oder einer Rechenaufgabe bestehen und stellt ganz allgemein die Situation dar, in der die Person ihr Testverhalten zeigt."
Das *Antwortformat* stellt die unterschiedlichen Möglichkeiten zur Erfassung des Testverhaltens dar, wie zum Beispiel:

- das Auswählen und Ankreuzen der vorgegebenen Antwortalternativen bei Multiple-Choice-Format,
- die wörtliche, schriftliche oder zeichnerische Formulierung einer Antwort bei freiem Antwortformat,
- oder das Markieren einer für die Testperson zutreffenden Abstufung bei Rating- beziehungsweise Analogskalen (Rost, 2004).

Die Itemauswahl muss einen bestimmten Geltungsbereich beschreiben, auf den die Itemantworten der Testperson verallgemeinert werden sollen (Rost, 2004). Welche Fragestellung soll durch die Vorgabe eines psychologisch-diagnostischen Verfahrens beantwortet werden? Beispiele für den Geltungsbereich sind z.B. die Diagnostik klinischer Störungsbilder, Berufseignung oder die Feststellung von Lern-, bzw. Leistungsproblemen (Kubinger, 2009a).

(2) Itemhomogenität

Bei homogenen Items handelt es sich um Testaufgaben, die dasselbe Personenmerkmal erfassen. Die Homogenität lässt sich bei Rasch-Modell-konformen Items zum Beispiel mit dem Martin-Löf-Test überprüfen. Dieser verwendet „[…] jedoch nicht die geschätzten Personenparameter […] sondern deren erschöpfende Statistiken, die Summenscores für beide Testhälften." (Rost, 2004, S. 351).
Erschöpfende Statistiken bieten eine unproblematische Möglichkeit zur Schätzung von Modellparametern. Sie geben auch an, „[…] welche Information aus den Testdaten ‚herangezogen' wird." (Rost 2004, S. 114)

2.5.3) Das Rasch-Modell

Das *Rasch*-Modell beschreibt laut Kubinger (2003) „ […] die Wahrscheinlichkeit, dass Testperson (Tp) v Item i löst (‚+'), in Abhängigkeit eines Personenparameters ξ_v, das ist die (wahre) Fähigkeit von v, und eines Itemparameters σ_i, das ist die (wahre) Schwierigkeit von *i*."

Die Lösungswahrscheinlichkeit eines Items lässt sich anhand folgender Formel herleiten:

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

aus Kubinger, 2009a, S. 89

Dem Rasch-Modell liegen folgende Annahmen zugrunde (vgl. Rost, 2004 & Kubinger, 2009a):

(1) Eindimensionalität:

Misst ein psychologisch-diagnostisches Verfahren eindimensional, so erfasst es ein einziges Personenmerkmal. Sowohl die Personenfähigkeit als auch die Itemschwierigkeit lassen sich somit an je einem Parameter festmachen. (frei nach Kubinger, 2009a)

(2) Lokale stochastische Unabhängigkeit:

Für jede Testaufgabe ist es unerheblich, welche Items die Testperson im Verlauf der Testung schon gelöst hat oder welche sie noch lösen wird – lediglich ihre Fähigkeit und die Schwierigkeit des Items sind für das Lösen oder Nicht-Lösen ausschlaggebend. (vgl. Kubinger, 2003)

(3) Logistische Wahrscheinlichkeitsfunktion,

zum Treffen von Annahmen bezüglich des Zusammenhangs zwischen manifestem Verhalten und latenten Persönlichkeitseigenschaften. Ausgangspunkt sind dichotom zu verrechnende Items – zum Beispiel wie beim AID 3 nach „richtig“ oder „falsch“. Im dichotom logistischen Rasch-Modell wird für die Testperson v mit dem Personenfähigkeitsparameter ξ_v in Abhängigkeit vom Itemschwierigkeitsparameter σ_i eine Antwortwahrscheinlichkeit bestimmt, nämlich $P(x_{vi}=0)$ für eine falsche Lösung oder $P(x_{vi}=1)$ für eine richtige Lösung.

Übersteigt der Personenfähigkeitsparameter ξ_v den Itemschwierigkeitsparameter σ_i, so geht die Lösungswahrscheinlichkeit gegen 1. Übertrifft die Itemschwierigkeit die Personenfähigkeit, so strebt die Lösungswahrscheinlichkeit gegen 0. Entsprechen σ_i und ξ_v einander, dann ist die Wahrscheinlichkeit der Lösung gleich 0,5. Items, deren Schwierigkeit gleich dem Personenfähigkeitsparameter ist, stellen die höchste Anforderung an die zugrundeliegende Merkmalsausprägung. Hier liegt der Schluss nahe, dass sie die größte diagnostische Information liefern, die Lösungswahrscheinlichkeit ist gleich der Gegenwahrscheinlichkeit (Moosbrugger & Kelava, 2008).

(4) Stichprobenunabhängigkeit:

Die resultierenden Testwerte können *spezifisch objektiv* interpretiert werden. Spezifische Objektivität bedeutet, dass der „ […] Unterschied in den Fähigkeiten ξ_v und ξ_w zwischen je zwei Personen *v* und *w* kann unabhängig davon bestimmt werden, welche Aufgaben des Tests dafür herangezogen werden; bzw. umgekehrt und wichtiger, der Vergleich je zweier Aufgaben *i* und *j* bezüglich σ_i und σ_j ist unabhängig davon möglich, welche Stichprobe dafür verwendet wird.“ (Kubinger 2009a, S. 89). Sofern das Rasch-Modell gilt ist die Schätzung des Itemparameters auch ohne Personenparameter möglich, denn die Summe gelöster Testaufgaben stellt dann einen aussagekräftigen Testwert dar und gleiche Testwerte bedeuten gleiche Fähigkeit (Kubinger, 2009a).

Das dichotom-logistische Modell nach Rasch muss nach Kubinger (2003) notwendiger weise gelten, wenn der Rohwert nur die relevante Information über die Testperson liefern soll. Es muss *notwendigerweise* gelten, ansonsten ist die Summierung der richtigen Lösungen kein fairer Verrechnungsmodus (Kubinger, 2000).

3) Adaptives Intelligenz Diagnostikum

3.1) Allgemeines

Das Adaptive Intelligenz Diagnostikum 3 (AID 3) von Kubinger und Holocher-Ertl (2014, in Druck) ist eine Intelligenz-Testbatterie zur Erfassung der kognitiven Leistungsfähigkeit. Hierbei handelt es sich um ein Verfahren zur Einzeltestung von Kindern und Jugendlichen im Alter von 6 bis 15,11 Jahren.
Das Adaptive Intelligenz Diagnostikum kann in verschiedenen Bereichen, wie zum Beispiel Eignungs-, Rehabilitations-, Entwicklungs- oder auch neuropsychologischer Diagnostik eingesetzt werden (Kubinger, 2009a).

3.2) Neuerungen

Die Unterschiede des AID 3 (Kubinger & Holocher-Ertl, 2014, in Druck) im Vergleich zum AID 2.2 (Kubinger, 2009b) werden kurz umrissen:

3.2.1) Items

In einzelnen Untertests des AID 2.2 (Kubinger, 2009b) fanden sich Items, die im AID 3 (Kubinger und Holocher-Ertl, 2014, in Druck) nicht mehr zur Anwendung kommen. Es wurden auch neue, aktuellere Items entwickelt. Diplomarbeiten, die die Items des AID 3 zum Thema haben, stammen zum Beispiel von Hellebart (2013), Görner (2013), Weber (2011) und Hagenmüller (2011).

3.2.2) Zusatztests

Für die Untertests 5 „Unmittelbares Reproduzieren“ und 6 „Synonyme finden“ wurden neue Zusatztests hinzugefügt.

Neben 5a „Unmittelbares Reproduzieren“ und 5b „Merken und Einprägen“ erfasst nun Zusatztest 5c „Lernen und langfristiges Merken“ eine weitere Domäne des Gedächtnisses und zusätzlich der Lernkompetenz der Testperson.

Zusatztest 6a „Antonyme finden“ kann zusätzlich zum Untertest 6 „Synonyme finden“ vorgegeben werden. Hier wird das elementare Wortverständnis der Testperson erfasst. Zusatztest 6a wird in der Diplomarbeit von Weber (2011) näher besprochen.

3.2.3) Untertest 12 „Formale Folgerichtigkeit“

Die Items dieses Untertests bestehen aus geometrischen Formen, die es in zwei Farben (grün und gelb) und in zwei verschiedenen Größen gibt. Aufgabe der Testperson ist es, eine logische Reihe geometrischer Formen mit einer geometrischen Figur auf dem dafür vorgesehenen Platz im Testheft durch Hinlegen zu ergänzen. (frei nach Hagenmüller, 2011).

3.2.4) AID-Gruppe

Ebenso neu ist die Entwicklung einer gruppentauglichen Version des AID, mit dem Arbeitstitel AID-Gruppe (Kubinger & Hagenmüller, in Vorb.).

Zur Machbarkeit gruppentauglicher AID-Untertests wurden verschiedene Diplomarbeiten behandelt, zum Beispiel Böck (2010), Hofmayer (2012), Eiter (2011), Wagner (in Vorb.), sowie Schock (in Vorb.).

3.2.5) Untertests des AID 3

Im Folgenden werden die Untertests des AID 3 tabellarisch dargestellt, die optionalen Zusatztests wurden bereits unter Abschnitt 2.2 „Neuerungen“ dargestellt.
Die Informationen für die tabellarische Darstellung der Untertests wurden bis auf Untertest 12 „Formale Folgerichtigkeit“ (Hagenmüller, 2011) dem Manual des AID 2 (Kubinger, 2009a) entnommen.

Tabelle 1

Untertest	Messintention	Vorgabe
1 Alltagswissen	Fähigkeit, sich Sachkenntnisse über Inhalte anzueignen, die in der heutigen Gesellschaft alltäglich sind	adaptiv
2 Realitätssicherheit	erfasst, inwieweit die Wirklichkeit um Dinge des Alltags verstanden wird	adaptiv
3 Angewandtes Rechnen	unabhängig von schulischen Rechenfertigkeiten, werden bei der Problemlösung alltäglicher Aufgabenstellungen durch Schlussfolgerungen die passenden	adaptiv
4 Soziale und Sachliche Folgerichtigkeit	Fähigkeit, die Abfolge sozialen Geschehens bzw. alltäglicher Sachgegebenheiten zu verstehen und zu kontrollieren	adaptiv
5 Unmittelbares Reproduzieren	Kapazität der seriellen Informationsverarbeitung (verbal-akustisch)	konventionell
6 Synonyme Finden	elementares Sprachverständnis, erfassen der Bedeutung sprachgebundener Begriffe, Wortschatz	adaptiv
7 Kodieren und Assoziieren	zwei voneinander z.T. unabhängige Fähigkeiten: die Informationsverarbeitungsschnelligkeit und die Fähigkeit inzidentellen Lernens	konventionell
8 Antizipieren und Kombinieren	schlussfolgerndes Denken, Teile eins Ganzen erkennen und dieses Ganze gestalten zu können	adaptiv
9 Funktionen Abstrahieren	Fähigkeit, durch Abstraktion zu einer Begriffsbildung zu gelangen.	adaptiv
10 Analysieren und Synthetisieren - abstrakt	Fähigkeit, komplexe Gestalten durch eine geeignete Strukturierung reproduzieren zu können	adaptiv
11 Soziales Erfassen und Sachliches Reflektieren	inwieweit begreift die Testperson Sachzusammenhänge der „gesellschaftlichen“ Umwelt begreift, Wissen um soziale Bedingungen und angepasstes Verhalten	adaptiv
12 Formale Folgerichtigkeit*	logisch-schlussfolgerndes Denken	adaptiv

Das Besondere an der Testvorgabe ist neben der Realisierung des adaptiven Testens (siehe Abschnitt 2.3), die Kombination aus Untertests mit Speed-, Power- und Speed-and-Power-Komponenten (vgl. Kubinger, 2009a).

Die Items werden bei 9 von 12 Untertests dichotom nach „gelöst“ und „nicht gelöst“ verrechnet und entsprechen dem Rasch-Modell.

Die Untertests 5 „Unmittelbares Reproduzieren“, die Zusatztests 5a, 5b und 5c sowie 7 „Kodieren und Assoziieren“ sind aufgrund der Gestaltung an eine konventionelle Vorgabe gebunden.

Das Adaptive Intelligenz Diagnostikum ermöglicht eine förderorientierte Diagnostik, bei der die intellektuellen Stärken und Schwächen einer Testperson miteinander in Beziehung gesetzt werden. Zu diesem Zweck ist eine hohe Messgenauigkeit unumgänglich. Diese wird durch das adaptive Testen gewährleistet (näheres siehe Abschnitt 2.2) (Kubinger & Holocher-Ertl, 2012).

3.3) Besonderheiten des Adaptiven Intelligenz Diagnostikums

Das Besondere am AID ist, dass spezielle Testkennwerte bestimmt werden können, die eine förderorientierte Diagnostik ermöglichen (Kubinger, 2009a):

3.3.1) Range

„Der ***Range* der "Intelligenz"** beschreibt die Schwankungsbreite der Untertestleistungen. Ein sehr hoher (und somit überdurchschnittlicher) Range kann auf spezifische Begabungen und/oder Schwächen der betreffenden Testperson hindeuten." (Kubinger, 2009b).

Der Wert ergibt sich aus der schlechtesten und der besten Untertestleistung im AID.

Es wird auch die zweitniedrigste Untertestleistung angegeben, um die intellektuelle Mindestfähigkeit abgesehen von einer Teilleistungsstörung interpretieren zu können (vgl. Kubinger, 2009b).

3.3.2) Intelligenzquantität

Die untere Grenze der Intelligenzquantität kann als Äquivalent zum IQ betrachtet werden. Sie ergibt sich aus der schlechtesten Untertestleistung. Sofern keine Teilleistungsstörung besteht – großer Unterschied zwischen niedrigster und zweitniedrigster Testleistung – kann die Intelligenzquantität als äquivalentes Maß für den herkömmlichen Intelligenzquotienten herangezogen werden und durch Transformation von einem *T*-Wert in einen IQ-Wert umgerechnet werden (Kubinger, 2009b).

3.3.3) Testprofil

Anhand des sogenannten „Testprofils" werden die Leistungshochs und –tiefs einer Testperson graphisch dargestellt. Die *T*-Werte der einzelnen Untertests werden in der Grafik (unter-, über-, durchschnittlicher Bereich) markiert und stellen dann ein Leistungsprofil dar, anhand dessen sich die Leistungsspitzen und Schwächen der Testperson verdeutlichen lassen (Kubinger, 2009b).

„Der Vorteil dieser Methode liegt in der Anschaulichkeit der Ergebnisse, da man auf einen Blick einen Eindruck vermittelt bekommt.“ (Kubinger, 2003, S.337)

3.3.4) Primär-IQ

Dieser ergibt sich aus den Untertestscores 1 „Alltagswissen“, 3 „Angewandtes Rechnen“, 6 „Synonyme Finden“, 9 „Funktionen Abstrahieren“, 11 „Soziales Erfassen und Sachliches Reflektieren“ und stellt den Mittelwert dieser fünf Untertests dar (Kubinger, 2009b). Das Heranziehen eben dieser 5 Untertests zur Berechnung des Primärintelligenzquotienten, lässt sich faktorenanalytisch begründen. (Kubinger, 2009b)

Dieser so errechnete Intelligenzquotient bildet somit nur eine einzige Intelligenzdimension ab, was den Informationsgewinn im Vergleich zu herkömmlichen Intelligenz-Testbatterien beträchtlich erhöht. In den handelsüblichen Intelligenztests werden nämlich alle Untertests zur Berechnung des Quotienten herangezogen, was eine kompensatorische Wirkung kognitiver Leistungsdomänen untereinander voraussetzen würde (Kubinger & Holocher-Ertl, 2012).

3.4) Untertest 9 „Funktionen Abstrahieren“

„Funktionen Abstrahieren“ gehört zu der Gruppe von Untertests, die verbal-akustische Fertigkeiten erfassen, genauso wie „Alltagswissen“, „Angewandtes Rechnen“, „Unmittelbares Reproduzieren“, „Antonyme finden“ sowie „Soziales und sachliches Reflektieren“ (Kubinger, 2009a).

„Mit dem Untertest **9 Funktionen Abstrahieren** soll die Fähigkeit erfasst werden, durch Abstraktion zu einer Begriffsbildung zu gelangen“ (Kubinger, 2009b, S. 9, 11). Hier werden der Testperson pro Item zwei Begriffe genannt, die eine gemeinsame Funktion haben. Diese Gemeinsamkeit soll von der Testperson benannt werden.

„Funktionen Abstrahieren“ erfasst eine kognitive Leistungsdomäne, die sich laut Kubinger (2009a) wie „Gemeinsamkeiten“ und „Analogien“ aus dem IST 2000 R und „Gleiche Wortbedeutungen“ aus dem WIT-2, mehr oder weniger dem Leistungsfaktor „logisches schlussfolgerndes Denken anhand verbalen Materials“ unterordnen lässt (Kubinger, 2009a).

3.5) Adaptives Testen

3.5.1) Idee

Laut Kubinger (2003) geht die Idee der adaptiven Testvorgabe von der Kritik am konventionellen Testen aus:

Bei der konventionellen Testvorgabe zielt der Testautor darauf ab, mit einem einzigen psychologisch-diagnostischen Verfahren eine möglichst große Personengruppe testen können, es soll also für Personen unterschiedlichen Alters oder unterschiedlicher Fähigkeiten geeignet sein. Um dies zu gewährleisten, muss ein solches Verfahren viele und vor allem unterschiedlich schwierige Testaufgaben beinhalten. Die Vorgabe beginnt üblicherweise mit dem einfachsten und endet mit dem schwierigsten Item.

Das Problem der konventionellen Testvorgabe liegt einerseits in der Testlänge, die

erforderlich ist, damit die Itemschwierigkeiten den unterschiedlichen Fähigkeiten der Testpersonen gerecht werden können. Andererseits darin, allen Testpersonen dieselben Items vorzugeben. „Kritisch ist dabei, dass leistungsfähigen Personen einige Items regelmäßig zu leicht, zumindest sehr leicht fallen, leistungsschwachen dagegen wieder andere Items zu schwer, zumindest sehr schwer.“ (Schlüsselbegriffe, Kubinger 2003, S. 1)

Diese Items liefern keine diagnostisch wertvolle Information, da der Testleiter schon im Vorhinein davon ausgehen kann, welche Aufgaben eine bestimmte Person wahrscheinlich lösen kann und welche wahrscheinlich nicht. Die vorgegebene Reihenfolge der Testaufgaben beeinflusst aber auch die Motivation und das Testverhalten: die zu einfachen Items können die Testperson unter Umständen langweilen, die zu schwierigen können dagegen frustrierend wirken (Kubinger, 2009a).

Aus diesen Kritikpunkten ergibt sich die Forderung nach einem psychologisch-diagnostischen Verfahren, das der Leistungsfähigkeit einer Testperson angemessen ist, ökonomisch testet und trotz einer kürzeren Testdauer einen großen Informationsgewinn liefert, zumindest gleich groß wie der Informationsgewinn aus der konventionellen Vorgabe.

3.5.2) Grundlagen

Eine simple Auswertungsstrategie ist, als Testwert die Anzahl richtig gelöster Testaufgaben heranzuziehen. Dabei erhält man jedoch keine Information darüber, welche Items genau eine Testperson lösen konnte, was die Vergleichbarkeit der Testleistun-

gen erheblich einschränkt, da somit keine adäquaten Rückschlüsse auf Fähigkeitsunterschiede der Testpersonen möglich sind. Beim adaptiven Testen werden den Testpersonen nur solche Aufgaben vorgegeben, die deren Fähigkeitsniveau entsprechen (vgl. Kubinger, 2009a).

Adaptives Testen beruht auf den Grundlagen der Item-Response-Theorie. Bei diesem speziellen Testvorgehen richtet sich die Vorgabe der Testaufgaben nach dem zuvor gezeigten Antwortverhalten der Testperson. (Kubinger, 2003)
Die Summenscores als Testwerte heranzuziehen ist dann möglich, wenn das psychologisch-diagnostische Verfahren Rasch-Modell-konform ist, da nur dann ein Zusammenhang zwischen Testwert und der zu messen beabsichtigten Variable besteht. Die Rasch-Modell-Konformität eines Verfahrens ist die Voraussetzung dafür, dass der Summenscore die Merkmalsausprägung einer Testperson adäquat abbildet und dieser somit als Testwert herangezogen werden kann. Nur wenn das Rasch-Modell gilt, liefern Items, deren Schwierigkeit der Personenfähigkeit entspricht, die meiste Information über die Fähigkeit der Testperson, da die Lösungswahrscheinlichkeit dann 50% beträgt (vgl. Kubinger, 2009a).

3.5.3) Arten adaptiven Testens

(1) Tailored-Testing

Eine rasche Schätzung des Personenparameters wird mithilfe einer maßgeschneiderten („tailored“) Vorgabe der Testitems ermöglicht. Hierbei wird der Person jenes Item zuerst vorgegeben, welches einen mittleren Schwierigkeitsgrad aufweist. Je nachdem, ob sie dieses lösen kann oder nicht, soll sie als nächstes das schwierigste

bzw. das leichteste Item des Itempools bearbeiten. Abhängig von der Lösung dieser Testaufgabe, wird als nächstes ein Item ausgewählt, welches die Testperson wahrscheinlich lösen kann. Ab diesem Punkt kann ein Fähigkeitsparameter geschätzt werden und eine dazu entsprechend schwierige Aufgabe vorgegeben werden. Das Prozedere des tailored-testing, sprich das Schätzen des Personenfähigkeitsparameters und die Auswahl eines informativen Items, ist nur durch aufwändige Rechenoperationen zu bewerkstelligen, weshalb diese Art des adaptiven Testens computergebunden ist (vgl. Kubinger, 2009a).

(2) Branched-Testing:

Diese Form des adaptiven Testens ist nicht an eine computerisierte Vorgabe gebunden, dafür ist die Genauigkeit der Parameterschätzung geringer. Wie der Name schon sagt, lässt sich branched-testing durch ein Verzweigungsschema zwischen Itemgruppen bewerkstelligen (Kubinger, 2009a).
Beispielsweise sind im AID 3 die Items in Gruppen zusammengefasst. Die Testung wird mit jener Aufgabengruppe gestartet, die altersgemäße Ansprüche an die Testperson stellt. Anhand der Testleistung in den 5 zusammengefassten Items wird die Testperson zu neuen Items „weiterverzweigt“ – dafür gibt es 3 Möglichkeiten: Löst die Testperson keine oder nur eine Aufgabe einer Gruppe, so wird die nächste Gruppe einfachere Items enthalten. Löst sie 2 oder 3 Aufgaben, so wird die nächste Aufgabengruppe aus etwa gleich schwierigen Items bestehen und bei 4 oder 5 richtigen Lösungen wird die nächste Aufgabengruppe schwierigere Items enthalten (vgl. Kubinger, 2003; Kubinger, 2009a).

II Empirieteil

4) Methoden

4.1) Studiendesign

Geplant war die Testung von 200 Schülerinnen und Schülern der 3. bis 9. Schulstufe aus Ober- und Niederösterreich. Ihnen sollte eine Itemauswahl der Untertests 4 und 9 des AID 3 (Kubinger & Holocher-Ertl, in Druck) und AID-Gruppe (Kubinger & Hagenmüller, in Vorb.) vorgegeben werden.

Zur Vorgabe der beiden Testbedingungen wurde ein balanciertes Studiendesign mit zwei Gruppen gewählt. Eine Hälfte der Schülerinnen und Schüler bearbeitete zuerst die Itemauswahl des AID-Gruppe, der anderen Hälfte wurde zuerst die Itemauswahl des AID 3 vorgegeben (siehe auch Abb. 1).

	Häufigkeit	**Prozent**
Gruppentestung zuerst	111	52,35
Einzeltestung zuerst	101	47,64
Gesamt	**212**	**100,00**

Tabelle 2 Verteilung der Häufigkeiten im Studiendesign

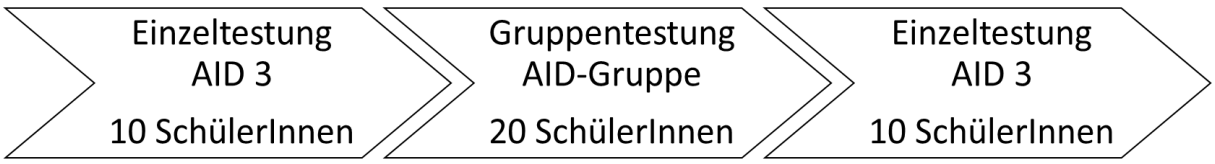

Abbildung 1 Illustration des Vorgabeschemas am Beispiel eine Klasse mit 20 SchülerInnen

Damit keine motivationalen Effekte die Datenqualität beeinflussen, wurden die Schülerinnen und Schüler in Oberösterreich im Juni getestet, in Niederösterreich fanden die Testungen im September statt. Somit können Motivationseffekte, die vor beziehungsweise nach den Sommerferien auftreten, als ausbalanciert betrachtet werden.

4.2) Deskriptivstatistik

4.2.1) Stichprobe

Da der Martin-Löf-Test im Statistikprogramm „R“ nicht mit unvollständigen Daten berechnet werden kann, wurden 7 Testpersonen von den Berechnungen ausgeschlossen, da sie es nicht schafften, die Itemauswahl des AID-Gruppe fertig zu bearbeiten.

Stichprobe			
Niederösterreich		Geschlecht	
		männlich	Weiblich
		Anzahl	Anzahl
Schulstufe	3. Klasse VS	7	3
	4. Klasse VS	8	6
	1. Klasse HS/AHS	15	11
	2. Klasse HS/AHS	2	9
	3. Klasse HS/AHS	5	5
	4. Klasse HS/AHS	1	15
	5. Klasse AHS / 1. Klasse BHS	2	10

Tabelle 3 Häufigkeiten NÖ

Stichprobe			
Oberösterreich		Geschlecht	
		Männlich	Weiblich
		Anzahl	Anzahl
Schulstufe	3. Klasse VS	10	8
	4. Klasse VS	10	7
	1. Klasse HS/AHS	5	10
	2. Klasse HS/AHS	5	10
	3. Klasse HS/AHS	7	9
	4. Klasse HS/AHS	4	11
	5. Klasse AHS / 1. Klasse BHS	2	15

Tabelle 4 Häufigkeiten OÖ

Getestet wurden 219 Schülerinnen und Schüler im Alter zwischen 8 und 16 Jahren aus Schulen in Ober- und Niederösterreich. Darunter waren 60,7% weiblichen Geschlechts, 39,3% männlich (die Häufigkeiten männlicher/weiblicher Testpersonen verteilt über Schulstufen finden sich in Tabelle 3 für Niederösterreich und in Tabelle 4 für Oberösterreich).

	Schultyp				
	Volksschule	Hauptschule	Gymnasium	HWL	BAKIP
Anzahl	59	31	75	30	17

Tabelle 5 Verteilung über Schultypen

Über die Schultypen verteilen sich die Testpersonen folgendermaßen: 28,3% besucht zum Testzeitpunkt die Volksschule, 14,2% die Hauptschule, 34,3% das Gym-

nasium und 23,3% besuchten eine neue Mittelschule (die Häufigkeiten der Testpersonen verteilt über die Schultypen finden sich in Tabelle 5).

4.3) Testsituation

Die Itemauswahl der beiden Untertests wurde als Power-Testung durchgeführt, es wurde keine begrenzte Bearbeitungszeit vorgegeben. Die Schülerinnen und Schüler hatten eine Unterrichtseinheit lang für die Gruppentestungen Zeit, die Dauer der Einzeltestungen variierte individuell und hing von den Leistungen der Schüler selbst ab. Im Durchschnitt dauerte eine Einzeltestung etwa 30 Minuten.

Kannten Schülerinnen und Schüler ein Wort beziehungsweise dessen genaue Bedeutung nicht, so wurden von den Testleiterinnen, wie es im Manual vorgeschlagen wird, keine Auskunft darüber gegeben. Die Testpersonen wurden darauf aufmerksam gemacht, dass sie später eine Lehrerin oder die Eltern dazu fragen können.
Bevor näher auf die verschiedenen Testsituationen eingegangen wird, werden zum besseren Verständnis erst die Antwortformate beider Testformen beschrieben.

4.3.1) Einzeltestungen

Vorgegeben wurde je eine Itemauswahl der Untertests 4 („Soziale und sachliche Folgerichtigkeit“) und 9 („Funktionen Abstrahieren“) des AID 3 sowie des neuen Gruppenverfahrens AID Gruppe. Die Hälfte jeder Klasse bearbeitete vor der Gruppentestung die Aufgaben des AID 3, die andere Hälfte danach.
Für die Einzeltestungen wurde in den Schulen ein eigener Raum zur Verfügung ge-

stellt, um ein ungestörtes Arbeiten zu gewährleisten. Jeder Testperson wurde ein Code zugeteilt, der sich aus den Anfangsbuchstaben des Vor- und Nachnamen sowie aus dem Tag und dem Monat des Geburtsdatums zusammensetzt. Außerdem wurde auf dem Protokollbogen die Leistung des Kindes aufgezeichnet. Dafür waren 19 Kästchen für UT4 und 30 Kästchen für UT 9 in Fünferblöcken angebracht.

Zuerst wurde der UT 4 „Soziale und Sachliche Folgerichtigkeit“ vorgegeben. Testmaterial sind Bildgeschichten auf deren Rückseite die Itemnummer, sowie je nach Anzahl der Bilder pro Item die Buchstaben a – g und die Zahlen 1 – 8 vermerkt sind. Die Testleiterin bzw. der Testleiter legt die Bilder anhand der Buchstaben aufsteigend auf – sodass die Testsituation für Schülerinnen und Schüler immer die gleiche ist. Aufgabe der Testperson ist es, die Bilder in die richtige Reihenfolge zu bringen. Ein Item gilt als richtig gelöst, wenn die Geschichte anhand der aufsteigenden Zahlen an der Bildrückseite richtig geordnet wurde.
Nach der Bearbeitung der 19 Bildgeschichten wurde der UT 9 „Funktionen Abstrahieren“, aus 30 Items bestehend, bearbeitet. Hier werden je Item zwei Begriffe vorgelesen. Aufgabe der Testperson ist es, die Gemeinsamkeit der beiden Begriffe zu nennen.

4.3.2) Gruppentestung

Die Gruppentestungen durften im Rahmen von Unterrichtseinheiten stattfinden. Jene Schüler, deren Eltern die Teilnahme an den Testungen nicht erlaubten, verließen mit der Lehrerin bzw. dem Lehrer den Klassenraum.
Zuerst wurden Testhefte ausgeteilt – abwechselnd nach Testheft A und B, die sich

durch eine abgeänderte Itemreihenfolge unterschieden, damit kein Abschreiben möglich war.

Auf dem Deckblatt sollten die Schülerinnen und Schüler ihr Alter in Jahren, ihren Schultyp und Klasse, ihr Geschlecht sowie das Testdatum vermerken. Anschließend wurde UT 4 bearbeitet. Für die Instruktion wurden zwei Beispielitems gemeinsam bearbeitet, anschließend sollten die Schülerinnen und Schüler die 30 Testitems in Stillarbeit bearbeiten.
Sobald alle Testpersonen fertig waren, wurde Untertest 9 wieder mit zwei Beispielitems instruiert und die 30 Testitems waren wieder in Stillarbeit zu bearbeiten.

4.3.3) Antwortformate bei AID 3 und AID-Gruppe

Die Antwortformate von AID 3 und AID Gruppe unterscheiden sich dahingehend, dass die Einzelversion auch von Kindern ab einem Alter von 6 Jahren, also auch ohne Lese- oder Schreibkompetenz bearbeitet werden kann, und zwar in einer interaktiven Frage-Antwortsituation zwischen Testperson und Testleiterin beziehungsweise Testleiter.

Für diese Diplomarbeit wurde den Kindern eine Itemauswahl aus UT 4 „Soziale und Sachliche Folgerichtigkeit“ sowie UT 9 „Funktionen Abstrahieren“ vorgegeben.
Den Testpersonen wurden im Untertest 4 Bilder vorgelegt, die richtig zu ordnen sind und im Untertest 9 Fragen zur Gemeinsamkeit zweier Begriffe gestellt (genaueres siehe Abschnitt 3.3).

Die Vorgabe eines Gruppenverfahrens ist an eine Verschriftlichung der Testitems und an Lesekompetenz seitens der Testpersonen gebunden:

Im Untertest 4 sind je nach Item unterschiedlich viele Begriffe in eine logische Reihenfolge zu bringen; unter den Begriffen befinden sich Kästchen, in die die Testperson Ziffern zur Ordnung der Begriffe eintragen soll.

Untertest 9 ist im Gruppenverfahren so gestaltet, dass unter einem fettgedruckten Begriff fünf weitere Begriffe angeordnet sind, von denen zwei dieselbe Funktion wie der obenstehende haben und die restlichen drei als sogenannte Distraktoren dienen.

Der Zweck eines *Distraktors* ist es laut Kubinger (2009a), die Testperson abzulenken. Es sind jene Antwortalternativen eines Items, die keine richtige Antwort auf die gestellte Frage darstellen. Im Falle des Untertests 9 sind die richtigen beiden Begriffe von den fünf möglichen herauszufinden und anzukreuzen (Kubinger, 2009a).

Untertest 9 entspricht demnach dem Multiple Choice Format 2 aus 5. Ein Item gilt nur dann als „richtig gelöst“, wenn die beiden richtigen Begriffe und keiner der drei Distraktoren markiert wurden.
Sollte eine Testperson die richtige Antwort nicht wissen und raten, so besteht in diesem Fall nur eine Wahrscheinlichkeit von $\binom{5}{2}$= 1/10=0,10, oder anders ausgedrückt

10% das Item auch ohne entsprechendes Können richtig zu lösen (Kubinger, 2009a).

4.4) Hypothesen

Anhand der Fragestellung dieser Diplomarbeit lässt sich folgendes Hypothesenpaar ableiten:
H0: Untertest 9 „Funktionen Abstrahieren“ misst in beiden Vorgabemodi dieselbe Personenfähigkeit; die Items in AID 3 und AID-Gruppe erfassen dieselbe Dimension.

H1: Untertest 9 „Funktionen Abstrahieren“ misst in beiden Vorgabemodi verschiedene Personenfähigkeiten; die Items in AID 3 und AID-Gruppe erfassen verschiedene Dimensionen.

4.4) Martin-Löf-Test

Der Martin-Löf-Test ist ein Signifikanztest für Rasch-Modelle, der „[…] nicht die geschätzten Personenparameter zum Gegenstand hat, sondern deren erschöpfende Statistiken, die Summenscores für beide Testhälften“ (Rost, 2004, S. 351).
Diese Diplomarbeit behandelt die Dimensionalität der beiden Testbedingungen Einzel- und Gruppenvorgabe, es soll also „die Gleichwertigkeit der Parameter der Item- und Skalenwerte sowie der Reliabilitäts- und Validitätskennwerte bei verschiedenen Vorgabebedingungen“ (Kubinger, 2003, S.33) getestet werden.

Zur Überprüfung der Dimensionalität wird die Itemhomogenität der beiden Vorgabemodi wird mittels Martin-Löf-Test berechnet.

Teststatistik des Martin-Löf-Tests lautet:

$$\frac{\prod_{t} \pi_t^{n_t} \prod_{\mathrm{x}} L(\mathrm{x}\,|\,t)}{\prod_{\mathrm{t}} \pi_{\mathrm{t}}^{n_{\mathrm{t}}} \prod_{(\mathrm{x}_1,\mathrm{x}_2)} L(\mathrm{x}_1,\mathrm{x}_2\,|\,t_1,t_2)}.$$

aus: Methods of Psychological Research Online 2000, Vol.5, No.1

t … Testscore für den gesamten Test

t_1, t_2 … Scores in den Testhälften 1 und 2

n_t … Anzahl der Testpersonen am Test t

L( ) … zeigt die Likelihoodfunktionen für Antwortmuster x

π … Parameter, theoretisches Maß für verschiedene Scores

x_1, x_2 … Variablen, partielle Antwortmuster in 2 Testteilen

df … $k_1 * k_2 - 1$

Π … Produktzeichen

Unter Annahme der Nullhypothese, dass beide Testhälften dieselbe latente Eigenschaft messen, ist die dem Martin-Löf-Test zugrundeliegende Statistik χ^2-verteilt, mit $k_1 * k_2 - 1$ Parametern (k_1 und k_2 … Anzahl der Items pro Testhälfte) (Verguts & De Boeck, 2000).

Nach Rost (2004) ist der Signifikanztest „ […]ein modifizierter Likelihoodquotiententest beziehungsweise χ^2–Test, der auf den bedingten Likelihoods beider Testteile

beruht“ (Rost, 2004, S. 352). Das bedeutet: im Gegensatz zu Andersen’s Likelihood-Ratio-Test vergleicht der Matin-Löf-Test nicht Personen sondern Items.

Ein Signifikanztest ist nach Rost (2004, S. 333) „[…] ein Verfahren, mit dem man eine statistische Hypothese prüfen kann.“ Signifikanz besagt, dass die Abweichung einer Prüfgröße (aus den Daten geschätzter Parameter) von einem kritischen Wert (anhand der Freiheitsgrade und Irrtumswahrscheinlichkeit erwarteter Wert) bedeutsam ist. Der kritische Wert wird bei einer Irrtumswahrscheinlichkeit nur mit 5%iger Wahrscheinlichkeit überschritten.

Bei einem χ^2–Test lässt sich der kritische Wert anhand der sogenannten Freiheitsgrade und der Irrtumswahrscheinlichkeit bestimmen (Rost, 2004).

Berechnet wurde der Martin-Löf-Test mit Hilfe der Statistiksoftware „R“, Version 3.0.2. Dazu wurde das Package „eRm“ (extended Rasch modeling) verwendet.

Zur Berechnung der Itemhomogenität wurden jene Testpersonen gänzlich aus weiteren Berechnungen ausgeschlossen, die es in der Bearbeitungszeit nicht schafften, die Gruppenversion fertig zu bearbeiten. Dies war notwendig, da das Programm den Martin-Löf Test mit fehlenden Daten nicht berechnen kann. 212 Datensätze wurden letztendlich in die Prüfung der Äquivalenz des UT 9 in AID 3 und AID-Gruppe miteinbezogen.

Als Teilungskriterium wurde die Zugehörigkeit der Items zum AID 3 oder AID-Gruppe verwendet.

Der Wert „LR-value“ stellt den empirischen χ^2–Wert dar. Der kritische Wert, χ^2_{krit}, wurde anhand der Freiheitsgrade, „df“, und der Irrtumswahrscheinlichkeit von α=0,05 ermittelt. Wenn der empirische Wert den kritischen Wert überschreitet, dann ist der Unterschied zwischen beiden Bedingungen signifikant.

Der Wert der Freiheitsgrade (df) ergibt sich aus der Itemanzahl der Einzeltestungen multipliziert mit der Itemanzahl der Gruppentestungen minus eins (30*30-1) (Verguts, De Boeck, 2000).

5) Ergebnisse

5.1) Deskriptivstatistiken

5.1.1) Geschlechtsunterschiede in der Lösungshäufigkeit

Innerhalb der Klassenstufen ist die Lösungshäufigkeit zwischen den Geschlechtern relativ gleichverteilt (siehe Abb. 2 und 3).

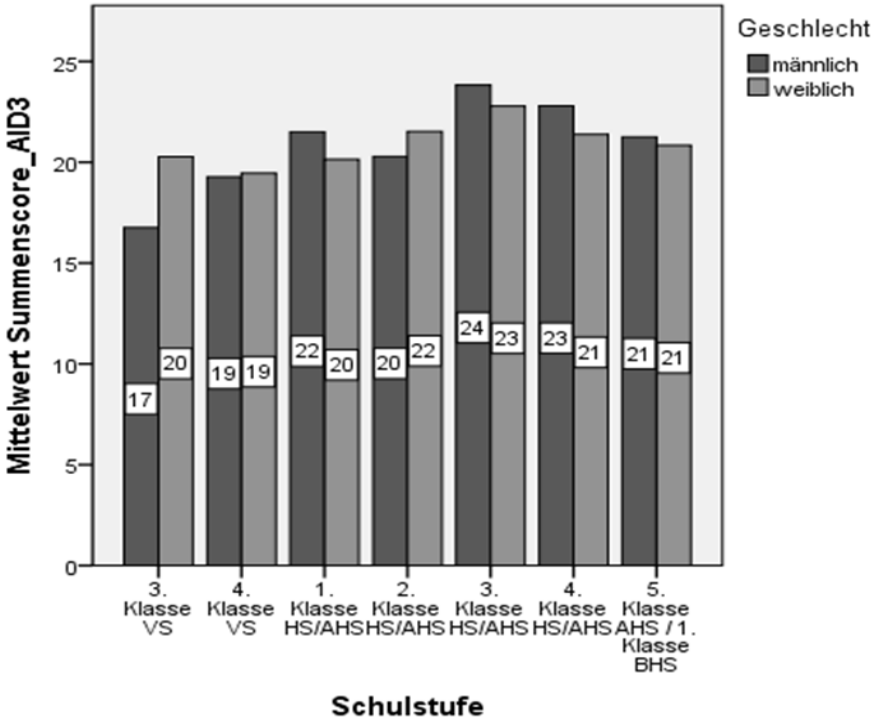

Abbildung 2 Geschlechterverteilung der Lösungshäufigkeiten in AID 3

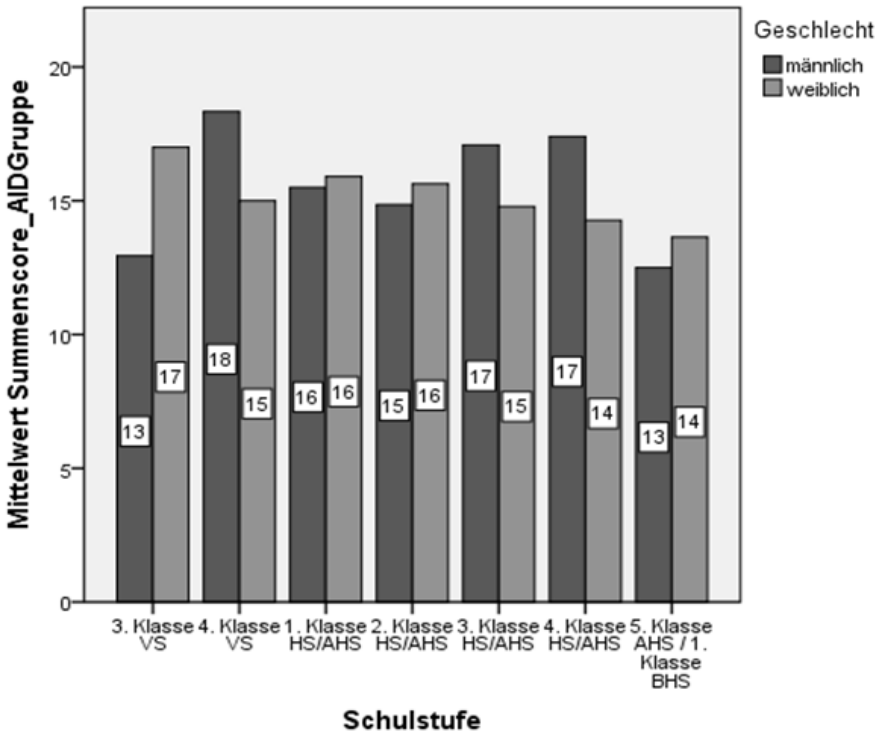

Abbildung 3 Geschlechterverteilungen der Lösungshäufigkeiten in AID-Grppe

5.1.2) Lösungshäufigkeit der Items pro Schulstufen

In den untenstehenden Tabellen 6 und 7 finden sich die durchschnittlichen Lösungshäufigkeiten für AID 3, AID-Gruppe und insgesamt, die Leistungsmaxima und –minima pro Klassenstufe, die Mittelwerte und Standardabweichungen.

AID 3: Die Schülerinnen und Schüler lösten durchschnittlich 20,72 Items in der Itemauswahl des AID 3. Eine Schülerin bzw. ein Schüler bearbeitete nur 9 Aufgaben richtig und eine Schülerin bzw. ein Schüler löste alle 30.

AID-Gruppe: In der Gruppenversion wurden durchschnittlich 15,27 Items richtig bearbeitet. 2 Schülerinnen bzw. Schüler lösten keine der gestellten Aufgaben, eine Schülerin bzw. ein Schüler löste 27.

Insgesamt wurden im Durchschnitt 36 von 60 Aufgaben gelöst. 2 Schülerinnen bzw. Schüler lösten in beiden Vorgabemodi gemeinsam nur 15 Items, 53 richtig gelöste Items wurden nur von einer Schülerin bzw. einem Schüler gelöst.

Lösungshäufigkeit der Items pro Schulstufe						
Schulstufe		N	Min.	Max.	$\bar{x}$	Stdabw.
3. Klasse VS	Summenscore_AID Gruppe	28	6	24	14,54	4,694
	Summenscore_AID3	28	9	26	18,14	4,536
	Summenscore_insgesamt	28	15	49	32,68	8,697
4. Klasse VS	Summenscore_AID Gruppe	31	0	23	16,94	5,099
	Summenscore_AID3	31	12	29	19,35	4,095
	Summenscore_insgesamt	31	18	52	36,29	8,174
1. Klasse HS/AHS	Summenscore_AID Gruppe	41	6	23	15,71	3,926
	Summenscore_AID3	41	12	28	20,80	4,057
	Summenscore_insgesamt	41	22	47	36,51	7,163
2. Klasse HS/AHS	Summenscore_AID Gruppe	26	5	21	15,42	4,319
	Summenscore_AID3	26	10	29	21,19	5,099
	Summenscore_insgesamt	26	23	50	36,62	7,430
3. Klasse HS/AHS	Summenscore_AID Gruppe	26	0	23	15,85	5,136
	Summenscore_AID3	26	17	29	23,27	3,505
	Summenscore_insgesamt	26	20	48	39,12	7,448
4. Klasse HS/AHS	Summenscore_AID Gruppe	31	5	27	14,77	5,130
	Summenscore_AID3	31	12	30	21,61	4,303
	Summenscore_insgesamt	31	21	53	36,39	8,184
5. Klasse AHS / 1. Klasse BHS	Summenscore_AID Gruppe	29	2	23	13,48	5,736
	Summenscore_AID3	29	12	28	20,90	4,655
	Summenscore_insgesamt	29	15	49	34,38	9,712

Tabelle 6 Lösungshäufigkeit der Items pro Schulstufe

Summenscores – Anzahl richtiger Lösungen

		Summenscore_AID 3	Summenscore_AID Gruppe	Summen-score_insgesamt
N	Gültig	212	212	212
	Fehlend	0	0	0
Mittelwert		**20,72**	**15,27**	**36,00**
Standardabweichung		4,504	4,890	8,198
Minimum		**9**	**0**	**15**
Maximum		**30**	**27**	**53**

Tabelle 7

5.2) Martin-Löf-Test

5.2.1) Tabellarische Darstellung

Martin-Löf-Test	
LR-value (χ^2_{emp})	453,739
χ^2_{df}	899[2]
χ^2_{krit}	969,865

Tabelle 8 Ergebnisse des Martin-Löf-Tests

In Tabelle 8 sind die Ergebnisse des Martin-Löf-Tests dargestellt. Der empirische χ^2-Wert (LR-value) ist das anhand des Datensatzes berechnete Ergebnis des Martin-Löf-Tests. Der kritische χ^2-Wert wird anhand der Anzahl der Freiheitsgrade und der Irrtumswahrscheinlichkeit α=0,05 berechnet.

[2] df = 30*30-1; Anzahl der Freiheitsgrade ist gleich der Itemanzahl in beiden Vorgabemodi minus 1 (Verguts & De Boeck, 2000)

5.2.2) Interpretation der Ergebnisse

Der kritische χ^2_{krit} –Wert überschreitet den empirischen χ^2–Wert, die Nullhypothese,

Untertest 9 „Funktionen Abstrahieren“ misst in beiden Vorgabemodi dieselbe Eigenschaftsdimension, wird beibehalten.

Der empirische und der kritische χ^2_{krit} –Wert weichen derart stark voneinander ab,

dass das erhaltene Ergebnis für die Eindimensionalität beider Vorgabemodi des Untertests 9 „Funktionen Abstrahieren“ spricht. Sie erfassen beide dasselbe Konstrukt.

6) Diskussion

Mit dieser Diplomarbeit sollte überprüft werden, ob der Untertest 9 „Funktionen Abstrahieren“ in den beiden Intelligenz-Testbatterien AID 3 (Kubinger & Holocher-Ertl, 2014, in Druck) und AID – Gruppe (Kubinger & Hagenmüller, in Vorb.) dasselbe Fähigkeitskonstrukt eindimensional erfassen.

Eine Dimensionalitätsprüfung ist in diesem Fall relevant, da es keinen Unterschied machen soll, ob einer Testperson die Individualversion oder die Gruppenversion eines psychologisch-diagnostischen Verfahrens vorgegeben wird.
Der Untertest 9 „Funktionen Abstrahieren“ soll trotz einer unterschiedlichen Itemgestaltung in der Einzel- und Gruppenversion dasselbe zugrundeliegende Fähigkeitskonstrukt, hier „Begriffsbildung durch Abstraktion“, erfassen.

Die Ergebnisse des Martin-Löf-Tests machen deutlich, dass die Items des Untertests 9 aus AID 3 (Kubinger & Holocher-Ertl, 2014, in Druck.) und AID-Gruppe (Kubinger & Hagenmüller, in Arbeit) homogen sind und dieselbe Fähigkeitsdimension „Begriffsbildung durch Abstraktion“ erfassen.

Im Nachhinein stellt sich die Frage, ob es Einflüsse auf den Verlauf der Studie gab (Auswahl der Items, Stichprobengröße, Testleitereffekte), die die Ergebnisse verzerren könnten:
Zum Beispiel könnte man annehmen, dass bei einer anderen Itemauswahl andere Ergebnisse möglich gewesen wären.
Die Itemauswahl wurde für die Untertests 4 „Soziale und Sachliche Folgerichtigkeit“ und 9 „Funktionen Abstrahieren“ vorab anhand der geschätzten Schwierigkeitsparameter getroffen. So ergaben sich folgende Itemanordnungen für den Untertest 9

„Funktionen Abstrahieren“ in AID 3 und AID-Gruppe:

- leicht (7 Items)
- leicht bis mittelschwierig (3 Items)
- mittelschwierig (10 Items)
- mittelschwierig bis schwierig (3 Items)
- schwierig (7 Items)

Durch das auf den geschätzten Schwierigkeitsparametern beruhende Itemsampling kann davon ausgegangen werden, dass es keine systematischen Einflüsse auf das Testverhalten der Schülerinnen und Schüler gab.

Sowohl bei der Instruktion zur Itemauswahl aus AID-Gruppe als auch bei der Vorgabe der Items in der Einzelsituation gibt es Gelegenheiten, in denen sich Testleitereffekte ergeben können. Da beide Testleiterinnen durch den AID-Zertifizierungskurs auf die Vorgabe beider Vorgabemodi geschult worden sind, sind keine signifikanten Effekte durch Unterschiede in der Vorgabe oder der Instruktion zu erwarten. Zusätzlich gibt es eine Verschriftlichung der Instruktionen der beiden Untertests „Soziale und Sachliche Folgerichtigkeit“ und „Funktionen Abstrahieren“ in beiden Vorgabemodi, an die sich jede der Testleiterinnen hielt, um Testleitereffekte weitestgehend zu vermeiden.

Bezüglich der Stichprobengröße kann angenommen werden, dass diese ausreichend groß war. Insgesamt wurden 219 Schüler und Schülerinnen getestet, 212 wurden in die Berechnungen miteinbezogen. Erläuterungen zur Stichprobengröße finden sich bei Kubinger, Rasch & Yanagida (2011).

Rückblickend sind in Bezug auf die Vorbereitungen der Studie, die Datenerhebung und die Datenauswertung keine systematischen Einflüsse auf die Ergebnisse zu erwarten.

Die Wichtigkeit der Dimensionalitätsprüfung in dieser Diplomarbeit besteht darin, dass beide Administrationsformen des Untertests 9 „Funktionen Abstrahieren“ in den Intelligenz-Testbatterien AID 3 (Kubinger & Holocher-Ertl, 2014, in Druck) und AID-Gruppe (Kubinger & Hagenmüller, in Arbeit) dasselbe zugrundeliegende Fähigkeitskonstrukt „Begriffsbildung durch Abstraktion“ eindimensional erfassen und als einander gleichwertig gelten. Somit kann davon ausgegangen werden, dass es in der Erfassung dieses Fähigkeitskonstruktes und in weiterer Folge auch für die Interpretation des Untertestergebnisses keinen Unterschied macht, ob eine Testperson Untertest 9 in der Einzel- oder in der Gruppenversion bearbeitet hat.

7) Zusammenfassung

Der AID 3 (Kubinger & Holocher-Ertl, 2014, in Druck) stellt die dritte Generation des Adaptiven Intelligenz Diagnostikums dar. Des Weiteren wird eine gruppentaugliche Intelligenz-Testbatterie des Adaptiven Intelligenz Diagnostikums erscheinen (Kubinger & Hagenmüller, in Arbeit).

Das Ziel dieser Diplomarbeit war, die Konstruktäquivalenz des Untertests 9 „Funktionen Abstrahieren“ in beiden Intelligenz-Testbatterien, AID 3 und AID-Gruppe zu prüfen.

Um die Forschungsfrage, ob Untertest 9 „Funktionen Abstrahieren“ in AID 3 und AID-Gruppe dieselbe zugrundeliegende Fähigkeitsdimension „Begriffsbildung durch Abstraktion“ messen, zu beantworten, wurde eine Itemauswahl von je 30 Items beider Vorgabemodi 212 Schülerinnen und Schülern zwischen 3. und 9. Schulstufe in Ober- und Niederösterreich vorgegeben.

Für die Studie wurde ein balanciertes 2-Gruppendesign zur Erhebung der Daten verwendet. Es wurde einer Hälfte der Schülerinnen und Schüler pro Klasse erst die Itemauswahl des AID-Gruppe vorgegeben, die andere Hälfte wurde zuerst in der Einzelsituation mit den Items des AID 3 getestet und anschließend mit der Gruppenversion.

Die Prüfung der Dimensionalität der Items erfolgte mittels Martin-Löf-Test. Das ist ein modifizierter Likelihood-Ratio-Test, der Items miteinander vergleicht, nicht Testpersonen. Für die Berechnungen wurde das Statistikprogramm „R“, Version 3.0.2, sowie

das Package „eRm“ verwendet.

Die Ergebnisse des Martin-Löf-Tests, sprechen dafür, dass die Items beider Untertests eindimensional messen. Das heißt, sie erfassen beide die Fähigkeitsdimension „Begriffsbildung durch Abstraktion“. Das bedeutet, dass die Ergebnisse des Untertest 9 in AID 3 oder AID-Gruppe gleich interpretiert werden können. Beide Administrationsformen sind trotz unterschiedlicher Gestaltungsweise einander gleichwertig.

8) Literatur

Böck, J. (2010). *AID 2 als Gruppentestung? Eine Machbarkeitsstudie.* Diplomarbeit, Universität Wien

Eiter, A. (2011). *AID 2 als Gruppen- oder Computertestung – Eine Machbarkeitsstudie*. Diplomarbeit, Universität Wien.

Gerrig, R.J., Zimbardo, P.G., (2008). *Psychologie* (18. aktualisierte Auflage). München: Pearson Studium

Görner, N. (2013). *Aktualisierung und Itemkonstruktion zu den Untertests U3 "Ange wandtes Rechnen" und U9 "Funktionen Abstrahieren" der Intelligenztestbatte rie AID 3*. Diplomarbeit, Universität Wien

Hagenmüller, B. (2011). *Entwicklung des Untertests "Formale Folgerichtigkeit" zur Erfassung von Reasoning in der Intelligenz-Testbatterie AID 3*. Unveröffentlichte Diplomarbeit, Universität Wien.

Hellebart, M. (2013). *Revision des Untertests "Alltagswissen" und "Soziales Erfassen und sachliches Reflektieren" für die Intelligenzbatterie AID 3*. Diplomarbeit, Universität Wien

Hofmayer, P.N. (2012). *Pilotstudie zu vier Untertests des AID als Gruppenversion*. Diplomarbeit, Universität Wien

Kubinger, K.D. (2000). Replik auf Jürgen Rost „Was ist aus dem Rasch-Modell geworden?“: Und für die Psychologische Diagnostik hat es doch revolutionäre Bedeutung. *Psychologische Rundschau, 51*(1), 33-34.

Kubinger, K.D. (2004). *On a Practitioner’s Need of Further Development of Wechsler Scales. Adaptive Intelligence Diagnosticum (AID 2).* Vol. 7, No. 2 (101-111)

Kubinger, K.D. (2009a). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens* (2., überarb. u. erweiterte Aufl.). Göttingen: Hogrefe.

Kubinger, K.D., (2009b). *Adaptives Intelligenzdiagnostikum – Version 2.2 (AID 2) samt AID 2-Türkisch*. Göttingen: Beltz Test.

Kubinger, K.D. & Jäger, R.S. (Hrsg.)(2003). *Schlüsselbegriffe der Psychologischen Diagnostik*. Weinheim: Beltz/PVU.

Kubinger, K.D. & Hagenmüller, B. (in Vorb.). *AID-Gruppe*.

Kubinger, K.D. & Holocher-Ertl, S. (Hrsg.) (2012). *Fallbuch AID*. Göttingen: Hogrefe.

Kubinger, K.D. & Holocher-Ertl, S. (in Druck). *Adaptives Intelligenzdiagnostikum 3 (AID 3)*. Göttingen: Beltz Test.

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

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

Moosbrugger, H., Kelava, A. (2008). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer Medizin Verlag

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (zweite, vollständig überarbeitete und erweiterte Auflage). Bern: Verlag Hans Huber

Schlagheck, W. & Petermann, F. (2006). Hochbegabungsdiagnostik mit dem HAWIK-III und AID 2. *Kindheit und Entwicklung*, 15, 93-99

Tewes, U. & Wildgrube, K. (Hrsg.) (1999). *Psychologie-Lexikon* (2. Aufl.). München: Oldenbourg

Verguts, T., De Boeck, P. (2000). A note on the Martin-Löf-Test for unidimensionality. In: *Methods of Psychological Research Online 2000*, 5, 77-82. Vol.5, No.1 (Download am 07.11.2013)

Wagner, M. (in Arbeit). *Konstruktion von Items für UT 4 – Soziale und Sachliche Fol gerichtigkeit und UT 11 - Soziales Erfassen und Sachliches Reflektieren für die Testbatterie AID-Gruppe*. Diplomarbeit, Universität Wien

Weber, B. (2011) *Konstruktion des sprachlichen Untertests "Antonyme finden" für die Intelligenztestbatterie AID 3*. Diplomarbeit, Universität Wien

Wechsler, D. (1956). *Die Messung der Intelligenz Erwachsener*. Bern: Huber.

Weichselbaum, S. (in Arbeit). *Dimensionalitätsprüfung des Untertests 4 Soziale und Sachliche Folgerichtigkeit der beiden Intelligenz-Testbatterien AID 3 und AID-Gruppe.* (in Arbeit)

9) Anhang

9.1) Personenparameter abhängig vom Summenscore:

Raw Score	Estimate	Std.Error
15	-1.57405314	0.3587838
18	-1.20857958	0.3403409
19	-1.09439335	0.3355735
20	-0.98321664	0.3313681
21	-0.87465791	0.3276676
22	-0.76837125	0.3244302
23	-0.66404726	0.3216154
24	-0.56140628	0.3191949
25	-0.46019235	0.3171444
26	-0.36016813	0.3154422
27	-0.26111130	0.3140759
28	-0.16281052	0.3130346
29	-0.06506222	0.3123111
30	0.03233210	0.3119018
31	0.12956886	0.3118067
32	0.22684519	0.3120302
33	0.32436117	0.3125770
34	0.42232298	0.3134592
35	0.52094558	0.3146884
36	0.62045586	0.3162827
37	0.72109624	0.3182655
38	0.82312862	0.3206613
39	0.92683902	0.3235032
40	1.03254309	0.3268291
41	1.14059261	0.3306860
42	1.25138382	0.3351273
43	1.36536748	0.3402214
44	1.48306185	0.3460453
45	1.60506940	0.3526962
46	1.73209802	0.3602954
47	1.86499048	0.3689916
48	2.00476505	0.3789775
49	2.15267509	0.3905000
50	2.31029795	0.4038930
52	2.66346551	0.4383188
53	2.86528925	0.4609727

9.2) Itemschwierigkeitsparameter

	Estimate	Std. Error	lower CI	upper CI
Gruppe_U9_it1	- 2.531	0.312	-3.142	-1.921
Gruppe_U9_it2	- 2.265	0.280	-2.814	-1.717
Gruppe_U9_it3	- 1.918	0.245	-2.398	-1.437
Gruppe_U9_it4	- 0.892	0.179	-1.243	-0.540
Gruppe_U9_it5	- 0.428	0.163	-0.747	-0.109
Gruppe_U9_it6	- 0.649	0.170	-0.981	-0.316
Gruppe_U9_it7	0.287	0.150	0.006	0.581
Gruppe_U9_it8	0.593	0.148	0.303	0.883
Gruppe_U9_it9	0.062	0.152	0.237	-0.360
Gruppe_U9_it10	- 0.892	0.179	-1.243	-0.540
Gruppe_U9_it11	- 0.224	0.158	-0.533	-0.085
Gruppe_U9_it12	- 0.481	0.164	-0.804	-0.159
Gruppe_U9_it13	0.507	0.148	0.216	0.797
Gruppe_U9_it14	0.984	0.149	0.692	1.276
Gruppe_U9_it15	- 0.008	0.153	-0.309	0.293
Gruppe_U9_it16	1.253	0.152	0.954	1.551
Gruppe_U9_it17	0.680	0.148	0.390	0.970
Gruppe_U9_it18	1.006	0.149	0.713	1.299
Gruppe_U9_it19	1.028	0.149	0.735	1.321
Gruppe_U9_it20	1.162	0.151	0.866	1.458
Gruppe_U9_it21	0.723	0.148	0.433	1.013
Gruppe_U9_it22	0.918	0.149	0.627	1.210
Gruppe_U9_it23	1.928	0.168	1.598	2.257
Gruppe_U9_it24	1.050	0.150	0.757	1.344
Gruppe_U9_it25	1.463	0.156	1.158	1.768
Gruppe_U9_it26	2.328	0.185	1.966	2.690
Gruppe_U9_it27	2.103	0.175	1.760	2.445
Gruppe_U9_it28	2.986	0.226	2.544	3.428
Gruppe_U9_it29	4.032	0.341	3.364	4.699
Gruppe_U9_it30	3.208	0.244	2.729	3.687

	Estimate	Std. Error	lower CI	upper CI
Einzel_U9_it1	-2.875	0.361	-3.582	-2.168
Einzel_U9_it2	-5.012	0.988	-6.949	-3.076
Einzel_U9_it3	-3.601	0.501	-4.583	-2.619
Einzel_U9_it4	-1.858	0.240	-2.328	-1.387
Einzel_U9_it5	-1.245	0.197	-1.631	-0.859
Einzel_U9_it6	-2.635	0.325	-3.273	-1.997
Einzel_U9_it7	-1.453	0.210	-1.864	-1.042
Einzel_U9_it8	-0.828	0.177	-1.175	-0.482
Einzel_U9_it9	-0.455	0.164	-0.775	-0.134
Einzel_U9_it10	-0.957	0.182	-1.315	-0.600
Einzel_U9_it11	-1.169	0.193	-1.547	-0.791
Einzel_U9_it12	-2.115	0.264	-2.632	-1.597
Einzel_U9_it13	0.130	0.151	0.166	0.427
Einzel_U9_it14	0.572	0.148	0.282	0.862
Einzel_U9_it15	-0.737	0.173	-1.076	-0.397
Einzel_U9_it16	-0.767	0.174	-1.108	-0.425
Einzel_U9_it17	0.354	0.149	0.061	0.646
Einzel_U9_it18	-0.150	0.156	-0.456	-0.156
Einzel_U9_it19	-1.409	0.207	-1.814	-1.004
Einzel_U9_it20	-1.453	0.210	-1.864	-1.042
Einzel_U9_it21	-0.175	0.157	-0.482	-0.132
Einzel_U9_it22	-0.350	0.161	-0.665	-0.035
Einzel_U9_it23	1.139	0.151	0.844	1.435
Einzel_U9_it24	0.463	0.148	0.172	0.754
Einzel_U9_it25	0.593	0.148	0.303	0.883
Einzel_U9_it26	0.875	0.148	0.584	1.166
Einzel_U9_it27	1.184	0.151	0.888	1.481
Einzel_U9_it28	2.362	0.186	1.997	2.727
Einzel_U9_it29	0.572	0.148	0.282	0.862
Einzel_U9_it30	2.986	0.226	2.544	3.428

Lebenslauf

Mayerhofer Doris

Buchegg 18 | 3632 Bad Traunstein

Staatsbürgerschaft: Österreich

Schulbildung

1996 – 2000	Volkschule Bad Traunstein
2000 – 2004	Hauptschule Schönbach
2004 – 2008	Realgymnasium Zwettl; Abschluss: Matura am 4. Juni 2008
WS 2008 – SS 2009	Bachelorstudium Biologie an der Universität Wien
Seit SS 2009	Diplomstudium Psychologie an der Universität Wien

Berufserfahrungen

6-Wochen-Praktikum am LKH Waidhofen an der Thaya (Station 6 Sozialpsychiatrie)

Freies Praktikum im Anton-Proksch-Institut 1230 Wien (Abteilung III)