



universität
wien

Diplomarbeit

Titel der Arbeit

Empirischer Vergleich der Messgenauigkeit zwischen branched-adaptivem und konventionellem Testen

Am Beispiel einer Population mit extremen Fähigkeiten

Verfasserin

Anna Lammert

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, 2014

Studienkennzahl: A 298

Studienrichtung: Psychologie

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

Hinweis zu Genderformulierungen:

Aufgrund der höheren Komfortabilität beim Lesen werden geschlechtsspezifische Differenzierungen, wie beispielsweise Schülerinnen und Schüler oder Schüler/innen unterlassen. Entsprechende Formulierungen sind als geschlechtsneutral zu verstehen.

Inhaltsverzeichnis

ABSTRACT	9
I EINLEITUNG	
II THEORETISCHER TEIL	
1 KONVENTIONELLES UND ADAPTIVES TESTEN.....	16
1.1 Konventionelles Testen	16
1.2 Adaptives Testen.....	17
1.2.1 Item-Response-Theorie	18
1.2.1.1 Das 1-PL Modell: „Dichotomes logistisches Testmodell“ von Rasch.....	19
1.2.2 Varianten des adaptiven Testens: tailored- und branched testing	21
1.3 Messgenauigkeit nach der Item-Response-Theorie.....	23
1.3.1 Berechnung der Messgenauigkeit bei branched-adaptiver Vorgabe.....	24
1.4 Stand der Forschung	25
2 ADAPTIVES INTELLIGENZ DIAGNOSTIKUM 3	27
III EMPIRISCHER TEIL	
3 ZIELSETZUNG UND FRAGESTELLUNG	30
4 METHODE	31
4.1 Untersuchungsdesign.....	31
4.2 Durchführung der Untersuchung.....	32
4.3 Untersuchungsmaterial	33
4.4 Statistische Methoden.....	34
5 ERGEBNISSE	35
5.1 Beschreibung der Stichprobe.....	35
5.2 Ergebnisse der Personenparameter im extremen Fähigkeitsbereich	37
5.2.1 Graphische Darstellung der empirisch festgestellten Fähigkeitsparameter	38
5.3 Ergebnisse der Standardschätzfehler im extremen Fähigkeitsbereich	41
5.3.1 Graphische Darstellung der empirisch festgestellten Standardschätzfehler	48

6	INTERPRETATION UND DISKUSSION	51
7	ZUSAMMENFASSUNG	53
8	LITERATURVERZEICHNIS.....	54
9	ABBILDUNGSVERZEICHNIS	56
10	FORMELVERZEICHNIS	56
11	TABELLENVERZEICHNIS.....	57
	ANHANG	58
	CURRICULUM VITAE.....	77

Abstract - deutsch

Für die in der psychologisch-diagnostischen Praxis relevanten Fragestellungen bezüglich Extremfähigkeiten, wie beispielsweise im Falle von Hochbegabungen, wird in der vorliegenden Arbeit die Messgenauigkeit zwischen branched-adaptiver und konventioneller Testvorgabe an einer Stichprobe mit hohen Fähigkeiten untersucht. Um diese hohen Fähigkeiten zu provozieren, erfolgte die Untersuchung an einer Stichprobe über der regulären Altersnorm, an 54 Abiturienten deutscher Gymnasien ($M = 18.3$ Jahre). Gemäß einem experimentellen Design wurden die Testpersonen randomisiert zu einer von drei Untersuchungsbedingungen zugewiesen: Die erste Bedingung beinhaltete eine branched-adaptive Testvorgabe, die zweite und dritte eine konventionelle in zwei unterschiedlichen Schwierigkeitsstufen. Das Ziel dieser Arbeit ist es, die Messgenauigkeit zwischen den Vorgabemodi anhand des AID 3 (Adaptives Intelligenz Diagnostikum, Version 3.1, Kubinger & Holocher-Ertl, 2014) zu analysieren. Die Ergebnisse der empirisch festgestellten Standardschätzfehler legen nahe, dass sich die Messgenauigkeit im Untertest 1 "Alltagswissen" ($p < .01$) und Untertest 11 „*Soziales Erfassen und Sachliches Reflektieren*“ ($p = .000$) in den Vorgabemodi signifikant unterscheidet, wobei die konventionell Bedingung mit der höheren Schwierigkeit hier die höchste Messgenauigkeit aufweist. Die konventionelle Bedingung mit der mittelschwierigeren Stufe misst über alle Untertests hinweg am unpräzisesten.

Schlagwörter: adaptives Testen, branched-adaptives Testen, konventionelles Testen, Messgenauigkeit, Standardschätzfehler

Abstract - english

With respect to the relevant issues of extreme competencies at psychological-diagnostical practice, e.g. in case of intellectual giftedness, this paper examines the difference between branched-adaptive and conventional testing with regard to the measurement efficiency on a sample with high abilities. To simulate this high competence, for the study an older sample as usual ($M = 18.3$ years) that consisted of 54 high-school graduates of German schools was used. The subjects were randomly assigned to one of three conditions: there was a branched-adaptive and two conventional conditions with two different degrees of difficulty with the aim of an empiric comparison of measurement efficiency. The testings were made with the AID 3 (Adaptives Intelligenz Diagnostikum, version 3.1, Kubinger & Holocher-Ertl, 2014). Results for high abilities reveal a significant statistical difference in measurement efficiency of subtest 1 (*Alltagswissen*) and subtest 11 (*Soziales Erfassen und Sachliches Reflektieren*) for which conventional testing with higher difficulty showed a higher measurement efficiency. However, branched-adaptive testing exceeded conventional testing with a medium difficulty in all five subtests.

Keywords: adaptive testing, branched-adaptive testing, conventional testing, measurement efficiency, standard error of estimation

I Einleitung

Viele Kinder und Jugendliche werden in ihrer schulischen Laufbahn aus den unterschiedlichsten Gründen mit psychologischen Tests konfrontiert. Die Fragestellungen hierbei sind vielfältig, jedoch ist die Vorgabe eines Leistungstests meist wesentlich, um differential-diagnostisch die kognitiven Fähigkeiten einschätzen zu können. Bei der Durchführung solcher leistungsbezogenen Untersuchungen ist ein besonderes Augenmerk auf die Effizienz dieser Verfahren zu richten (Kubinger & Wild, 1989). Mit der Entwicklung computerisierter adaptiver Verfahren in den 70er Jahren hat sich der Fokus auf eine ökonomische Testvorgabe gelegt. Ziel dieser adaptiven Vorgaben ist es, zu mehr Informationen über das zu interessierende Merkmal einer Person zu gelangen oder die Anzahl der benötigten Items zu verringern, jedoch ohne einen Informationsverlust einbüßen zu müssen. Lord (1977) und Weiss (2004) belegen eindrucksvoll die Effizienz solcher adaptiver Verfahren. Eine neue Variante des adaptiven Testens stellt, auch wenn als eine bloß suboptimale Lösung verstanden (Kubinger & Wild, 1989), die „branched-adaptive“ Testvorgabe dar. Diese verlangt keine Computertestung, sondern ist relativ unkompliziert in Papier-Bleistift-Form nach bestimmten Regeln vom Testleiter vorzugeben. Das Ziel dieser Diplomarbeit bezieht sich auf den Vergleich der Messgenauigkeit zwischen dieser branched-adaptiven und den theoretisch möglichen konventionellen Anwendungen. Des Weiteren konzentriert sich diese Untersuchung auf eine bestimmte Zielgruppe, nämlich auf Personen mit sehr hohen Fähigkeiten. Der Grund hierfür liegt darin, dass sich die Psychologische Diagnostik doch zu einem entscheidenden Teil auf Fragestellungen in Grenzbereichen bezieht, wie beispielsweise eine Abklärung von Minder- oder Hochbegabung (vgl. Kubinger & Wild, 1989). Konkret wird in der folgenden Untersuchung das AID 3 (Adaptives Intelligenz Diagnostikum, Version 3.1, Kubinger & Holocher-Ertl, 2014) in unterschiedlichen Varianten vorgegeben. Die Durchführung bezieht sich auf insgesamt 54 Abiturienten der 12. und 13. Schulstufe deutscher Gymnasien.

Die Testpersonen werden gemäß einem experimentellen Untersuchungsdesign zufällig einer von insgesamt drei Untersuchungsbedingungen zugewiesen. Unterschieden werden eine Bedingung, bei der das Verfahren standardgemäß branched-adaptiv und zwei Bedingungen, die nicht standardgemäß und zwar konventionell mit zwei unterschiedlichen Schwierigkeitsgraden vorgegeben werden. Die Ergebnisse der einzelnen Untertests werden zwischen den drei Untersuchungsbedingungen einer multivariaten Varianzanalyse bzw. dem Kruskal-Wallis-Test unterzogen, um herauszufinden, ob und in wie weit Unterschiede in der Messgenauigkeit zwischen den drei Untersuchungsbedingungen im extremen Fähigkeitsbereich existieren.

II THEORETISCHER TEIL

1 Konventionelles und adaptives Testen

Im Folgenden werden zwei unterschiedliche Testvorgaben näher erläutert und im Vergleich zu einander beschrieben, wobei auf das adaptive Testen genauer eingegangen wird. Im nächsten Schritt folgen die theoretische Herleitung der Begrifflichkeit *Messgenauigkeit* sowie ein Abriss des bisherigen Forschungsstandes bezüglich der Messgenauigkeit zwischen beiden Vorgabemodi.

1.1 Konventionelles Testen

Konventionelle Tests werden jene genannt, die mit einer festen Anzahl von Items allen Testpersonen vorgegeben werden. Alle erhalten den gleichen Test mit genau den gleichen Items in einer festgelegten Reihenfolge (vgl. Kubinger, 2003). Diese Variante wird üblicherweise in schriftlicher Form (Papier-Bleistift-Test) vorgegeben. Die Mehrheit aller psychologischen Instrumente stützt sich heutzutage auf solche konventionelle Testkonstruktionen (Weiss, 2004). Kubinger (2003) verweist in dem Zusammenhang darauf, dass beispielsweise leistungsfähigeren oder leistungsschwächeren Personen nach der konventionellen Testvorgabe manche Items regelmäßig zu leicht bzw. zu schwer fallen. Diese Items seien in diesem Fall wenig informativ, weil ihr Ausgang schon zu erwarten ist und daher der Sinn ihrer Vorgabe in Frage gestellt wird. Eine mögliche Alternative bietet das adaptive Testen.

1.2 Adaptives Testen

Adaptive Tests sind so konstruiert, dass Ihre Items adaptiv, also angepasst an das gezeigte Leistungsverhalten einer Testperson, vorgegeben werden. Die Reihenfolge dieser Items ist also nicht fix, wie bei konventioneller Vorgabe, sondern variiert entsprechend der Leistung (vgl. Kubinger, 2003). Dieser „neue“ testtheoretische Ansatz entwickelte sich in den 60er Jahren zu der sogenannten probabilistischen Testtheorie. Parallel dazu entstand vermehrt die Forderung nach höherer Testökonomie. So sollten, orientiert am diagnostischen Informationsgewinn, vergleichsweise wenige Ressourcen beansprucht werden (vgl. Kubinger, 2003; Lienert & Raatz, 2001). Dies beinhaltet u.a. die Durchführungszeit, das Testmaterial sowie die leichtere Handhabung. Die Tests sollten dementsprechend zwar kürzer werden, aber nicht an Messgenauigkeit einbüßen. Genaugenommen würden den Testpersonen nur diejenigen Items vorgelegt werden, die auch diagnostisch relevante Informationen über ein Merkmal enthalten. In diesem Fall käme es zu keinem Informationsverlust und der Test wäre gemäß der geforderten Testökonomie kürzer. Die Vorgabe der Items (vgl. AID 2, Version 2.2, Kubinger, 2009), die wenig Information liefern, seien „... weitgehend unerheblich. Somit ist es vorzuziehen, an ihrer Stelle einige andere, sehr wohl informative Aufgaben zu administrieren, die die gemessene Fähigkeit in fein abgestufte Grade zu differenzieren vermögen“ (S. 36). Diese Aussage beinhaltet den Kerngedanken der Theorie des adaptiven Testens.

Damit das adaptive Testen überhaupt gelingt, bedarf es mathematisch-stochastischer Techniken. Diese ermöglichen eine wahre, aber unbekannte Fähigkeit einer beliebigen Person über eine Modellgleichung zu schätzen, sofern, und das ist die Grundvoraussetzung, das zugrundeliegende mathematische Modell gilt (Fischer, 1974; vgl. Kubinger, 2003). Ist diese Voraussetzung erfüllt, können diejenigen Items identifiziert werden, die auch tatsächlich diagnostisch relevante Informationen enthalten. Andere weniger informative Items können nach Fischer und Molenaar (1995) durch andere bessere ersetzt oder verändert werden. Das adaptive Testen ist dadurch aber

stets an die Anwendung der *Item-Response*-Theorie (IRT), im Speziellen an die des *Rasch*-Modells, gebunden.

1.2.1 *Item-Response*-Theorie

Die *Item-Response*-Theorie (IRT), auch als Probabilistische Testtheorie bezeichnet, stellt einen neuen testtheoretischen Ansatz dar, der die Schwächen und Grenzen der Klassischen Testtheorie überwindet. Nach Kubinger (2003) basiert die Klassische Testtheorie im Wesentlichen auf einem korrelationsspezifischen, stichprobenabhängigen Zusammenhang zwischen dem Testwert und einem Merkmal einer Testperson. Dabei wird angenommen, dass ein Testwert eine latente Ausprägung dieser Eigenschaft widerspiegelt. Dieser „deterministische“ Ansatz lässt sich nicht überprüfen und wird als gegeben erachtet (Fischer, 1974; Kubinger, 2003). Die *Item-Response*-Theorie dagegen betrachtet einen Testwert als einen Indikator für eine latente Dimension und beruht auf einem wahrscheinlichkeitsbedingten Verhältnis, welches auf überprüfbaren Annahmen basiert (Fischer, 1974; vgl. Kap. 1.2.1.1.).

Die Wurzeln der *Item-Response*-Theorie liegen bei Rasch und Birnbaum. Innerhalb dieser wird vor allem zwischen diesen drei Modellen unterschieden (vgl. Kubinger, 2003):

- 1-PL Modell „Dichotomes logistisches Testmodell“ von Rasch
- 2-PL Modell „Zwei-parametrisches logistisches Modell“ von Birnbaum
- 3-PL Modell „Drei-parametrisches logistisches Modell“ von Birnbaum

Diese Modelle setzen sich aus den Parametern ξ_v und σ_i zusammen, wobei die beiden Birnbaum Modelle zusätzliche Itemparameter, in Form von Deskriminations- bzw. Rateparameter beinhalten. Alle gemeinsam basieren auf einer Berechnung von Wahrscheinlichkeiten, dass es bei einem Item zu einer Lösung kommt. Im Rahmen dieser Diplomarbeit soll aber nur auf das „Dichotome logistische Testmodell“ von Rasch näher eingegangen werden.

1.2.1.1 Das 1-PL Modell: „Dichotomes logistisches Testmodell“ von Rasch

Das sogenannte „Dichotome logistische Testmodell“ (kurz *Rasch-Modell*) von Rasch (1960) geht von einer probabilistischen Beziehung zwischen der Fähigkeit ξ_v einer Person, der Aufgabenschwierigkeit σ_i und dem gezeigten Antwortverhalten dieser Testperson v aus. Das bedeutet nicht, dass eine bestimmte Fähigkeit unmittelbar zu einer bestimmten Lösung führt. Vielmehr hängt die Lösungswahrscheinlichkeit vom Verhältnis der (wahren) Fähigkeit dieser Person und der (wahren) Schwierigkeit der Aufgabe ab. So erhöht sich nach Kubinger (2003) beispielsweise die Lösungswahrscheinlichkeit bei steigender Fähigkeit und konstant gehaltener Itemschwierigkeit und reduziert sich bei sinkender Fähigkeit. Die Formel dazu lautet:

Formel 1: Berechnung für die Lösungswahrscheinlichkeit eines Items

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

Konkret wird die Wahrscheinlichkeit erfasst, dass eine Person v mit der Fähigkeit von ξ_v und der Aufgabe i mit der Schwierigkeit von σ_i , eine besagte Aufgabe löst („+“). Bevor die Lösungswahrscheinlich aber berechnet werden kann, müssen die beobachteten Daten gewissen Modellvoraussetzungen genügen, die die Gültigkeit des *Rasch-Modells* darstellen. Sofern das Modell für die empirischen Daten gilt, sind die Daten *Rasch-Modell* konform und ermöglichen „spezifische objektive Vergleiche“. Testpersonen können dann in ihrer Leistung miteinander verglichen werden, unabhängig davon, welche Items dafür verwendet oder welche Personen sonst noch getestet wurden (Stichprobenunabhängigkeit). Sie bildet damit die methodische Grundlage für das adaptive Testen (vgl. Fischer, 1974).

Die formalen Annahmen für die Gültigkeit des *Rasch*-Modells (Fischer, 1974; Kubinger, 2003) werden im Folgenden zusammengefasst:

- **Eindimensionalität**

Die Items eines Tests müssen homogen sein. Das bedeutet, die Beantwortung der Items ist von einer bestimmten Fähigkeit abhängig und andere Kompetenzen beeinflussen diese nicht. So sollte die Lösung mathematischer Items nur aufgrund von mathematischen Fähigkeiten und nicht durch andere Fähigkeiten, wie z.B. der Sprachkompetenz, zustande kommen.

- **Lokale stochastische Unabhängigkeit der Antworten**

Die Wahrscheinlichkeit ein bestimmtes Item zu lösen, hängt nur von diesem Item ab und nicht von der Lösungswahrscheinlichkeit eines anderen Items. Beispielsweise wäre im Falle von erfolgsabhängigen Lernprozessen oder bei Aufgaben, die aufeinander aufbauen, die lokale stochastische Unabhängigkeit verletzt.

- **Erschöpfende Statistik**

Die Addition zu den Summenwerten schöpft die gesamte Information für die Parameterschätzung aus. Das bedeutet, dass die Testwerte, sowohl für die Personenscores als auch für die Itemscores sämtliche relevante Informationen enthalten. Es ist daher nicht von Bedeutung, welche Items genau gelöst (oder von welchen Personen diese gelöst) wurden. Es sind nur die Summenwerte ausschlaggebend. Dadurch werden „objektive Vergleiche“ möglich.

Unter der Voraussetzung, dass diese Annahmen gelten, können anhand von Schätzverfahren, wie beispielsweise über die Maximum-Likelihood-Funktion, die Item- und Fähigkeitsparameter einer bestimmten Testperson aus den empirischen Daten geschätzt werden. Ob ein Itempool dann tatsächlich unter diesen Voraussetzungen *Rasch*-Modell konform ist, wird anhand von Modelltests geprüft. Entspricht der beobachtete Datensatz nicht dem Modell, bedeutet das in weiterer Folge, dass das *Rasch*-Modell nicht gilt. Mindestens eine der beschriebenen Modellannahmen ist dann verletzt.

Das heißt, die Summe der gelösten Aufgaben stellt keinen aussagekräftigen Testwert dar (Kubinger, 2003). Gleiche Testwerte sind nicht gleichbedeutend mit gleichen Fähigkeiten.

1.2.2 Varianten des adaptiven Testens: tailored- und branched testing

Kubinger (2003) differenziert beim adaptiven Testen zwischen zwei unterschiedlichen Vorgabevarianten: Dem *tailored-testing* und *branched-testing*.

- **Tailored-testing**

Beim tailored-testing erfolgt die Durchführung an einem Computer, daher auch die Bezeichnung „CAT, computerized adaptive testing“. Dabei wird eine „maßgeschneiderte“ Auswahl an Items vorgenommen, die mit der Fortdauer der Durchführung eine verbesserte bzw. genauere Schätzung aufgrund des modellierten Zusammenhangs der Item- und Personenparameter ermöglicht. Die Abweichung vom geschätzten $\hat{\xi}_v$ zum wahren Fähigkeitsparameter ξ_v wird dadurch immer geringer bis eine bestimmte, tolerierbare Abweichung erreicht ist. Werden zwei gleiche Werte oder zumindest in einem tolerierbaren Abstand zueinander gemessen, stellt dies das Abbruchkriterium dar und die Testung kann beendet werden. Diese Art der Vorgabe ist stets an die Nutzung eines Computers gebunden, weil das tailored-testing eine fortlaufende Parameterschätzung erfordert, die nach jedem einzelnen Item durchgeführt wird.

- **Branched-testing**

Laut Kubinger (2003) stellt jede vom tailored-testing abweichende Variante nur eine suboptimale Lösung adaptiven Testens dar. Allerdings hat das branched-adaptive Testen den besonderen Vorteil, dass die Durchführung auch ohne die Nutzung eines Computers, nur mit Papier und Bleistift erfolgen kann.

Das branched-adaptive Testen wird z.B. anhand des Leistungstests AID (Adaptives Intelligenz Diagnostikum) von Kubinger & Wurst (1985) genauer beschrieben (vgl. Kap. 2, Adaptives Intelligenz Diagnostikum 3). Hier werden fünf Items zu einer Itemgruppe zusammengefasst, die unterschiedlichen Niveaustufen entsprechen. Die Vorgabe erfolgt gemäß diesen Itemgruppen

leistungsabhängig nach bestimmten, vorher festgelegten Regeln. Diese Regeln können als eine Verzweigung gesehen werden, anhand derer die Testpersonen vom Testleiter zur nächsten Itemgruppe administriert werden. Begonnen wird anhand von Vorinformationen bei einer bestimmten Itemgruppe. Löst eine Testperson in dieser Gruppe maximal ein Item, wird sie im nächsten Schritt zu einer leichteren Itemgruppe verwiesen. Löst sie mindestens vier Items, dann zu einer schwierigeren. Wenn die Testperson zwei bis drei Items löst, was in etwa 50% entspricht, bleibt sie in diesem Schwierigkeitsniveau. Aufgrund der extrem reduzierten Itemkombinationen und möglichen Testleistungen ist die Schätzung des Personenparameters nicht wie beim tailored-testing unmittelbar nach jedem Item notwendig, aber auch nicht nach jeder Itemgruppe. Die Personenparameter können beim branched-testing nach der Testdurchführung in einer vorher aufbereiteten Tabelle, die normalerweise im Manual festgehalten ist, abgelesen werden. Sie enthält sämtliche Parameterschätzungen (vgl. Kubinger, 2003).

1.3 Messgenauigkeit nach der *Item-Response*-Theorie

Die Messgenauigkeit beschreibt die formale Exaktheit einer Merkmals-erfassung (vgl. Kubinger, 2003). Nach Lienert und Raatz (2001) wird sie als Grad der Genauigkeit bezeichnet, mit der ein Test ein bestimmtes Merkmal misst.

Im Rahmen der *Item-Response*-Theorie wird sie durch die Bestimmung des Standardschätzfehlers (auch SEE, „standard error of estimation“) erfasst. Die Standardschätzfehler dienen also als Basis für die Berechnung der Messgenauigkeit und können für die Bestimmung des Konfidenzintervalls dienen, innerhalb dessen der wahre schätzfehlerbereinigte Testwert einer Testperson liegt (Formel 3: Berechnung des Konfidenzintervalls). Gemäß der *Item-Response*-Theorie wird die Messgenauigkeit als kein einziger globaler Wert angesehen, sondern hängt von der Höhe der Testleistung ab (Fischer, 1974; Kubinger, 2003). Um jeden Testwert resultiert ein unterschiedlich breites Intervall des wahren Personenparameters. Der Grund hierfür liegt in der Tatsache, dass der Standardschätzfehler von den einzelnen Items abhängt und diese nach einem *Rasch*-Modell konstruierten Test unterschiedliche Parameter aufweisen können. Der Standardschätzfehler wird aus dem Kehrwert der Informationsfunktion $I(\nu)$ der aufsummierten Items bestimmt. Aus der folgenden Formel können die Standardschätzfehler über alle vorgegebenen Items abgeleitet und somit die Messgenauigkeit bestimmt werden:

Formel 2: Berechnung der Standardschätzfehler (SEE)

$$S(\xi_{\nu}) = \sqrt{\frac{1}{I(\nu)}} = \left[\sum_{j=f_1(\nu)}^{f_{k_1}(\nu)} \frac{e^{\xi_{\nu} - \sigma_j}}{1 + e^{\xi_{\nu} - \sigma_j}} \frac{1}{1 + e^{\xi_{\nu} - \sigma_j}} \right]^{-\frac{1}{2}}$$

Um die Schätzfehler gering zu halten und somit die Messgenauigkeit zu erhöhen, ist es bedeutsam, die am meisten informativen Items auszuwählen. Denn der Schätzfehler wird bei der Wahl der ideal informativen Items in

Bezug auf die Fähigkeit ξ_v minimal (vgl. Fischer, 1974; Kubinger, 2003). Aus der Formel 1 (Berechnung für die Lösungswahrscheinlichkeit eines Items, S. 19) wird deutlich, dass ihr Maximum bei 0,50 liegt, was einer Lösungswahrscheinlichkeit von $p = 0,5$ entspricht. Das ist immer dann der Fall, wenn die Aufgabe (σ_i) so schwierig wie die Person (ξ_v) fähig ist ($\xi_v = \sigma_i$). In diesem Fall ergibt die Exponentialfunktion von der Differenz ($\xi_v = \sigma_i$) eine Lösungswahrscheinlichkeit von 0,5. Leitet man die Informationsfunktion $I(v)$ ab und setzt diese in die oben genannte Formel 2 ein, wird der Standardschätzfehler minimal.

Durch Berücksichtigung der gewünschten Irrtumswahrscheinlichkeit kann ein Konfidenzintervall um den geschätzten Personenparameter berechnet werden. Bei höherer Messgenauigkeit und somit kleiner Standardschätzfehler ergibt sich ein schmäleres Konfidenzintervall.

Formel 3: Berechnung des Konfidenzintervalls

$$\hat{\xi}_v^{1,2} = \hat{\xi}_v \pm z_\alpha \cdot S(\xi_v)$$

1.3.1 Berechnung der Messgenauigkeit bei branched-adaptiver Vorgabe

Bei branched-adaptiver Vorgabe muss der Standardschätzfehler nach der Formel 2 (S. 23) für jeden Personenparameter ξ_v und für jede Itemkombination, die beim gegebenen Verzweigungsschema möglich ist, berechnet werden. Diese sind dann mit der Wahrscheinlichkeit zu gewichten, dass es beim bestimmten ξ_v gerade zu dieser Itemkombination kommt (Kubinger & Wild, 1989) und letztendlich noch aufsummiert werden.

Sind Tests nach dem *Rasch*-Modell konstruiert, sollten die jeweiligen Standardschätzfehler für jeden einzelnen Testwert im Manual tabelliert sein.

1.4 Stand der Forschung: Messgenauigkeit zwischen adaptiven und konventionellen Testvorgaben

Laut Wild (1986) zeigen praktisch alle Studien, die sich auf den Vergleich der Messgenauigkeit zwischen adaptivem und konventionellem Testen konzentrieren, eine deutliche Überlegenheit der adaptiven Tests. Die Autorin stellt fest, dass adaptive Verfahren besonders in den extremen Bereichen der Fähigkeitsverteilung genauer messen als konventionelle. Auch Lord (1977) bestätigt die höhere Messgenauigkeit beim adaptiven Testen und betont, dass diese gerade in extremen Bereichen angebracht seien. Frey und Ehmke (2007) beschreiben im Rahmen von Untersuchungen über die Erfüllung deutscher Bildungsstandards, dass bei adaptiver Vorgabe bei gleicher Messgenauigkeit meistens 40-60 Prozent weniger Items benötigt werden. Diese Analysen beziehen sich zwar auf die Variante des tailored-testings, allerdings haben Kubinger und Wild (1989) belegt, dass der Unterschied in der Messgenauigkeit zwischen tailored- und branched-testing nicht so groß ist, wie früher vermutet wurde. Die Untersuchungen haben gezeigt, dass die Standardschätzfehler bei branched-adaptiver Vorgabe der tailored-adaptiven relativ nahe kommen, ausgenommen im engumliegenden mittleren Fähigkeitsbereich. Wohingegen der gesamt erfasste Bereich zwischen -10 und +10 lag. Wie aber bereits erwähnt, ist dieser mittlere Bereich für die diagnostische Praxis und daher auch für diese Untersuchung weniger relevant. Diese Analysen beziehen sich auf den branched-adaptiven Leistungstest AID (Kubinger & Wurst, 1985) und hatten das Ziel passende Verzweigungsschritte und Startgruppen zu identifizieren, aber auch die Messgenauigkeit zwischen den Vorgabebedingungen des branched-adaptiven und konventionellen Testens zu untersuchen. In Abhängigkeit der Fähigkeitsparameter wurden die Standardschätzfehler verglichen. Die Ergebnisse haben gezeigt, dass ca. 15 Items ausreichen, um den Fähigkeitsparameter hinreichend zu schätzen (vgl. Wild, 1986). Desweiteren konnte beim Vergleich zwischen konventionellem und branched-adaptivem Testen belegt werden, dass bei gleicher Testlänge von 15 Items die konventionelle Vorgabe der branched-adaptiven Variante klar unterlegen ist.

Allerdings zeigt sich auch, dass der Unterschied im Fähigkeitsbereich von 1.5 bis 3.0 und ab -1.5 in der Messgenauigkeit am größten ist. Die branched-adaptive Vorgabe schneidet besser ab. Der Unterschied wird jedoch geringer je mehr es in den Extrembereich, also zu den sehr hohen Fähigkeitsparametern geht (im Bereich ab +4). In sehr niedrigem Bereich schneidet das branched-adaptive Testen besser ab (im Bereich ab -2). Hier ergibt die Differenz der Messgenauigkeit ca. 0.2 an Standardschätzfehlern. Ebenfalls ist zu erkennen, dass die Messgenauigkeit in der branched-adaptiven Vorgabe im Bereich von -5 bis +3 relativ stabil bleibt. Ab einem Fähigkeitsparameter von +3 büßt die adaptive Vorgabe einen Verlust an Messgenauigkeit ein.

Eine konventionelle Vorgabe aller 53 Items wies im Vergleich zu der branched-adaptiven mit nur 15 Items eine bessere Messgenauigkeit auf. Es zeigt sich zwar ein Anstieg der Schätzfehler in dem Bereich ab dem Fähigkeitsparameter von +3, allerdings bleiben die Schätzfehler kleiner als bei der branched-adaptiven Vorgabe. Kubinger und Wild (1989) betonen in diesem Zusammenhang, dass die Entscheidung über den Bedarf für einen um mehr als dreifach längeren Test, gemessen an der ökonomischen Relevanz, subjektiv dem Testleiter obliegt.

Die meisten Forschungsergebnisse in Bezug auf das adaptive Testen beziehen sich allerdings auf das tailored-testing (Bloxom, 1989; Frey & Ehmke, 2007; Lord, 1977; Wild, 1986; Weiss, 2004). Bezüglich des branched-testings existieren nur die Untersuchungen von Kubinger und Wild (1989). Diese beziehen sich, wie oben beschrieben, vordergründig auf die Identifizierung der passenden Startgruppen und verteilen sich auf einem von niedrigen bis hohen Fähigkeitsbereich. Für den interessierenden extremen Bereich können tendenzielle Werte abgelesen werden, nur fehlt es hier an Forschung, die sich explizit diesem, für die Praxis relevanten Bereich widmet.

2 Adaptives Intelligenz Diagnostikum 3

Das AID 3 (Adaptives Intelligenz Diagnostikum, Version 3.1, Kubinger & Holocher-Ertl, 2014) ist eine Testbatterie, die mit 12 Untertests der Erfassung komplexer und basaler Kognitionen bei Kinder und Jugendlichen im Alter von 6;00 bis 15;11 Jahren dient. Als ein Individualverfahren vorgesehen, welches sich auf förderungsorientierte Diagnostik spezialisiert, hat das AID sich speziell im schulischen Kontext, aber auch in der Berufs- und Bildungsberatung bewährt (Kubinger & Wurst, 1985).

Inhaltlich basiert die Testbatterie auf dem Testkonzept von David Wechsler und der Investmenttheorie von Raymond B. Cattell. Kubinger definiert Intelligenz als "...die Gesamtheit aller kognitiven Voraussetzungen, die notwendig sind, um Wissen zu erwerben und Handlungskompetenzen zu entwickeln" (2009, S. 23).

Das AID ist im Jahr 1985 erstmals veröffentlicht worden, gefolgt von überarbeiteten Auflagen und neuen Versionen, auch in türkischer Sprache (Adaptives Intelligenz Diagnostikum 2, Version 2.2, Kubinger, 2009). Aktuell befindet sich die dritte Version AID 3 (Adaptives Intelligenz Diagnostikum, Version 3.1, Kubinger & Holocher-Ertl, 2014) im baldigen Einsatz.

Erfasst werden die verbal-akustischen Fähigkeiten, wie in den Untertests „Alltagswissen“, „Angewandtes Rechnen“, „Unmittelbares Reproduzieren- numerisch-numerisch“, „Synonyme Finden“, „Funktionen Abstrahieren“, „Soziales Erfassen und Sachliches Reflektieren“ und die manuell-visuellen Fähigkeiten, wie in den Untertests „Realitätssicherheit“, „Soziale und Sachliche Folgerichtigkeit“, „Kodieren und Assoziieren“, „Antizipieren und Kombinieren - figural“ sowie „Analysieren und Synthetisieren - abstrakt“. Bei der Version 3.1 kommt noch ein zusätzlicher Untertest „Formale Folgerichtigkeit“ als neuer Untertest hinzu.

Das Manual des AID 3 (Adaptives Intelligenz Diagnostikum, Version 3.1, Kubinger & Holocher-Ertl, 2014) enthält, in Abhängigkeit von der Altersstufe und der gezeigten Leistung, aufbereitete Tabellen mit sämtlichen Personenparametern. Zusätzlich sind die Standardschätzfehler für jede

Konstellation der Itemgruppen für die entsprechenden Testwerte tabelliert und können ebenfalls abgelesen werden.

III EMPIRISCHER TEIL

3 Zielsetzung und Fragestellung

Ziel dieser Untersuchung ist es, die Messgenauigkeit zwischen der branched-adaptiven und konventionellen Testvorgabe im hohen Fähigkeitsbereich zu analysieren.

Die zugrunde liegende Forschungsfrage lautet: Bestehen Unterschiede zwischen branched-adaptiver und konventioneller Testvorgabe in der Messgenauigkeit im extremen Fähigkeitsbereich?

Bezugnehmend auf die bisherigen Forschungsergebnisse sollen in der vorliegenden Untersuchung folgende zwei Hypothesen überprüft werden.

Hypothese 1

H_1^0 : Die durchschnittlichen Fähigkeitsparameter unterscheiden sich nicht zwischen den drei Untersuchungsbedingungen.

H_1^1 : Die durchschnittlichen Fähigkeitsparameter unterscheiden sich zwischen den drei Untersuchungsbedingungen.

Hypothese 2

H_2^0 : Die durchschnittlichen Standardschätzfehler unterscheiden sich nicht zwischen den drei Untersuchungsbedingungen.

H_2^1 : Die durchschnittlichen Standardschätzfehler unterscheiden sich zwischen den drei Untersuchungsbedingungen.

4 Methode

Im Folgenden soll das Untersuchungsdesign, die genaue Durchführung der Untersuchung, das Erhebungsinstrument und die statistischen Methoden im Einzelnen vorgestellt werden.

4.1 Untersuchungsdesign

Die Untersuchung erfolgte gemäß einem experimentellen Design, wobei die Testpersonen randomisiert drei unterschiedlichen Untersuchungsbedingungen zugewiesen wurden. Dabei erfolgte die erste Untersuchungsbedingung nach der standardmäßigen branched-adaptiven Vorgabe, die zweite und dritte nach konventioneller Testvorgabe. Konventionell wurde in zwei unterschiedlichen Schwierigkeitsstufen getestet. Die einzelnen Bedingungen sind in der Tabelle 1 übersichtlich dargestellt.

Tabelle 1: Beschreibung der Untersuchungsbedingungen

Untersuchungsbedingung	Vorgabevariante	Itemgruppen	durchschnittliche Itemparameter
1. Bedingung	adaptiv	6 - variierend	variierend
2. Bedingung	konventionell	6 - 7- 8	3.46
3. Bedingung	konventionell	6- 13 - 18	2.14

In der konventionellen Untersuchungsbedingungen unterschieden sich die Items im Grad der Schwierigkeit (Schwierigkeits- oder Itemparameter) und wurden nach einer festen Reihenfolge jeweils der Itemgruppen 6-7-8 bzw. 6-13-18 vorgegeben. Um gerade die sehr hohen Fähigkeiten der Testpersonen zu erfassen, wurde eine Stichprobe gewählt, die die normale Altersnorm überschreitet und vermutlich hohe Fähigkeiten besitzt. So wurden für diese Untersuchung angehende Abiturienten der 12. und 13. Schulklassen deutscher Gymnasien gewählt.

Die Durchführung der Testungen erfolgte von der Verfasserin, mit Unterstützung von drei Studenten, die dem Arbeitsbereich für Psychologische Diagnostik der Universität Wien angehören und im Zuge einer Zertifizierung für das AID Testpersonen testen müssen. Es standen somit insgesamt vier Testleiter zur Verfügung.

Bei der telefonischen Kontaktaufnahme wurden in erster Linie direkt dem Schulleiter das Vorhaben und der Zweck der Untersuchung geschildert. Zwei Schulen zeigten auf Anhieb Interesse, sodass per E-Mail nähere Informationen (siehe Anhang) versendet wurden. Nach der mündlichen Zusage der Schulleitung, erfolgte der weitere Kontakt jeweils mit den Oberstufenkoordinatoren, mit denen die genaue Durchführung und Terminplanung folgte.

Da bereits beide Schulleitungen mündlich zusagten, wurde aus zeitlichen und ökonomischen Gründen auf einen Antrag beim Kultusministerium verzichtet. Im Zuge der Testungen, die an minderjährigen Schülern erfolgte, wurden im Vorfeld Teilnahmebestätigungen an die Schulen versendet. Diese sollten von den Erziehungsberechtigten unterschrieben und vor der jeweiligen Testung von den Testleitern entgegengenommen werden.

4.2 Durchführung der Untersuchung

Die Testungen erfolgten im Zeitraum November 2012 bis Februar 2013 in den deutschen Bundesländern Bayern und Schleswig-Holstein. Am 06. November 2012 wurden diese in Passau (Bayern) am Gisela-Gymnasium und vom 18. bis 21. Dezember 2012 am Leibniz-Gymnasium in Bad Schwartau (Schleswig-Holstein) durchgeführt. Die Rückmeldungen für beide Schulen erfolgten am 21. Jänner in Passau und am 06. Februar in Bad Schwartau. Die genauen Termine befinden sich im Anhang (S. 61).

Ein kurzes standardisiertes Einleitungsgespräch diente der Gewinnung von Informationen bezüglich den Zweck und den Ablauf der Untersuchung. In diesem wurden demographische Daten, wie das Alter und das Geschlecht, erhoben und das mündliche Einverständnis für die Teilnahme eingeholt. Im

Falle von Minderjährigkeit wurde eine unterschriebene Teilnahmeerklärung seitens der Eltern vor der eigentlichen Untersuchung entgegen genommen (Anhang, S. 60).

Die Untersuchung ist ohne besondere Zwischenfälle oder problematische Situationen verlaufen, allerdings gab es einige Ausfälle zu verzeichnen. Eine angemeldete Testperson beim Gisela-Gymnasium ist aufgrund einer Verspätung nicht zu ihrem Termin erschienen, sodass der Test aus Zeitgründen mit ihr nicht durchgeführt werden konnte. Am Leibniz-Gymnasium sind zwei Schüler aufgrund einer Erkrankung bzw. ebenfalls einer Verspätung der Untersuchung ferngeblieben. Sowohl am Gisela-Gymnasium, als auch am Leibniz-Gymnasium erfolgten die Testungen in eigenen Räumen, sodass eine ruhige und ablenkungsfreie Durchführung gewährleistet werden konnte.

Die Ergebnisse der Testuntersuchungen wurden auf Wunsch mit dem jeweiligen Schüler besprochen. Hier wurde den Testpersonen mitgeteilt, welche relativen Schwächen bzw. Stärken im Leistungsprofil zu erkennen sind (Anhang, S. 64). Um Fehlinterpretationen zu vermeiden, erfolgte die Rückmeldung stets persönlich.

4.3 Untersuchungsmaterial

Die Untersuchung erfolgte mit dem AID 3 (Adaptives Intelligenz Diagnostikum, Version 3.1, Kubinger & Holocher-Ertl, 2014) und beinhaltete fünf Untertests:

- Untertest 1: „*Alltagswissen*“
- Untertest 3: „*Angewandtes Rechnen*“
- Untertest 6: „*Synonyme Finden*“
- Untertest 9: „*Funktionen Abstrahieren*“
- Untertest 11: „*Soziales Erfassen und Sachliches Reflektieren*“

4.4 Statistische Methoden

Als unabhängige Variable werden die in Kapitel 4.1 (Tabelle 1) aufgeführten drei Untersuchungsbedingungen nach der Art der Testvorgabe gebildet: adaptiv oder konventionell (in zwei unterschiedlichen Schwierigkeitsstufen). Die abhängigen Variablen stellen die Ergebnisse der Testpersonen in den einzelnen Untertests dar und beschreiben, abhängig von der jeweiligen Hypothese, die Ergebnisse der Fähigkeitsparameter (Hypothese 1) bzw. die der Standardschätzfehler (Hypothese 2). Die erhobenen Scores pro Untertest wurden mit Hilfe des Manuals den dazugehörigen Fähigkeitsparametern zugewiesen. Ebenfalls im Manual befindet sich die tabellierte Auflistung der Standardschätzfehler, die von der Anzahl der richtigen Lösungen (Scores) abhängt.

Die Unterschiedsprüfung sowohl der Fähigkeitsparameter als auch der Standardschätzfehler wird anhand der multivariaten Varianzanalyse durchgeführt. Die Voraussetzungen dafür liegen in intervallskalierten Variablen und in der Homogenität der Kovarianz-Matrizen. Ist eine dieser Voraussetzungen nicht erfüllt, wird in weiterer Folge der Kruskal-Wallis-Test für Mehrgruppen ($k > 2$) verwendet. Um die Verteilungsunterschiede besser aufzeigen zu können, werden die Ergebnisse graphisch dargestellt.

Für alle in der vorliegenden Arbeit durchgeführten inferenzstatistischen Analysen wurde ein zweiseitiges Signifikanzniveau von $\alpha = 0.05$ gewählt.

Die Auswertung der Daten erfolgte mit dem Programm IBM SPSS Statistics Version 20.

5 Ergebnisse

In diesem Kapitel sollen die Auswertungsergebnisse dargestellt werden. Im ersten Schritt wird die Stichprobe näher vorgestellt, im zweiten erfolgt die Darstellung der Ergebnisse bezüglich der Fähigkeitsparameter, gefolgt von den Ergebnissen der Standardschätzfehler.

5.1 Beschreibung der Stichprobe

Der Tabelle 2 ist zu entnehmen, dass der adaptiven Bedingung 19 Abiturienten, der konventionell-schwierigen Bedingung 18 und der konventionell-mittelschwierigen Bedingung 17 Personen zugewiesen waren.

Tabelle 2: Verteilung über die Untersuchungsbedingungen

Vorgabebedingung	Absolute Häufigkeit	Anteil in %
branched- adaptiv	19	35,2
konventionell-schwierig	18	33,3
konventionell-mittel	17	31,5

Im Bundesland Bayern wurden 11 Abiturienten getestet und in Schleswig-Holstein 43 (siehe Tabelle 3). Die Besonderheit hier ist, dass alle 11 Schüler in Bayern aus der 12. Schulstufe kamen. Das hat den Grund, dass seit 2011 mit der Einführung des achtjährigen Gymnasiums (kurz: „G8“) der Lehrplan auf acht Schulstufen komprimiert wurde. In Schleswig-Holstein soll die Einführung dieser bis zum Jahre 2016 abgeschlossen sein. Nähere Informationen zur Vereinbarung und Gestaltung der gymnasialen Oberstufe in der Sekundarstufe II ist unter dem gleichnamigen Titel der Kultusministerkonferenz nachzulesen (2013) .

Tabelle 3: Anzahl der Testpersonen pro Bundesland

Bundesland	Absolute Häufigkeit	Anteil in %
Bayern	11	20,4
Schleswig-Holstein	43	79,6

Das hat zu Folge, dass diese kleine Teilstichprobe die jüngsten Teilnehmer beinhaltet (Tabelle 4: Altersverteilung). Insgesamt waren 15 Schüler aus der 12. Schulklasse und 39 aus der 13. Schulklasse. Dementsprechend gab es 7 Personen im Alter von 17 Jahren, 24 im Alter von 18 Jahren und 23 Schüler im Alter von 19 Jahren ($M = 18.3$; $SD = 0.7$)

Tabelle 4: Altersverteilung

Alter in Jahren	Absolute Häufigkeit	Anteil in %
17	7	13,0
18	24	44,4
19	23	42,6

Die Geschlechterverteilung basiert auf 24 weiblichen und 30 männlichen Testpersonen (Tabelle 5).

Tabelle 5: Geschlechterverteilung

Geschlecht	Absolute Häufigkeit	Anteil in %
weiblich	24	44,4
männlich	30	55,6

5.2 Ergebnisse der Personenparameter im extremen Fähigkeitsbereich

Die Fähigkeitsparameter verfügen über eine Intervallskala. Die Voraussetzungen für die Anwendung einer multivariaten Varianzanalyse sind aufgrund eines nicht signifikanten Box-*M*-Test ($F(30,8101.57) = 1.31, p = .12$) erfüllt. Die Tabelle 6 zeigt die Ergebnisse der mehrdimensionalen Varianzanalyse. Es werden keine signifikanten Unterschiede zwischen den drei Untersuchungsbedingungen in den fünf Untertests nachgewiesen.

Tabelle 6: Ergebnisse der multivariaten Varianzanalyse (Fähigkeitsparameter)

Wilks-Lambda	F-Wert	df	Signifikanz
0.72	1.70	10	0.09
			n. sign.

Die Mittelwerte und die Standardabweichungen sind Tabelle 7 für jeden Untertest zu entnehmen.

Tabelle 7: Mittelwerte und Standardabweichungen der Fähigkeitsparameter

Skala	Untersuchungsbedingungen		
	branched-adaptiv (n = 19) <i>M(SD)</i>	konv. schwierig (n = 18) <i>M(SD)</i>	konv. mittel (n = 17) <i>M(SD)</i>
Untertest 1	6.05 (1.34)	5.89 (1.53)	5.33 (1.05)
Untertest 3	4.61 (1.22)	4.67 (1.28)	3.95 (1.33)
Untertest 6	2.76 (0.94)	3.17 (0.80)	2.99 (0.90)
Untertest 9	3.63 (1.29)	4.16 (1.23)	3.07 (0.79)
Untertest 11	5.34 (1.13)	5.93 (1.43)	5.02 (0.71)

5.2.1 Graphische Darstellung der empirisch festgestellten Fähigkeitsparameter

Im Folgenden sollen hier die empirisch festgestellten Fähigkeitsparameter in den drei Bedingungen graphisch dargestellt werden (Abbildungen 1-5).

Abbildung 1: UT 1 - Verteilung der Fähigkeitsparameter

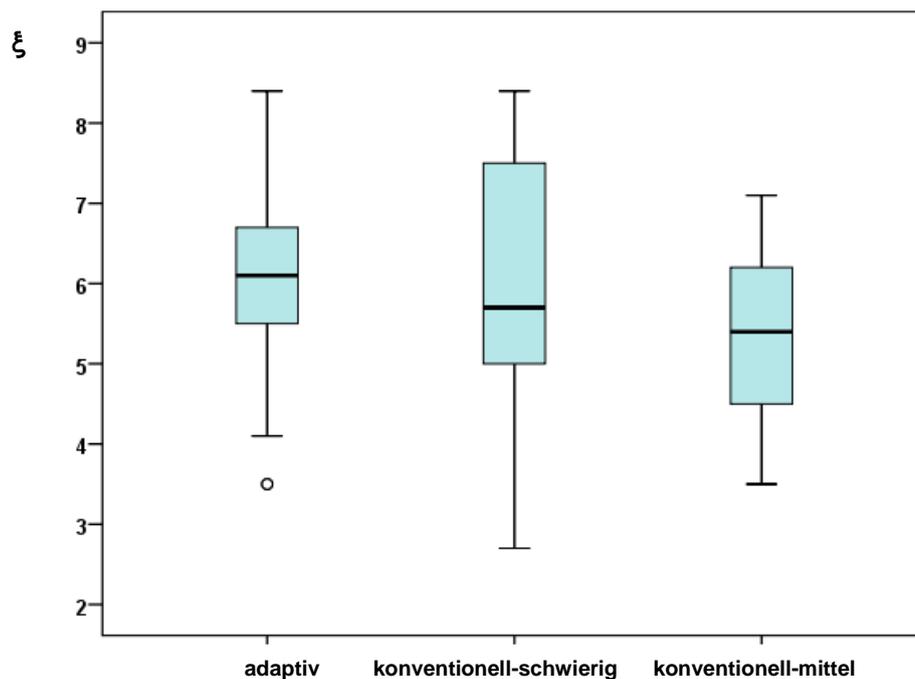


Abbildung 2: UT 3 - Verteilung der Fähigkeitsparameter

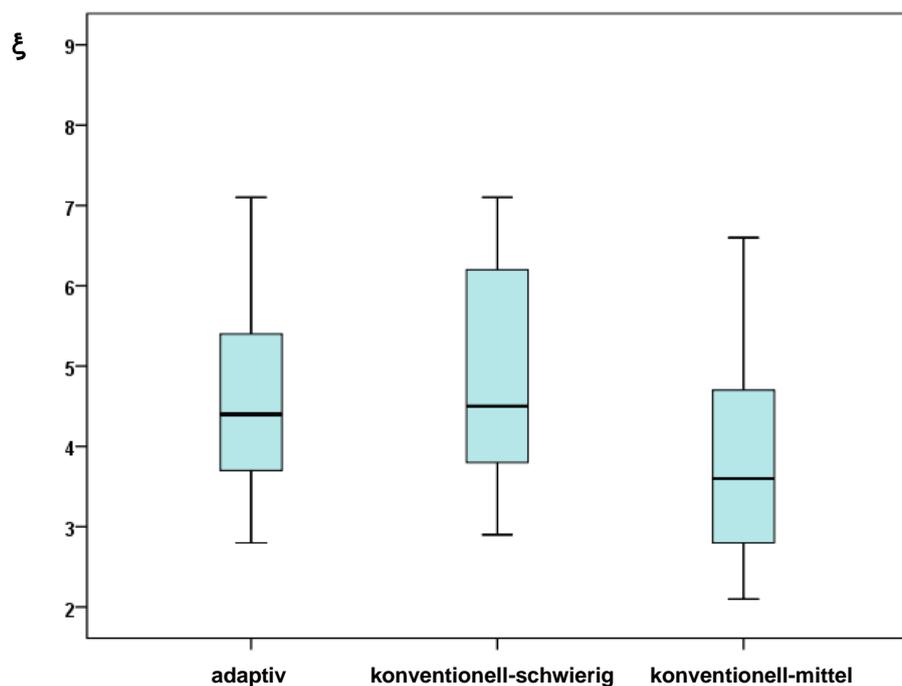


Abbildung 3: UT 6 - Verteilung der Fähigkeitsparameter

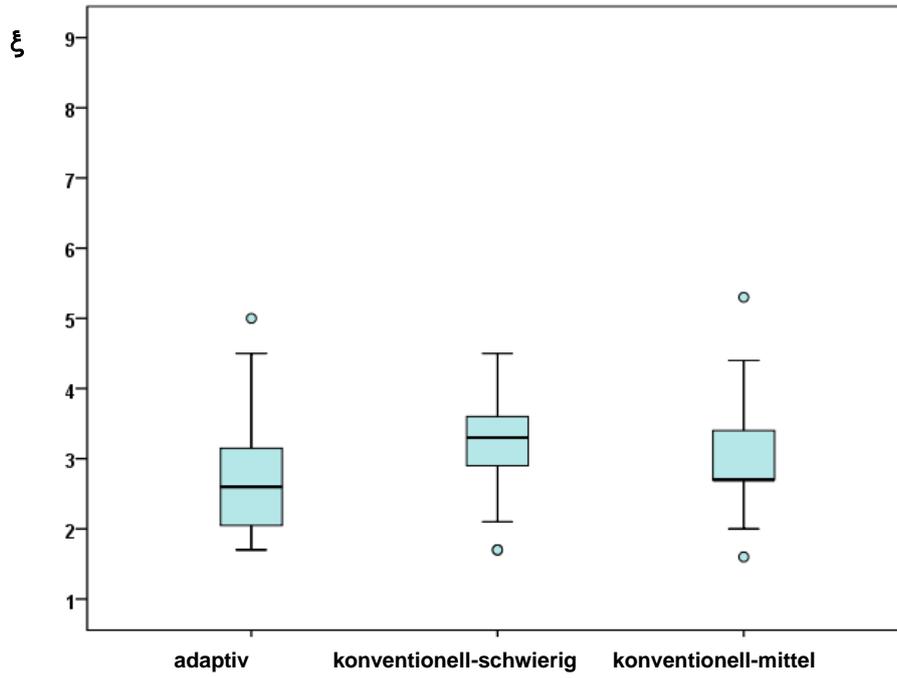


Abbildung 4: UT 9 - Verteilung der Fähigkeitsparameter

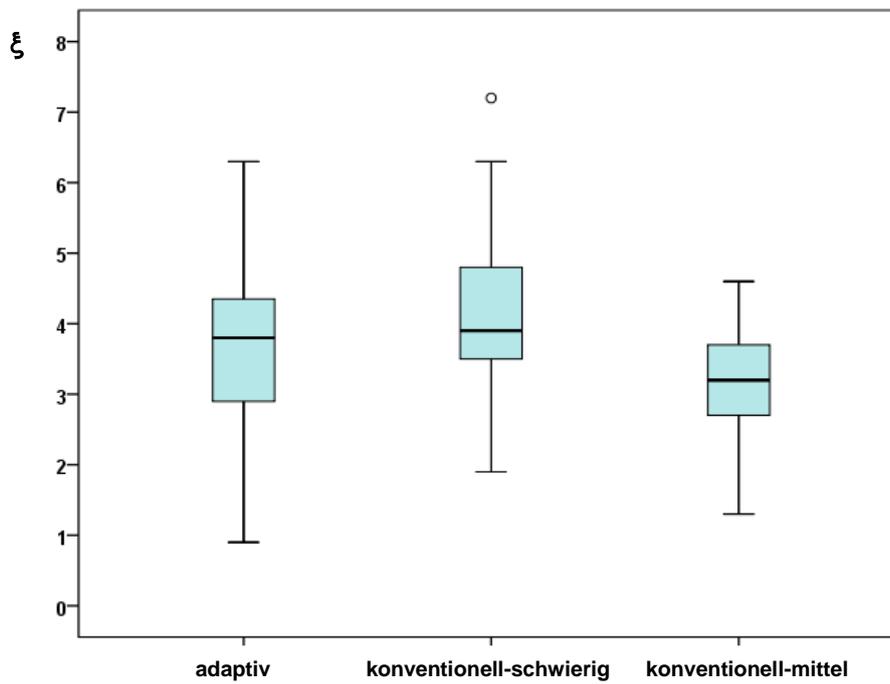
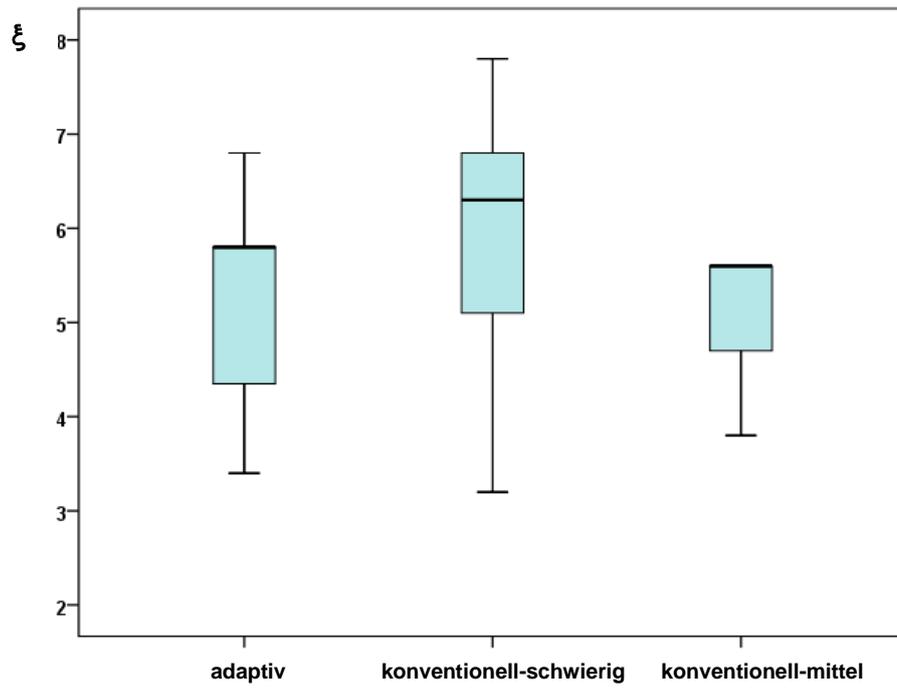


Abbildung 5: UT 11 - Verteilung der Fähigkeitsparameter



5.3 Ergebnisse der Standardschätzfehler im extremen Fähigkeitsbereich

Im Zuge der Auswertung für die Standardschätzfehler resultiert eine spezielle Problematik. Items, bei denen Testpersonen gar keinen, also 0 Punkte oder volle Punktzahl (15 Punkte) erreichten, liefern nach der Konzeptionierung der *Item-Response*-Theorie keine Informationen über die Fähigkeit einer Testperson und weisen daher im Manual keine entsprechenden Standardschätzfehler. Insgesamt war dies bei 20 Fällen (von 54x5) vorgekommen. Wie die Tabelle 8 zeigt, besonders beim Untertest 11.

Tabelle 8: Absolute Häufigkeiten für das Erreichen maximaler Testscores

	Absolute Häufigkeit pro Bedingung			Gesamt
	adaptiv	konventionell-schwierig	konventionell-mittel	
Untertest 1	1	2	2	5
Untertest 3	0	0	2	2
Untertest 6	0	0	0	0
Untertest 9	0	1	0	1
Untertest 11	0	3	9	12
Gesamt	1	6	13	20

Dort wurde 12 Mal ein maximaler Testscore von 15 Punkten erreicht. Dies macht einen prozentuellen Anteil von 52,9 % aus. Maximale Scores wurden ebenfalls im Untertest 1 (5 Mal), im Untertest 3 (2 Mal) sowie im Untertest 9 (1 Mal) gemessen. Nur im Untertest 6 („*Synonyme Abstrahieren*“) sind keine Fälle mit maximalen Testscores beobachtet worden. In der Regel würden jene Fälle, die volle Punkte erhalten haben und dadurch keine Angabe über die Standardschätzfehler aufweisen in die Auswertung nicht mit einfließen. Um dieses Problem, besonders den damit einhergehenden Informationsverlust, zu umgehen, waren die Standardschätzfehler somit zu schätzen. Die fehlenden Werte wurden anhand der unten aufgeführten Formel 4 linear

extrapoliert, wie z. B. Schreiner (2011) erläutert. Damit wird allen 20 Fällen der maximale Endwert zugeordnet. Die genauen Fälle mit den linear extrapolierten Standardschätzfehlern sind im Anhang tabelliert (S. 67).

Formel 4: Lineare Extrapolation

$$y(x') = y_1 + \frac{x' - x_1}{x_2 - x_1} (y_2 - y_1)$$

Dabei beschreibt x' den maximalen Score von 15 Punkten zum dazugehörigen gesuchten Fähigkeitsparameter $y(x')$. x_1 und x_2 bilden die beiden letzten Summenscores (13 und 14 Punkte), während y_1 und y_2 die dazugehörigen Standardschätzfehler beschreiben.

Um genau ermitteln zu können, in welcher Gruppe, welche Standardschätzfehler vorkommen, sollen auf den beiden nächsten Seiten Kreuztabellen (Tab. 9-13) für die einzelnen Untertests pro Untersuchungsbedingung dargestellt werden. In diesen Kreuztabellen sind jeweils die absoluten Häufigkeiten angegeben, die in der jeweiligen Untersuchungsbedingung des Untertests erreicht wurden.

Tabelle 9: UT 1 - Absolute Häufigkeit der SEE in den Untersuchungsbedingungen

Bedingung	Standardschätzfehler Untertest 1					Gesamt
	0,6	0,7	0,8	1,1	1,4	
adaptiv	6	3	6	3	1	19
konventionell 6-7-8	9	4	0	3	2	18
konventionell 6-13-18	5	3	3	4	2	17
Gesamt	20	10	9	10	5	54

Tabelle 10: UT 3 - Absolute Häufigkeit der SEE in den Untersuchungsbedingungen

Bedingung	Standardschätzfehler Untertest 3						Gesamt
	0,6	0,7	0,8	0,9	1,0	1,4	
adaptiv	7	5	3	3	1	0	19
konventionell 6-7-8	4	8	1	4	1	0	18
konventionell 6-13-18	7	3	5	0	0	2	17
Gesamt	18	16	9	7	2	2	54

Tabelle 11: UT 6 - Absolute Häufigkeit der SEE in den Untersuchungsbedingungen

Bedingung	Standardschätzfehler Untertest 6				Gesamt
	0,6	0,7	0,8	1,1	
adaptiv	17	2	0	0	19
konventionell 6-7-8	15	3	0	0	18
konventionell 6-13-18	14	1	1	1	17
Gesamt	46	6	1	1	54

Tabelle 12: UT 9 - Absolute Häufigkeit der SEE in den Untersuchungsbedingungen

Bedingung	Standardschätzfehler Untertest 9					Gesamt
	0,6	0,7	0,8	1,1	1,4	
adaptiv	7	9	2	1	0	19
konventionell 6-7-8	9	6	1	1	1	18
konventionell 6-13-18	7	4	5	1	0	17
Gesamt	23	19	8	3	1	54

Tabelle 13: UT 11 - Absolute Häufigkeit der SEE in den Untersuchungsbedingungen

Bedingung	Standardschätzfehler Untertest 11							Gesamt
	0,6	0,7	0,8	0,9	1,1	1,3	1,4	
adaptiv	1	5	2	7	4	0	0	19
konventionell 6-7-8	1	3	3	2	6	3	0	18
konventionell 6-13-18	0	0	3	0	5	0	9	17
Gesamt	2	8	8	9	15	3	9	54

Anhand dieser Kreuztabellen ist ersichtlich, dass gerade im Untertest 1 („Alltagswissen“), Untertest 6 („Synonyme Finden“) sowie Untertest 9 („Funktionen Abstrahieren“) eine Verteilung resultiert, die nicht der erwarteten entspricht. Werden die geringen Standardschätzfehler von 0,6 und 0,7 betrachtet, zeigt sich, dass in der konventionell-schwierigen Vorgabe mindestens ähnlich häufig diese niedrigen Standardschätzfehler erreicht wurden als in der adaptiven Bedingung, im Untertest 1 sogar häufiger. Die konventionell-schwierige Bedingung sollte einen zu hohen Anspruch simulieren, jedoch zeigt die Verteilung eine für die gewählte Stichprobe „passenden“ Leistungsanspruch. Die Ergebnisse legen nahe, dass mit der konventionell-schwierigen Bedingung die gewählte Stichprobe mit den hohen Fähigkeiten unabsichtlich ideal getestet wurde, sodass die Testpersonen in der „schwierigen“ Bedingung die Leistung zeigten, die der branched-adaptiven entspricht. Als Konsequenz werden all jene Personen, die gerade diese Itemgruppen erreicht haben, die sie in der branched-adaptiven erreicht hätten, in die branched-adaptive Bedingung umgeordnet. Die neue resultierende Verteilung wird wiederum in Kreuztabellen (Tab.14-18) dargestellt.

Tabelle 14: UT 1 - Absolute Häufigkeit der SEE in den Untersuchungsbedingungen (nach Umordnung)

Bedingung	Standardschätzfehler Untertest 1					Gesamt
	0,6	0,7	0,8	1,1	1,4	
adaptiv	8	4	6	6	3	27
konventionell 6-7-8	8	3	0	0	0	11
konventionell 6-13-18	4	3	3	4	2	16
Gesamt	20	10	9	10	5	54

Tabelle 15: UT 3 - Absolute Häufigkeit der SEE in den Untersuchungsbedingungen (nach Umordnung)

Bedingung	Standardschätzfehler Untertest 3						Gesamt
	0,6	0,7	0,8	0,9	1,0	1,4	
adaptiv	10	9	3	7	2	0	31
konventionell 6-7-8	4	4	1	0	0	0	9
konventionell 6-13-18	4	3	5	0	0	2	14
Gesamt	18	16	9	7	2	2	54

Tabelle 16: UT 6 - Absolute Häufigkeit der SEE in den Untersuchungsbedingungen (nach Umordnung)

Bedingung	Standardschätzfehler Untertest 6				Gesamt
	0,6	0,7	0,8	1,1	
adaptiv	26	2	0	0	28
konventionell 6-7-8	13	3	0	0	16
konventionell 6-13-18	7	1	1	1	10
Gesamt	46	6	1	1	54

Tabelle 17: UT 9 - Absolute Häufigkeit der SEE in den Untersuchungsbedingungen (nach Umordnung)

Bedingung	Standardschätzfehler Untertest 9					Gesamt
	0,6	0,7	0,8	1,1	1,4	
adaptiv	10	9	3	2	1	25
konventionell 6-7-8	9	6	0	0	0	15
konventionell 6-13-18	4	4	5	1	0	14
Gesamt	23	19	8	3	1	54

Tabelle 18: UT 11 - Absolute Häufigkeit der SEE in den Untersuchungsbedingungen (nach Umordnung)

Bedingung	Standardschätzfehler Untertest 11							Gesamt
	0,6	0,7	0,8	0,9	1,1	1,3	1,4	
adaptiv	1	7	4	9	10	3	0	34
konventionell 6-7-8	1	1	1	0	0	0	0	3
konventionell 6-13-18	0	0	3	0	5	0	9	17
Gesamt	2	8	8	9	15	3	9	54

Aufgrund der Umordnung und der damit einhergehenden starken Gruppengrößenunterschiede scheidet die multivariate Varianzanalyse als Verfahren aus. Es wird stattdessen der Kruskal-Wallis-Test gerechnet.

Der Kruskal-Wallis-Test (Tabelle 19) fällt für den Untertests 1 ($H(2) = 10.06$, $p < .01$) und Untertest 11 ($H(2) = 17.46$, $p = .000$) signifikant aus. Die Untertests 3, 6 und 9 zeigen keinen signifikanten Unterschied zwischen den drei Untersuchungsbedingungen.

Tabelle 19: Kruskal-Wallis für Unterschiedsprüfung der SEE zwischen den drei Untersuchungsbedingungen

Untertest	χ^2	df	Signifikanz	
Untertest 1	10.06	2	< .01	sign.
Untertest 3	2.50	2	0.29	n. sign.
Untertest 6	3.74	2	0.15	n. sign.
Untertest 9	5.70	2	0.06	n. sign.
Untertest 11	17.46	2	< .001	sign.

Tabelle 20: Mittlere Ränge für die Unterschiedsprüfung der SEE zwischen den drei Untersuchungsbedingungen

Untertest	Bedingung	N	Mittlerer Rang
Untertest 1	adaptiv	27	30,33
	konventionell schwierig	11	14,59
	konventionell mittel	16	31,59
Untertest 3	adaptiv	31	28,47
	konventionell schwierig	9	20,33
	konventionell mittel	14	29,96
Untertest 6	adaptiv	28	25,36
	konventionell schwierig	16	28,38
	konventionell mittel	10	32,10
Untertest 9	adaptiv	25	28,58
	konventionell schwierig	15	20,40
	konventionell mittel	14	33,18
Untertest 11	adaptiv	34	23,35
	konventionell schwierig	3	7,50
	konventionell mittel	17	39,32

5.3.1 Graphische Darstellung der empirisch festgestellten Standardschätzfehler

Abbildung 6: UT 1 - Verteilung der Standardschätzfehler in absoluten Häufigkeiten

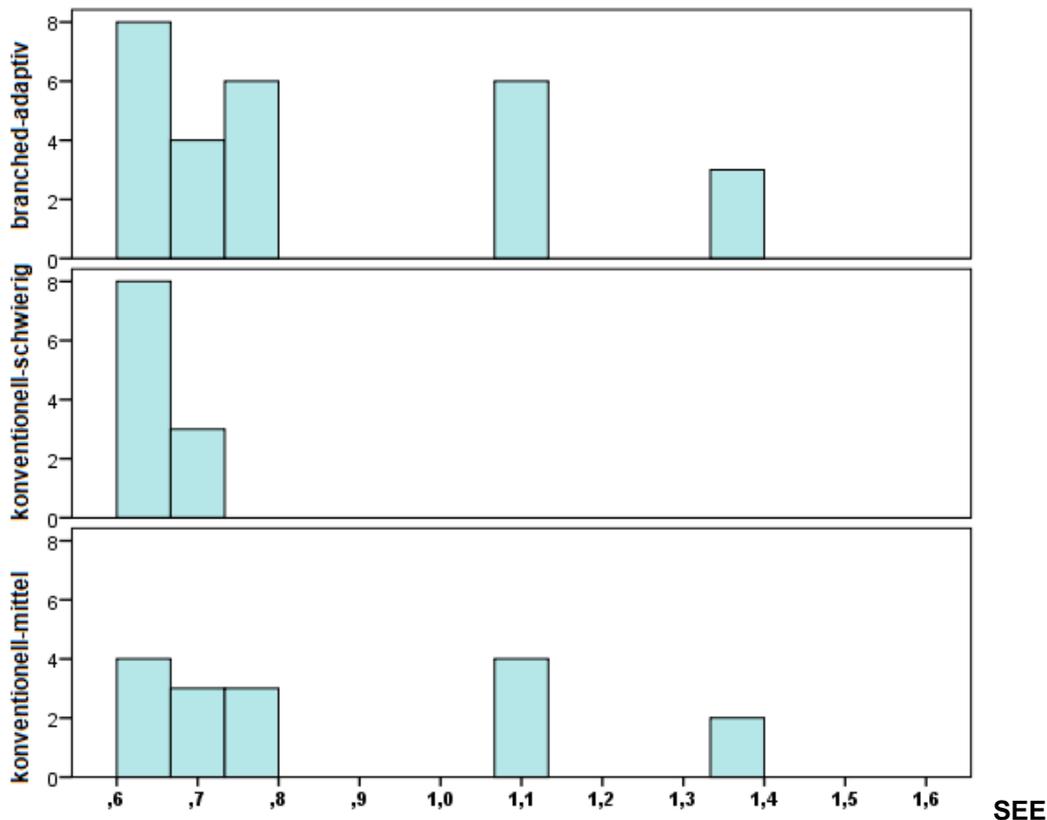


Abbildung 7: UT 3 - Verteilung der Standardschätzfehler in absoluten Häufigkeiten

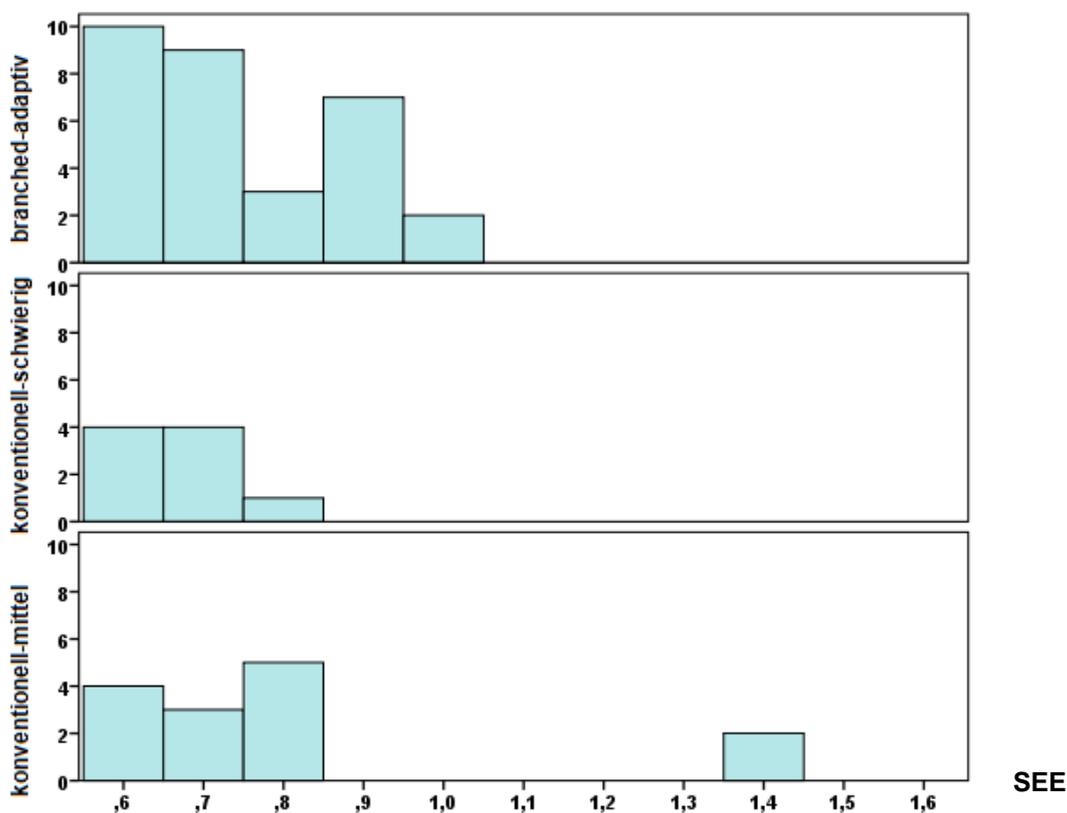
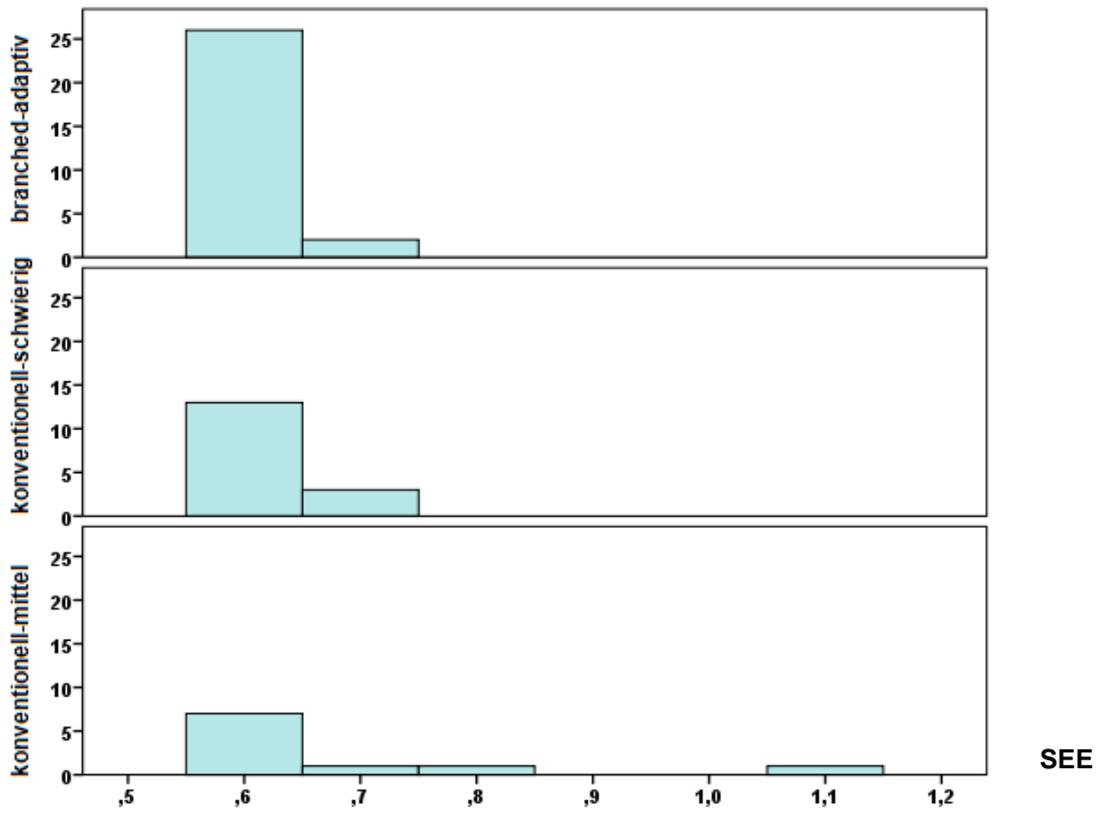
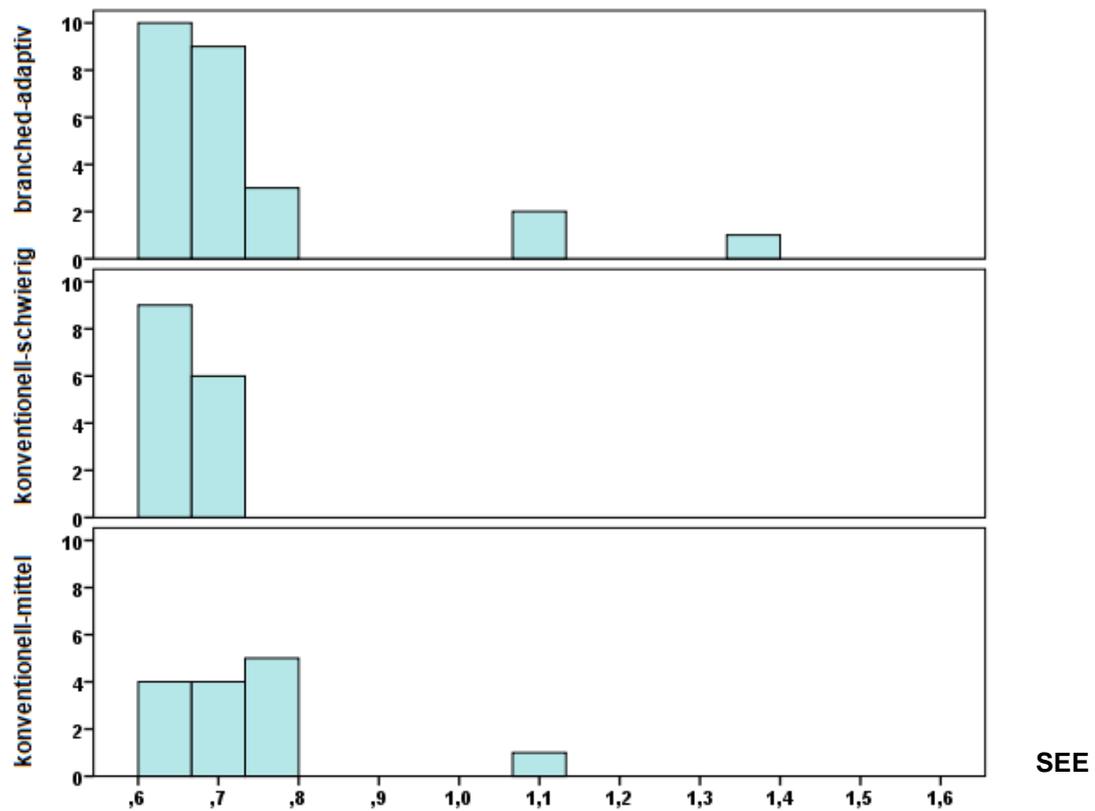


Abbildung 8: UT 6 - Verteilung der Standardschätzfehler in absoluten Häufigkeiten



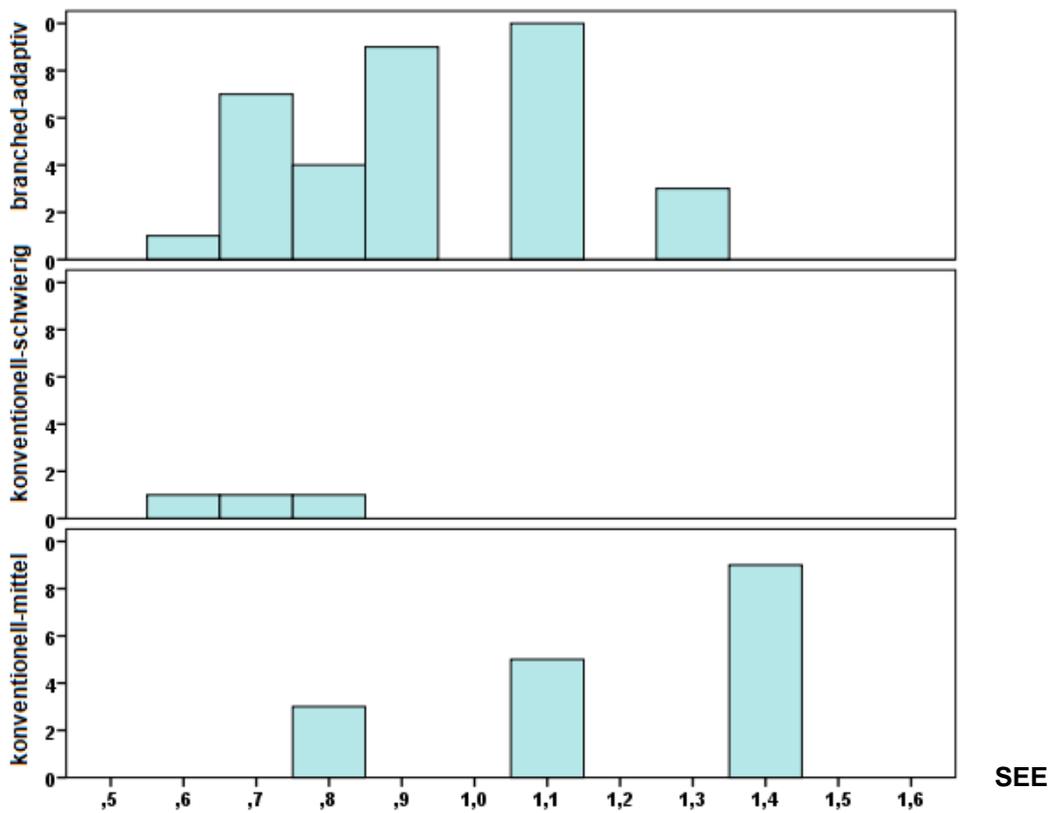
SEE

Abbildung 9: UT 9 - Verteilung der Standardschätzfehler in absoluten Häufigkeiten



SEE

Abbildung 10: UT 11 - Verteilung der Standardschätzfehler in absoluten Häufigkeiten



SEE

6 Interpretation und Diskussion

Beim Vergleich der Messgenauigkeit zwischen adaptivem und konventionellem Testen ergeben sich relevante Unterschiede in den Untertests 1 und 11 zugunsten der konventionell-schwierigen Versuchsbedingung. Daraus kann aber keine Schlussfolgerung allgemein auf das konventionelle Testen vorgenommen werden, weil die Ergebnisse, abhängig von der Schwierigkeitsstufe, unterschiedlich ausfallen. So zeigt das konventionelle Testen mit dem mittleren Schwierigkeitsgrad über alle Untertests hinweg die geringste Messgenauigkeit, während das branched-adaptive Testen nur im Untertest 6 geringfügig genauer misst als in den anderen beiden Bedingungen. Die Ergebnisse von Kubinger und Wild (1989), bei denen das konventionelle Testen der branched-adaptiven Vorgabe unterlegen war, können nur für die Vorgabe mit den mittelschwierigen Aufgaben bestätigt werden, nicht jedoch für die mit den schwierigen. Werden die Ergebnisse im Detail deskriptiv betrachtet, können gewisse Auffälligkeiten festgestellt werden. Bei den Verteilungen über die einzelnen Standardschätzfehler ist eine Tendenz dahingehend zu erkennen, dass die adaptive Bedingung über alle Untertests, mit Ausnahme des Untertests 1, häufiger die niedrigeren Standardschätzfehler aufweist (wie z.B. SEE von 0.6 und 0.7). Die dann doch messgenaueren Ergebnisse beim konventionell-schwierigeren Testen können wie folgt erklärt werden. Die oft erhaltenen hohen Punkte der Testpersonen ziehen hohe Standardschätzfehler mit sich und somit eine ungenauere Aussagekraft in diesem hohen Bereich. Die Umordnung der ideal getesteten Personen führte unmittelbar dazu, dass die Testpersonen mit den hohen Standardschätzfehlern aus der konventionell-schwierigen in die adaptive Bedingung überführt wurden und somit im Mittel die Messgenauigkeit beim adaptiven Testen reduziert haben. Dies ist z. B. dann der Fall, wenn eine Testperson beim konventionellen Testen in den Itemgruppen 6-7-8 immer mindestens vier Punkte erreicht, also ihrer Leistung entsprechend getestet wurde und somit in die adaptive Bedingung überführt wird. Dieser hohe Score bringt allerdings auch mit sich, dass der

zugewiesene Standardschätzfehler einen negativen Einfluss auf die Messgenauigkeit in der adaptiven Bedingung aufweist.

Des Weiteren kann die Vermutung angenommen werden, dass besonders der Untertest 11 für die gewählte Stichprobe, speziell unter der konventionell-mittleren Bedingung, zu einfach war. Es konnte eindeutig ein Deckeneffekt gezeigt werden. Alle Testpersonen haben mindestens 13 Punkte und 82% aller Testpersonen sogar mindestens 14 von insgesamt 15 Punkten erreicht.

Sehr wahrscheinlich war das gewählte Alter der Testpersonen zu hoch. Das AID 3 ist für den Altersbereich bis einschließlich dem 15. Lebensjahr konzipiert, während das durchschnittliche Alter der Stichprobe bei 18,3 Jahre lag und sogar 23 Testpersonen mit 19 Jahren enthielt. Es hätten sich differenziertere Ergebnisse in den Untertests ergeben, wären jüngere Testpersonen gewählt worden. Der bereits erwähnte Deckeneffekt sowie die Problematik der häufig erreichter sehr hoher Werte unterstreicht die Annahme, dass die Stichprobe zu hohe Fähigkeiten aufwies. In dem Zusammenhang ist auch anzumerken, dass es sinnvoller wäre, am Leibniz-Gymnasium ebenfalls die 12. Schulklasse, entsprechend der Stichprobe am Gisela-Gymnasium, zu untersuchen. Der Fokus auf Abiturienten ist nicht notwendig gewesen.

Für die Psychologische Diagnostik im Bereich der Hochbegabung ist es notwendig hier weitere Untersuchungen durchzuführen. Aus den oben genannten Gründen ist für Nachfolgestudien allerdings eine jüngere und somit wenig fähigere Stichprobe anzuraten. Des Weiteren ist anzumerken, dass die Wahl der drei Bedingungen für diese Fragestellung zwar ausreichend gewesen ist, allerdings für Nachfolgestudie mit einer jüngeren Stichprobe, eine zusätzliche Bedingung mit einer leichteren Form günstig wäre.

7 Zusammenfassung

Eine Variante des adaptiven Testens stellt die branched-adaptive Vorgabe dar, welche im Vergleich zu konventionellen Tests antwortabhängig, nur jene Items vorgibt, die auch tatsächlich Rückschlüsse über die messende Eigenschaft geben (Kubinger, 2003). Für die in der psychologisch-diagnostischen Praxis häufig interessierende Fragestellung bezüglich Extremfähigkeiten (Kubinger & Wild, 1989), wie beispielsweise bei Minder- oder Hochbegabung, untersucht die vorliegende Arbeit die Messgenauigkeit zwischen der branched-adaptiven und konventionellen Testvorgabe an einer Stichprobe mit hohen Fähigkeiten. Um diese hohen Fähigkeiten zu provozieren, erfolgte die Untersuchung an 54 Abiturienten deutscher Gymnasien über der regulären Altersnorm ($M = 18.3$, $SD = 0.7$). Diese Testpersonen wurden gemäß einem experimentellen Design randomisiert drei Untersuchungsbedingungen zugeteilt: Die erste Bedingung beinhaltete eine branched-adaptive Testvorgabe, die zweite und dritte eine konventionelle in zwei unterschiedlichen Schwierigkeitsstufen. Als Verfahren wurde das AID 3 (Adaptives Intelligenz Diagnostikum, Version 3.1, Kubinger & Holoher-Ertl, 2014) angewandt.

Die Ergebnisse über die Messgenauigkeit zeigen signifikante Unterschiede in dem Untertest 1 „Alltagswissen“ ($p < .01$) und Untertest 11 „Soziales Erfassen und Sachliches Reflektieren“ ($p = .000$) zugunsten der konventionell-schwierigen Versuchsbedingung. Die konventionell-mittlere Bedingung dagegen misst über alle Untertests hinweg am unpräzisesten.

8 Literaturverzeichnis

- Bloxom, B. (1989). Adaptive testing: A review of recent results. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 10(1), 1-17.
- Fischer, G. H. (1974): *Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen*. Bern: Huber.
- Fischer, G.H. (1995). Derivations of the Rasch Model. In G.H. Fischer & L.W. Molenaar (Eds.), *Rasch models* (pp. 15-38). New York: Springer.
- Frey, A. & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. *Zeitschrift für Erziehungswissenschaft*, 8, 169–184.
- Kubinger, K. D. (2003). *Schlüsselbegriffe der Psychologischen Diagnostik*. Berlin: Beltz.
- Kubinger, K. D. (2009). *Adaptives Intelligenz Diagnostikum - Version 2.2 (AID 2) samt AID 2-Türkisch*. Göttingen: Beltz.
- Kubinger, K. D. & Holoher-Ertl, S. (2014). *Adaptives Intelligenz Diagnostikum (Version 3.1)*. Göttingen: Hogrefe.
- Kubinger, K. D., Rasch, D. & Yanagida, T. (2011). *Statistik in der Psychologie: Vom Einführungskurs bis zur Dissertation*. Göttingen: Hogrefe.
- Kubinger, K.D. & Wild, B. (1989). Die Optimierung der Meßgenauigkeit beim "branched"-adaptiven Testen. In K.D. Kubinger (Hrsg.), *Moderne Testtheorie - Ein Abriß samt neuesten Beiträgen* (S. 187-218). München: Psychologie Verlags Union.
- Kubinger, K. D. & Wurst, E. (1985). *Adaptives Intelligenz Diagnostikum (AID)*. Weinheim: Beltz.
- Lienert, G. & Ratz, A. (2001): *Testanalyse und Testkonstruktion*. Weinheim: Beltz.

- Lord, F. M. (1977). Broad-Range Tailored Test of Verbal Ability. *Psychological Measurement*, 1(1), 95–100.
- Schreiner, K. (2011). *Basiswissen Verbrennungsmotor: fragen-rechnen-verstehen-bestehen*. Wiesbaden: Springer.
- Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2013). Vereinbarung zur Gestaltung der gymnasialen Oberstufe in der Sekundarstufe II [online]. URL: <http://www.kmk.org/bildung-schule/allgemeine-bildung/sekundarstufe-ii-gymnasiale-oberstufe.html> [03.06.2013].
- Weiss, D. J. (2004). Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education. *Measurement and Evaluation in Counseling and Development*, 37(2), 70–84.
- Wild, B. (1986). *Der Einsatz adaptiver Teststrategien in der Fähigkeitsmessung*. Unveröffentlichte Dissertationsschrift. Institut für Psychologie der Universität Wien.

9 Abbildungsverzeichnis

Abbildung 1: UT 1 - Verteilung der Fähigkeitsparameter	38
Abbildung 2: UT 3 - Verteilung der Fähigkeitsparameter	38
Abbildung 3: UT 6 - Verteilung der Fähigkeitsparameter	39
Abbildung 4: UT 9 - Verteilung der Fähigkeitsparameter	39
Abbildung 5: UT 11 - Verteilung der Fähigkeitsparameter	40
Abbildung 6: Verteilung der Standardschätzfehler (Untertest 1)	48
Abbildung 7: Verteilung der Standardschätzfehler (Untertest 3)	48
Abbildung 8: Verteilung der Standardschätzfehler (Untertest 6)	49
Abbildung 9: Verteilung der Standardschätzfehler (Untertest 9)	49
Abbildung 10: Verteilung der Standardschätzfehler (Untertest 11)	50

10 Formelverzeichnis

Formel 1: Berechnung für die Lösungswahrscheinlichkeit eines Items.....	19
Formel 2: Berechnung der Standardschätzfehler (SEE).....	23
Formel 3: Berechnung des Konfidenzintervalls	24
Formel 4: Lineare Extrapolation	42

11 Tabellenverzeichnis

Tabelle 1: Beschreibung der Untersuchungsbedingungen	31
Tabelle 2: Verteilung über die Untersuchungsbedingungen	35
Tabelle 3: Anzahl der Testpersonen pro Bundesland	36
Tabelle 4: Altersverteilung.....	36
Tabelle 5: Geschlechterverteilung.....	36
Tabelle 6: Ergebnisse der multivariaten Varianzanalyse (Fähigkeitsparameter).....	37
Tabelle 7: Mittelwerte und Standardabweichungen der Fähigkeitsparameter	37
Tabelle 8: Absolute Häufigkeiten für das Erreichen maximaler Testscores.....	41
Tabelle 9: UT 1 - Absolute Häufigkeit der SEE.....	42
Tabelle 10: UT 3 - Absolute Häufigkeit der SEE.....	43
Tabelle 11: UT 6 - Absolute Häufigkeit der SEE	43
Tabelle 12: UT 9 - Absolute Häufigkeit der SEE.....	43
Tabelle 13: UT 11 - Absolute Häufigkeit der SEE	44
Tabelle 14: UT 1 - Absolute Häufigkeit der SEE.....	45
Tabelle 15: UT 3 - Absolute Häufigkeit der SEE.....	45
Tabelle 16: UT 6 - Absolute Häufigkeit der SEE	45
Tabelle 17: UT 9 - Absolute Häufigkeit der SEE.....	46
Tabelle 18: UT 11 - Absolute Häufigkeit der SEE	46
Tabelle 19: Kruskal-Wallis für Unterschiedsprüfung der SEE	47
Tabelle 20: Mittlere Ränge für die Unterschiedsprüfung der SEE	47

Anhang

Kontaktaufnahme Schule



universität
wien

Fakultät für Psychologie

Institut für Angewandte Psychologie:
Gesundheit, Entwicklung und Förderung
Sekretariat Psychologische Diagnostik
Mag. Suzana Canji
Liebiggasse 5
A- 1010 Wien

T +43 (1) 4277-472 05
F +43 (1) 4277-472 06
suzana.canji@univie.ac.at

Wien, am 26.10.2012

Wissenschaftliche Untersuchung an deutschen Gymnasien

Sehr geehrte Damen und Herren,

viele Kinder und Jugendliche werden im Laufe ihrer schulischen Laufbahn aus den unterschiedlichsten Gründen mit psychologischen Tests konfrontiert. Die Fragestellungen hierbei sind vielfältig, jedoch wird meist ein Leistungstest vorgegeben, um die intellektuellen bzw. kognitiven Fähigkeiten abschätzen zu können. Diese Leistungstests müssen durch wissenschaftliche Untersuchungen auf ihre Aussagekraft geprüft werden.

Konkret wird im Rahmen der folgenden Diplomarbeit das AID 3 (Adaptive Intelligenz Diagnostikum, Version 3.1, Kubinger & Holocher-Ertl, in Vorb.) vorgegeben. Die Testergebnisse werden in anonymisierter Form quantitativen Analysen unterzogen.

Die dafür notwendige Stichprobe bezieht sich auf Schüler und Schülerinnen der 12. bzw. 13. Schulklasse auf deutschen Gymnasien.

Die Untersuchung selbst wird von einer Diplomandin der Universität Wien (Fakultät für Psychologie), Frau Anna Lammert, durchgeführt.

Die gewonnenen Daten werden im Sinne des Datenschutzes ausschließlich für wissenschaftliche Zwecke verwendet. Die Teilnahme an der Untersuchung ist selbstverständlich freiwillig.

Wir ersuchen Sie um Unterstützung dieses wissenschaftlichen Vorhabens.

Mit freundlichen Grüßen

A handwritten signature in blue ink, appearing to read 'K. Kubinger'.

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

Einverständniserklärung Eltern minderjähriger Schüler

Einverständniserklärung – Eltern

*Ich bin damit einverstanden, dass mein/e minderjähriger/r Sohn bzw. Tochter,
_____, an der psychologischen Testung teilnimmt.*

Ort, Datum

Unterschrift

Terminliste Gisela-Gymnasium (Bayern)

06. November 2012

09:30 Uhr	1. Testperson	Testleiter 1
09:30 Uhr	2. Testperson	Testleiter 2
09:30 Uhr	3. Testperson	Testleiter 3
11:00 Uhr	4. Testperson	Testleiter 1
11:00 Uhr	5. Testperson	Testleiter 2
11:00 Uhr	6. Testperson	Testleiter 3
12:45 Uhr	7. Testperson	Testleiter 1
12:45 Uhr	8. Testperson	Testleiter 2
12:45 Uhr	9. Testperson	Testleiter 3
14:15 Uhr	10. Testperson	Testleiter 2
14:15 Uhr	11. Testperson	Testleiter 3

Termine für die Rückmeldegespräche

Gisela-Gymnasium in Passau (Bayern): 21. Jänner 2013 von 08:00 bis 14:00 Uhr

Terminliste Leibniz-Gymnasium (Schleswig-Holstein)

18. Dezember 2012

07:30 Uhr	1. Testperson	Testleiter 1
09:00 Uhr	2. Testperson	Testleiter 1
10:00 Uhr	3. Testperson	Testleiter 1
11:00 Uhr	4. Testperson	Testleiter 1
12:10 Uhr	5. Testperson	Testleiter 1
13:05 Uhr	6. Testperson	Testleiter 1
14:05 Uhr	7. Testperson	Testleiter 1
15:05 Uhr	8. Testperson	Testleiter 1

19. Dezember 2012

09:00 Uhr	9. Testperson	Testleiter 1
10:00 Uhr	10. Testperson	Testleiter 1
11:00 Uhr	11. Testperson	Testleiter 1
12:10 Uhr	12. Testperson	Testleiter 1
13:05 Uhr	13. Testperson	Testleiter 1
14:05 Uhr	14. Testperson	Testleiter 1
15:05 Uhr	15. Testperson	Testleiter 1
13:05 Uhr	16. Testperson	Testleiter 1
14:05 Uhr	17. Testperson	Testleiter 1
15:05 Uhr	18. Testperson	Testleiter 1

20. Dezember 2012

07:30 Uhr	19. Testperson	Testleiter 1
09:00 Uhr	20. Testperson	Testleiter 1
10:00 Uhr	21. Testperson	Testleiter 4
10:00 Uhr	22. Testperson	Testleiter 1
11:00 Uhr	23. Testperson	Testleiter 4
11:00 Uhr	24. Testperson	Testleiter 1
12:10 Uhr	25. Testperson	Testleiter 4
12:10 Uhr	26. Testperson	Testleiter 1
13:00 Uhr	27. Testperson	Testleiter 4

13:00 Uhr	28. Testperson	Testleiter 1
14:05 Uhr	29. Testperson	Testleiter 4
14:05 Uhr	30. Testperson	Testleiter 1
15:05 Uhr	31. Testperson	Testleiter 4
15:05 Uhr	32. Testperson	Testleiter 1

20. Dezember 2012

07:50 Uhr	33. Testperson	Testleiter 1
09:00 Uhr	35. Testperson	Testleiter 1
10:00 Uhr	36. Testperson	Testleiter 1
11:00 Uhr	37. Testperson	Testleiter 1
12:10 Uhr	38. Testperson	Testleiter 1
13:05 Uhr	39. Testperson	Testleiter 1
14:05 Uhr	40. Testperson	Testleiter 1
15:05 Uhr	41. Testperson	Testleiter 1
12:30 Uhr	42. Testperson	Testleiter 1
14:00 Uhr	43. Testperson	Testleiter 1
15:00 Uhr	44. Testperson	Testleiter 1

Termine für die Rückmeldegespräche

Bad Schwartau (Schleswig-Holstein): 06. Februar 2013 von 08:00 bis 12:30 Uhr

Muster Protokollbogen (Rückmeldung der Ergebnisse)

Protokollbogen		AID 3	
1. Alltagswissen	19 27 34 39 45 50 55 62 66 73 81	Name:	_____
3. Angewandtes Rechnen		Alter:	_____
6. Synonyme finden		Geschlecht: m <input type="checkbox"/> w <input type="checkbox"/>	
9. Funktionen abstrahieren		Testleiter:	_____
11. Soziales Erfassen und Sachliches Reflektieren		Datum:	_____

Tabellen - Ausgabe SPSS

Deskriptive Statistik

Age

	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig 17	7	13,0	13,0	13,0
18	24	44,4	44,4	57,4
19	23	42,6	42,6	100,0
Gesamt	54	100,0	100,0	

Geschlecht

	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig männlich	24	44,4	44,4	44,4
weiblich	30	55,6	55,6	100,0
Gesamt	54	100,0	100,0	

Bundesland

	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig Bayern	11	20,4	20,4	20,4
Schleswig-Holstein	43	79,6	79,6	100,0
Gesamt	54	100,0	100,0	

Schulklasse

	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig 12	15	27,8	27,8	27,8
13	39	72,2	72,2	100,0
Gesamt	54	100,0	100,0	

Bedingung

	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig adaptiv	19	35,2	35,2	35,2
konventionell schwierig	18	33,3	33,3	68,5
konventionell mittel	17	31,5	31,5	100,0
Gesamt	54	100,0	100,0	

Box-M-Test Fähigkeitsparameter alle Untertests

Box-Test auf Gleichheit der Kovarianzenmatrizen^a

Box-M-Test	46,003
F	1,313
df1	30
df2	8101,573
Sig.	,118

Prüft die Nullhypothese, daß die beobachteten Kovarianzen- matrizen der abhängigen Variablen über die Gruppen gleich sind.

a. Design: Konstanter Term + BED

Multivariate Tests^a

Effekt	Wert	F	Hypothese df	Fehler df	Sig.
Pillai-Spur	,300	1,697	10,000	96,000	,092
Wilks-Lambda	,717	1,699 ^b	10,000	94,000	,092
BED					
Hotelling-Spur	,369	1,699	10,000	92,000	,093
Größte charakteristische Wurzel nach Roy	,282	2,703 ^c	5,000	48,000	,031

a. Design: Konstanter Term + BED

b. Exakte Statistik

c. Die Statistik ist eine Obergrenze auf F, die eine Untergrenze auf dem Signifikanzniveau ergibt.

Lineare Extrapolation

	Testpersonen ID	Itemgruppen (Bedingung)	Linear extrapolierte SEE
Untertest 1	104 A	adaptiv (1)	1.4
	210 A	6-7-8 (2)	1.4
	212 A	6-13-18 (3)	1.4
	216 A	6-7-8 (2)	1.4
	236 A	6-13-18 (3)	1.4
Untertest 3	222 A	6-13-18 (3)	1.4
	236 A	6-13-18 (3)	1.4
Untertest 9	210 A	6-7-8 (2)	1.4
Untertest 11	202 A	6-13-18 (3)	1.4
	207 A	6-13-18 (3)	1.4
	212 A	6-13-18 (3)	1.4
	215 A	6-13-18 (3)	1.4
	216 A	6-7-8 (2)	1.3
	221 A	6-7-8 (2)	1.3
	224 A	6-13-18 (3)	1.4
	227 A	6-13-18 (3)	1.4
	228 A	6-7-8 (2)	1.3
	229 A	6-13-18 (3)	1.4
	232 A	6-13-18 (3)	1.4
	236 A	6-13-18 (3)	1.4

Ergebnisse Kreuztabellen (vor der Umordnung)

Bedingung * Standardschätzfehler 1 Kreuztabelle

		Standardschätzfehler 1					Gesamt
		,6	,7	,8	1,1	1,4	
Bedingung	adaptiv	6	3	6	3	1	19
	konventionell schwierig	9	4	0	3	2	18
	konventionell mittel	5	3	3	4	2	17
Gesamt		20	10	9	10	5	54

Bedingung * Standardschätzfehler 3 Kreuztabelle

		Standardschätzfehler 3						Gesamt
		,6	,7	,8	,9	1,0	1,4	
Bedingung	adaptiv	7	5	3	3	1	0	19
	konventionell schwierig	4	8	1	4	1	0	18
	konventionell mittel	7	3	5	0	0	2	17
Gesamt		18	16	9	7	2	2	54

Bedingung * Standardschätzfehler 6 Kreuztabelle

		Standardschätzfehler 6				Gesamt
		,6	,7	,8	1,1	
Bedingung	adaptiv	17	2	0	0	19
	konventionell schwierig	15	3	0	0	18
	konventionell mittel	14	1	1	1	17
Gesamt		46	6	1	1	54

Bedingung * Standardschätzfehler 9 Kreuztabelle

		Standardschätzfehler 9					Gesamt
		,6	,7	,8	1,1	1,4	
Bedingung	adaptiv	7	9	2	1	0	19
	konventionell schwierig	9	6	1	1	1	18
	konventionell mittel	7	4	5	1	0	17
Gesamt		23	19	8	3	1	54

Bedingung * Standardschätzfehler 11 Kreuztabelle

		Standardschätzfehler 11						Gesamt	
		,6	,7	,8	,9	1,1	1,3		1,4
Bedingung	adaptiv	1	5	2	7	4	0	0	19
	konventionell	1	3	3	2	6	3	0	18
	schwierig								
	konventionell mittel	0	0	3	0	5	0	9	17
Gesamt		2	8	8	9	15	3	9	54

Mittelwertvergleich der Standardschätzfehler alle Untertests (vor der Umordnung)

Bericht

Bedingung		SEE 1	SEE 3	SEE 6	SEE 9	SEE 11
adaptiv	Mittelwert	,800	,726	,611	,695	,863
	N	19	19	19	19	19
	Standardabweichung	,2236	,1284	,0315	,1177	,1571
konventionell schwierig	Mittelwert	,794	,744	,617	,717	,967
	N	18	18	18	18	18
	Standardabweichung	,2859	,1247	,0383	,2093	,2275
konventionell mit- tel	Mittelwert	,865	,771	,647	,712	1,206
	N	17	17	17	17	17
	Standardabweichung	,2783	,2519	,1281	,1317	,2358
Insgesamt	Mittelwert	,819	,746	,624	,707	1,006
	N	54	54	54	54	54
	Standardabweichung	,2600	,1734	,0775	,1552	,2498

Mittelwertvergleich der Standardschätzfehler aller Untertests (nach der Umordnung)

Untertest 1- adaptiv

Deskriptive Statistik ^a					
	N	Mittelwert		Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 1	27	,859	,0518	,2693	,073
Gültige Werte (Listenweise)	27				

a. Bedingung = adaptiv

Untertest 1- konventionell schwierig

Deskriptive Statistik ^a					
	N	Mittelwert		Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 1	11	,627	,0141	,0467	,002
Gültige Werte (Listenweise)	11				

a. Bedingung = konventionell schwierig

Untertest 1- konventionell mittel

Deskriptive Statistik ^a					
	N	Mittelwert		Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 1	16	,881	,0697	,2786	,078
Gültige Werte (Listenweise)	16				

a. Bedingung = konventionell mittel

Untertest 3- adaptiv

Deskriptive Statistik ^a					
	N	Mittelwert		Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 3	31	,742	,0240	,1336	,018
Gültige Werte (Listenweise)	31				

a. Bedingung = adaptiv

Untertest 3 - konventionell schwierig

Deskriptive Statistik ^a					
	N	Mittelwert	Standardabweichung	Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 3	9	,667	,0236	,0707	,005
Gültige Werte (Listenweise)	9				

a. Bedingung = konventionell schwierig

Untertest 3 - konventionell mittel

Deskriptive Statistik ^a					
	N	Mittelwert	Standardabweichung	Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 3	14	,807	,0707	,2645	,070
Gültige Werte (Listenweise)	14				

a. Bedingung = konventionell mittel

Untertest 6 - adaptiv

Deskriptive Statistik ^a					
	N	Mittelwert	Standardabweichung	Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 6	28	,607	,0050	,0262	,001
Gültige Werte (Listenweise)	28				

a. Bedingung = adaptiv

Untertest 6 - konventionell schwierig

Deskriptive Statistik ^a					
	N	Mittelwert	Standardabweichung	Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 6	16	,619	,0101	,0403	,002
Gültige Werte (Listenweise)	16				

a. Bedingung = konventionell schwierig

Untertest 6 - konventionell mittel

Deskriptive Statistik ^a					
	N	Mittelwert		Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 6	10	,680	,0512	,1619	,026
Gültige Werte (Listenweise)	10				

a. Bedingung = konventionell mittel

Untertest 9 - adaptiv

Deskriptive Statistik ^a					
	N	Mittelwert		Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 9	25	,732	,0390	,1952	,038
Gültige Werte (Listenweise)	25				

a. Bedingung = adaptiv

Untertest 9 - konventionell schwierig

Deskriptive Statistik ^a					
	N	Mittelwert		Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 9	15	,640	,0131	,0507	,003
Gültige Werte (Listenweise)	15				

a. Bedingung = konventionell schwierig

Untertest 9 - konventionell mittel

Deskriptive Statistik ^a					
	N	Mittelwert		Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 9	14	,736	,0357	,1336	,018
Gültige Werte (Listenweise)	14				

a. Bedingung = konventionell mittel

Untertest 11 - adaptiv**Deskriptive Statistik**

	N	Mittelwert		Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 11	34	,932	,0334	,1950	,038
Gültige Werte (Listenweise)	34				

Untertest 11 - konventionell schwierig**Deskriptive Statistik**

	N	Mittelwert		Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 11	3	,700	,0577	,1000	,010
Gültige Werte (Listenweise)	3				

Untertest 11 - konventionell mittel**Deskriptive Statistik**

	N	Mittelwert		Standardabweichung	Varianz
	Statistik	Statistik	Standardfehler	Statistik	Statistik
Standardschätzfehler 11	17	1,206	,0572	,2358	,056
Gültige Werte (Listenweise)	17				

Kruskal-Wallis Standardschätzfehler (nach der Umordnung)

Ränge

	Bedingung	N	Mittlerer Rang
Standardschätzfehler 1	adaptiv	27	30,33
	konventionell schwierig	11	14,59
	konventionell mittel	16	31,59
	Gesamt	54	

Statistik für Test^{a,b}

Standardschätzfehler 1	
Chi-Quadrat	10,055
df	2
Asymptotische Signifikanz	,007

a. Kruskal-Wallis-Test

b. Gruppenvariable: Bedingung

Ränge

	Bedingung	N	Mittlerer Rang
Standardschätzfehler 3	adaptiv	31	28,47
	konventionell schwierig	9	20,33
	konventionell mittel	14	29,96
	Gesamt	54	

Statistik für Test^{a,b}

Standardschätzfehler 3	
Chi-Quadrat	2,503
df	2
Asymptotische Signifikanz	,286

a. Kruskal-Wallis-Test

b. Gruppenvariable: Bedingung

Ränge

	Bedingung	N	Mittlerer Rang
Standardschätzfehler 6	adaptiv	28	25,36
	konventionell schwierig	16	28,38
	konventionell mittel	10	32,10
	Gesamt	54	

Statistik für Test^{a,b}

Standardschätzfehler 6	
Chi-Quadrat	3,741
df	2
Asymptotische Signifikanz	,154

a. Kruskal-Wallis-Test

b. Gruppenvariable: Bedingung

Ränge

	Bedingung	N	Mittlerer Rang
Standardschätzfehler 9	adaptiv	25	28,58
	konventionell schwierig	15	20,40
	konventionell mittel	14	33,18
	Gesamt	54	

Statistik für Test^{a,b}

Standardschätzfehler 9	
Chi-Quadrat	5,704
df	2
Asymptotische Signifikanz	,058

a. Kruskal-Wallis-Test

b. Gruppenvariable: Bedingung

Ränge

	Bedingung	N	Mittlerer Rang
Standardschätzfehler 11	adaptiv	34	23,35
	konventionell schwierig	3	7,50
	konventionell mittel	17	39,32
	Gesamt	54	

Statistik für Test

	Standardschätzfehler 11
Chi-Quadrat	17,461
df	2
Asymptotische Signifikanz	,000

Curriculum Vitae

Persönliche Angaben

Name	Lammert, Anna
e-Mail	lammert.anna@outlook.com
Geburtsdatum	15.11.1984
Geburtsort, Geburtsland	Tursunsade, Tadschikistan

Ausbildung

10/2007 - dato	Universität Wien Dipl. Psychologie Schwerpunkte: Psychologische Diagnostik sowie Arbeits-, Organisations- und Wirtschaftspsychologie
08/2005 - 07/2007	Hanse-Schule für Wirtschaft und Verwaltung - Lübeck (DE) Allgemeine Hochschulreife - Schwerpunkt Wirtschaft
08/2002 - 06/2005	Autokraft GmbH - Neustadt i. H. (DE) Lehre: Kauffrau für Bürokommunikation

Berufserfahrung

01/2014 - dato	Erste Bank AG - Wien Personalentwicklung und Recruiting
10/2013 - 12/2013	UniCredit Bank Austria AG - Wien Human Resources Management & Strategy
05/2011- dato	Health Consult Ges.m.b.H. - Wien Arbeitsmedizinisches und Arbeitspsychologisches Zentrum
04/2012 - 01/2013	Haus der Barmherzigkeit Wirtschaftsbetriebe GmbH - Wien Projekt: Mitarbeiterbefragung
08/2012 - 12/2012	HILL - AMC Management GmbH - Wien Pflichtpraktikum: Forschung & Entwicklung