universität wien

Diplomarbeit

Titel der Diplomarbeit

Testtheoretische Analyse des Zahlenreihentests: Effekte bei wiederholter Vorgabe

Verfasser

Gottfried Berndl

angestrebter akademischer Grad

Magister der Naturwissenschaften (Mag. rer. nat.)

Wien, 2015

Studienkennzahl lt. Studienblatt:	A 298
Studienrichtung lt. Studienblatt:	Diplomstudium Psychologie
Betreuer:	Univ.-Prof. i.R. Mag. Dr. Klaus Kubinger

Danksagung

Wieder geht ein Abschnitt zu Ende und das ist schön. Zahlreiche Unterstützer bereicherten diesen Weg und halfen, Hürden zu überwinden. All diesen Personen möchte ich an dieser Stelle danken.

Insbesondere bedanke ich mich bei Herrn Univ.-Prof. i.R. Mag. Dr. Klaus Kubinger für seine wertvollen Anmerkungen und die Betreuung dieser Arbeit.

Ein großer Dank gebührt auch Herrn Mag. Herbert Poinstingl für seine umfangreiche fachliche Unterstützung.

Ein herzliches Dankeschön möchte ich auch an die lieben Kolleginnen und Kollegen richten, die mich bei den Testungen begleiteten. In diesem Zusammenhang sei auch den Direktorinnen und Direktoren sowie den Schuladministratorinnen und –administratoren für ihren beherzten Einsatz gedankt und natürlich auch allen teilnehmenden Schülerinnen und Schülern.

Für die geduldige „Rundumbetreuung“ danke ich meinen Eltern, Schwiegereltern und Großeltern. Vor allem meine Frau Romana war mir in dieser Zeit in jeder Hinsicht eine große Hilfe. Vielen Dank, meine Liebe!

Zuletzt sei auch meine Tochter Livia erwähnt. Sie hat die letzte Phase dieser Arbeit zwar lediglich im Bauch meiner Frau begleitet, vor allem unsere taktilen Kommunikationen haben mich aber täglich zum Strahlen gebracht. Nun liegt sie neben mir und ich bin schwer verliebt.

Kurzzusammenfassung

Im Rahmen der Selektionsdiagnostik ist es üblich, Bewerberinnen und Bewerbern die Möglichkeit zur Wiederholung eines Aufnahmeverfahrens zu geben. Neben simplen Retest-Effekten besteht allerdings die Gefahr, dass es dabei zu einer Veränderung substantieller Messeigenschaften eines Tests kommt.

In dieser Arbeit wird untersucht, ob der Reasoning-Test ZART (Poinstingl, 2009c) für den wiederholten Einsatz geeignet ist und mit welchen Veränderungen zu rechnen ist. In einem Messwiederholungsdesign wurde bei einem Retestungsintervall von zwei Wochen eine repräsentative Itemstichprobe in identer Form vorgegeben. In die Auswertung gingen Daten von $N = 241$ Schülerinnen und Schülern höherer Schulen im Alter von 14-20 Jahren ein.

Nach Analysen mit dem 1PL-Modell wurde nach Ausschluss eines Items in der Retestung dieselbe latente Dimension wie in der Initialtestung gemessen. Modellierungen mit dem linearen logistischen Testmodell von Fischer sprechen für einen vom Antwortformat unabhängigen Retest-Effekt, um den sich die Schwierigkeiten der Aufgaben in der Retestung gleichermaßen verringern. Das Ausmaß dieses Effektes dürfte in den meisten Fällen aber praktisch uninteressant sein. Zudem wird auf Veränderungen bei der Bearbeitungszeit eingegangen und ein erster Befund zur Kriteriumsvalidität des ZART dargestellt.

Abstract

In selection settings it is common to allow applicants to retake an assessment procedure. In addition to simple retest effects, there is a risk to that it will trigger a change of the psychometric properties of a test across test sessions.

This study focuses on the ZART reasoning test (Poinstingl, 2009c). Is it an appropriate test for retest situations, and which changes in psychometric properties should be taken into account? For this reason, students of academic secondary schools and a college for higher vocational education were tested in two sessions at an interval of two weeks with an identical test form, consisting of a representative sample of items from the itempool of the ZART. Data of $N = 241$ students aged between 14 and 20 has been analysed.

After the exclusion of one item, Rasch model analysis indicated that the same latent trait was measured at both test sessions. The application of Fischer's Linear Logistic Test Model reveals a general retest effect, by which the difficulties of the items are reduced in the retest. The retest effect was independent of the response format used. However, in most cases the size of the retest effect should be negligible. Changes in test time across the sessions and a first evidence for the criterion-related validity of the ZART are also reported.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

I Einleitung

Der von Poinstingl (2009c) entwickelte *Zahlenreihentest* (ZART; siehe auch Poinstingl, 2010a, 2010b) zielt auf die Erfassung *Schlussfolgernden Denkens* ab. Aufgabe ist, die Regeln, nach der eine Reihe von Zahlen aufgebaut ist, zu erkennen, um ein fehlendes Glied zu bestimmen. Der Hauptanwendungsbereich des bislang unpublizierten Computertests ist die berufs- und ausbildungsbezogene Eignungsdiagnostik und somit liegt auch eine Verwendung im Rahmen von psychologischen Auswahlverfahren nahe.

Vor allem im akademischen Kontext ist es im Allgemeinen möglich, bei einem Auswahlverfahren mehrmals anzutreten. Beispielsweise haben nach Lievens, Buyse und Sackett (2005) etwa 40% der Personen, welche von 2000 bis 2003 an einem Zulassungsverfahren für Medizinstudien in Belgien teilnahmen, die Testungen zumindest einmal wiederholt, wobei durchschnittlich ca. 30% je Durchgang aufgenommen wurden. Im beruflichen Kontext dürfte dieser Anteil geringer ausfallen (vgl. Hausknecht, Halpert, Di Paolo & Moriarty Gerrard, 2007). Seitens der Bewerber und Bewerberinnen liegt der Grund für die Wiederholung eines Auswahlverfahrens auf der Hand, sie hoffen durch eine neuerliche Teilnahme in der gewünschten Ausbildung zugelassen oder in dem angestrebten Beruf eingestellt zu werden. Auch die *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2011, S. 146) und andere Qualitätsstandards (siehe dazu Lievens et al., 2005; Lievens, Reeve & Heggestad, 2007) empfehlen, den Interessenten die Möglichkeit zur Wiederholung von Aufnahmetests zu geben. Begründungen für diese Empfehlung dürften nach Lievens et al. (2005, S. 982) eventuelle Messfehler in der Ersttestung, z.B. infolge von Krankheit oder durch Abweichungen von der standardisierten Testsituation, sowie die Chance, dass sich eine Testperson zwischenzeitlich im interessierenden *Konstrukt*[1] verbessern konnte, sein.

Obwohl die Möglichkeit zur Testwiederholung im Rahmen der Selektionsdiagnostik üblich scheint, fehlt es nach Lievens et al. (2007) an einem umfassenden

[1] „Konstrukte sind … allgemein anerkannte, aber eben nicht direkt beobachtbare ‚Phänomene', wie z.B. Intelligenz, Angst oder Stress" (Kubinger, 2009, S. 57).

Verständnis über *Retest-Effekte* – „the psychometric ‚costs' of retesting are currently unknown" (S. 1672). Häufiger untersucht ist die Veränderung von Testwerten infolge der Retestung (vgl. die Metaanalysen von Hausknecht et al., 2007; Kulik, Kulik & Bangert, 1984). Würden eventuelle Verbesserungen bei einem Aufnahmeverfahren unberücksichtigt bleiben, würden all jene, die zum ersten Mal teilnehmen, systematisch benachteiligt werden – das Gütekriterium *Fairness* des Verfahrens wäre verletzt. In den letzen Jahren wird häufiger ein grundsätzliches Problem angesprochen (vgl. z.B. Lievens et al., 2005; Lievens et al., 2007; Matton, Vautier & Raufaste, 2009; Reeve & Lam, 2005): Wird in der Retestung überhaupt noch dasselbe, z.B. das durch die Anforderungsanalyse intendierte Merkmal erfasst? Hierbei geht es, allgemein ausgedrückt, um die Veränderung substantieller Messeigenschaften, insbesondere in Hinblick auf das Gütekriterium *Validität*.

Abgesehen von der Selektionsdiagnostik erfordern auch wiederholte Testungen mit dem Ziel der Verlaufskontrolle Verfahren, deren Kennwerte weitgehend unabhängig von den Vorgaben ihrer selbst oder von Paralleltests, also der „völlig gleichwertigen Nachahmung des Tests" (Kubinger, 2009, S. 50) sind.

Nach den *International Guidelines for Test Use* (International Test Commission, 2001, S. 21) sollen bei der Interpretation von Testleistungen Vorerfahrungen mit einem Test berücksichtigt werden, sofern Daten über Effekte auf die Testleistung verfügbar sind. Hinsichtlich der berufsbezogenen Eignungsbeurteilung regelt schließlich die DIN 33430 (DIN Deutsches Institut für Normung e. V., 2002), dass „bei wiederholtem Einsatz von Verfahren in gleichartigen Eignungsuntersuchungen[,] ... die bei der entsprechenden Fragestellung erreichte Gültigkeit der Verfahren ermittelt werden [sollte]" (S. 7). Ziel der vorliegenden Arbeit ist es, entsprechende Daten für den ZART zu sammeln. Es wird untersucht, ob bzw. wie sich die Messeigenschaften dieses Tests verändern, wenn er von denselben Personen ein zweites Mal bearbeitet wird. Zudem wird der Einfluss des Antwortformates auf das Ausmaß eventuell vorhandener Effekte betrachtet.

Kapitel 1 widmet sich einer genaueren Beschreibung des Tests, welcher in dieser Arbeit Betrachtung findet, dem ZART. Dabei wird an geeigneten Stellen auf die Vorzüge des Testmodells, nach dem dieser Test konstruiert wurde, eingegangen, dem *dichotomen logistischen Testmodell* von Rasch (auch 1PL-Modell genannt oder kurz *Rasch*-Modell; 1960). Zum besseren Verständnis und da auch in

der restlichen Arbeit immer wieder darauf zurückgegriffen werden wird, erfolgen in Kapitel 1.1 ergänzende Ausführungen zu testtheoretischen Grundlagen.

Kapitel 2 setzt sich mit der Literatur zu Retest-Effekten bei Tests zur Ermittlung intellektueller Fähigkeiten auseinander. Nach einem Überblick, welche psychometrischen Eigenschaften im Rahmen einer wiederholten Testung bzw. Retestung im Fokus der Forschung stehen und den nötigen Begriffsbestimmungen, wird in Kapitel 2.1 auf Faktoren eingegangen, welche das Ausmaß von Retest-Effekten beeinflussen. In Kapitel 2.2 werden verschiedene Erklärungen über das Zustandekommen von Retest-Effekten, samt deren Auswirkung auf die Validität von Initial- und Retestung, gegeben. Es geht also darum, wie Retest-Effekte zu interpretieren sind. Kapitel 2.3 wirft einen Blick auf die Messung von Veränderung. Ein Schwerpunkt liegt hierbei auf den Möglichkeiten, die das Rasch-Modell und dessen Spezialfall, das *linear logistische Testmodell* von Fischer (LLTM; 1973), in diesem Zusammenhang bieten. In Kapitel 2.4 werden schlussendlich die in der Literatur gefundenen Befunde zu Retest-Effekten bei Tests mit Zahlenfolgen bzw. Zahlenreihen dargestellt.

II Theoretischer Teil

1 Der Zahlenreihentest (ZART)

Neben weiteren Eckdaten wurde in der Einleitung bereits erwähnt, dass der ZART auf die Erfassung von *Reasoning* bzw. *Schlussfolgerndem Denken* mithilfe von numerischem Material abzielt. „*Reasoning* ist die Fähigkeit, Gesetzmäßigkeiten oder logisch zwingende Zusammenhänge erkennen und zweckentsprechend verwerten zu können“ (Kubinger, 2009, S. 206). Zunächst wird in diesem Kapitel die bisherige Entwicklungsarbeit des Testautors zur Erfassung dieses Konstruktes beschrieben. Danach wird die konkrete Gestaltung der einzelnen Aufgaben sowie des Tests insgesamt dargestellt. Zuletzt wird auf die Gütekriterien des ZART eingegangen. Da es sich beim ZART noch um ein relativ neues psychologisch-diagnostisches Verfahren handelt, welches sich noch in Entwicklung befindet, sind dazu noch keine umfassenden Befunde verfügbar.

Die Entwicklung des ZART erfolgte nach zwei Gesichtspunkten (vgl. Poinstingl, 2009e, 2010a, 2010b): Die Aufgaben, also die Zahlenreihen wurden nach einem zuvor erstellten Konstruktionsrational generiert und sollten konform mit dem Rasch-Modell sein.

Die Formulierung grundlegender Regeln, vor der Itemgenerierung, verfolgt das Ziel, dass verschiedene Testentwickler Aufgaben erstellen können, die dasselbe Konstrukt messen (Feger, 1984). Die rationale bzw. regelgeleitete Itemkonstruktion dient somit vor allem dazu, *inhaltliche Gültigkeit*[2] für ein Verfahren zu erreichen (Embretson, 1999; Schott & Wieberg, 1984). „Bei der Erstellung des Konstruktionsrationals des [ZART] … wurde besonders darauf geachtet, dass die Rechenfertigkeit der Testperson kein zu starkes Gewicht bekommt, um allein bzw. überwiegend das Konstrukt des *Schlussfolgernden Denkens* (eben im numerischen Bereich) zu erfassen“ (Berndl, Steinfeld & Poinstingl, 2012, S. 162). Die konkreten Regeln finden sich bei Poinstingl (2009c).

[2] „Inhaltliche Gültigkeit im Sinne von logischer Validität ist … dadurch zu erreichen, dass bei der Konstruktion der einzelnen Items eines Tests ganz bestimmte, definitorisch festgelegte Regeln Anwendung finden. ... (nur) die richtige Anwendung dieser Regeln [führt] zur Lösung … . Von denjenigen Tpn [Testpersonen], welche viele Aufgaben lösen, ist – bis auf gegenteilige empirische Belege – anzunehmen, dass sie die fraglichen Regeln beherrschen“ (Kubinger, 2009, S. 56).

Die Konstruktion nach dem Rasch-Modell bringt einige Vorzüge hinsichtlich der Güte eines Verfahrens, welche im Rahmen der Gütekriterien dargestellt werden. Für eine anwendungsbezogene Einführung sei auf Kubinger, Rasch und Yanagida (2011) verwiesen. Umfangreiche Darstellungen finden Interessierte bei Fischer (1974) oder Kubinger (1989). Zumindest die Grundidee dieses probabilistischen Testmodells ist in Kapitel 1.1 skizziert. Erste Informationen zum ZART sammelte Poinstingl (2009d, 2009e) in einer Vortestung, in der 30 Items einer kleinen Stichprobe von 35 Studierenden, gemeinsam mit einem Fragebogen zu Testsoftware und Test, vorgegeben wurden. Die Ergebnisse und Rückmeldungen führten zur Überarbeitung des Konstruktionsrationals und der Generierung von neuen Items, von denen 68 einer Kalibrierung unterzogen wurden (Poinstingl, 2010a, 2010b). Dabei bearbeiteten etwa 600 Testpersonen jeweils eine von sechs Testformen mit jeweils 10 – 20 Aufgaben. Jeweils idente Items, sogenannte Brücken- oder Verlinkungsitems, waren in den Testformen derart verteilt, dass der Bezug zwischen allen Personen und allen Items nicht verloren gehen konnte. Nach dem Ausschluss von „nicht real arbeitenden“ Testpersonen, aufgrund von unrealistischen Antwortzeiten und sehr hohen Personfit-Indices, gingen Daten von ca. 420 Personen in die psychometrischen Analysen ein. Nach dem Ausschluss von drei Items konnte a posteriori die Gültigkeit des Rasch-Modells angenommen werden, d.h. für die verbliebenen 65 Items kann, solange keine gegenteiligen empirischen Belege vorliegen, angenommen werden, dass sie dieselbe *latente Eigenschafts*- oder *Fähigkeitsdimension*[3] erfassen.

Im beschriebenen Fall der a posteriori erreichten Modellgeltung schlägt Kubinger (2005) eine „Art Kreuzvalidierung“ vor, indem die verbliebenen Items an einem neuen Datensatz auf ihre Rasch-Modell-Konformität hin überprüft werden.

Die Geltung des Rasch-Modells ist auch die Voraussetzung für weitere Analysen mit dem linear logistischen Testmodell (LLTM). Die Grundidee des LLTM ist wieder in Kapitel 1.1 dargestellt. Darüber hinausgehende Informationen sind ebenfalls in Fischer (1974) oder Kubinger (1989) nachlesbar. In einem nächsten Schritt möchte Poinstingl (vgl. z.B. 2010a) unter Anwendung des LLTM das Konstruktionsrational des ZART einer empirischen Prüfung unterziehen. Falls das gelingt,

[3] Während der Begriff Konstrukt keine testtheoretischen Eigenschaften impliziert, gehen die genannten Begriffe von der Eindimensionalität der Messung aus, was eine Voraussetzung für die Geltung des Rasch-Modells darstellt (siehe z.B. Kubinger et al., 2011).

wären die Schwierigkeiten der jeweiligen kognitiven Operationen, die zur Lösung der Problemstellungen nötig sind, bekannt (vgl. z.B. Hornke & Habon, 1986; Poinstingl, 2009a). Dadurch ließen sich neue Aufgaben in erforderlicher Komplexität erstellen, ohne aufwendige Kalibrierungsstudien durchzuführen.

Die Zahlenreihen des ZART setzen sich aus sieben Gliedern zusammen. Jeweils ein beliebiges Glied wurde durch ein Fragezeichen ersetzt. Zur Illustration soll folgendes Beispiel dienen: 17 13 8 2 (?) -13 -22[4]. Aufgabe der Testperson ist es, die Regeln nach der eine Aufgabe aufgebaut ist, zu erkennen, um das fehlende Glied zu bestimmen. Die Bearbeitung erfolgt ohne zeitliche Beschränkung, Notizen sind nicht gestattet. In dem genannten Beispiel wird eine Subtraktion mit einer schrittweise um eins ansteigenden Zahl durchgeführt. Somit lautet die Lösung: -5. Durch die Verwendung negativer Zahlen ist die Vorgabe des ZART erst ab der achten bis neunten Schulstufe sinnvoll.

Es konnte gezeigt werden (vgl. Kubinger & Gottschall, 2007; Kubinger, Holocher-Ertl, Reif, Hohensinn & Frebort, 2010), dass durch die Wahl des Antwortformates durchaus relevante Rateeffekte provoziert werden können. Kubinger und Gottschall (2007) beispielsweise fanden beim Multiple-Choice-Format „1 aus 6“ im Schnitt deutlich geringere Item-Schwierigkeitsparameter als beim Multiple-Choice-Format „x aus 5“ oder dem freien Antwortformat. Über die Software des ZART (Poinstingl, 2009b) können die Aufgaben wahlweise im Multiple-Choice-Format mit sechs Antwortmöglichkeiten (eine Lösung und fünf Distraktoren), im sequenziellen Antwortformat (siehe Kubinger, 2009, S. 140), in dem die sechs Möglichkeiten nacheinander vorgegeben werden, oder im freien Antwortformat vorgebeben werden. Je nach Auswahl sinkt die a priori Ratewahrscheinlichkeit von einem Sechstel bis gegen Null (vgl. Kubinger, 2009). Einschränkend muss aber erwähnt werden, dass in der Kalibrierung das Multiple-Choice-Format verwendet wurde (vgl. Poinstingl, 2010a, 2010b).

Mit der von Poinstingl erstellen Software ist es auch möglich, Items, mit den verschiedenen Antwortformaten in einer Testung vorzugeben, und zwar indem Itemblöcke mit den jeweiligen Antwortformaten mit den entsprechenden Instruktionen hintereinander vorgegeben werden, was vordergründig für wissenschaftliche

[4] Um zu verhindern, dass Aufgaben des ZART an die Öffentlichkeit gelangen, werden in dieser Arbeit keine Testitems wiedergegeben.

Zwecke interessant ist. Mit der unter aktuellen Versionen von Microsoft Windows lauffähigen Software ist grundsätzlich auch *adaptives Testen* (zum Begriff siehe Kubinger, 2009) möglich.

Um die Testpersonen mit dem Test vertraut zu machen, wird ihnen die Problemstellung anhand von zwei Beispielaufgaben in der Instruktion erklärt. Danach folgen Übungsaufgaben, deren Anzahl vom Testleiter im Vorfeld festgelegt wird. Je nachdem ob eine Testperson eine Übungsaufgabe lösen kann, erhält sie eine entsprechende Rückmeldung. Bei einer negativen Rückmeldung hat sie die Möglichkeit, die Übungsaufgabe nochmals zu bearbeiten. Während den Übungs- und Testaufgaben hat eine Testperson jederzeit die Möglichkeit, sich die Problemstellung über die Hilfe nochmals in Erinnerung zu rufen.

Neben den Eingaben einer Testperson beim Aufruf des Programms (Alter, Geschlecht etc.) und den Antworten bei den Testitems, registriert die Software auch die Bearbeitungsdauer je Item sowie die Bearbeitungsdauer je Block, welcher sich aus Instruktion, Beispielitem(s) und Testaufgaben zusammensetzt.

Der von Poinstingl entwickelte ZART besteht also im Wesentlichen aus einem Rasch-Modell konformen Itempool (vgl. 2010a, 2010b) und einer Software (2009b), mit der eine beliebige Auswahl daraus sehr flexibel vorgegeben werden kann. Als Testkennwert Schlussfolgerndes Denken erhält man die (blockweise) Schätzung der Personenparameter aus dem Rasch-Modell.

Durch die Geltung des Rasch-Modells für den ZART (Poinstingl, 2010a, 2010b) soll die *Skalierung*[5] als Ausgangspunkt bei der Betrachtung der Gütekriterien dienen. Über die Eindimensionalität der Messung hinaus ist es dabei notwendig, dass die Verrechnung leistungsadäquat ist. Aufgrund der Charakteristika des Rasch-Modells erfolgt die Schätzung der Personenparameter im ZART[6] auf Basis der Anzahl der gelösten Items. Gleichgültig, welche Aufgaben eine Testperson löst, enthält die Anzahl gelöster Items sämtliche Information über den Leistungsgrad einer Person, man spricht in diesem Zusammenhang von einer *erschöpfenden Statistik* (Fischer, 1974, S. 195). „Das *Rasch*-Modell [muss] notwendigerweise gel-

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

[6] Die Ausführungen gelten selbstverständlich nur bei einer festen Anzahl an Aufgaben, also bei konventioneller Vorgabe und nicht beim grundsätzlich möglichen adaptiven Testen. Wie die Schätzung hier funktioniert, ist bei Kubinger (2009, S. 102-107) ersichtlich.

ten … , wenn die Anzahl gelöster Aufgaben ein faires Maß für die erbrachte Testleistung sein soll“ (Kubinger, 2009, S. 88).

Nachdem die Items aufgrund der Konformität mit dem Rasch-Modell nachweisbar eindimensional messen, kann nach Kubinger (2009) die Messgenauigkeit in Bezug auf die *innere Konsistenz* als gegeben angesehen werden (S. 96). Das Konfidenzintervall für ein Testergebnis, zur Beurteilung von dessen Genauigkeit, lässt sich anhand des *Standardschätzfehlers* errechnen. Der Standardschätzfehler ist mit der Mathematischen Statistik individuell bestimmbar (S. 95-96), im Gegensatz zum *Standardmessfehler,* der in der *klassischen Testtheorie* dafür Verwendung findet und für die Testergebnisse aller Personen gleichermaßen gilt. Die Annahme des klassischen Reliabilitätskonzeptes, dass ein Test jede Leistungsausprägung gleich genau erfassen kann, ist auch „nicht unmittelbar plausibel“ (S. 54).

Die *inhaltliche Gültigkeit* ist aufgrund der eingangs beschriebenen Konstruktion des ZART per Definition gegeben. An der Objektivität im Sinne der *Testleiterunabhängigkeit* und *Verrechnungssicherheit* besteht ebenfalls kein Zweifel, da es sich um einen Computertest handelt. Normierungsstudien sind derzeit keine vorhanden. Insofern fehlen auf eine Eichung bezogene Prozentränge zur Gewährleistung der *Interpretationseindeutigkeit*.

Soweit die generellen Befunde zur Güte des ZART. Einen wesentlichen Einfluss auf die Gütekriterien hat natürlich die letztendliche Gestaltung eines Tests bei der Anwendung. Zur Beurteilung der Ökonomie eines Verfahrens macht es beispielsweise einen Unterschied, ob die Items konventionell oder adaptiv vorgegeben werden (vgl. Kubinger, 2009).

18 Aufgaben des ZART fanden innerhalb des *Wiener Self-Assessments Architektur*© 2011[7] (siehe Khorramdel, Maurer, Frebort & Kubinger, 2012) Verwendung im *Wiener Zahlenreihentest* (siehe Berndl, Steinfeld & Poinstingl, 2012). Auf Basis ihrer Testergebnisse erhalten die am Architekturstudium interessierten Testpersonen eine „quasi-individualisierte“ Rückmeldung, wodurch die Interpretationseindeutigkeit erreicht wird. Zur *prognostischen Validität* des Wiener Zahlenreihentests zeigte sich bei Müller (2011), die das Self-Assessment einer ersten Evaluation unterzog, kein positiver Befund. Dabei ist aber auf die grundsätzlichen Probleme

[7] http://studienwahl.tuwien.ac.at

bei prognostischen Validierungsversuchen hinzuweisen (siehe Kubinger, Frebort & Müller, 2012).

Informationen zur Nützlichkeit des Wiener Zahlenreihentests ergaben weitere Analysen der Daten von Müller (2011) durch Berndl et al. (2012). Dabei scheint der Wiener Zahlenreihentest mit 0,95 eine hohe Sensitivität aufzuweisen, die Spezifität, also die Wahrscheinlichkeit, eine vorliegende Schwäche entsprechend zu prognostizieren, war hingegen problematisch. Bezüglich Self-Assessments kann man aber argumentieren, dass es „angeraten erscheint, ein geringes Risiko 1. Art einzugehen, d.h. Studiumsinteressierten eine Schwäche zu attestieren, obwohl sie gar keine solche haben; das deshalb, weil davon ausgegangen werden kann, dass bei einem nicht unerheblichen Anteil die Rückmeldung einen Einfluss auf die Entscheidung für oder gegen das Studium hat“ (S. 167).

Ferner ist zu erwähnen, dass beim Wiener Zahlenreihentest lediglich neun Aufgaben für den Testkennwert Schlussfolgerndes Denken verrechnet werden. Dies ist einem speziellen Vorgabemodus geschuldet, wodurch auch die *Belastbarkeit* einer Testperson erhoben werden kann.

1.1 Testtheoretische Grundlagen

In Kapitel 1 wurde mehrfach auf das *dichotome logistische Testmodell* von Rasch (auch 1-PL Modell genannt oder kurz *Rasch*-Modell; 1960) eingegangen und auch dessen Spezialfall, das *linear logistische Testmodell* von Fischer (LLTM; 1973), fand bereits Erwähnung. In diesem Kapitel werden die Ideen hinter diesen Modellen skizziert und noch nicht genannte wichtige Charakteristika dargestellt.

Das Rasch-Modell ist ein Vertreter der sogenannten *Item Response Theory* (IRT), welche sich als testtheoretischer Zugang neben der klassischen Testtheorie etabliert hat. Gegenstand der IRT ist es, „das Zustandekommen einer Reaktion (Antwort) auf eine Aufgabe, Frage oder Feststellung … eines psychologischen Tests oder Fragebogens modellhaft zu beschreiben“ (Kubinger et al., 2011, S. 555).

Beim Rasch-Modell hängt, wie in Formel 1 ersichtlich, die Wahrscheinlichkeit, dass eine Testperson v ein Item i löst, von der Fähigkeit der Testperson ξ_v und der Schwierigkeit der Aufgabe σ_i ab (siehe z.B. Kubinger, 1989, S. 22).

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

Aus Formel 1 ist auch erkennbar, dass dieses Modell die *lokale stochastische Unabhängigkeit* der Antworten voraussetzt, d.h. die Wahrscheinlichkeit ein Item zu lösen, ist unabhängig davon, welche anderen Aufgaben eine Person bereits gelöst hat (siehe z.B. Fischer, 1974, S. 211).

Im Zusammenhang mit dem Gütekriterium Skalierung im letzten Kapitel wurden bereits einige Vorzüge des Rasch-Modells gegenüber Modellen der klassischen Testtheorie genannt: der mögliche Notwendigkeitsbeweis um festzustellen, ob die Anzahl gelöster Aufgaben ein faires Testmaß darstellt, die Existenz erschöpfender Statistiken und die nachweisbare Eindimensionalität der Messung. Zudem ermöglicht das Rasch-Modell *spezifisch objektive Vergleiche*, woraus folgt, dass die Parameterschätzungen *stichprobenunabhängig* sind und, ebenfalls bereits angeklungen, dass mit *Modelltests* geprüft werden kann, ob ein Itempool Rasch-Modell-konform ist.

Spezifisch objektiv meint, dass der Vergleich zweier Personen hinsichtlich ihrer Fähigkeitsparameter unabhängig von den Parametern der eingesetzten Items möglich ist (Fischer, 1974, S. 210).

Umgekehrt bedeutet das nach Kubinger (1989) auch, „daß der Vergleich je zweier Items bezüglich σ_i und σ_j unabhängig davon ist, welche Stichprobe dafür verwendet wird: Die Schätzungen der Parameter sind ‚stichprobenunabhängig', weil die Wahl der Stichprobe aus einer bestimmten Population für die statistische Inferenz dieser Parameter keine Rolle spielt" (S. 23). Dadurch lassen sich, mithilfe der *Conditioned-Maximum-Likelihood*-Schätzung (CML), die Strukturparameter σ_i unabhängig von den Personenparametern ξ_v schätzen. In der in Formel 1 wiedergegebenen Modelldarstellung mit logarithmierten Parametern liegen Personen- und Itemparameter dabei auf einer Differenzenskala (Rost, 2004, S. 121).

Die Stichprobenunabhängigkeit bildet auch die Grundlage für die Prüfbarkeit des Modells (siehe z.B. Fischer, 1974, S. 281-300). Unterscheiden sich die Schätzungen der Strukturparameter in zwei, nach beliebigen Kriterien geteilten Teilstichproben, so wären die Modellannahmen verletzt. Nach diesem Prinzip geht der (*bedingte*) *Likelihood-Quotienten-Test* (*Likelihood Ratio Test*; LRT) von Andersen (1973) vor, wobei von einer Konformität des Itempools nach dem Rasch-Modell

ausgegangen werden kann, solange die H_0: $\sigma_i^{(1)} = \sigma_i^{(2)} = \sigma_i$ nicht widerlegt wurde. Als Teilungskriterium bietet sich neben sogenannten externen Kriterien, beispielweise Geschlecht oder Alter, das interne Kriterium Testscore, also die Anzahl gelöster Aufgaben, an (Kubinger, 1989, S. 24). Eine große Trennschärfe des LRT erwartet man sich vor allem bei letztgenanntem.

Um einen Eindruck über die Passung der einzelnen Aufgaben an das Modell zu erlangen, um gegebenenfalls welche auszuschließen, bietet sich eine *Grafische Modellkontrolle* (siehe z.B. Kubinger et al., 2011, S. 556) an. Dabei werden die zuvor für beide Substichproben auf $\sum_{i=1}^{k} \hat{\sigma}_i = 0$ standardisierten Schätzungen der Itemparameter in ein kartesisches Koordinatensystem eingetragen. Wird jede Aufgabe als Punkt, mit den beiden Schätzungen als Koordinaten, eingezeichnet, so sollten bei Geltung des Rasch-Modells alle Punkte nahe der durch den Ursprung verlaufenden 45° Geraden liegen. Weitere Methoden, um die empirische Geltung des Rasch-Modells zu testen, finden sich bei Glas und Verhelst (1995).

Im LLTM werden, wie in Formel 2 dargestellt, die Itemparameter σ_i des Rasch-Modells durch eine Linearkombination sogenannter Basisparameter η_j ersetzt (Fischer, 1973; siehe aber auch Fischer, 1974; Kubinger, 1989).

$$\sigma_i = \sum_j^p q_{ij}\eta_j \tag{2}$$

Es können damit also Basisparameter η_j geschätzt werden, die, gewichtet nach ihrem Vorkommen im Item i, die Schwierigkeiten der Items σ_i erklären sollen. Die Modellierung der zuvor postulierten Annahmen erfolgt anhand der Strukturmatrix Q, mit den Gewichten q_{ij}. Für die Schätzung der Strukturparameter η_j im LLTM muss deren Anzahl p geringer sein als die Anzahl k der Strukturparameter σ_i im zugrundeliegenden Rasch-Modell. Die Schätzung der Basisparameter η_j kann wieder mit der CML-Methode erfolgen.

Die Basisparameter können die Schwierigkeit der kognitiven Operationen widerspiegeln, die zur Lösung einer Aufgabe nötig sind, wie es in Bezug auf Konstruktionsrationale in Kapitel 1 angesprochen wurde (vgl. z.B. Hornke & Habon, 1986; Poinstingl, 2009a; Sonnleitner, 2008). Mit Hilfe des LLTM können aber auch Effekte der Testvorgabe, z.B. Positionseffekte (vgl. Hahne, 2008; Hohensinn,

Kubinger, Reif, Holocher-Ertl, Khorramdel & Frebort, 2008), untersucht werden. Ein umfangreicher Einblick zu den Möglichkeiten des LLTM findet sich bei Kubinger (2008).

Ebenso wie das Rasch-Modell ist auch das LLTM ein prüfbares Modell (Kubinger, 1989, S. 34). Zunächst muss die Prüfung des zugrundeliegenden Rasch-Modells mit den oben beschriebenen Methoden erfolgen. Gilt dieses, so dient es wie ein saturiertes Modell dazu, die Annahmen im LLTM zu prüfen. Dazu wird untersucht, ob die Daten durch das LLTM statistisch genauso gut wie durch das Rasch-Modell erklärt werden. Dies kann ebenfalls mit einem (bedingten) LRT erfolgen. Die aus Formel 3 resultierende Prüfgröße Z ist asymptotisch χ^2-verteilt mit $df = (k-1) - p$.

$$Z = -2ln\left\{\frac{L_{LLTM}}{L_{RM}}\right\} \tag{3}$$

Auch unterschiedliche Modellannahmen können über diesen Ansatz untersucht werden, indem alternative LLTM geschätzt und deren Passung über Formel 3 getestet wird. Über hierarchisch angewandte LRT lassen sich dann damit sehr elegant spezifische Hypothesen über einzelne Vorgabeeffekte prüfen (vgl. Hahne, 2008; Hohensinn et al., 2008; Kubinger, 2008). Formel 4 testet beispielsweise die Nullhypothese $\lambda = 0$. Da die beiden Modelle nur in einem Parameter differieren, ist die resultierende Prüfgröße Z asymptotisch χ^2-verteilt mit $df = 1$.

$$Z = -2ln\left\{\frac{L_{LLTM(\lambda=0)}}{L_{LLTM(\lambda\neq0)}}\right\} \tag{4}$$

In Kapitel 2.3.1 werden zu einem konkreten Anwendungsfall die einzelnen Schritte dazu dargestellt.

2 Effekte bei wiederholter Vorgabe

In der einschlägigen Literatur werden die Effekte bei wiederholter Vorgabe von identen- oder Paralleltests unter den Begriffen *retest effects* und *practice effects* diskutiert. Beide Begriffe stehen im engeren Sinn für Veränderungen des Testwertes einer Person infolge der wiederholten Vorgabe eines Tests (Hausknecht et al., 2007, S. 374; Lievens et al., 2005, S. 982). Im weiteren Sinn werden mit diesen Begriffen alle psychometrischen Veränderungen bei der Retestung umschrieben. Wie in der Einleitung schon angesprochen, werden zunehmend häufiger auch das *Validitätsproblem der Veränderungsmessung* (siehe z.B. Rost, 2004, S. 280) und dessen Ursachen berücksichtigt. Bei Personen, die einen Test oder eine Testbatterie schon einmal durchgeführt haben, könnte in der Retestung nicht dasselbe Konstrukt wie in der Initialtestung erfasst werden. Wie Testwiederholungen die Validität eines Tests beeinflussen können, wird in Kapitel 2.2 näher ausgeführt.

Zu den Veränderungen bei der Retestung sind auch jene der Bearbeitungszeit zu zählen. In der angesprochenen Literatur spielen diese aber kaum eine Rolle, da die Tests aufgrund der Gruppentestungen meist mit Zeitbegrenzung vorgegeben wurden. Mit der Computerdiagnostik besteht die Möglichkeit, Tests auch ohne *Speed*-Komponente in der Gruppe vorzugeben und dennoch eine ungestörte Testsituation zu gewährleisten (Kubinger, 2009, S. 152). Beispielsweise stellten Raymond, Neustel und Anderson (2009) bei einem Zertifizierungstest fest, dass die durchschnittliche Bearbeitungsdauer je Item in der Retestung geringer war.

Die Definitionen von Hausknecht et al. (2007) und Lievens et al. (2005) von *retest effects* bzw. *practice effects* schließen auch Effekte durch zwischenzeitliche Interventionen (z.B. Coaching) mit ein. In der vorliegenden Arbeit werden hingegen lediglich Retest-Effekte betrachtet, welche allein durch die wiederholte Bearbeitung eines Tests bzw. einer Testbatterie oder dessen bzw. deren Parallelform auftreten. Diese Eingrenzung verfolgt selbstverständlich nur didaktische Zwecke, da die Durchführung eines Tests durchaus als eine spezielle Form der Übung anzusehen ist. Gewisse Ähnlichkeiten bestehen zu einem Phänomen, welches im Zusammenhang mit Tests bzw. Prüfungen aus dem Ausbildungskontext diskutiert wird, dem sogenannten *testing effect*. Dabei kommt es infolge des Kontakts mit einem Test, der eine Gedächtniskomponente erfasst, zu einer langfristigen Leistungssteigerung im getesteten Material gegenüber dem restlichen Lernstoff

(Roediger & Karpicke, 2006, S. 181). Von Retest-Effekten sind zudem *Positionseffekte* (siehe dazu Hahne, 2008; Hohensinn et al., 2008; Kubinger, 2008) abzugrenzen, das sind Lern- als auch Ermüdungseffekte die bereits während der Testbearbeitung auftreten können.

Durch die mannigfaltige Verwendung des Begriffes *Test* ist noch konkreter darzustellen, welche Tests in dieser Arbeit im Zusammenhang mit Retest-Effekten diskutiert werden: In der Literatur zu Retest-Effekten erfolgt vor allem eine differenzierte Betrachtung gegenüber Tests, wie sie im Rahmen von Examen oder Zertifizierungen eingesetzt werden. „Typical achievement tests are employed primarily to assess terminal performance, as for example in a licensing examination ... . Typical aptitude tests, on the other hand, are specially designed for predictive purposes“ (Anastasi, 1981, S. 1087). Die Übergänge zwischen diesen Polen sind aber durchaus fließend (vgl. Anastasi, 1981, S. 1087), so hat ein Test, der die Lehrinhalte aus Chemie in einem Aufnahmetest prüft, eher den Sinn, den späteren Lernerfolg für diese Inhalte abzuschätzen. Der Grund für diese Unterscheidung liegt in den in Kapitel 2.2 diskutierten Theorien über das Zustandekommen von Retest-Effekten. Während überdauernde Eigenschaften, wie die Intelligenz, „cannot be substantially and permanently raised by special training“ (Spitz, 1986, S. 219, zitiert nach Carroll, 1993, S. 669) und sich „erfahrungsgemäß bloß in Folge gravierender *Life-Events* entscheidend verändern“ (Kubinger, 2009, S. 14), können leichter erwerbbare Fähigkeiten (z.B. deklaratives Wissen) einfacher modifiziert werden. Aufgrund des Reasoning-Tests, der Ausgangspunkt dieser Arbeit ist, liegt der Schwerpunkt auf den genannten „aptitude tests“, welche durch die Prüfung der Leistungshöhe[8] auf die Ermittlung kognitiver Fähigkeiten abzielen. Nicht angesprochen sind zudem Verfahren, bei denen bereits durch deren Gestaltung oder Inhalte deutliche Retest-Effekte zu erwarten sind. Beispielhaft seien Tests zur Messung der Merkfähigkeit bei identem Retest sowie Verfahren genannt, die irgendeine Art von Rückmeldung bieten (z.B. *Lerntests*, siehe Guthke & Wiedl, 1996).

In einem methodischen Zusammenhang mit Retest-Effekten steht die *Retest-Reliabilität* bzw. die *Stabilität*, sofern zwischen Test und Retest ein längerer Zeitraum besteht (zu den Begriffen siehe Kubinger, 2009, S. 52). Die Korrelation zwi-

[8] im Gegensatz zur Beurteilung der Bearbeitungsgeschwindigkeit bei Speed-Tests.

schen Initial- und Retestung wird in der klassischen Testtheorie dazu verwendet, um sich dem angesprochenen Validitätsproblem zu nähern (Rost, 2004, S. 280). Näheres dazu wird in Kapitel 2.3 beschrieben.

2.1 Faktoren und deren Einfluss auf Retest-Effekte

Die Darstellungen in diesem Kapitel beruhen zu einem guten Teil auf den Metaanalysen von Kulik et al. (1984) und Hausknecht et al. (2007). Im Gegensatz zu den Letztgenannten nahmen Kulik et al. (1984) nur Studien in ihre Analyse auf, deren Ergebnisse unabhängig von zwischenzeitlichen Interventionen waren. Zudem bezogen die Autoren, im Gegensatz zu Hausknecht et al. (2007), auch Untersuchungen mit ein, die Retest-Effekte bei „achievement tests“ untersuchten. Insgesamt zeigte sich in den Metaanalysen ein durchschnittlicher Anstieg in den Testwerten, von etwa einem Fünftel bis Viertel der Standardabweichung[9,10], wobei weitere Testwiederholungen mit erneuten Anstiegen einher gingen (Hausknecht et al., 2007; Kulik et al., 1984). Dabei fand die Forschung mehrere Faktoren bzw. Moderatorvariablen, die einen Einfluss auf Verbesserungen im Retest haben und somit einen differenzierteren Einblick in Retest-Effekte erlauben.

Im Folgenden wird zunächst auf Faktoren eingegangen, die mit einem Test selbst bzw. dessen Gestaltung in Verbindung stehen, gefolgt von Effekten, die auf Eigenschaften und Zuständen der Testpersonen selbst beruhen. Zuletzt werden Variablen dargestellt, die sich im Zusammenhang mit dem Studienkontext finden.

Einen wesentlichen Einfluss auf das Ausmaß von Retest-Effekten dürften die verwendeten Testformen haben. Für idente Testformen in der Initial- und Retestung ermittelte Hausknecht et al. (2007) eine mittlere Effektgröße von $0{,}40$, bei Paralleltests zeigten sich im Schnitt lediglich etwa halb so große Retest-Effekte. Die einschlägigen Effektgrößen bei Kulik et al. (1984) hatten ein ähnliches Ausmaß. Es erscheint einleuchtend, dass mit der Länge des Retestungsintervalls Retest-Effekte abnehmen müssten (vgl. Amelang & Bartussek, 2001, S. 616). Diesen Zu-

[9] Wenn nicht anders gekennzeichnet, beziehen sich die wiedergegebenen Schätzungen der (relativen) Effektgrößen $\hat{E}$ auf die Standardabweichung im Initialtest: $\hat{E} = (\bar{x}_{\text{Retest}} - \bar{x}_{\text{Test}})/s_{\text{Test}}$.

[10] Hausknecht et al. (2007) ermittelten die Effektgrößen mit einer Schätzung der Bezugsgröße aus den Standartabweichungen aus beiden Durchgängen (Cohen´s d).

sammenhang konnten Hausknecht et al. (2007) nur bei identen Testformen nachweisen, was sie auf Erinnerungseffekte zurück führten. Die Autoren errechneten, dass Retest-Effekte bei identen Testformen erst nach ca. zwei Jahren das Niveau bei den Parallelen erreicht haben. Weitere nötige Informationen, um das Resultat einordnen zu können, etwa deskriptive Statistiken zur Variable Testintervall, sind in dem Artikel aber leider nicht ersichtlich.

Zum Einfluss der Itemgestaltung auf Retest-Effekte wurde insgesamt relativ wenig geforscht. Powers (1986) untersuchte in einer Zusammenschau von 10 Studien, wie sich verschiedene Itemcharakteristiken auf das Ausmaß der Retest-Effekte auswirken. Dabei gingen längere Instruktionen und komplexere Aufgaben mit höheren Effekten und längere Antwortzeiten je Item mit geringeren einher. Erwähnenswert, da relativ einfach zu adaptieren, ohne grundlegende Eigenschaften eines Tests zu ändern, ist die Auswirkung des Antwortformates. Deutlichere Retest-Effekte zeigten sich bei Tests, welche über alle Aufgaben hinweg dieselben Antwortmöglichkeiten boten sowie bei Items mit vier Antwortmöglichkeiten gegenüber jenen mit fünf. Einschränkend merkt Powers aber an, dass alle Aufgaben mit vier Antwortoptionen gleichzeitig über den Test hinweg unverändert blieben.

Ebenfalls diskutiert wird der Einfluss der g[11]-Sättigung des von einem Test erfassten *spezifischen* Konstruktes. Die Metaanalyse von te Nijenhuis, van Vianen, und van der Flier (2007), welche die Steigerung der Testwerte bei Intelligenz-Testbatterien im Retest (ohne zwischenzeitliche Interventionen) anhand von 64 Studien untersuchte, fand hierzu einen korrigierten Korrelationskoeffizienten von $-1{,}00$ für den Zusammenhang zwischen den g-Faktor Ladungen und dem Ausmaß der Retest-Effekte. Demnach fallen Retest-Effekte bei Tests mit hoher g-Sättigung geringer aus. Welche Schlussfolgerung dieses Ergebnis zur Validität von Retestungen zulässt, wird im folgenden Kapitel weiter ausgeführt.

Zu den Eigenschaften der Testpersonen, welche das Ausmaß von Retest-Effekten moderieren, ist vor allem die allgemeine kognitive Leistungsfähigkeit g untersucht worden. Nach Kulik et al. (1984) scheinen Personen mit hohem Leistungsvermögen mehr von der Retestung zu profitieren. In ihrer Metastudie reichten die mittleren Effektgrößen von $0{,}17$ bei Stichproben mit leistungsschwachen Personen bis

[11] Die Literatur bezieht sich hier weitestgehend auf den allgemeinen Faktor der Intelligenz (g) nach Spearman (1927; siehe auch Carroll, 1993).

zu $0{,}82$ bei leistungsstarken, in der ersten Retestung bei Verwendung identer Testformen.

Außer zu kognitivem Leistungsvermögen existiert wenig Forschung, die sich mit dem Einfluss von individuellen Unterschieden auf Retest-Effekte befasst. Reeve und Lam (2007) untersuchten den Einfluss von generellen Einstellungen gegenüber Tests und der Motivation, den Test durchzuführen. Über die verwendeten Subtests hinweg zeigte sich aber kein konsistentes Ergebnis.

In einem Zusammenhang mit der Motivation ist auch der Studienkontext zusehen. So wird angenommen, dass bei Aufnahmeverfahren, also im praktischen Kontext, ein höheres Ausmaß an Retest-Effekten zu beobachten sein müsste als bei Stichproben, welche rein zum Forschungszweck getestet wurden (Hausknecht et al., 2007; Reeve & Lam, 2007). Ausgehend von der in Aufnahmeverfahren gegebenen Selbst-Selektion der Teilnehmer, dürfte der Anreiz im wissenschaftlichen Kontext, selbst wenn das Studiendesign diverse Motivatoren enthält, kaum an die lohnenden Folgen im praktischen Kontext heran reichen. Grundsätzlich ist den Bewerberinnen und Bewerbern in Aufnahmeverfahren zudem häufig im Vorfeld bekannt, welche Aufgabentypen verwendet werden. Auch wenn Vorbereitungsmöglichkeiten somit bereits für den ersten Antritt genutzt werden können, nehmen Hausknecht et al. (2007) an, dass dies aufgrund der höheren Motivation im verstärktem Maße für die Retestung stattfindet. Ein Einfluss des Studienkontextes konnte in ihrer Metaanalyse allerdings nicht nachgewiesen werden. Dass im reinen Forschungskontext die Retestungsintervalle üblicherweise geringer sind, könnte das Ergebnis möglicherweise beeinflusst haben, wurde von den Autoren aber nicht angesprochen.

Werden Retest-Effekte im praktischen Kontext erforscht, dann setzt sich die Stichprobe in der Retestung in der Regel aus Personen zusammen, welche in vorangegangenen Durchgängen das Selektionskriterium nicht erfüllt haben. Da die Stichprobe nicht nach dem Zufall, sondern aufgrund von Selektionen in Vortests gebildet wird, macht dieses Design anfällig für ein methodisches Artefakt, den *Regressionseffekt* (Guthke & Wiedl, 1996; Hausknecht et al., 2007; Stelzl, 1982). Die Wahrscheinlichkeit, dass eine Steigerung der Mittelwerte der Testwerte auch auf Regressionsartefakte zurückzuführen ist, steigt mit dem Anteil der Selektierten in den vorangegangenen Tests. Für zwei Untersuchungen konnten Hausknecht et

al. (2007) zeigen, dass etwas weniger als 10% der ermittelten Effektgrößen auf Regressionseffekte zurückgeführt werden können.

2.2 Theorien zur Interpretation von Retest-Effekten

Dieses Kapitel widmet sich der Frage, ob die wiederholte Durchführung eines Tests mit quantitativen und bzw. oder qualitativen Veränderungen im gemessenen Konstrukt einher geht. Dies ist wesentlich, um einen adäquaten Umgang mit Retest-Effekten zu finden.

Rein quantitative Veränderungen nimmt Erklärung 1 an (vgl. z.B. Lievens et al., 2007, S. 1674-1675). Ein höherer Testwert ist somit auf Verbesserungen in dem durch den Test bzw. der Testbatterie erfassten Konstrukt zurückzuführen. Das infrage stehende Konstrukt wird somit auf dieselbe Art und Weise zu den verschiedenen Zeitpunkten erfasst, die substantiellen Messeigenschaften des Tests bleiben also stabil. In diesem Fall könnten Retest-Effekte durch entsprechende Abzüge berücksichtigt werden, um ein Aufnahmeverfahren diesbezüglich fair zu gestalten (Anastasi, 1981, S. 1087).

Die postulierten konstruktrelevanten Lerneffekte infolge der Ersttestung sind konsistent mit dem oben angesprochenen testing effect (vgl. Lievens et al., 2007, S. 1674-1675). Im Gegensatz zu Tests, die leichter veränderbare Fähigkeiten erfassen (z.B. Wissenstests), scheint es unwahrscheinlich, dass die Durchführung eines Tests zu einer Förderung überdauernder *breiter* Konstrukte (z.B. g, siehe oben) führt. So konnten Matton et al. (2009) zeigen, dass die Verbesserungen von Bewerberinnen und Bewerbern beim zweiten Antritt für ein Pilotentraining auf konstruktirrelevante Veränderungen zurückgeführt werden können. Auch die in Kapitel 2.1 bereits angesprochene Metaanalyse von te Nijenhuis et al. (2007) führte zu dem Ergebnis, dass die Steigerungen im Retest bei Intelligenz-Testbatterien keine g-Sättigung enthielten.

Die alternativen Erklärungen gehen alle von qualitativen Veränderungen aus, welche sich in diversen psychometrischen Variationen des Tests oder der Testbatterie manifestieren:

Dass in der Initialtestung hemmende, konstruktirrelevante Einflüsse stärker vorhanden sind, meint Erklärung 2. Anastasi (1981) nimmt an, dass allgemein durch die Erfahrungen mit standardisierten Tests Unruhe und Ängstlichkeit vermindert und Selbstvertrauen gestärkt werden, was zu genaueren und valideren Testergebnissen führt. Im Gegensatz zu erfahrenen Testpersonen können unerfahrene ihr Potential somit erst in einem weiteren Durchgang ausschöpfen.

Nach diesem Ansatz erscheint es sinnvoll, den Bewerbern und Bewerberinnen bei Auswahlverfahren im Vorfeld Informationsmaterial und Beispielitems zu Verfügung zu stellen, um eventuelle Benachteiligungen auszugleichen.

Ein Teil dieser Erklärung deckt sich mit dem *Warming-up* Effekt, der aber nur zu Beginn eines Tests auftritt. Ansonsten finden sich keine empirischen Bestätigungen dieser Erklärung. Reeve und Bonaccio (2008) untersuchten, ob Ängstlichkeit in Bezug auf einen Test dessen Messeigenschaften beeinflusst, konnten dies aber nicht nachweisen.

Erklärung 3 geht von einer Erhöhung des Testwertes im Retest aufgrund der Förderung konstruktirrelevanter Anteile aus (Lievens et al., 2007, S. 1675). Alles Wissen im Zusammenhang mit Tests, also von der allgemeinen Vertrautheit mit standardisierten Testsituationen, über die Kenntnis von spezifischen Aufgaben, bis zur Erinnerung an einzelne Lösungen, führt demnach, überspitzt formuliert, zu einem Verfall der Messeigenschaften eines Tests. Darunter fallen somit alle Erkenntnisse, die vom intendierten Konstrukt abweichen. Beispielsweise auch die Lösung durch alleinige Analyse der Antwortmöglichkeiten, wie sie Mittring und Rost (2008) untersuchten, wenngleich dies nur in Einzelfällen relevant sein dürfte. Bei einer wiederholten Testung werden dadurch spezifische und in weiterer Folge auch breite Konstrukte weniger valide erfasst.

Bei einem Assessment wäre nach Erklärung 3 in der Retestung z.B. mit einer geringeren *prognostischen Validität* zu rechnen (Lievens et al., 2007, S. 1675). Somit wäre bei deutlichen Retest-Effekten eine wiederholte Zulassung zu einem Aufnahmeverfahren, aus messtechnischer Sicht, kritisch.

Lievens et al. (2007) untersuchten, wie sich die Messeigenschaften einer Reasoning-Testbatterie bei einem Retestungsintervall von etwa zwei Monaten im praktischen Kontext änderten. Die Ergebnisse waren konform mit Erklärung 3.

Aufgrund weiterer Analysen führten die Autoren die Retest-Effekte zum Teil auf Erinnerungseffekte zurück.

Erklärung 4 nimmt an, dass die Differenzen der Testwerte in den einzelnen Durchgängen auf tatsächlichen Veränderungen in den zugrundeliegenden spezifischen Konstrukten beruhen (Arendasy & Sommer, 2013, S. 183-184). Im Gegensatz zu Erklärung 1 sind die Faktorenladungen zwischen den Subtestwerten und den breiteren Konstrukten bzw. Faktoren strukturellen Veränderungen unterworfen. Die aufgabenspezifischen Verbesserungen generalisieren nicht auf die breiteren Konstrukte und sind in Bezug auf g letztendlich unbedeutend.

Neben den aktuelleren Ergebnissen von Arendasy und Sommer (2013) sprechen auch jene von Reeve und Lam (2005) für diese Annahmen. Bei Aufnahmeverfahren sollten somit die Faktor-Score-basierten Schätzungen breiterer Konstrukte zur Interpretation heran gezogen werden, anstatt die Rohwerte gleichwertig zu verrechnen (vgl. Reeve & Lam, 2005). Letzteres ist aber grundsätzlich schlicht falsch (Kubinger, 2009, S. 59).

In der Literatur zu Retest-Effekten wurden häufig Verfahren mit Zeitbegrenzung eingesetzt (z.B. Matton et al., 2009; Reeve & Lam, 2005; 2007; bei Lievens et al., 2007 sind dazu keine Informationen verfügbar). Dies könnte in zweierlei Hinsicht ein Artefakt begünstigen, was aber in keiner der Studien Berücksichtigung fand:
Einerseits wäre es nicht verwunderlich, dass die Erfahrung mit einem Aufgabentyp zu einer effizienteren Bearbeitung im Retest führt. Somit könnten die Testpersonen mehr Items in der zu Verfügung stehenden Zeit bearbeiten, was letztendlich höhere Testwerte im Retest begünstigen würde. Damit konsistent sind die bereits genannten, geringeren Effekte bei längeren Antwortzeiten je Item bei Powers (1986).

Je nach Ausgestaltung der Zeitbegrenzung führen diese Bedingungen zur Erfassung von zwei vermengten Eigenschaften, „nämlich die Fähigkeit, bestimmte Anforderungen – auch schwierige – grundsätzlich zu erfüllen, mit der Fähigkeit, dies auch (hinreichend) schnell zu können“ (Kubinger, 2009, S. 84). Aufgrund der Erfahrung mit einem Aufgabentyp wäre es somit andererseits wahrscheinlich, dass diese beiden Anteile im Retest in einem anderen Verhältnis als im ersten Durchgang erfasst werden würden. Dazu konnte Mollenkopf (1950, 1960) für Auf-

gaben eines Test zur Erfassung von arithmetischem Reasoning zeigen, dass sich bei der Gabe von zusätzlicher Bearbeitungszeit, direkt im Anschluss an den Test, deutliche Unterschiede in den Rangordnungen der Testpersonen unter den beiden Zeitbedingungen ergaben.

2.3 Methoden der Veränderungsmessung

Entsprechend der Ausrichtung dieser Arbeit liegt der Schwerpunkt hinsichtlich der Methoden darauf, welche Möglichkeiten das Rasch-Modell und das LLTM (siehe Kapitel 1.1) zur Modellierung von Veränderung bieten. Diese werden unter Punkt 2.3.1 dargestellt. Hier wird zunächst darauf eingegangen, welche grundsätzlichen Probleme der klassischen Testtheorie mithilfe dieser probabilistischen Modelle hinsichtlich der Veränderungsmessung gelöst werden können.

Da zur Beurteilung der Dimensionalität von Testbatterien am häufigsten die Faktorenanalyse eingesetzt wird, wird infolge skizziert, wie im Rahmen dieses Messmodells die Invarianz der Messungen zu verschiedenen Zeitpunkten untersucht werden kann. Eine umfangreiche Darstellung dazu findet sich bei Vandenberg und Lance (2000). Dies soll vor allem dazu dienen, die Ergebnisse einiger Studien, welche in dieser Arbeit zitiert wurden, beispielsweise jene von Lievens et al. (2007) und Reeve und Laam (2005), besser reflektieren zu können.

Bezüglich der klassischen Probleme der Veränderungsmessung, der mangelhaften Reliabilität der Differenzwerte zwischen Post- und Prätest, der künstlichen negativen Korrelation zwischen Ausgansniveau und Differenzwert und dem bereits angesprochenen Regressionseffekt sei auf Guthke und Wiedl (1996) oder Stelzl (1982) verwiesen.

Bevor das Ausmaß eventueller Retest-Effekte ermittelt wird, sollte geklärt sein, dass zu allen Zeitpunkten dieselbe Variable gemessen wurde (vgl. Rost, 2004, S. 280). Ansonsten macht die Bildung von Differenzwerten schlichtweg keinen Sinn, würden nach Rost doch „Äpfel und Birnen“ (2004, S. 280) miteinander verrechnet werden. Dies beschreibt das *Validitätsproblem der Veränderungsmessung*.

Im Rahmen der klassischen Testtheorie wird die Frage, ob zu zwei Zeitpunkten in einem Test dasselbe erfasst wird, anhand der Korrelation der beiden Testwerte untersucht. Erfolgt die Erhebung der beiden Testwerte unter gleichen Be-

dingungen, so entspricht die Korrelation bei kurzfristiger Wiederholung der *Retest-Reliabilität*, bei einem längeren Zeitraum zwischen Test und Retest der *Stabilität* (vgl. Kubinger, 2009, S. 52).

Eine geringe Stabilität, also eine deutliche Veränderung der Rangordnung der Testpersonen im Retest, kann nun zweierlei bedeuten: Eben, dass sich die Messeigenschaften des Tests substantiell geändert haben oder dass differentielle Veränderungen der Testpersonen dahinter stehen (Guthke & Wiedl, 1996, S. 344). Im Rahmen der klassischen Testtheorie kann nicht entschieden werden, welche dieser Alternativen zutrifft.

Mit Hilfe des Rasch-Modells ist das Validitätsproblem hingegen sehr wohl lösbar (Fischer, 1974, 1995; Rost, 2004). Zwei praktikable Möglichkeiten, das Rasch-Modell als Veränderungsmodell umzuformulieren sind in Kapitel 2.3.1 dargestellt. Die Möglichkeiten sind insofern praktikabel, als dass bei jenen Modellen auch die Vorzüge des Rasch-Modells in vollem Umfang gelten (siehe Kapitel 1 und 1.1). Neben der Lösung des Validitätsproblems bietet das Rasch-Modell gegenüber der klassischen Testtheorie somit auch Kennwerte, die unabhängig von der Zusammensetzung der Stichprobe sind, wobei all diese Annahmen grundsätzlich einer Prüfung unterzogen werden können. Darauf aufbauende Analysen durch das LLTM sind ebenfalls möglich (siehe vor allem Fischer, 1974, 1995). Weitere probabilistische Modelle für die Veränderungsmessung sind bei Fischer (1977) dargestellt.

Da die Faktorenanalyse ebenfalls der klassischen Testtheorie zuzurechnen ist, können damit deren grundsätzliche Probleme nicht überwunden werden: Die Grundannahme, dass sich jede beobachtete Variable aus einer Linearkombination latenter Eigenschaften (Faktoren) zusammensetzt, ist nicht prüfbar und alle Ergebnisse sind abhängig von der Zusammensetzung der Stichprobe (Kubinger, 2009, S. 58-59). Hinzu kommen verfahrenstechnische Probleme, die Bestimmung der Faktorenanzahl und die inhaltliche Beschreibung der Faktoren (ebd.).

Auch dem Validitätsproblem kann man sich mit diesem Modell lediglich nähern. Grundgedanke ist, die Invarianz der einzelnen Kennwerte zu den verschiedenen Zeitpunkten zu untersuchen (vgl. Vandenberg & Lance, 2000). Dies geschieht im Rahmen einer konfirmatorischen Faktorenanalyse für Multi-Gruppen-Vergleiche (*multigroup confirmatory factor analysis*), welche durch lineare Struk-

turgleichungsmodelle modelliert werden kann. Zur Prüfung der Invarianz der Modellkennwerte werden sogenannte genestete Modelle, das sind Modelle mit ineinander verschachtelten Spezifikationen, erstellt. Um beispielsweise zu prüfen, ob sich die Ladungen zu zwei Zeitpunkten unterscheiden, wird ein Modell, in dem die Ladungen gleich gesetzt wurden, mit einem Modell ohne diese Restriktion verglichen. Der genaue Ablauf ist bei Vandenberg und Lance (2000) beschrieben.

Ob ein erstelltes Modell zur beobachteten Datenstruktur passt, kann mit dem χ^2-Test geprüft werden (vgl. Schermelleh-Engel, Moosbrugger & Müller, 2003, S. 31-33). Für den Vergleich zweier genesteter Modelle steht der χ^2-Differenz-Test zu Verfügung (S. 33-35). Die Signifikanztests haben aber einige Schwächen (siehe S. 32-34). Bei geringem Stichprobenumfang steigt beispielsweise beim χ^2-Test die Wahrscheinlichkeit, fälschlicherweise die Passung eines Modells anzunehmen (Fehler 2. Art). Bei großen Stichproben kommt es hingegen zu einer Überhöhung der Irrtumswahrscheinlichkeit (Fehler 1. Art). Deshalb stehen zur Beurteilung der Passung eines Modells als auch zum Modellvergleich zusätzlich eine Reihe deskriptiver Maße, sogenannte Fit-Indizes, zur Verfügung (siehe Schermelleh-Engel et al., 2003). Auch wenn Empfehlungen zur Beurteilung vorhanden sind (S. 51-53), bleibt in diesem Punkt ein Ermessensspielraum.

Die Stichprobengröße ist auch hinsichtlich der Parameterschätzung ein heikler Punkt (vgl. Schermelleh-Engel et al., 2003, S. 48-51). Sind nur wenige manifeste Indikatoren je Faktor vorhanden, so sollte das mit einer größeren Stichprobe kompensiert werden (S. 50). Deshalb teilten Lievens et al. (2007) und Reeve und Laam (2005) die Subtests in mehrere Teile auf. Somit werden zwar Schätzprobleme durch die höhere Anzahl manifester Indikatoren vermieden, die Reliabilität der Indikatoren sinkt aber entsprechend.

Ein interessanter Methodenmix findet sich neuerdings bei Arendasy und Sommer (2013). Um Veränderungen der Messeigenschaften der eingesetzten Testbatterie zu beurteilen, wurde ebenfalls eine konfirmatorischen Faktorenanalyse für Multi-Gruppen-Vergleiche angewandt. Dass zu beiden Zeitpunkten die gleichen latenten Dimensionen erfasst wurden, konnte auf Itemebene durch die Geltung des Rasch-Modells für die vier Subtests nachgewiesen werden.

2.3.1 Veränderungsmessung im Rasch-Modell und im LLTM

Die Überlegungen in diesem Kapitel stehen alle im Kontext der Messung von Retest-Effekten. Der Einfachheit halber erfolgen alle Darstellungen für eine Testwiederholung, also zwei Testzeitpunkte und anhand von vollständigen Designs. Davon ausgehend sind Verallgemeinerungen leicht möglich, wenngleich z.B. unterschiedliche Aufgaben in den einzelnen Durchgängen, etwa für die Klärung der Validitätsproblematik, keinen Beitrag leisten können.

Eine rein formale Möglichkeit, das Rasch-Modell zur Messung von Retest-Effekten zu modifizieren, bestehet darin, den Personenparameter ξ in einen Ausgangsparameter ξ_v und einen Wiederholungsparameter λ_v zu zerlegen (Kubinger, 1979, 2008). $\xi_v - \sigma_i$ aus Formel 1 wird somit zu $(\xi_v + \lambda_v) - \sigma_i$, wobei zu Testzeitpunkt eins $\lambda_v = 0$ gilt. λ_v stellt somit die personenspezifischen Veränderungen in der Eigenschaftsdimension dar, welche sich aufgrund der vorangegangenen Bearbeitung des Tests mit den Items i zu Zeitpunkt zwei zeigen. Da je Person nur ein Personenparameter geschätzt werden kann, reduziert sich $(\xi_v + \lambda_v) - \sigma_i$ zu $\xi_{vt} - \sigma_i$, wobei ξ_{vt} die Fähigkeit der Testperson v zu Testzeitpunkt t darstellt. Zur Schätzung von ξ_{vt} ist die Datenmatrix wie in Abbildung 1 gezeigt zu organisieren. Jede Person, die ein zweites Mal den Test bearbeitet, wird dabei als „virtuelle“ neue Person angeschrieben, welche die gleichen Items bearbeitet. Anders ausgedrückt bilden sich in diesem Modell, unter der Annahme, dass die Itemparameter konstant sind, jegliche Veränderungen in den Personenparametern ab.

Diese Annahme kann nun geprüft werden, indem, wie in Kapitel 1.1 beschrieben, die Geltung des Rasch-Modells für diesen Datensatz untersucht wird. Dabei dürfte das Teilungskriterium Testzeitpunkt hinsichtlich der Validitätsproblematik besonders trennscharf sein.

Kann die Konformität mit dem Rasch-Modell angenommen werden, so ist das ein Beleg dafür, dass sich die substantiellen Messeigenschaften des Testmodells nicht verändert haben. Veränderungen finden somit auf der Fähigkeitsdimension, welche der Test erfasst, statt.

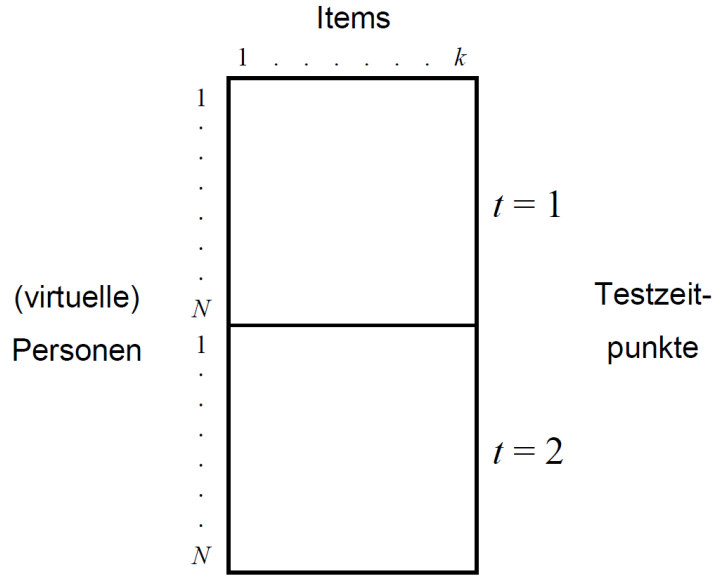


Abbildung 1. Datenmatrix zur Modellierung personenspezifischer Veränderung. Zu Testzeitpunkt $t = 2$ werden die N Personen so angeschrieben, als ob sie die k Items als „virtuelle“ neue Personen bearbeitet hätten.

Die geschätzten Personenparameter $\hat{\xi}_{vt}$ können nun weiteren Analysen unterzogen werden. Um zu prüfen, ob ein Retest-Effekt in statistisch signifikantem Ausmaß vorliegt, kann nun wie üblich mit einem t-Test für abhängige Stichproben untersucht werden.

Hinsichtlich der oben aufgezählten klassischen Probleme der Veränderungsmessung ist bei der Analyse der Differenzwerte nach Guthke und Wiedl (1996, S. 368-369) aber ebenso Vorsicht geboten.

Der Wiederholungsparameter kann aber auch in den Schwierigkeitsparameter integriert werden. Als Rasch-Modell angeschrieben wird $\xi_v - \sigma_i$ aus Formel 1 dabei zu $\xi_v - \sigma_{it}$, wobei σ_{it} die Schwierigkeit eines Items i zu Zeitpunkt t darstellt (vgl. Kubinger, 1979, 2008). In diesem Modell werden die Fähigkeiten der Personen über die Zeitpunkte hinweg als konstant angenommen. Die Schwierigkeiten der Aufgaben zu Zeitpunkt zwei verändern sich aber für alle Personen infolge der erneuten Bearbeitung des Tests.

In der entsprechenden Datenmatrix (siehe Abbildung 2) ist jedes Item aus dem zweiten Durchgang als virtuelles neues Item anzuschreiben, welches von der gleichen Person bearbeitet wurde (vgl. Fischer, 1974, 1995; Kubinger, 1979, 2008).

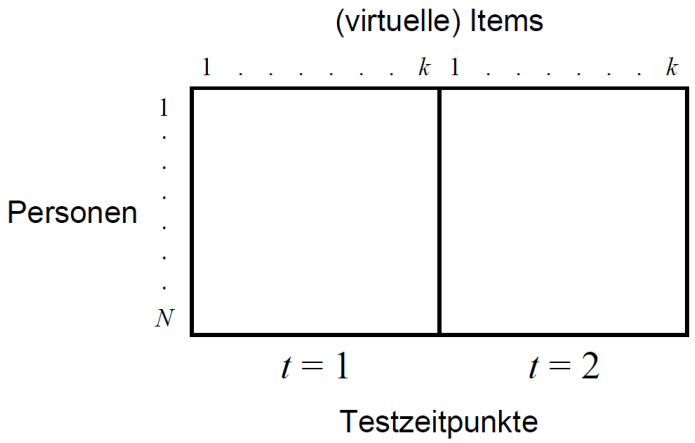


Abbildung 2. Datenmatrix zur Modellierung itemspezifischer Veränderung. Die k Items aus Testzeitpunkt $t = 2$ werden so angeschrieben, als ob sie von den N Personen als „virtuelle“ neue Items bearbeitet worden wären.

Da beim LRT nach Andersen das Teilungskriterium Testzeitpunkt in diesem Fall nicht verfügbar ist, kann zur Überprüfung der Modellgeltung zusätzlich der Martin-Löf-Test (Martin-Löf, 1973) als diesbezüglich sensitiver Test eingesetzt werden. Dieser prüft im Wesentlichen, ob zwei Testteile dieselbe latente Dimension erfassen.

Kann die Geltung des Rasch-Modells angenommen werden, so ist das ein Beleg dafür, dass alle (virtuellen) Items dieselbe latente Dimension auf einer Differenzenskala messen. Der Bezug zwischen den Testzeitpunkten geht für die Items zunächst verloren, kann aber durch entsprechende Annahmen in einem LLTM wieder hergestellt werden. So kann für weitere Analysen direkt in diesem Ansatz verblieben werden.

Im Kapitel 1.1 wurde bereits darauf eingegangen, wie das LLTM zur Untersuchung von Vorgabeeffekten eingesetzt werden kann. Das zuvor beschriebene Rasch-Modell dient dabei wie ein saturiertes Modell dazu, die Annahmen im LLTM mit Formel 3 zu prüfen.

Damit kann z.B. die einfache Annahme untersucht werden, ob sich die Schwierigkeiten der Aufgaben alle gleichermaßen um den Retest-Effekt λ zu Testzeitpunkt zwei verändert haben. Die $2k$ Strukturparameter σ_{it} des geltenden Rasch-Modells werden dabei durch Modell 1: $\sigma_i + q_t\lambda$ mit $k + 1$ Basisparametern ersetzt, wobei σ_i die Ausgangsschwierigkeit der Items darstellt und q_t zu Zeitpunkt eins gleich null und zu Zeitpunkt zwei gleich eins ist. Bei der zu erwartenden Erleichterung der Aufgaben müsste ein negativer Wert für λ resultieren. Eine nicht-

signifikante Prüfgröße aus Formel 3 würde für die Annahmen in Modell 1 sprechen.

Zur Überprüfung der Alternativhypothese, dass $\lambda \neq 0$ ist, muss aber noch ein weiteres LLTM unter der Annahme der $H_0: \lambda = 0$ erstellt werden. Modell 2 nimmt lediglich an, dass die Itemschwierigkeiten zu beiden Zeitpunkten gleich sind: $\sigma_{i1} = \sigma_{i2} = \sigma_i$. Spricht ein signifikanter (bedingter) LRT gemäß Formel 3 gegen die weitere Reduktion auf k Strukturparameter, wäre die H_0 abzulehnen und das sparsamere Modell 1 anzunehmen.

Erklärt Modell 2 die Daten aus dem zugrunde gelegten Rasch-Modell hingegen statistisch genauso gut, so kann mit einem (bedingten) LRT gemäß Formel 4 die H_0 spezifisch getestet werden. Ein signifikantes Ergebnis würde in diesem Fall wieder auf die Annahmen unter Modell 1 hindeuten.

Zu den Vorzügen der dargestellten Methoden sei noch erwähnt, dass die Größe des geschätzten Retest-Effektes $\hat{\lambda}$ auch unabhängig von Boden- oder Deckeneffekten ist, da die Stichprobenzusammenstellung zur Schätzung der Strukturparameter im Rasch-Modell und im LLTM keine Rolle spielt (siehe Kapitel 1.1). Zudem liegt λ auf einer absoluten Skala (Fischer, 1974, S. 384).

2.4 Retest-Effekte bei Tests mit Zahlenreihen

Wenngleich Tests mit Zahlenreihen immer wieder als Bestandteil von Testbatterien in der Literatur zu Retest-Effekten Verwendung fanden, finden sich kaum standardisierte Effektgrößen zu deren Ausmaß bzw. die nötigen Angaben, um diese zu errechnen. Die Stichproben, auf denen die im Folgenden genannten Resultate basieren, wurden alle im wissenschaftlichen Kontext erhoben.

Nach der Handanweisung des *Zahlenfolgentests* (ZF; Weiß, 1987) zeigten sich bei der zweimaligen Vorgabe der identen Form einer Erprobungsfassung in zwei Untersuchungen keine statistisch signifikanten Übungseffekte. Die Retestungsintervalle betrugen zwei bis drei Monate für Grundschüler und -schülerinnen ($N = 42$) und viereinhalb Monate bei Hauptschülern bzw. -schülerinnen ($N = 34$). Analysiert

wurde jeweils die mittlere Differenz der Test-Scores der Testpersonen mithilfe eines t-Tests für abhängige Stichproben.

Zahlenreihen werden in der Psychologischen Diagnostik meist zur Messung von Reasoning oder verwandten Konstrukten eingesetzt. Bei der Konstruktion des ZF stand aber vor allem im Vordergrund, dass Zahlenreihen immer auch die Rechenfähigkeit und den Umgang mit Zahlen miterfassen (vgl. Weiß, 1987). Da diese Aspekte förderabhängig sind, kann der ZF in Verbindung mit anderen Verfahren (siehe ebenfalls Weiß, 1987) differential-diagnostisch relevante Informationen liefern. Die Bearbeitung erfolgt unter einer von zwei wählbaren Zeitbegrenzungen, eine Speed-Komponente ist somit vorhanden.

Die von Reeve und Lam (2005, 2007) verwendeten Daten wurden anhand von sechs Subskalen der *Employee Aptitude Survey* (EAS; Ruch, Stang, McKillip & Dye, 1994) erhoben. Dabei wurden mit dem Subtest „numerical reasoning“ auch Zahlenreihen eingesetzt. Die EAS-Auswahl wurde von Reeve und Lam (2005, 2007) dreimal, in identer Form, mit je einer Woche Abstand einer Gruppe von Studierenden vorgegeben. Bei den Zahlenreihen ergab sich ein statistisch signifikanter Anstieg erst im zweiten Retest, wobei die Effektgröße $0{,}5$ betrug.

Wesentlich für einen Vergleich mit Tests mit ähnlichen Aufgaben ist sicher das der EAS zugrundegelegte Entwicklungskonzept: „Maximum validity per minute of testing time“ (Ruch et al., 1994, S. 3), weshalb von einer deutlichen Speed-Komponente ausgegangen werden kann. Aus ihren Analysen, basierend auf den Antworten von 123 Studierenden, folgerten Reeve und Lam (2005), dass die latenten Fähigkeiten der EAS-Subtests in den einzelnen Durchgängen messtechnisch invariant erfasst wurden. Mit den genutzten Methoden kann man sich diesem Problem, insbesondere mit dieser Stichprobe, aber bestenfalls nähern (vgl. Kapitel 2.3).

In der aktuelleren Untersuchung von Arendasy und Sommer (2013) wurden ebenfalls Zahlenreihen eingesetzt, und zwar im Rasch-Modell-konformen Subtest „numerical inductive reasoning“ (NID), welcher als *Power*-Test vorgegeben wurde. Im Retest kamen drei Testformen zum Einsatz: (1) mit identen, (2) isomorphen und (3) psychometrisch parallelisierten Items, verglichen mit den Aufgaben des Initialtests. Die isomorphen Items waren nach derselben Regelstruktur wie jene im Initi-

altest aufgebaut, wobei in Vorstudien die psychometrische Äquivalenz belegt werden konnte. Die psychometrisch parallelisierten Items teilten sich mit jenen der Initialtestung lediglich die psychometrischen Eigenschaften, also die Item-Schwierigkeitsparameter.

Auf Basis des Rasch-Modells (vgl. Kapitel 2.3.1) konnte nachgewiesen werden, dass nach einem Retestungsintervall von rund zweieinhalb Wochen dieselbe latente Dimension erfasst wurde. Die Stichprobe umfasste 358 Personen. Bei allen Testformen zeigte sich ein statistisch signifikanter Anstieg, mit Ausnahme der Gruppe von Personen mit geringem g in dem psychometrisch parallelisierten Retest. Bei den identen und isomorphen Testformen fiel das Ausmaß der Retest-Effekte generell höher aus als bei den psychometrisch parallelisierten. Effektgrößen, die einen ungefähren Vergleich mit bisher genannten Maßen erlauben, werden bei Arendasy und Sommer (2013) nicht genannt.

III Empirischer Teil

3 Ziel der Untersuchung und Hypothesen

Nach den eingangs erwähnten Richtlinien möchten potentielle Anwender wissen: (1) ob ein psychologisch-diagnostisches Verfahren überhaupt für einen wiederholten Einsatz geeignet ist und (2) ob bzw. mit welchen Veränderungen zu rechnen ist. Ziel dieser Untersuchung ist es, bei zweimaliger Vorgabe erste Befunde zu diesen Fragen für den ZART zu sammeln.

Kommt es infolge der Retestung zu substantiellen Veränderungen der Messeigenschaften, so wäre eine Eignung für diesen Einsatz nicht gegeben. Wie in Kapitel 2.2 dargestellt, sind generell für Tests zur Ermittlung kognitiver Fähigkeiten die Befunde dazu widersprüchlich. Hinsichtlich Zahlenreihen deuten die Ergebnisse von Reeve und Lam (2005) auf keine substantiellen Veränderungen hin. Methodische Schwächen relativieren diese Aussage allerdings[12].

Es gilt zu klären, ob bei erneuter Durchführung des ZART noch dieselbe latente Dimension wie in der Initialtestung erfasst wird. Die Nullhypothese dazu lautet:

H_0^1: Wenn der ZART denselben Testpersonen ein zweites Mal vorgegeben wird, erfassen die Items zu beiden Zeitpunkten dieselbe latente Dimension.

Sofern H_0^1 nicht abzulehnen ist, kann der zweite Teil der obigen Hauptfragestellung bearbeitet werden. In der Literatur fanden sich für Tests mit Zahlenreihen Anstiege in den Testwerten bis zur Hälfte der Standardabweichung im Initialtest (siehe Kapitel 2.4). Ein Vergleich mit dem ZART ist aufgrund der unterschiedlichen Zielsetzungen jedoch nicht sinnvoll.

Die vorliegende Untersuchung möchte einen Schritt weiter gehen, als bloß das durchschnittliche Ausmaß eventueller Verbesserungen der Testwerte zu errechnen, und den Retest-Effekt, wie oben hinsichtlich itemspezifischer Veränderung dargestellt, im LLTM modellieren. Durch die Unabhängigkeit des Wiederholungs-

[12] Die Studie von Arendasy und Sommer (2013) wurde nach den Erhebungen für diese Untersuchung publiziert und war für die Planung somit nicht relevant.

parameters von der Stichprobenzusammensetzung könnte dieser dann zumindest bei ähnlichen Bedingungen als Richtwert dienen und vor allem auch bei einer anderen Itemauswahl des ZART angewendet werden. Die entsprechende Nullhypothese lautet:

H_0^2: Wenn der ZART denselben Testpersonen ein zweites Mal vorgegeben wird, tritt kein Retest-Effekt auf.

Falls ein Effekt vorhanden ist, wird erwartet, dass sich die Testpersonen in ihrer Fähigkeitsdimension durch die vorangegangene Bearbeitung verbessern (personenspezifische Veränderung) bzw. dass die Items in der Retestung für die Personen gleichermaßen leichter werden (itemspezifische Veränderung).

In einer Nebenfragestellung soll in diesem Zusammenhang untersucht werden, ob durch das Antwortformat der Aufgaben im ZART, konkret durch das Multiple-Choice-Format (1 aus 6) und das freie Antwortformat, das Ausmaß eventueller Retest-Effekte verringert werden kann. Zum Einfluss des Antwortformates konnte lediglich die Studie von Powers (1986) gefunden werden, wobei hier mehrere Variablen vermengt waren (siehe Kapitel 2.1). Die Nullhypothese lautet:

H_0^3: Das Ausmaß der Retest-Effekte ist unabhängig vom Antwortformat der Items des ZART.

Die entsprechende Alternativhypothese dazu ist ungerichtet. Denn einerseits könnte das offene Antwortformat zu einer elaborierteren Verarbeitung und dadurch zu einem größeren Übungseffekt führen. Andererseits könnten Antwortmöglichkeiten im Multiple-Choice-Format kleine Hilfestellungen bieten, welche Lernerfahrungen anregen. Ein eventueller Rateeffekt bestünde in beiden Durchgängen und hat auf einen Retest-Effekt vermutlich keinen Einfluss.

Da die Aufgaben des ZART grundsätzlich ohne Zeitbegrenzung vorgegeben werden, könnte sich die Bearbeitungszeit der Items in den Durchgängen ebenfalls ändern. Es wird angenommen, dass die Testpersonen, infolge einer effizienteren Bearbeitung, in der Retestung weniger Zeit zum Bearbeiten der Items benötigen. Die Nullhypothese dazu lautet:

H_0^4: Die Testpersonen benötigen zu beiden Zeitpunkten gleich lange, die Items des ZART zu bearbeiten.

Da eine Erhebung an Schulen geplant ist, soll auch ein erster Befund zur Kriteriumsvalidität des ZART erhoben werden und zwar der Zusammenhang mit der Leistung im Fach Mathematik. Auch wenn vor allem das deskriptive Maß interessiert, soll der Vollständigkeit halber auch hier die Nullhypothese wiedergegeben werden, wobei von einem negativen Zusammenhang ausgegangen wird:

H_0^5: Zwischen den Leistungen im ZART und der Note im Fach Mathematik besteht kein Zusammenhang.

4 Methode

4.1 Untersuchungsplan

Zunächst werden die groben Rahmenbedingungen der Untersuchung angesprochen, die auch Überlegungen zur Stichprobenakquirierung enthalten. Darauf aufbauend erfolgt eine konkretere Festlegung der Konfiguration des Erhebungsinstrumentes im Zusammenhang mit organisatorischen Überlegungen. Im Anschluss wird grob auf die geplante Auswertung eingegangen. Zuletzt wird der nötige Stichprobenumfang in Bezug auf die Genauigkeitsanforderungen der Untersuchung (vgl. Kubinger, Rasch & Yanagida, 2011, S. 196-207) festgelegt.

Nachdem bei der ersten bis zur vierten Hypothese die Initialtestung selbst das Treatment darstellt, ist deren Untersuchung im Rahmen eines Messwiederholungsdesigns für zwei Zeitpunkte notwendig. Für die fünfte Hypothese reichen die Daten der Initialtestung.

Zunächst sind die Rahmenbedingungen der Untersuchung festzulegen. Für die Anwendung des ZART wäre es natürlich am günstigsten, wenn die Nullhypothesen eins und zwei beibehalten werden können, wenn also die erfasste Dimension zu den beiden Zeitpunkten dieselbe wäre und keine Retest-Effekte vorhanden sein würden. Die Untersuchungsbedingungen sollen derart gestaltet werden, es diesem „Wunsch“ möglichst zu erschweren und eventuelle Retest-Effekte, im Rahmen der Möglichkeiten, zu maximieren. Bedingungen die, entsprechend der Literatur, einen Einfluss haben könnten, sind das Retestungsintervall, ob idente

oder parallele Testformen verwendet werden und die Motivation der Testpersonen.

Zu beiden Durchgängen sollen idente Items des ZART vorgegeben werden und das Retestungsintervall soll etwa zwei Wochen betragen. Kürzere Retestungsintervalle werden in praktischen Anwendungen kaum von Relevanz sein. Der geringe Abstand könnte natürlich Erinnerungseffekte provozieren, wie sie z.B. bei Lievens et al. (2007) diskutiert werden. Für den ZART wird aber nicht angenommen, dass nach zwei Wochen Erinnerungen an eine Antwort eine Rolle bei der Bearbeitung spielen. Bei der Untersuchung von H_0^1 müssten derartige Effekte allerdings auffallen.

Die Vorgabe in einem praktischen Kontext schien leider nicht realisierbar zu sein. Deshalb sollen, wie bereits angesprochen, Schüler und Schülerinnen ab der zehnten Schulstufe für die Untersuchung akquiriert werden. In dieser Schulstufe ist der Umgang mit negativen Zahlen bereits zur Gewohnheit geworden. Zur Förderung der Motivation soll hervorgehoben werden, dass die Testungen eine gute Lernerfahrung für künftige Eignungstests sind. Zudem soll eine individuelle Rückmeldung der Testergebnisse angeboten werden.

Zu jedem Zeitpunkt wird dieselbe Auswahl an 12 Items des ZART vorgegeben. Die Anzahl ist bewusst etwas geringer gehalten, da für einen Durchgang vermutlich nur eine Schulstunde, also 50 Minuten, zur Verfügung stehen wird. Zudem ist mit Abschlägen für Organisatorisches, Vorstellung und den gemeinsamen Einstieg in die Testsoftware zu rechnen. Wichtig ist auch, dass zum Ende der Testung kein Zeitdruck entsteht, da beim ZART keine Zeitbegrenzung vorgesehen ist. Deshalb soll nach der Bearbeitung der 12 Items noch Zeit übrig bleiben.

Derzeit führt Herr Mag. Poinstingl eine Kalibrierungsstudie für den ZART durch. Die Auswahl der Items sollte also bereits an einem Rasch-Modell-konformen Itempool erfolgen können. Die Itemstichprobe soll zwei Gesichtspunkten genügen. Die Itemparameter sollen einerseits breit streuen, andererseits sollen aber keine zu schwierigen Aufgaben vorgegeben werden, um Frustration möglichst zu vermeiden.

Die Vorgabe der Aufgaben erfolgt in Blöcken, ein Block mit sechs Items im Multiple-Choice-Antwortformat (1 aus 6) und ein Block mit sechs Items im freien Antwortformat. Beide Blöcke werden in beiden Durchgängen inhaltlich ident vor-

gegeben. Bezüglich der unabhängigen Variable (UV) Testzeitpunkt liegt dadurch eine abhängige (Item-)Stichprobe vor, für die UV Antwortformat eine unabhängige.

Die Reihenfolge der beiden Blöcke soll in beiden Durchgängen zufällig variieren. Vorrangig deshalb, um den Testpersonen ein Abschreiben zu erschweren, aber auch um eventuelle Effekte der Reihenfolge auszugleichen.

Auch die Position der Items innerhalb der Blöcke soll im zweiten Durchgang verändert werden, um keine Erinnerungseffekte aufgrund der Abfolge zu provozieren.

Um alle Testpersonen ausreichend lange zu beschäftigen und somit eine störungsfreie Testsituation zu erreichen, wird in beiden Durchgängen, immer nach den 12 Items, ein dritter Block mit weiteren Items aus dem Itempool des ZART vorgegeben. Dieser ist für die Untersuchung der genannten Hypothesen aber nicht weiter von Relevanz.

Für die Hypothesen eins bis drei ist geplant, dass sich die Auswertung an dem orientiert, was in Kapitel 2.3.1 hinsichtlich itemspezifischer Veränderung dargestellt wurde. Sofern die Voraussetzungen nicht verletzt sind, erfolgt die Analyse somit anhand des dichotomen logistischen Testmodells von Rasch (Rasch-Modell; 1960) und des linearen logistischen Testmodells von Fischer (LLTM; 1973). In diesem Ansatz kann auch der Einfluss des Antwortformates modelliert werden.

Neben der Modellierung von Retest-Effekten im LLTM soll bei signifikanten Ergebnissen auch die durchschnittliche Veränderung der Personenparameter bei konstanten Itemparametern bei beiden Durchgängen geschätzt werden. Die durchschnittliche personenspezifische Veränderung wäre aufgrund der Modelleigenschaften betragsmäßig gleich groß wie ein Wiederholungseffekt bei der itemspezifischen Modellierung. Lediglich das Vorzeichen ändert sich. Dazu kann auch das auf die zu erhebende Stichprobe bezogene eingegangene Risiko 2. Art angegeben werden.

Zunächst ist aber für Testzeitpunkt eins festzustellen, ob für die verwendeten ZART-Items die Geltung des Rasch-Modells angenommen werden kann. Das ist notwendig, da teilweise ein anderes Antwortformat als in der laufenden Kalibrierungsstudie verwendet werden wird und eben die weitere Vorgehensweise bei der Auswertung davon abhängt.

Die Auswertung zu den restlichen Hypothesen wird mit konventionellen statistischen Verfahren vorgenommen, wobei aufgrund der Skaleneigenschaften der abhängigen Variable Testzeit (H_0^4) und der UV Mathematik-Note (H_0^5) verteilungsfreie Verfahren Anwendung finden werden.

Eine Planung der Stichprobengröße in Bezug auf die praktische Relevanz von Effekten kann nicht direkt für die Berechnungen im LLTM stattfinden. Wie in Kapitel 2.3.1 dargestellt, ist für den allgemeinen Retest-Effekt in der Hauptfragestellung der t-Test für abhängige Stichproben das zum LLTM äquivalente Verfahren bei personenspezifischer Veränderung. Daran kann eine Abschätzung der erforderlichen Stichprobengröße erfolgen.

In den geplanten Modellierungsversuchen können einseitige Alternativhypothesen (H_1^2) und Mindestgrößen für Effekte aber nicht integriert werden. Die Beurteilung der Relevanz modellierter Parameter kann erst im Nachhinein erfolgen. Ab welchem Wert ein Modellparameter als praktisch relevant gilt, soll zumindest für den allgemeinen Retest-Effekt in der Hauptfragestellung im Folgenden festgelegt werden. Darauf aufbauend wird dann die nötige Stichprobengröße ermittelt, um einen entsprechenden Effekt hinsichtlich personenspezifischer Veränderung, bei noch festzulegenden Risiken 1. und 2. Art und einseitiger Fragestellung, mit der zu erhebenden Stichprobe auffinden zu können.

Die Beurteilung der praktischen Relevanz eines Retest-Effektes geschieht, aufgrund der besseren Anschaulichkeit, anhand der an der Standardabweichung in der Initialtestung relativierten Effektgröße hinsichtlich personenspezifischer Veränderung.

Grundsätzlich kann eine Beurteilung der praktischen Relevanz nur in Bezug auf die jeweilige Anwendung erfolgen. Bei einer Verlaufskontrolle beispielsweise wären standardisierte Effekte selbst bei einem hohen Reliabilitätskoeffizienten von $0{,}95$ und einseitig gerechneten Konfidenzintervallen ($\alpha = 0{,}05$) erst bei einem Betrag größer $0{,}736$ als praktisch relevant einzustufen, da Abweichungen bis zu diesem Wert auch auf Messfehler zurückzuführen wären. Bei Aufnahmeverfahren hingegen könnten auch geringere Effekte zu Veränderungen der Rangordnung führen, wodurch die Gruppe derer, die zum ersten Mal teilnehmen, systematisch benachteiligt werden würde. Mit Hinblick auf die genannten Rahmenbedingungen

der Untersuchung dürften relative Effektgrößen kleiner einem Betrag von $0{,}2$ aber praktisch uninteressant sein. Das würde einem durchschnittlichen Anstieg der Testwerte an Testzeitpunkt zwei von nur 2 T-Werten entsprechen.

Da zum ZART noch keine Normierungsuntersuchungen vorliegen, ist eine Abschätzung der Standardabweichung σ der Personenparameter nötig, um die praktisch relevante Mindestgröße für den absoluten Retest-Effekt zu ermitteln. Dazu ist lediglich die Standardabweichung von $1{,}11$ aus der Vortestung (Poinstingl, 2009d, 2009e) bekannt. Die Stichprobe in der Vortestung war sehr homogen in der Zusammensetzung, was der Schätzung einer „unteren Schranke“ für die relevante Mindestgröße aber entgegenkommt. Als „ungünstiger“ Wert für σ (siehe Kubinger et al., 2011, S. 201) wird somit $1{,}11$ angenommen. Das entspricht einer durchschnittlichen Verbesserung der Testpersonen an Testzeitpunkt zwei von zumindest $0{,}222$ bei personenspezifischer Veränderung bzw. einer Verringerung der Schwierigkeit der Items an Testzeitpunkt zwei bei itemspezifischer Veränderung um denselben Betrag.

Für den t-Test für abhängige Stichproben ist zur Berechnung der erforderlichen Stichprobengröße eine Schätzung der Standardabweichung der Differenzwerte σ_D der Personenparameter an beiden Zeitpunkten über Formel 5 (vgl. Erdfelder, Lang & Buchner, 2007, S. 182) nötig. Gemäß Kubinger et al. (2011, S. 201) soll wiederum ein ungünstiges Maß für σ_D angenommen werden. Insofern wird der Zusammenhang der Testergebnisse an beiden Zeitpunkten ρ mit $0{,}7$ geschätzt.

$$\sigma_D = \sigma * \sqrt{2 * (1 - \rho)} \tag{5}$$

Die Auswirkungen eines Fehlers 1. und 2. Art werden für einen allgemeinen Retest-Effekt als gleichermaßen ungünstig angenommen, da einerseits bei der Korrektur von (in Wahrheit nicht vorhandenen) Retest-Effekten diejenigen benachteiligt werden würden, die wiederholt antreten. Bei einem Fehler 2. Art andererseits würden Korrekturen ausbleiben und jene, die zum ersten Mal antreten, benachteiligt werden. Insofern wird die Irrtumswahrscheinlichkeit α mit $0{,}05$ und die Teststärke $1 - \beta$ mit $0{,}95$ festgelegt.

Um einen absoluten Effekt in der Größe von $0{,}222$, unter den Annahmen für ρ und σ, bei einseitiger Alternativhypothese und mit den genannten Risiken entde-

cken zu können, wird eine Stichprobe im Ausmaß von 164 Personen benötigt. Die Berechnung erfolgte mit der Freeware G*Power, Version 3.1 (Faul et al., 2007).

Um eine entsprechende Genauigkeit der Parameterschätzungen im Rasch-Modell und im LLTM zu garantieren, sollten zur Analyse aber zumindest Daten von 200 Testpersonen zu Verfügung stehen (vgl. Kubinger, Rasch & Yanagida, 2009, S. 371). Deswegen soll für die Auswertung eine Stichprobe im Umfang von mindestens 200 Personen vorhanden sein.

Es ist davon auszugehen, dass einige Schüler und Schülerinnen an einem der Testzeitpunkte nicht anwesend sein werden und auch, dass die eingesetzten Motivatoren bei manchen nicht reichen werden, die entsprechende Anstrengungsbereitschaft für eine instruktionskonforme und „reale“ Bearbeitung der Aufgaben aufzubringen. Auch eine potenzielle Verweigerung der Teilnahme seitens der Eltern oder der Testpersonen ist zu berücksichtigen. Da die Zielsetzungen der Untersuchung nur mit den Daten von beiden Durchgängen bearbeitet werden können, ist also mit einem erhöhten Ausschuss zu rechnen. Insofern soll an den Schulen für die Testung von zumindest 300 Schülern bzw. Schülerinnen angefragt werden.

4.2 Erhebungsinstrument

Allgemeine Informationen zum ZART (Poinstingl, 2009c) wurden bereits in Kapitel 1 wiedergegeben. In diesem Kapitel wird die konkrete Konfiguration für diese Untersuchung dargestellt, wobei soweit nötig auch auf technische Aspekte eingegangen werden wird. Zudem wird beschrieben, welche Begleitinformationen erhoben werden.

Gemäß den Vorgaben aus der Planung wurden zunächst die 12 auswertungsrelevanten Items, die nach den Auswertungen von Herrn Mag. Poinstingl im Rahmen der Kalibrierungsstudie als Rasch-Modell-konform gelten, für die beiden ersten Blöcke ausgewählt. Für den dritten Block wurden 13 weitere Items ausgewählt. Diese werden ebenfalls im Multiple-Choice-Format (1 aus 6) vorgegeben. Für den zweiten Durchgang wird die Reihenfolge der Aufgaben in Block drei umgekehrt. Die konkreten Positionen der Items sind in Tabelle 1 ersichtlich. Darin ist auch die

Itemnummer aus der Kalibrierung angegeben, um einen Bezug herstellen zu können, sowie auch die Nummerierung für die Ergebnisdarstellung.

Tabelle 1

Position der ZART-Items innerhalb der Vorgabeblöcke zu den Testzeitpunkten (TZP), Nummerierung der Items für die Ergebnisdarstellung und Nummer der Items bei der Kalibrierung

Itemnummer	Nr.-Kalibrierung	Block	Position TZP1	Position TZP2
I1	1	MC	1	4
I2	77	MC	2	2
I3	24	MC	3	5
I4	46	MC	4	1
I5	41	MC	5	6
I6	94	MC	6	3
I7	23	AF	1	4
I8	75	AF	2	2
I9	18	AF	3	5
I10	43	AF	4	1
I11	54	AF	5	6
I12	29	AF	6	3
	79, 50, 98, 92, 105, 70, 42, 60, 16, 32, 30, 66, 39	3	1-13	13-1

Anmerkungen. Die Reihenfolge der Blöcke mit den Aufgaben im Multiple-Choice-Antwortformat (MC) und im freien Antwortformat (AF) variiert zu beiden Testzeitpunkten. Der dritte Block wird immer als letzter vorgegeben, vorausgesetzt eine Testperson kommt soweit.

Die Vorgabe dieser ZART-Version erfolgt mit der testeigenen Software (Poinstingl, 2009b). Eine Installation auf den Schulrechnern wäre aus organisatorischen und Datenschutzgründen problematisch. Deshalb soll das sogenannte *Remote Testing* (siehe dazu Poinstingl, 2009d, 2009e) zum Einsatz kommen. Dabei befindet sich die Testsoftware auf einem Server, auf dem auch die Testbearbeitung und schließlich die Speicherung durchgeführt werden. Die Rechner in den PC-Räumen dienen lediglich als Eingabemedien. Die Verbindung zum Server wird über eine Remotedesktopverbindung hergestellt. Das nötige gleichnamige Programm ist in allen gängigen Versionen von Microsoft Windows enthalten.

Neben einem funktionierenden Netzwerk samt Internetanbindung seitens der Schulen ist somit ein leistungsfähiger Server notwendig, auf dem die Testungen stattfinden können. Für letzteres stehen drei Server des Arbeitsbereiches Psychologische Diagnostik der Fakultät für Psychologie zur Verfügung, auf welche die Testpersonen, bei entsprechend großer Anzahl, gleichzeitig aufgeteilt werden können. Bei technischen Problemen wird das mobile Mehrplatz-Testsystem des Arbeitsbereiches und der Firma Schuhfried zum Einsatz kommen. Dieses besteht aus 15 Notebooks, auf denen die ZART-Software in der gleichen Konfiguration lokal installiert wird.

Die Variation der Reihenfolge der ersten beiden Blöcke wird über die Software des ZART erreicht. Dabei wird mit jeder Anmeldung an einem Testserver oder bei jedem Start der Software auf einem Laptop abwechselnd eine von zwei Testformen gestartet. Eine Testform startet mit dem Block mit den Aufgaben im Multiple-Choice-Format (MC), die andere mit den Aufgaben im freien Antwortformat (AF).

Der Ablauf gestaltet sich bei einer Testung folgendermaßen. Nach Anmeldung auf einem der Server bzw. nach Start der Software am Laptop öffnet sich zunächst ein Datenformular, über welches personenspezifische Informationen erfragt werden.

Zu beiden Testzeitpunkten werden die Testpersonen nach ihrem Alter, ihrer Muttersprache, ihrem Geschlecht, einer Identifikation und einem Passwort gefragt. Bei der Identifikation handelt es sich um eine Kombination aus Buchstaben und Zahlen, welche sich aus personenspezifischen Informationen zusammensetzt, konkret: (1) den ersten drei Buchstaben des Vornamens der Mutter, (2) dem Geburtsmonat (zweistellig) und (3) den ersten drei Buchstaben der Straße bzw. der Anschrift der Testperson. Die Identifikation dient der Wahrung der Anonymität der Schülerinnen und Schüler sowie der Zuordnung der Testergebnisse von beiden Durchgängen. Gemeinsam mit dem selbst gewählten Passwort wird es nach den Testungen für die Testpersonen zudem möglich sein, ihre Testergebnisse individuell über eine Webseite abzurufen.

Zu Testzeitpunkt eins wird zusätzlich abgefragt, ob die Testpersonen bereits an Testungen an einem Computer teilgenommen haben und ob sie bereits einen Test oder eine Prüfung bearbeitet haben, in dem das Multiple-Choice-Format verwendet wurde. Diese Variablen können eventuell als weitere Teilungskriterien zur Prüfung des Testmodells dienen.

Im zweiten Durchgang werden die Testpersonen nach ihrer letzten Jahresnote im Fach Mathematik gefragt und, lediglich als Hilfsvariable für das Zusammenfügen der Daten, ob sie bereits an der ersten Testung teilgenommen haben.

Da das Datenformular sehr wenig Platz zum Formulieren der Fragen bietet, müssen diese im Vorfeld mit den Schülern genau besprochen werden.

Nach dem Formular folgen wie beschrieben die drei Vorgabeblöcke. Jeder Block wird mit einer antwortspezifischen Instruktion eingeleitet, welche auch zwei Beispielitems enthält. Vor den Testaufgaben ist dann jeweils noch eine Übungsaufgabe zu bearbeiten.

4.3 Durchführung der Untersuchung

Zunächst wird auf die Akquirierung der Schulen sowie der Schülerinnen und Schuler eingegangen. Dann wird der Ablauf der Untersuchung beschrieben, wobei auch Probleme bei der Durchführung angesprochen werden.

Im Oktober 2009 wurde mit der Akquirierung höherer Schulen in Niederösterreich begonnen. Nach einer telefonischen Kontaktaufnahme erhielten interessierte Direktorinnen und Direktoren schriftliche Informationen über die Fragestellungen der Untersuchung, den ZART und die beabsichtigte Testung. Schlussendlich konnte von sechs Schulen eine Zusage zur Testung von insgesamt 19 Klassen mit ungefähr 360 Schülern und Schülerinnen erhalten werden. Für diese sechs Schulen wurde beim Landesschultrat für Niederösterreich um eine Genehmigung zur Durchführung der Untersuchung angefragt. Der entsprechende Bescheid befindet sich im Anhang (siehe Abbildung 16).

In persönlichen Treffen mit den Direktorinnen und Direktoren und den jeweiligen Schuladministratorinnen und –administratoren wurden dann die organisatorischen Details besprochen. Dabei wurden auch Briefe für die Eltern verteilt (siehe Abbildung 17), welche Informationen zur Testung und die nötige Einverständniserklärung zur Teilnahme an der Untersuchung enthielten. Für die Klassenvorstände wurden ebenfalls Informationsschreiben aufgelegt.

Es stellte sich heraus, dass an fünf der sechs Schulen die Netzwerkeinstellungen ein Remote Testing nicht zuließen. Vier Schulen ließen entsprechende Änderungen durchführen. In einer Schule war dies aufgrund von Sicherheitsbedenken

nicht möglich, hier kamen von vornherein das mobile Mehrplatz-Testsystem und ein weiterer Laptop zum Einsatz.

Die Testungen fanden vom 25. Jänner bis zum 4. März 2010 statt. Die Termine konnten so eingeteilt werden, dass das Retestungsintervall für jede Testperson genau zwei Wochen betrug. Die Testpersonen wurden in Gruppen zwischen drei und 22 Personen getestet, im Schnitt waren es 14. Manchmal wurden aus zeitlichen Gründen zwei Gruppen parallel getestet. Waren zu wenige Rechner in den Computerräumen vorhanden, konnte die Infrastruktur flexibel mit den Notebooks erweitert werden.

Bei jeder Gruppentestung waren zwei Testleiter bzw. -leiterinnen anwesend. Lehrpersonal war nur bei einer Klasse gegenwärtig. Als Testleiter bzw. -leiterinnen wurden neben meiner Person Diplomanden und Diplomandinnen des Arbeitsbereiches Psychologische Diagnostik eingesetzt. Im Vorfeld erhielten alle Testleiterinnen und -leiter schriftlich alle relevanten Informationen sowie einen halbstrukturierten Leitfaden zur Durchführung der Testungen. Bei der gemeinsamen Anreise zu den Schulen wurde noch einmal alles besprochen und eventuelle Fragen geklärt. Um die Erläuterungen zum Datenformular und den Einstieg zu den Testservern zu veranschaulichen, wurde eine digitale Präsentation vorbereitet. Als Backup lagen die gleichen Informationen auch auf großen Papierbögen vor.

Zu den Testungen wurden nur jene minderjährigen Schüler und Schülerinnen zugelassen, von denen eine unterschriebene Einverständniserklärung der Eltern oder eines bzw. einer Erziehungsberechtigten vorlag. Die entsprechenden Abschnitte aus den Elternbriefen wurden von den Klassenvorständen eingesammelt und verwaltet. Von den an den beiden Zeitpunkten anwesenden Schülerinnen und Schülern sind aber lediglich vier Fälle bekannt, die keine Einverständniserklärung hatten. Diese wurden einstweilen in anderen Klassen untergebracht. Informationen zu dieser Gruppe waren nicht zugänglich, weswegen eine weitere Analyse nicht möglich war. Aufgrund der schlussendlich geringen Gruppengröße wäre eine Analyse dieser Non-Responder auch nicht aussagekräftig.

Nach der Vorstellung seitens der Testleiter bzw. -leiterinnen wurde nochmals auf die Freiwilligkeit der Teilnahme, die weitestgehende Anonymität und die Vertraulichkeit der Testung hingewiesen (durch die Rückmeldung der Testergebnisse

und den Bezug zwischen den beiden Durchgängen war eine kodierte Zuordnung notwendig). Ein Schüler machte von seinem Recht Gebrauch und lehnte die Testung ab.

Dann wurde ausführlich das Datenformular besprochen. Um ihre Identifikation und ihr Passwort notieren zu können und die Identifikation im Vorfeld schon einmal herzuleiten, wurden den Testpersonen Handzettel zur Verfügung gestellt. Die Schüler und Schülerinnen wurden darauf hingewiesen, dass sie für die Bearbeitung des Tests ausreichend Zeit haben, der Test aus drei Testteilen besteht und sie keine Notizen machen dürfen. Im Anschluss wurde der jeweilige Einstieg in die Testsoftware erläutert.

Vor Ende der Schulstunde wurde den Testpersonen mitgeteilt, dass es keinen Einfluss auf ihr Testergebnis hat, wenn sie nicht fertig geworden sind, um ein Durchklicken gegen Ende des Tests zu verhindern. Nicht beendete Testungen wurden beim Läuten der Glocke seitens der Testleiterinnen bzw. –leiter geschlossen.

Von der technischen Seite her liefen die Testungen weitgehend stabil. Selten musste sich eine Testperson ein zweites Mal am Testserver anmelden, da aufgrund der höheren Anzahl an gleichzeitigen Anmeldungen eine Fehlermeldung erschien. Eine Testperson flog zu Beginn der Testung aus dem System, weswegen ebenfalls von Neuem gestartet werden musste. Zudem stürzte einmal ein Notebook ab, wobei aber lediglich Daten aus Block drei verloren gingen. Bei einer Schule war im zweiten Durchgang keine Verbindung zum Internet verfügbar, was durch die geliehenen Notebooks aber gelöst werden konnte.

Eine Klasse war zu Zeitpunkt zwei überraschenderweise auf einem Ausflug. Etwa die Hälfte einer anderen Klasse fiel aufgrund von einem Projekt weg.

Nachdem die genannten Probleme keine weiteren Auswirkungen nach sich ziehen und nur die schlussendliche Stichprobengröße verringern, haben die folgenden Vorkommnisse Auswirkungen auf die weitere Vorgehensweise oder die Gültigkeit einzelner Häufigkeiten.

Hinsichtlich der Frage, ob bereits an Testungen am Computer teilgenommen wurde, bestand, entsprechend den aufgekommenen Fragen in manchen Gruppen,

Unklarheit darüber, welche Testungen hier gemeint sind. Insofern ist hier davon auszugehen, dass die Frage nicht von allen gleichermaßen verstanden wurde.

In allen Schulen war es vereinbart, dass die Schüler und Schülerinnen selbstständig in den Computerraum kommen. In manchen Fällen klappte das allerdings nicht. Bis die Schüler und Schülerinnen gefunden wurden und im Testraum waren, vergingen im extremsten Fall etwa 20 Minuten. Somit konnte der geplante Zeitpolster nach den ersten beiden Blöcken nicht immer gewährleistet werden. Dass jemand die ersten beiden Blöcke nicht abschließen konnte, war dennoch selten.

Einige zeigten nicht die nötige Motivation zur Bearbeitung des Tests bzw. die nötige Anstrengungsbereitschaft. Andere klickten sich durch Teile des Tests durch, sei es aufgrund von aufgekommenem Zeitdruck und dem Ignorieren der Hinweise der Testleiterinnen bzw. Testleiter gegen Ende der Testzeit oder auch zwischendurch, wiederum aufgrund der fehlenden Motivation zu Bearbeitung der jeweiligen Aufgaben. Für die Auswertung bedarf es somit entsprechender Bemühungen, die Daten zu bereinigen. Solange nur Zeitdruck im dritten Block entstanden ist, wäre das nicht relevant. Aufgrund der sicher nur lückenhaft möglichen Verhaltensbeobachtungen ist bei mindestens 20 Testpersonen von einer „nichtrealen“ Testbearbeitung auszugehen.

4.4 Stichprobe

Aufgrund der genannten Auffälligkeiten wird zunächst dargestellt, welche Anzahl an Personen aus welchem Grund von den weiteren Analysen ausgeschlossen wurden. Danach folgt eine Beschreibung der verbliebenen Schüler und Schülerinnen hinsichtlich der erhobenen Daten.

Insgesamt nahmen zu Testzeitpunkt eins 304 und zu Zeitpunkt zwei 303 Personen an den Testungen teil. Daten von beiden Durchgängen waren von 261 Testpersonen vorhanden.

Von den 261 wurden die Daten von einer Testperson entfernt, da sie den Test im ersten Durchgang ein zweites Mal bearbeitet hat. Fünf weitere wurden von den weiteren Analysen ausgeschlossen, da sie zumindest bei einem der beiden Durchgänge einen der beiden interessierenden Blöcke nicht abschließen konnten.

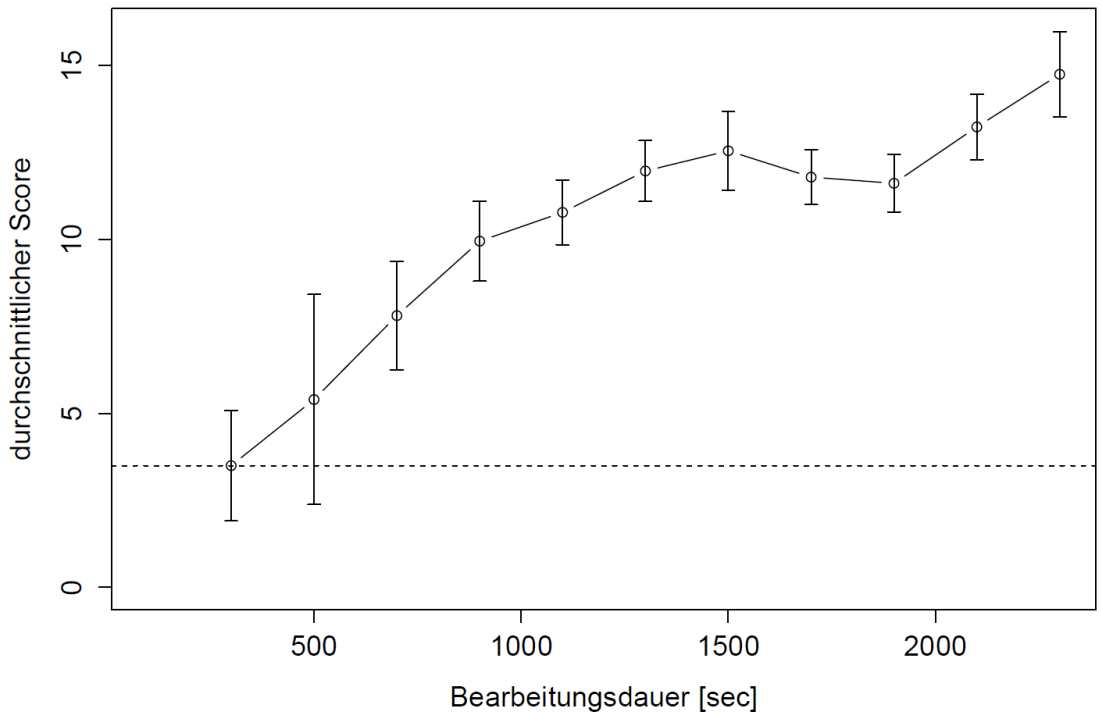


Abbildung 3. Durchschnittlicher Score in Abhängigkeit von der Bearbeitungsdauer. Der Grafik liegen die Daten aller 346 Testpersonen, die an einem der Testzeitpunkte anwesend waren, von höchstens 25 Items zugrunde. Bei 156 Testpersonen kam es zu einem Abbruch seitens der Testleitung, da die Schulstunde zu Ende war. Ein Abbruch erfolgte frühestens nach 818 Sekunden. Die Berechnung der eingezeichneten Konfidenzintervalle je Abschnitt erfolgte für $\alpha = 0{,}05$. Bearbeitungsdauer ist jeweils die benötigte Zeit für die Testaufgaben, exklusive Instruktionszeit und Zeit, die für das Übungsitem benötigt wurde. Die horizontale Linie kennzeichnet den durchschnittlichen Score, der rein durch Raten zu erwarten gewesen wäre.

Der Ausschluss „nicht-real“ arbeitender Testpersonen erfolgte in zwei Schritten. Zunächst wurde die Gesamtzeit analysiert, welche die Testpersonen für die Bearbeitung der Aufgaben benötigten. Dabei sollte, inspiriert durch Schnipke und Scrams (1997), jene Zeit ermittelt werden, bei welcher ein Lösungsverhalten seitens der Testpersonen gegenüber einem Rateverhalten in den Vordergrund tritt. Um eine ausreichend große Anzahl an Fällen zu erreichen, gingen alle Testpersonen, die an einem der beiden Testzeitpunkte anwesend waren, in die Analyse ein. Außerdem wurde die Bearbeitungsdauer aller durchgeführten Aufgaben, also auch der 13 MC-Aufgaben aus Block drei, berücksichtigt. Die Testpersonen konnten somit höchstens 25 Aufgaben bearbeiten, 19 davon im MC-Format. Rein durch Raten wäre somit ein durchschnittlicher Score von $3{,}5$ zu erwarten. Ein Abbruch seitens der Testleitung erfolgte im ungünstigsten Fall nach einer Bearbeitungszeit (exklusive Zeiten für Instruktionen und Übungsbeispiele) von 818 Sekunden. Aufgrund der grafischen Analyse in Abbildung 3 scheint ab einer Bearbeitungszeit von

500 Sekunden ein „reales“ Testverhalten in den Vordergrund zu treten. Testpersonen mit einer Bearbeitungszeit kleiner gleich 500 Sekunden an einem der beiden Durchgänge wurden deshalb von der Auswertung ausgeschlossen. Von den verbliebenen 255 Testpersonen waren davon sechs betroffen.

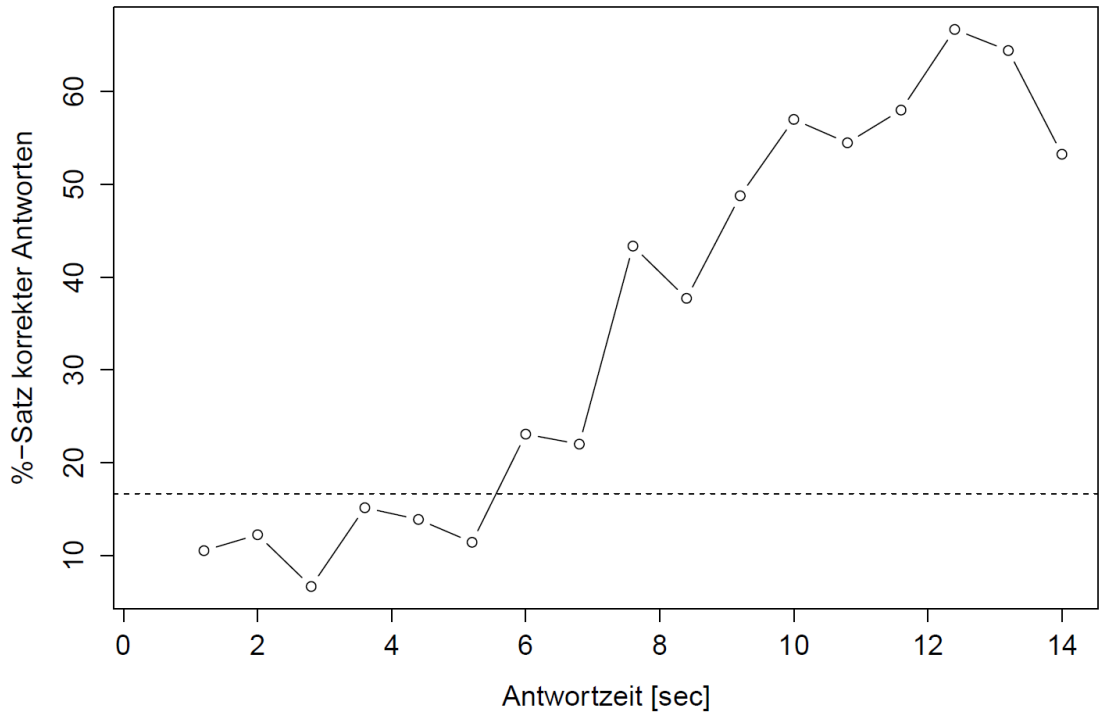


Abbildung 4. Prozentsatz korrekter Antworten in Abhängigkeit von der Antwortzeit. Der Grafik liegen die Daten aller 346 Testpersonen, die an einem der Testzeitpunkte anwesend waren, von allen Aufgaben, zu denen eine Antwort vorliegt, zugrunde. Für eine bessere Auflösung im relevanten Zeitbereich wurden in der Abbildung nur Antwortzeiten bis 15 Sekunden berücksichtigt. Die horizontale Linie stellt den Anteil korrekter Antworten für die Aufgaben im Multiple-Choice-Antwortformat dar, der aufgrund von reinem Rateverhalten zu erwarten gewesen wäre.

In einem zweiten Schritt wurden die Antwortzeiten analysiert. Analog zu vorhin sollte jene Antwortzeit ermittelt werden, bei der ein Lösungsverhalten gegenüber einem Rateverhalten in den Vordergrund tritt (vgl. Schnipke & Scrams, 1997). Um wiederum entsprechende Fallzahlen zur graphischen Analyse zu erhalten, flossen die Antwortzeiten und Antworten aller bearbeiteten Items, also auch jener aus Block drei und aller an einem von beiden Testzeitpunkten anwesenden Testpersonen ein. Für die Aufgaben im Multiple-Choice-Format (1 aus 6) ist zumindest bis zu einem durchschnittlichen Anteil von $1/6$ an korrekten Antworten mit reinem Rateverhalten zu rechnen. Nach Abbildung 4 scheint mit einer Antwortzeit von sechs Sekunden ein „reales“ Testverhalten in den Vordergrund zu treten. Testpersonen,

welche in den für die Auswertung relevanten Aufgabenblöcken (MC und AF) ein Item unter sechs Sekunden beantworteten, wurden ausgeschlossen. Von den verbliebenen Testpersonen waren davon acht betroffen.

Für die weitere Auswertung standen somit Daten von beiden Testzeitpunkten für $N = 241$ Testpersonen zur Verfügung. Die 113 Schüler und 128 Schülerinnen (53%) waren zwischen 14 und 20 Jahren alt (siehe Abbildung 5) und besuchten die 10. bis 13. Schulstufe (siehe Abbildung 6) einer höher bildenden Schule in Niederösterreich. 219 (91%) der Testpersonen gaben als Muttersprache Deutsch an, acht Personen gaben eine andere Sprache an. 14 Personen gaben neben Deutsch auch weitere Sprachen als Muttersprache an, was aber auch auf eine teilweise Fehlinterpretation der Frage hindeuten könnte.

Die Häufigkeiten der Antworten zur letzten Jahresnote in Mathematik sowie zur Erfahrung mit Testungen am Computer und mit dem Multiple-Choice-Format sind in Abbildung 7 und 8 dargestellt.

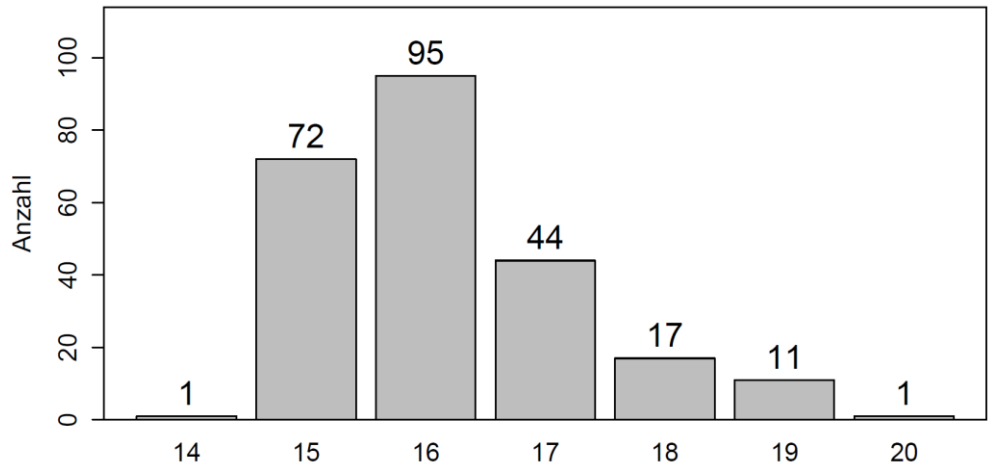


Abbildung 5. Verteilung der Variable Alter in Jahren.

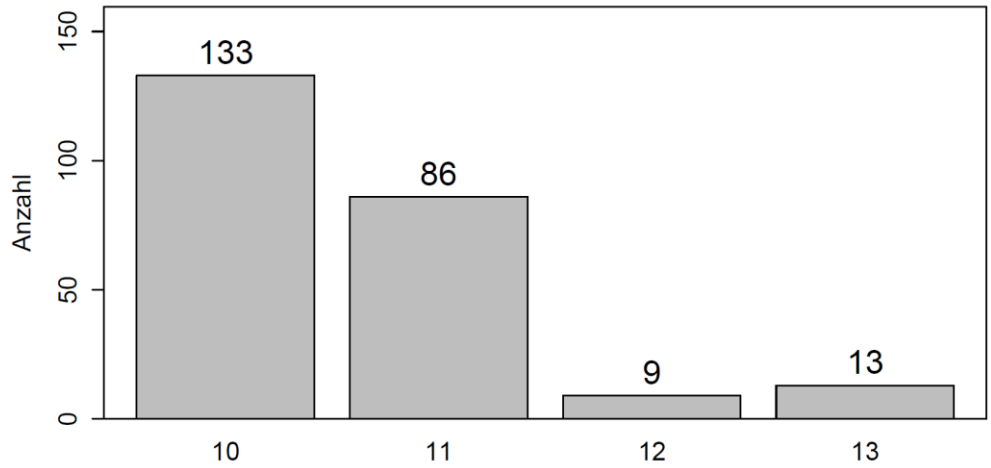


Abbildung 6. Verteilung der Variable Schulstufe.

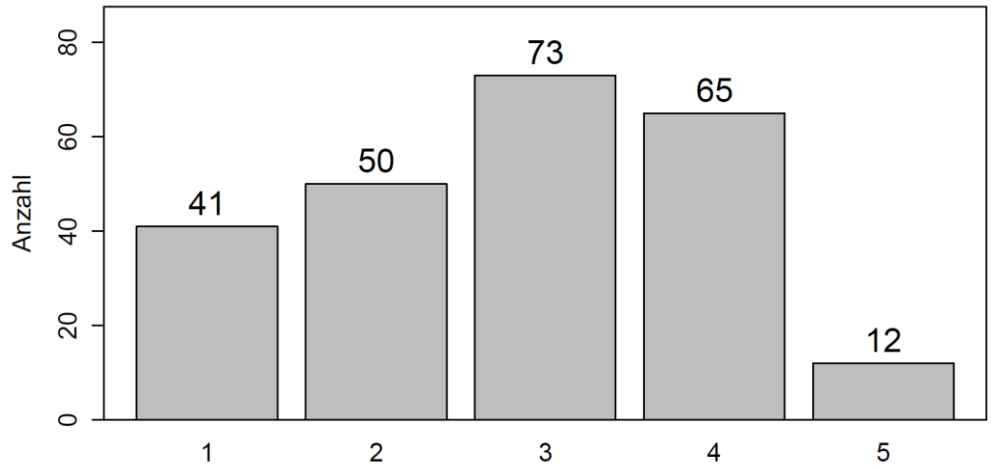


Abbildung 7. Verteilung der Variable Jahresnote in Mathematik.

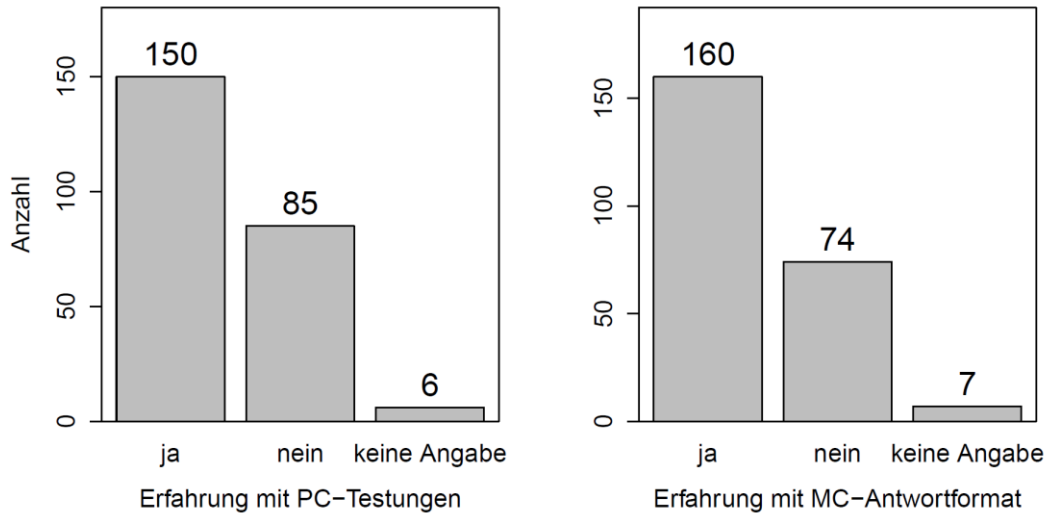


Abbildung 8. Erfahrung mit PC-Testungen und dem Multiple-Choice-Format.

4.5 Auswertung

Zunächst wird die Geltung des Rasch-Modells für die Items 1-12 zu Testzeitpunkt eins geprüft. Das Vorgehen ist in Kapitel 1.1 beschrieben. Spricht ein (bedingter) Likelihood-Quotienten-Tests (LRT) von Andersen gegen die Modellgeltung, wird versucht, durch Ausschluss einzelner Items a posteriori Modellgültigkeit zu erlangen. Dabei erfolgt die Beurteilung der Passung der Items (deskriptiv) mit Hilfe des Grafischen Modelltests.

Als Teilungskriterien werden verwendet: (1) der Median des Rohscores, (2) der Median der Bearbeitungszeit der Aufgaben, (3) das Geschlecht der Testpersonen, (4) der Median des Alters der Testpersonen zu Testzeitpunkt eins, (5) die Reihenfolge, in der die Blöcke AF und MC zu Testzeitpunkt eins vorgegeben wurden, (6) ob Erfahrung mit Testungen am Computer vorliegt und (7) ob Erfahrung mit dem MC-Antwortformat vorliegt. Für die letzten beiden Kriterien werden fehlende Angaben zu der Antwortkategorie „nein“ hinzugerechnet. Dieses Vorgehen schmälert natürlich die Aussagekraft der jeweiligen LRT, ein Ausschluss der betreffenden Testpersonen ist bezüglich der Fragestellungen aber nicht begründbar. Der Umfang der Teilstichproben ist in Tabelle 2 ersichtlich.

Tabelle 2

Teilungskriterien und Umfang der Teilstichproben zu Testzeitpunkt eins

	Rohscore		Bearbeitungszeit		Geschlecht		Alter	
	≤ 8	> 8	≤ 947s	> 947s	männl.	weibl.	≤ 16	> 16
n	145	96	121	120	113	128	168	73

	Reihenfolge		Erfahrung PC-Testung		Erfahrung MC-Format	
	MC-AF	AF-MC	ja	nein	ja	nein
n	122	119	150	91	160	81

Anmerkungen. Bei den Teilungskriterien Erfahrung mit PC-Testung und Erfahrung mit dem MC-Format wurden fehlende Angaben zur Antwortkategorie „nein“ hinzugerechnet. AF…freies Antwortformat, MC…Multiple-Choice-Antwortformat.

Sprechen die Belege nicht gegen das Gütekriterium Skalierung der ZART-Items, wird der Korrelationskoeffizient nach Spearman r_s, zwischen den Variablen Rohscore zu Testzeitpunkt eins und Note im Fach Mathematik als Maß für die Kriteri-

umsvalidität des ZART gerechnet und auf ihre Signifikanz geprüft ($H_0^5: r_s = 0$, $H_1^5: r_s \leq 0$).

Kann die Geltung des Rasch-Modells für die Initialtestung angenommen werden, wird gemäß den Darstellungen für itemspezifische Veränderung in Kapitel 2.3.1 mit der Untersuchung der Hypothesen H_0^1 bis H_0^3 fortgefahren.

Zuerst wird die Geltung des Rasch-Modells für die Items 1-12 aus beiden Testzeitpunkten geprüft, wobei die Items aus Testzeitpunkt zwei als virtuelle neue Items angeschrieben werden. Die Prüfung der Modellgeltung erfolgt analog zu vorhin. H_0^1 für jeden LRT von Andersen lautet $\sigma_{it}^{(1)} = \sigma_{it}^{(2)}$, $H_1^1: \sigma_{it}^{(1)} \neq \sigma_{it}^{(2)}$.

Als Teilungskriterien werden verwendet: (1) der Median des Rohscores, (2) der Median der durchschnittlichen Bearbeitungszeit zu beiden Testzeitpunkten, (3) das Geschlecht der Testpersonen und (4) der Median des Alters der Testpersonen zu Testzeitpunkt eins. Der Umfang der Teilstichproben ist in Tabelle 3 ersichtlich.

Tabelle 3

Teilungskriterien und Umfang der Teilstichproben für beide Testzeitpunkte

	Rohscore		Bearbeitungszeit		Geschlecht		Alter	
	≤ 16	> 16	≤ 829s	> 829s	männl.	weibl.	≤ 16	> 16
n	134	107	121	120	113	128	168	73

Zudem wird zur Prüfung der Modellgeltung der Martin-Löf-Test eingesetzt, wobei die Teilung der Items nach dem Testzeitpunkt erfolgt. Ein signifikantes Ergebnis würde wiederum H_0^1 widerlegen.

Sofern die Modellgeltung nicht widerlegt wurde, dient das Rasch-Modell für itemspezifische Veränderung mit 24 Itemparametern, ähnlich wie ein saturiertes Modell dazu, um Annahmen für die Untersuchung von H_0^2 und H_0^3 in linear-logistischen-Testmodellen (LLTM) aufzustellen und zu prüfen. Folgende Modelle werden geprüft:

Modell 1 besteht aus 14 Basisparametern: 12 Ausgangsschwierigkeiten σ_i der ZART-Items, einem allgemeinen Wiederholungparameter bzw. Retest-Effekt λ, um den sich die Schwierigkeiten aller Aufgaben zu Testzeitpunkt zwei gleichermaßen

verändern und einem für das MC-Format spezifischen Retest-Effekt γ. Die Linearkombination aus Formel 2 vereinfacht sich dabei zu $\sigma_i + q_t\lambda + q_{MC}\gamma$, wobei q_t zu Zeitpunkt eins gleich null und zu Zeitpunkt zwei gleich eins ist und q_{MC} für die Items im MC-Format zu Testzeitpunkt zwei eins und ansonsten null ist. Die Strukturmatrix für Modell 1 ist in Abbildung 9 dargestellt.

		σ_1	.	.	.	.	σ_{12}	λ	γ
$t = 1$	I_1	1							
	.		.						
	.			.					
	.				.				
	.					.			
	I_{12}						1		
$t = 2$	I_1	1						1	1
	.		.					.	.
	.			.				.	1
	.				.			.	
	.					.		.	
	I_{12}						1	1	

Abbildung 9. Strukturmatrix für Modell 1. σ_1 - σ_{12}…Ausgangsschwierigkeiten der Items I, λ…Wiederholungsparameter bzw. Retest-Effekt, γ…für das MC-Format spezifischer Retest-Effekt, t…Testzeitpunkt.

In Modell 2 fällt der für das MC-Format spezifische Retest-Effekt γ weg. Dieses LLTM setzt sich somit aus 13 Basisparametern zusammen. Formel 2 vereinfacht sich zu $\sigma_i + q_t\lambda$, wobei q_t zu Zeitpunkt eins gleich null und zu Zeitpunkt zwei gleich eins ist. Die Strukturmatrix ist ident mit jener in Abbildung 9, nur eben ohne die Spalte γ.

Modell 3 nimmt lediglich an, dass die Itemschwierigkeiten zu beiden Zeitpunkten gleich sind: $\sigma_{i1} = \sigma_{i2} = \sigma_i$. Dieses LLTM enthält die 12 Ausgangsschwierigkeiten σ_i als Basisparameter und entspricht H_0^2. Die Strukturmatrix für Modell 3 kann ebenfalls Abbildung 9 entnommen werden, die letzten beiden Spalten sind dabei auszublenden.

Ob die Daten durch die Modellannahmen statistisch genauso gut erklärt werden wie im zugrundeliegenden Rasch-Modell für itemspezifische Veränderung (Modell 0), wird für jedes Modell mit einem LRT gemäß Formel 3 geprüft. Kann dadurch noch keine Entscheidung gegen zumindest zwei Modelle getroffen werden, wird im Rahmen hierarchisch angewandter LRT über Formel 4 untersucht, welches Modell die beste Passung aufweist (siehe Kapitel 1.1 und 2.3.1). Erweisen sich alle drei Modelle als statistisch gleich passend, wird das sparsamste Modell drei angenommen und somit auch H_0^2.

Ein Vergleich von Modell eins und zwei mit Formel 4 bietet eine spezifische Aussage zu $H_0^3: \gamma = 0$. Eine signifikante Prüfgröße würde für $H_1^3: \gamma \neq 0$ sprechen.

Ein Vergleich von Modell zwei und drei mit Formel 4 bietet eine spezifische Aussage zu $H_0^2: \lambda = 0$. Eine signifikante Prüfgröße würde für $H_1^2: \lambda \neq 0$ sprechen.

Sollten dabei signifikante Schätzungen für γ und bzw. oder λ resultieren, werden, wie in der Untersuchungsplanung beschrieben, die jeweiligen Effekte $\hat{\gamma}_v$ und bzw. oder $\hat{\lambda}_v$ hinsichtlich personenspezifischer Veränderung geschätzt. Dazu wird dann auch die geschätzte (relative) Effektgröße (bezogen auf die Standardabweichung in der Initialtestung) $\hat{E}$ für die vorliegende Stichprobe angegeben und das eingegangene Risiko 2. Art β_{post} berechnet.

Zuletzt erfolgt die Untersuchung von H_0^4 unter Verwendung eines Vorzeichen-Rang-Tests von Wicoxon. Die statistische Alternativhypothese H_1^4 dazu lautet: $Mdn_{Bearbeitungszeit1} > Mdn_{Bearbeitungszeit2}$.

Für die statistischen Analysen wird die freie Statistiksoftware R, Version 3.0.3 (R Core Team, 2014), verwendet, die Berechnungen zum Rasch-Modell und zum LLTM erfolgen im Programmpaket eRm, Version 0.15-4 (Mair, Hatzinger & Maier, 2014). Die Berechnungen zum ergebnisbasierten Risiko 2. Art (siehe Kubinger et al., 2011, S. 205-206) erfolgen wiederum mit der Freeware G*Power, Version 3.1 (Faul et al., 2007).

Das Signifikanzniveau für inferenzstatistische Analysen wird mit $\alpha = 0{,}05$ festgesetzt. Bei der Prüfung auf Geltung des Rasch-Modells werden jeweils mehrere

Signifikanztests durchgeführt. Um eine untersuchungsbezogene Überhöhung des Risikos 1. Art zu berücksichtigen, wird hier $\alpha = 0{,}01$ gewählt (vgl. Kubinger, 2005).

5 Ergebnisse

5.1 Überprüfung des Testmodells für Testzeitpunkt 1

5.1.1 Rasch-Modell Analysen

In Tabelle 4 sind die Ergebnisse der LRT von Andersen zur Prüfung der Modellgültigkeit der Items 1-12 zu Testzeitpunkt eins dargestellt. Beim Teilungskriterium Rohscore ist Item 7 von der Berechnung ausgeschlossen worden, da es in der Gruppe der Leistungsstärkeren von allen gelöst wurde. In keinem der Teilungskriterien zeigt sich ein signifikantes Ergebnis. Die Grafischen Modelltests sind in Abbildung 18 im Anhang dargestellt.

Tabelle 4

Ergebnisse der Likelihood-Quotienten-Tests von Andersen für Testzeitpunkt eins

Teilungskriterium	χ^2	df	$\chi^2_{\alpha=0{,}01}$	
Rohscore	6,56	10	23,21	nicht signifikant
Bearbeitungszeit	7,11	11	24,72	nicht signifikant
Geschlecht	18,88	11	24,72	nicht signifikant
Alter	8,44	11	24,72	nicht signifikant
Reihenfolge	17,15	11	24,72	nicht signifikant
Erfahrung PC-Testung	6,69	11	24,72	nicht signifikant
Erfahrung MC-Format	15,47	11	24,72	nicht signifikant

Anmerkungen. Beim Teilungskriterium Rohscore wurde Item 7 aufgrund der Lösungshäufigkeiten in den Teilstichproben von den Berechnungen ausgeschlossen.

Für die 12 Items des ZART kann somit die Geltung des Rasch-Modells angenommen werden, was auch die Voraussetzung für die weiteren Analysen nach diesem Testmodell darstellt. Die geschätzten Item-Schwierigkeitsparameter $\hat{\sigma}_i$ können aus Tabelle 11 im Anhang entnommen werden, die geschätzten Personenparameter $\hat{\xi}_v$ aus Tabelle 12.

5.1.2 Deskriptive Statistiken und Kriteriumsvalidität

In Abbildung 10 ist die Verteilung der Anzahl gelöster Aufgaben, also des Scores, an Testzeitpunkt eins ersichtlich. Der Mittelwert beträgt $7{,}90$, die Standardabweichung $2{,}30$.

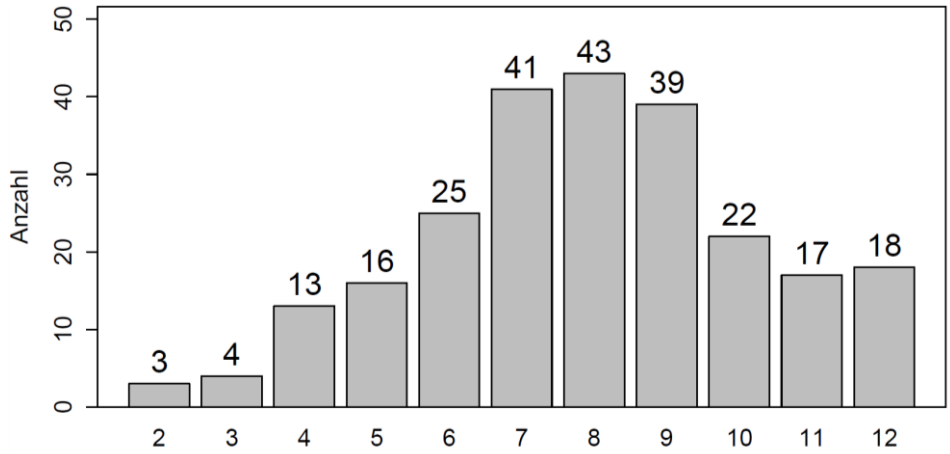


Abbildung 10. Verteilung der Anzahl gelöster Aufgaben zu Testzeitpunkt eins.

Um später die (relative) Effektgröße eventueller Retest-Effekte schätzen zu können, wird auch die Standardabweichung der Personenparameter angegeben, sie beträgt $1{,}46$.

Zwischen dem Score zu Testzeitpunkt eins und der letzten Jahresnote im Fach Mathematik besteht ein signifikanter negativer Zusammenhang. H_0^5 kann somit verworfen und H_1^5 angenommen werden. Der Korrelationskoeffizient nach Spearman r_s beträgt $-0{,}31$.

5.2 Effekte wiederholter Vorgabe

5.2.1 Rasch-Modell für itemspezifische Veränderung

Aus Tabelle 5 können die Ergebnisse der LRT von Andersen und des Martin-Löf-Tests zur Prüfung der Modellgültigkeit der Items 1-12 zu beiden Testzeitpunkten entnommen werden. Wie bei den Analysen zu Testzeitpunkt eins wurde Item 7 beim Teilungskriterium Rohscore von der Berechnung ausgeschlossen. Beim Teilungskriterium Geschlecht zeigt sich ein signifikantes Ergebnis. In Abbildung 11 ist

der Grafische Modelltest für das Teilungskriterium Geschlecht dargestellt. Die Nummerierung der Items orientiert sich an der Position zu Testzeitpunkt eins, wenn mit dem Itemblock gestartet wird, der die Aufgaben im Multiple-Choice-Antwortformat (MC) enthält (siehe Tabelle 1). Um zu kennzeichnen, zu welchem Zeitpunkt ein Item vorgegeben wurde, wird, getrennt durch einen Punkt, der Testzeitpunkt hinten an die Itemnummer angefügt, z.B. I9.2 für Item 9, das zu Testzeitpunkt zwei vorgegeben wird.

Tabelle 5

Ergebnisse der Likelihood-Quotienten-Tests von Andersen und des Martin-Löf-Tests für beide Testzeitpunkte

Teilungskriterium	χ^2	df	$\chi^2_{\alpha=0,01}$	
Rohscore	21,98	21	38,93	nicht signifikant
Bearbeitungszeit	22,93	23	41,64	nicht signifikant
Geschlecht	43,13	23	41,64	signifikant
Alter	19,61	23	41,64	nicht signifikant
Testzeitpunkt (M.Löf)	97,06	143	185,26	nicht signifikant

Anmerkungen. Beim Teilungskriterium Rohscore wurde Item 7 aufgrund der Lösungshäufigkeiten in den Teilstichproben von den Berechnungen ausgeschlossen.

A priori kann die Geltung des Rasch-Modells für itemspezifische Veränderung somit nicht angenommen werden. Die Ergebnisse sprechen aber nicht generell gegen H_0^1.

Im Grafischen Modelltest in Abbildung 11 ist das Item 4 auffällig. Item 4 weist zu beiden Zeitpunkten eine ähnliche Abweichung auf. Im Rasch-Modell für itemspezifische Veränderung, geht die Abweichung doppelt in die globale Teststatistik des LRT von Andersen ein, was die Signifikanz gegenüber der analogen Statistik für die Daten aus Testzeitpunkt eins bedingen dürfte.

Voraussetzung für die LLTM-Analysen ist aber jedenfalls die Rasch-Modell-Konformität des zugrundeliegenden Modells. Deswegen soll in einem zweiten Berechnungsdurchgang untersucht werden, ob durch Ausschluss von Item 4 (genauer: I4.1 und I4.2) zumindest a posteriori die Modellgeltung angenommen werden kann.

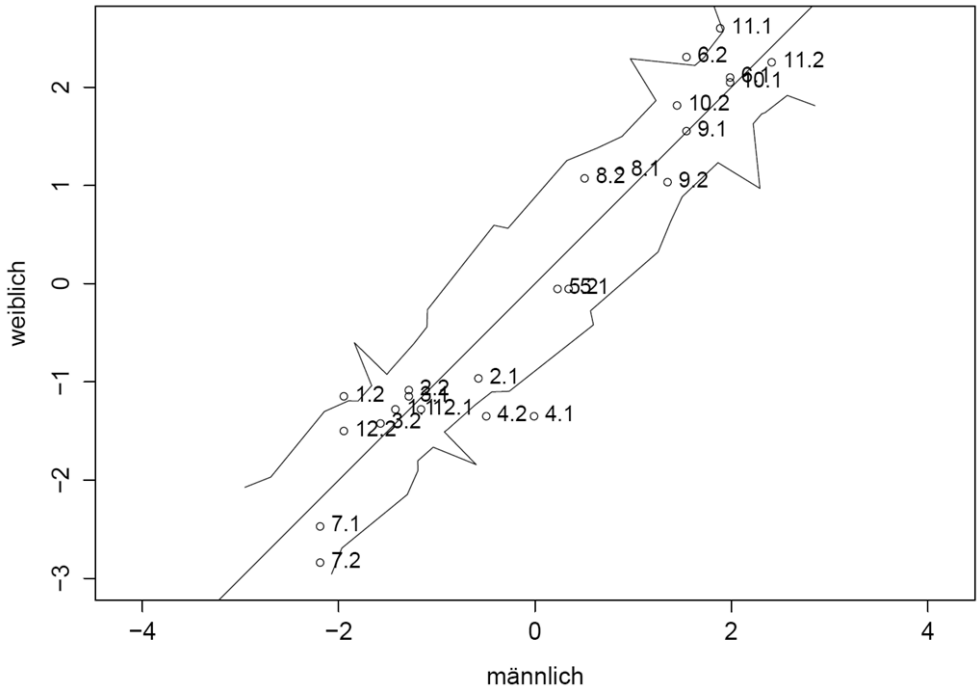


Abbildung 11. Grafischer Modelltest zum Rasch-Modell für itemspezifische Veränderung für das Teilungskriterium Geschlecht. Als visuelle Orientierungshilfe sind die Konfidenzbänder ($\alpha = 0{,}05$) eingezeichnet.

5.2.1.1 Zweiter Berechnungsdurchgang nach Ausschluss von Item 4

Durch den Ausschluss von Item 4 kommt es zu Veränderungen bei den abhängigen Variablen. In Tabelle 6 sind die aktualisierten Umfänge der Teilstichproben für die LRT von Andersen dargestellt.

Tabelle 6

Teilungskriterien und Umfang der Teilstichproben für beide Testzeitpunkte nach Ausschluss von Item 4

	Rohscore		Bearbeitungszeit		Geschlecht		Alter	
	≤ 14	> 14	≤ 761s	> 761s	männl.	weibl.	≤ 16	> 16
n	127	114	121	120	113	128	168	73

Aus Tabelle 7 können die Ergebnisse zur Modellgeltung entnommen werden. Wie bereits zuvor ist beim Teilungskriterium Rohscore Item 7 von der Berechnung ausgeschlossen worden. Nach Ausschluss von Item 4 zu beiden Testzeitpunkten zeigt sich für kein Teilungskriterium ein signifikantes Ergebnis. Die Grafischen Modelltests sind in Abbildung 19 im Anhang dargestellt.

Tabelle 7

Ergebnisse der Likelihood-Quotienten-Tests von Andersen und des Martin-Löf-Tests nach Ausschluss von Item 4 für beide Testzeitpunkte

Teilungskriterium	χ^2	df	$\chi^2_{\alpha=0,01}$	
Rohscore	14,15	19	36,19	nicht signifikant
Bearbeitungszeit	20,65	21	38,93	nicht signifikant
Geschlecht	23,06	21	38,93	nicht signifikant
Alter	16,15	21	38,93	nicht signifikant
Testzeitpunkt (M.Löf)	89,03	120	158,95	nicht signifikant

Anmerkungen. Beim Teilungskriterium Rohscore wurde Item 7 aufgrund der Lösungshäufigkeiten in den Teilstichproben von den Berechnungen ausgeschlossen.

Nach Ausschluss von Item 4 kann somit a posteriori die Geltung des Rasch-Modells für itemspezifische Veränderung angenommen werden, was auch die Voraussetzung für die darauf aufbauenden LLTM-Analysen darstellt. Zumindest für diese Daten kann auch die H_0^1 beibehalten werden. Die geschätzten Item-Schwierigkeitsparameter $\hat{\sigma}_{it}$ können aus Tabelle 13 im Anhang entnommen werden.

Aufgrund der Ergebnisse kann zumindest für 11 der 12 untersuchten Items des ZART angenommen werden, dass zu beiden Zeitpunkten dieselbe latente Dimension erfasst wurde.

5.2.2 LLTM Analysen zu Retest-Effekten

Durch den Ausschluss von Item 4 kommt es in diesem Kapitel zu geringfügigen Änderungen gegenüber den in Kapitel 4.5 dargestellten Vorhaben. Die LLTM-Modelle haben jeweils um einen Basisparameter weniger, nämlich σ_4. Die Bezeichnung der anderen Basisparameter σ_1 - σ_3 und σ_5 - σ_{12} (sowie λ und γ) bleibt der Einfachheit halber gleich.

In Tabelle 8 sind die logarithmierten Likelihoods L des zugrundeliegenden Rasch-Modells für itemspezifische Veränderung (Modell 0) und der LLTM-Modelle eins bis drei dargestellt. Die vollständigen Strukturmatrizen der LLTM-Modelle sind im Anhang in Tabelle 14 dargestellt.

Tabelle 8

Log-Likelihoods L des zugrundeliegenden Rasch-Modells für itemspezifische Veränderung (Modell 0) und der LLTM-Modelle 1-3

Modell	Parameter	df	$\ln L$
0	22 Itemparameter σ_{it}	21	-1715,59
1	11 x σ_i, λ und γ	13	-1717,78
2	11 x σ_i und λ	12	-1717,95
3	11 x σ_i	11	-1721,55

Anmerkungen. σ_i…Ausgangsschwierigkeiten der Items, λ…allgemeiner Wiederholungsparameter, γ…für das MC-Format spezifischer Retest-Effekt.

Die Vergleiche der Modelle eins bis drei mit Modell 0 gemäß Formel 3 als auch der Modelle eins bis drei untereinander nach Formel 4 finden sich in Tabelle 9.

Tabelle 9

Hierarchische Modellvergleiche

	χ^2	df	$\chi^2_{\alpha=0,05}$	
Modell 0 vs. Modell 1	4,38	8	15,51	nicht signifikant
Modell 0 vs. Modell 2	4,72	9	16,92	nicht signifikant
Modell 0 vs. Modell 3	11,92	10	18,31	nicht signifikant
Modell 1 vs. Modell 2	0,34	1	3,84	nicht signifikant
Modell 2 vs. Modell 3	7,20	1	3,84	signifikant

Die Modelle eins, zwei und drei erklären die Daten statistisch genauso gut wie Modell 0. Modell 2 erklärt die Daten signifikant besser als Modell 3. H_0^2 ist somit abzulehnen.

Modell 1 erklärt die Daten statistisch genauso gut wie Modell 2. Da Modell 2 das sparsamere von beiden darstellt, ist nach den vorliegenden Daten H_0^3 beizubehalten. Der Wiederholungsparameter λ wird in Modell 2 auf $-0{,}203$ geschätzt, bei einem Standardschätzfehler von $0{,}076$ und einem Konfidenzintervall bei $\alpha = 0{,}05$ von $[-0{,}352,\ -0{,}055]$. Eine Gegenüberstellung der Schätzungen der Itemparameter aus Modell 0 und Modell 2 kann der Tabelle 15 im Anhang entnommen werden.

Die Ergebnisse sprechen für einen vom verwendeten Antwortformat unabhängigen Retest-Effekt λ, um den sich die Schwierigkeiten der 11 ZART-Items zu Testzeitpunkt zwei gleichermaßen verringern.

Hinsichtlich personenspezifischer Veränderung (bei der Annahme identer Itemparameter zu beiden Zeitpunkten) kann aus diesem Modell eine durchschnittliche Verbesserung $\hat{\lambda}_v$ von $0{,}203$ zu Testzeitpunkt zwei abgeleitet werden.

5.2.3 Deskriptive Statistiken, Effektgröße und Bearbeitungszeit

Durch den Ausschluss von Item 4 werden im Folgenden die deskriptiven Statistiken für die abhängigen Variablen Rohscore und Bearbeitungszeit nach Ausschluss von Item 4 getrennt für die beiden Testzeitpunkte dargestellt.

In Abbildung 12 ist die Verteilung der Anzahl gelöster Aufgaben zu Testzeitpunkt eins ersichtlich. Der Mittelwert beträgt $7{,}08$, die Standardabweichung $2{,}18$.

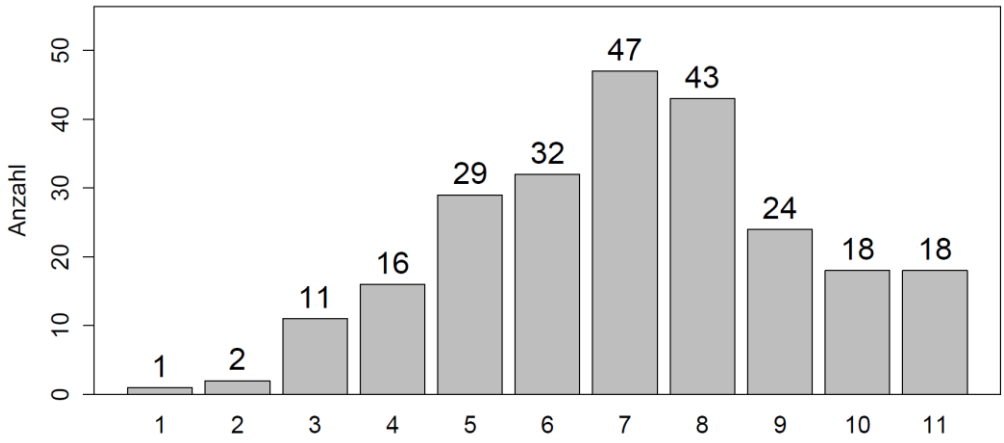


Abbildung 12. Verteilung der Anzahl gelöster Aufgaben zu Testzeitpunkt eins, nach Ausschluss von Item 4.

Abbildung 13 zeigt die Verteilung der Anzahl gelöster Aufgaben zu Testzeitpunkt zwei. Der Mittelwert beträgt $7{,}37$, die Standardabweichung $2{,}21$.

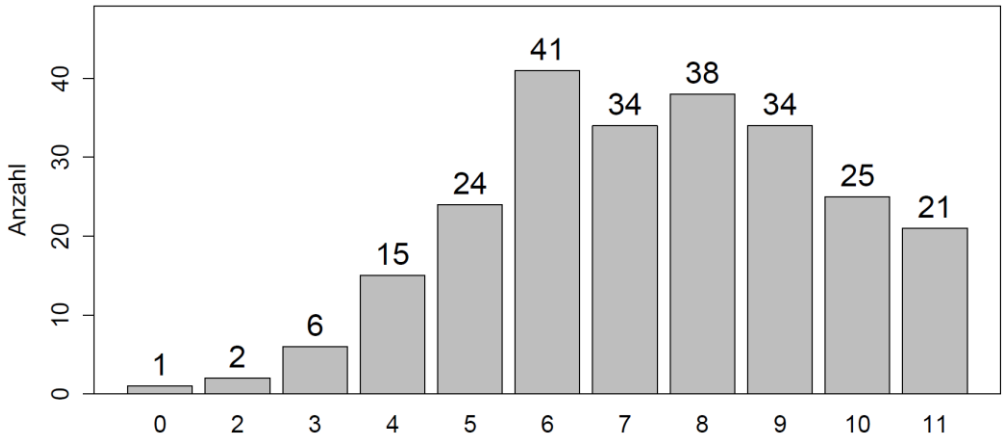


Abbildung 13. Verteilung der Anzahl gelöster Aufgaben zu Testzeitpunkt zwei, nach Ausschluss von Item 4.

Die Schätzung der auf die Stichprobe bezogenen Effektgröße $\hat{E}_\lambda$ für den im vorherigen Kapitel gefundenen Retest-Effekt λ wird auf die Schätzung der Standardabweichung der Personenparameter zu Testzeitpunkt eins ohne Ausschluss von Item 4 bezogen, da für Testzeitpunkt eins weiterhin das Rasch-Modell gilt: $\hat{E}_\lambda = 0{,}14$.

Das Risiko 2. Art β_{post} für $\hat{\lambda}_v$ wird für einen t-Test für abhängige Stichproben (vgl. Kapitel 4.1) bei einseitiger Fragestellung ermittelt. Die Schätzung der Standardabweichung der Differenzwerte $\hat{\sigma}_D$ erfolgt über Formel 5 aus der Standardabweichung der Personenparameter zu Testzeitpunkt eins. Der Zusammenhang ρ der beiden Testergebnisse wird über die Korrelation der Rohscores geschätzt, da Personenparameter und Rohscores im Allgemeinen sehr hoch miteinander korrelieren.

Die Rohscores, nach Ausschluss von Item 4, der beiden Testzeitpunkte korrelieren zu $0{,}66$. Das Risiko 2. Art β_{post} für $\hat{\lambda}_v$ in Bezug auf die vorliegende Stichprobe liegt bei $0{,}16$. Dies liegt über dem in der Planung festgesetzten β von $0{,}05$.

In Abbildung 14 ist das Histogramm für die Bearbeitungszeit zu Testzeitpunkt eins dargestellt. Das Histogramm für die Bearbeitungszeit zu Testzeitpunkt zwei kann Abbildung 15 entnommen werden. Die Quantile für die beiden Variablen sind in Tabelle 10 ersichtlich.

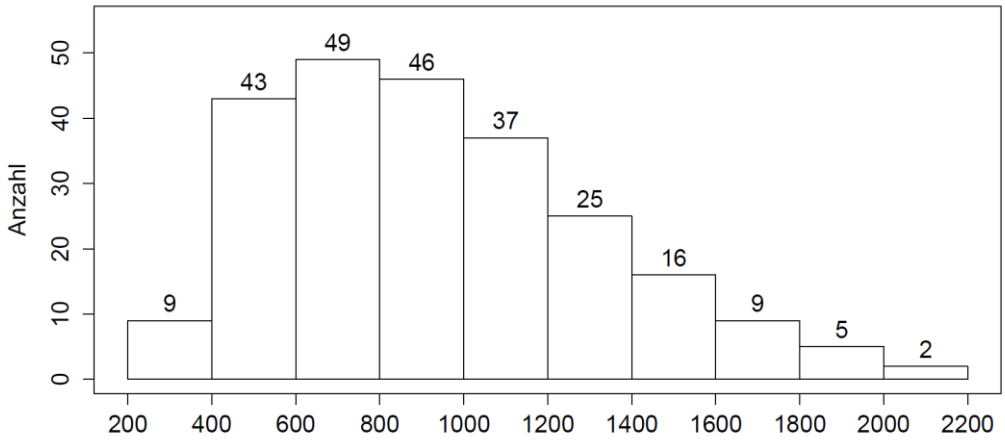


Abbildung 14. Histogramm der Bearbeitungszeit in Sekunden zu Testzeitpunkt eins nach Ausschluss von Item 4.

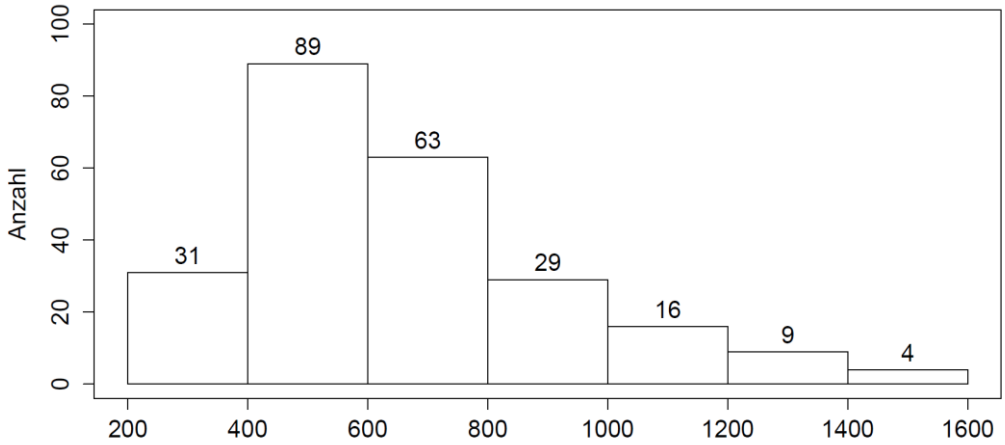


Abbildung 15. Histogramm der Bearbeitungszeit in Sekunden zu Testzeitpunkt zwei nach Ausschluss von Item 4.

Tabelle 10

Quantile für die Bearbeitungszeiten in Sekunden zu den beiden Testzeitpunkten nach Ausschluss von Item 4

	0%	25%	50%	75%	100%
TZP1	271	634	880	1181	2057
TZP2	232	464	605	788	1566

Die Testpersonen brauchen zur Bearbeitung der 11 ZART-Items zu Testzeitpunkt zwei signifikant kürzer als zu Testzeitpunkt eins. H_0^4 ist somit abzulehnen. Im Schnitt dauert die Bearbeitung im zweiten Durchgang etwa $7/10$ der Bearbeitungszeit von Testzeitpunkt eins.

6 Diskussion (und Ausblick)

Zunächst werden die Ergebnisse zu den Hauptfragestellungen diskutiert, gefolgt von den zu bedenkenden Einschränkungen bzw. Grenzen der Aussagen. Danach wird auf die Nebenfragestellungen eingegangen.

Aus dem, entsprechend der Kalibrierungsstudie von Poinstingl (2010a, 2010b), Rasch-Modell-konformen Itempool des ZART wurden in der vorliegenden Arbeit 12 Items untersucht. Da diese Aufgaben nach demselben Konstruktionsrational erstellt wurden, ist diese Auswahl repräsentativ für den Itempool des ZART.

Nach den vorliegenden Ergebnissen kann nur für 11 der 12 Items angenommen werden, dass bei identer Vorgabe nach zwei Wochen dieselbe latente Dimension wie in der Initialtestung gemessen wird. Schülerinnen scheint Item 4 bereits zu Testzeitpunkt eins etwas leichter zu fallen als ihren männlichen Kollegen. Zu Testzeitpunkt zwei zeigt sich dasselbe Bild.

Hierzu sei erwähnt, dass Item 4, im Gegensatz zu den anderen Aufgaben, auch durch die Analyse visueller Regelmäßigkeiten gelöst werden kann. Eine genauere Diskussion dieser Auffälligkeit muss aber an anderer Stelle erfolgen, da in dieser Arbeit keine Items dargestellt werden.

Insofern sprechen die Ergebnisse des Rasch-Modells für itemspezifische Veränderung nicht gegen die Eignung des ZART für den wiederholten Einsatz, sondern begründen einen Zweifel an der generellen Passung von Item 4, zumindest in der Population der Schülerinnen und Schüler in den betrachteten höheren Schulen ab der zehnten Schulstufe.

Da die Entwicklung von psychologisch-diagnostischen Verfahren sehr aufwendig ist, sollte der Ausschluss von Items aber auch nicht vorschnell erfolgen. In der Kalibrierungsstudie war das entsprechende Item mit der Nummer 46 schließlich unauffällig. Auch in den Analysen zu Testzeitpunkt eins in dieser Arbeit konnte die Geltung des Rasch-Modells für alle 12 Items angenommen werden und wie beschrieben, gehen beim Rasch-Modell für itemspezifische Veränderung die Abweichungen von Item 4 doppelt in die Teststatistik des LRT von Andersen ein.

Da die Geltung des Rasch-Modells in der Kalibrierungsstudie lediglich a posteriori angenommen werden konnte, ist nach Kubinger (2005) ohnehin eine neuerli-

che Überprüfung des verbliebenen Itempools notwendig, in der der Verbleib von Item 4 bzw. 46 zu klären ist.

Werden die Aufgaben nach zwei Wochen erneut den Testpersonen der untersuchten Population in identer Form vorgegeben, so verringern sich die Schwierigkeiten der Items nach den vorliegenden Ergebnissen um $0{,}203$. Arendasy und Sommer (2013) fanden bei einem ähnlichen Test in einem ähnlichem Setting absolute Effekte von $-0{,}265$ und $-0{,}301$, abhängig von der allgemeinen Intelligenz der Testpersonen. Ob diese Veränderung praktisch relevant ist, ist für eine jeweilige Anwendung zu ergründen.

Für die vorliegende Stichprobe kann daraus ein relativer Effekt durch die vorangegangene Testung von $0{,}14$, bezogen auf die durchschnittliche personenspezifische Verbesserung, abgeleitet werden. In dieser Aussage wurde allerdings ein überhöhtes Risiko 2. Art von $0{,}16$ (statt $\beta = 0{,}05$) eingegangen, da die Untersuchung nicht für derartig geringe Effekte ausgelegt wurde. Welche kognitiven Veränderungen infolge der Initialtestung auftreten, kann mit dieser Untersuchung selbstverständlich nicht erfasst werden, es kann lediglich von einem allgemeinen Lern- oder Übungseffekt ausgegangen werden.

Die zur Schätzung der Standardabweichung der Differenzwerte der Scores zu beiden Testzeitpunkten ermittelte Korrelation der Scores sollte nicht als Maß für die Stabilität des ZART interpretiert werden. Einerseits beruhen die Messungen zu einem Zeitpunkt lediglich auf 11 Items. Wesentlicher ist aber, dass es aufgrund der Deckeneffekte zu einer Unterschätzung des Zusammenhangs kommt. Im Gegensatz zum Wiederholungsparameter im LLTM ist das Korrelationsmaß abhängig von der Zusammensetzung der Stichprobe.

Ein Bezug des absoluten Effektes auf repräsentative Streuungsmaße ist derzeit noch nicht möglich, aber selbst in einer homogenen Population (vgl. Kapitel 4.1) dürften durchschnittliche Verbesserungen der Testpersonen im Retest um 2 T-Werte bereits äußerst unwahrscheinlich sein. In Anbetracht der Messgenauigkeit psychologisch-diagnostischer Verfahren wäre somit ein Retest-Effekt beim ZART bei einer Verlaufskontrolle praktisch uninteressant, wobei eine Anwendung der Ergebnisse auf andere Populationen und Rahmenbedingungen natürlich kritisch ist (siehe unten).

Auch die Fairness eines Aufnahmeverfahrens wäre bei so einem geringen Effekt wohl nicht beeinträchtigt. Hierbei ist zudem zu bedenken, dass die Retestungsintervalle in der Regel länger als zwei Wochen sind und es bei einen Rasch-Modell-konformen Test nicht notwendig ist, eine idente Testform als Paralleltest vorzugeben (vgl. Kubinger, 2009, S. 96). Als Paralleltest wäre auch jede andere Auswahl an Aufgaben aus dem Itempool möglich. Orientiert man sich an den Metastudien von Kulik et al. (1984) und Hausknecht et al. (2007), wäre bei Verwendung eines Paralleltests mit einer weiteren Halbierung des Retest-Effektes zu rechnen. Bei Einsatz des ZART in einem Auswahlverfahren sollte aber ohnehin eine spezifische Evaluation stattfinden. Die vorliegenden Ergebnisse sprechen jedenfalls nicht gegen einen solchen Einsatz.

In Kapitel 2.2 wurden die in der Literatur diskutierten Theorien zur Interpretation von Effekten bei wiederholter Vorgabe dargestellt. Die vorliegende Studie wurde nicht dazu designt, hier einen wesentlichen Beitrag zu leisten. Für den ZART kann auf der Basis von aussagekräftigen Methoden zumindest behauptet werden, dass die latente Dimension Schlussfolgerndes Denken im numerischen Bereich zu beiden Testzeitpunkten auf dieselbe Art und Weise erfasst wird. Dies schließt eine Veränderung in substantiellen Messeigenschaften, wie sie in den Erklärungen 2 und 3 postuliert werden, auf Itemebene aus.

Deutlich einzuschränken sind diese Aussagen durch die Testsituation bzw. den Studienkontext. Auch wenn Hausknecht et al. (2007) keinen Einfluss des Studienkontextes auf das Ausmaß von Retest-Effekten nachweisen konnten, ist nicht auszuschließen, dass dieselben Schülerinnen und Schüler in einem Aufnahmeverfahren (für das sie sich selbst gemeldet haben) ein anderes Testverhalten zeigen würden.

Aufgrund der Stichprobenunabhängigkeit der Parameterschätzung im Rasch-Modell kann eine Gültigkeit der Aussagen zumindest für die vorliegende Population angenommen werden. Eine Übertragung der Ergebnisse auf andere Populationen und Rahmenbedingungen ist ohne vorangegangene empirische Prüfung problematisch.

Natürlich kann für den verwendeten Untersuchungsplan nicht zur Gänze ausgeschlossen werden, dass die Ergebnisse durch unkontrollierte Störeinflüsse verursacht wurden. Zwischen den beiden Testzeitpunkten lagen für eine Mehrheit der

Testpersonen die Semesterferien. Eine zwischenzeitliche Beschäftigung mit Folgen und Reihen dürfte somit weniger plausibel sein. Nicht auszuschließen ist aber, dass nach der Initialtestung ein Austausch zwischen den Schülerinnen und Schülern stattfand, der den Retest-Effekt begünstigte.

Die teilweise fehlende Anstrengungsbereitschaft der Testpersonen, die in der Durchführung der Untersuchung auffiel, gemeinsam mit dem Vergleich der Verteilungen der Scores zu beiden Zeitpunkten in Abbildung 12 und 13, lässt zumindest eine Unterschätzung des Retest-Effektes aufgrund eines eventuellen Motivationsverlustes vermuten.

In diesem Zusammenhang sei auch kurz auf den Ausschluss von Testpersonen aus der Stichprobe eingegangen. Aufgrund der Beobachtungen während den Testungen war der Ausschluss nicht real arbeitender Testpersonen notwendig. Auch wenn in diesem Punkt immer ein Ermessensspielraum vorhanden ist, wurde zumindest im Vorfeld versucht, möglichst plausible Kriterien festzulegen. Dass aufgrund der Kriterien zur Bearbeitungs- und Antwortzeit lediglich Daten von 14 Testpersonen ausgeschlossen werden mussten, deutet insgesamt aber auf eine gute Bereitschaft seitens der Schülerinnen und Schüler zur Durchführung der Testungen hin.

Die genannten Einschränkungen gelten auch bei Interpretation der Ergebnisse zu den Nebenfragestellungen.

Die gefundenen Übungs- bzw. Lerneffekte sind unabhängig von den verwendeten Antwortformaten. Die Möglichkeit, das Ausmaß von Retest-Effekten zu verringern, indem man ein anderes Antwortformat verwendet, scheint auf Basis dieser Ergebnisse nicht gegeben zu sein.

Interessant, wenn auch nicht überraschend, ist, dass die durchschnittlich besseren Testergebnisse zu Testzeitpunkt zwei in deutlich geringeren Bearbeitungszeiten erreicht wurden. Zu Testzeitpunkt zwei wurden die Aufgaben somit effizienter bearbeitet. Auch hier sollte nicht unbedacht ein Vergleich mit einer Situation erfolgen, in der den Testpersonen von vornherein bekannt ist, dass eine Zeitbegrenzung besteht. Dennoch nährt dieses Ergebnis die Vermutung, dass manche in der Literatur berichteten Retest-Effekte bei Verfahren mit Zeitbegrenzung eben durch diese Zeitbegrenzung begünstigt worden sind.

Die Kriteriumsvalidität des ZART in Bezug auf die letzte Jahresnote im Fach Mathematik fällt eher mäßig aus. Eine möglichst exakte Erfassung des Zusammenhangs stand allerdings nicht im Fokus der Arbeit. So geht die Erhebung dieses Befundes mit einigen Ungenauigkeiten einher. Für die Schätzung der Personenparameter wurden lediglich 12 Items verwendet, die Standardschätzfehler sind entsprechend hoch. Auch die Genauigkeit der von den Schülern und Schülerinnen selbst berichteten Noten könnte eingeschränkt sein, insbesondere da die Jahresnoten mehr als ein halbes Jahr zuvor vergeben wurden. Hinzu kommen grundsätzliche Probleme, zum Beispiel dass die Notengebung in den einzelnen Klassen und vor allem in den einzelnen Schulen höchstwahrscheinlich nicht nach den gleichen Kriterien erfolgt. Deswegen stellt dieser Wert eben nicht mehr als einen ersten Befund dar.

Nach den vorliegenden Ergebnissen ist der von Poinstingl (2009c) entwickelte Zahlenreihentest (ZART) für den wiederholten Einsatz geeignet, wobei eine neuerliche Überprüfung des Itempools notwendig erscheint.

Die vorhandenen Lern- bzw. Übungseffekte dürften in den allermeisten Fällen nicht von praktischer Relevanz sein. Eine Korrektur der Testergebnisse von Personen, die den Test schon einmal durchgeführt haben, wäre auf Basis dieser Arbeit nicht zu begründen.

Zur Fundierung dieser Ergebnisse wären vor allem Untersuchungen zum ZART im Rahmen von praktischen Anwendungen wünschenswert.

7 Zusammenfassung

Der nach dem dichotomen logistischen Testmodell von Rasch (1960) entwickelte Zahlenreihentest (ZART; Poinstingl, 2009c) zielt auf die Erfassung von Schlussfolgerndem Denken mit Hilfe von numerischem Aufgabenmaterial ab. Hauptanwendungsbereich dieses Computertests ist die berufs- und ausbildungsbezogene Eignungsdiagnostik.

Im Rahmen der Selektionsdiagnostik ist es üblich, Bewerberinnen und Bewerbern die Möglichkeit zur Wiederholung von Aufnahmeverfahren zu geben. In der Literatur werden aber unterschiedliche Theorien über den psychometrischen Wert der Ergebnisse aus dem Retest diskutiert. Neben simplen Retest-Effekten besteht die Gefahr, dass es infolge einer Testwiederholung zu einer Veränderung substantieller Messeigenschaften kommt. Neben anderen Richtlinien regelt vor allem die DIN 33430 (DIN Deutsches Institut für Normung e. V., 2002), dass bei wiederholtem Einsatz von Verfahren deren Gültigkeit ermittelt werden sollte.

Ziel dieser Arbeit war es zu untersuchen, ob der ZART für den wiederholten Einsatz geeignet ist und mit welchen Veränderungen zu rechnen ist. Weiters wurde geprüft, ob durch die Gestaltung des Antwortformates das Ausmaß eventueller Retest-Effekte vermindert werden kann. Zudem sollte ein erster Befund zur Kriteriumsvalidität des Tests erhoben werden. Dazu wurde eine repräsentative Itemstichprobe in identer Form in einem Messwiederholungsdesign bei einem Retestungsintervall von zwei Wochen vorgegeben.

Die Erhebungen erfolgten mit Einverständnis des Landesschulrates für Niederösterreich, der Eltern und aller Teilhabenden an höheren Schulen. Die Testungen wurden in Gruppen unter Wahrung der Anonymität der Schülerinnen und Schüler durchgeführt. Von den 12 untersuchungsrelevanten Aufgaben wurde eine Hälfte im Multiple-Choice-Format (1 aus 6) und die andere im freien Antwortformat jeweils ohne Zeitbegrenzung vorgegeben.

Das Verhalten einiger Testpersonen bei der Bearbeitung des Tests ließ auf eine nicht ausreichende Anstrengungsbereitschaft schließen. Um „nicht-real“ arbeitende Testpersonen von der Auswertung auszuschließen, wurden, inspiriert durch Schnipke und Scrams (1997), in einem ersten Schritt die Gesamtzeiten, welche die Testpersonen für die Bearbeitung der Aufgaben benötigten, und in einem zweiten die Item-Antwortzeiten analysiert. Dabei wurden in graphischen Analysen je-

weils jene Zeiten ermittelt, an denen ein Lösungsverhalten gegenüber einem Rateverhalten in den Vordergrund tritt. In die Auswertung gingen schlussendlich Daten von $N = 241$ Schülerinnen und Schüler im Alter von 14-20 Jahren ein.

Analysen mit dem Rasch-Modell für itemspezifische Veränderung belegen, dass in der Retestung dieselbe latente Dimension wie in der Initialtestung gemessen wurde. Ein Item musste aufgrund der fehlenden Passung ausgeschlossen werden, was bei einer neuerlichen Überprüfung des Itempools berücksichtigt werden sollte. Grundsätzlich ist dem ZART auf Basis der Ergebnisse eine Eignung für den wiederholten Einsatz zuzusprechen.

Modellierungen mit dem linearen logistischen Testmodell von Fischer (1973) sprechen für einen vom Antwortformat unabhängigen Retest-Effekt, um den sich die Schwierigkeiten der Aufgaben in der Retestung gleichermaßen verringern. Das Ausmaß dieses Effektes dürfte in den meisten Fällen aber praktisch uninteressant sein. Hinsichtlich der Bearbeitungszeit wurde eine deutliche Verringerung in der Retestung beobachtet.

Die Kriteriumsvalidität des ZART in Bezug auf die Jahresnote in Mathematik fällt mäßig aus. Der Korrelationskoeffizient nach Spearman beträgt $-0{,}31$.

Literaturverzeichnis

Amelang, M. & Bartussek, D. (2001). *Differentielle Psychologie und Persönlichkeitsforschung* (5. Aufl.). Stuttgart: Kohlhammer.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2011). *Standards for Educational and Psychological Testing* (6th ed.). Washington, DC: American Educational Research Association.

Anastasi, A. (1981). Coaching, Test Sophistication, and Developed Abilities. *American Psychologist, 36*(10), 1086-1093. Retrieved from http://ovidsp.tx.ovid.com/sp-3.8.1a/ovidweb.cgi

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*(1), 123-140. doi:10.1007/BF02291180

Arendasy, M. E. & Sommer, M. (2013). Quantitative differences in retest effects across different methods used to construct alternate test forms. *Intelligence, 41*(3), 181-192. doi:10.1016/j.intell.2013.02.004

Berndl, G., Steinfeld, J. & Poinstingl, H. (2012). Schlussfolgerndes Denken numerisch: Der Wiener Zahlenreihentest. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 161-169). Lengerich: Pabst.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

DIN Deutsches Institut für Normung e. V. (2002). *DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*(4), 407-433. doi:10.1007/BF02294564

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. Retrieved from http://www.gpower.hhu.de/

Feger, B. (1984). Die Generierung von Testitems zu Lehrtexten. *Diagnostica, 30*(1), 24-46.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

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

Fischer, G. H. (1977). Some Probabilistic Models for the Description of Attidudinal and Behavioral Changes Under the Influence of Mass Communication. In W. F. Kempf & B. H. Repp (Eds.), *Mathematical Models for Social Psychology* (pp. 102-151). Bern: Huber.

Fischer, G. H. (1995). Linear Logistic Models for Change. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (pp. 157-180). New York: Springer.

Glas, A. W. & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 69-95). New York: Springer.

Guthke, J. & Wiedl, K. H. (1996). *Dynamisches Testen: Zur Psychodiagnostik der intraindividuellen Variabilität*. Göttingen: Hogrefe.

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly, 50*(3), 379-390. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/3-2008/05_Hahne.pdf

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T. & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*(2), 373-385. doi:10.1037/0021-9010.92.2.373

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L. & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly, 50*(3), 391-402. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/3-2008/06_Hohensinn.pdf

Hornke, L. F. & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10*(4), 369-380. doi:10.1177/014662168601000405

International Test Commission (2001). International Guidelines for Test Use. *International Journal of Testing, 1*(2), 93-114. Retrieved from http://www.intestcom.org/upload/sitefiles/41.pdf

Khorramdel, L., Maurer, M., Frebort, M. & Kubinger, K. D. (2012). Ein Anforderungsprofil als Voraussetzung eines Self-Assessments zur Studienberatung - am Beispiel "Architektur". In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 49-62). Lengerich: Pabst.

Kubinger, K. D. (1979). Das Problemlöseverhalten bei der statistischen Auswertung psychologischer Experimente. Ein Beispiel hochschuldidaktischer Forschung. *Zeitschrift für Experimentelle und Angewandte Psychologie, 26*(3), 467-495.

Kubinger, K. D. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie. In K. D. Kubinger (Hrsg.), *Moderne Testtheorie: ein Abriß samt neuesten Beiträgen* (2. Aufl., S. 19-83). München: PVU.

Kubinger, K. D. (2005). Psychological Test Calibration Using the Rasch Model - Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5*(4), 377-394. doi:10.1207/s15327574ijt0504_3

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly, 50*(3), 311-327.

Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/3-2008/01_Kubinger.pdf

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

Kubinger, K. D., Frebort, M. & Müller, C. E. (2012). Self-Assessment im Rahmen der Studienberatung: Möglichkeiten und Grenzen. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 4-29). Lengerich: Pabst.

Kubinger, K. D. & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on different item response formats--An experiment in fundamental research on psychological assessment. *Psychology Science, 49*(4), 361-374. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/4-2007/05_Kubinger.pdf

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C. & Frebort, M. (2010). On Minimizing Guessing Effects on Multiple-Choice Items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment, 18*(1), 111-115. doi:10.1111/j.1468-2389.2010.00493.x

Kubinger, K. D., Rasch, D. & Yanagida, T. (2009). On designing data-sampling for Rasch model calibrating an achievement test. *Psychology Science Quarterly, 51*(4), 370-384. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/4-2009/psq_4_2009_370-384.pdf

Kubinger, K. D., Rasch, D. & Yanagida, T. (2011). *Statistik in der Psychologie. Vom Einführungskurs bis zur Dissertation*. Göttingen: Hogrefe.

Kulik, J. A., Kulik, C.-L. C. & Bangert, R. L. (1984). Effects of Practice on Aptitude and Achievement Test Scores. *American Educational Research Journal, 21*(2), 435-447. Retrieved from http://www.jstor.org/stable/1162453

Lievens, F., Buyse, T. & Sackett, P. R. (2005). Retest effects in operational selection settings: development and test of a framework. *Personnel Psychology, 58*(4), 981-1007. doi:10.1111/j.1744-6570.2005.00713.x

Lievens, F., Reeve, C. L. & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92*(6), 1672-1682. doi:10.1037/0021-9010.92.6.1672

Mair, P., Hatzinger, R. & Maier, M. J. (2014). eRm: Extended Rasch Modeling (R package version 0.15-4) [Computer Software]. Retrieved from http://erm.r-forge.r-project.org/

Martin-Löf, P. (1973). *Statistika modeller*. Anteckningar fran seminarier läsaret 1969-70 utarbetade av Rolf Sundberg. 2: a uppl. Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistik vid Stockholms Universitet.

Matton, N., Vautier, S. & Raufaste, E. (2009). Situational Effects May Account for Gain Scores in Cognitive Ability Testing: A Longitudinal SEM Approach. *Intelligence, 37*, 412-421. doi:10.1016/j.intell.2009.03.011

Mittring, G. & Rost, D. H. (2008). Die verflixten Distraktoren. Über den Nutzen einer theoretischen Distraktorenanalyse bei Matrizentests (für besser Begabte und Hochbegabte). *Diagnostica, 54*(4), 193-201. doi:10.1026/0012-1924.54.4.193

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika, 15*(3), 291-315. Retrieved from http://link.springer.com/content/pdf/10.1007%2FBF02289044.pdf

Mollenkopf, W. G. (1960). Time Limits and the Behavior of Test Takers. *Educational and Psychological Measurement, 20*(2), 223-230. doi:10.1177/001316446002000203

Müller, C. E. (2011). *Ermittlung prototypischer Testkennwerte für das Wiener Self-Assessment für Architektur anhand erfolgreich und wenig(er) erfolgreich Studierender*. Unveröffentlichte Diplomarbeit, Universität Wien.

Poinstingl, H. (2009a). The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychology Science Quarterly, 51*(2), 123-134. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/2-2009/02_Poinstingl.pdf

Poinstingl, H. (2009b). Software zum Zahlenreihentest (ZART) [Computer Software]. Unpublizierte Software.

Poinstingl, H. (2009c). *Zahlenreihentest (ZART)*. Manuskript in Vorbereitung.

Poinstingl, H. (2009d). *Remote Testing*. Presentation at the General Online Research 09, Vienna, April 2009.

Poinstingl, H. (2009e). *Eignungsdiagnostik mittels Remote Testing: am Beispiel des ZART*. Poster präsentiert bei der 6. Tagung der Fachgruppe Arbeits- und Organisationspsychologie der Deutschen Gesellschaft für Psychologie, Wien, September 2009.

Poinstingl, H. (2010a). *Der Remote Testing Ansatz am Beispiel des Zahlenreihentest ZART*. Poster präsentiert bei der 9. Tagung der Österreichischen Gesellschaft für Psychologie, Salzburg, April 2010.

Poinstingl, H. (2010b). *Konstruktion des Zahlenreihentest ZART anhand von meta-itemgenerierenden Regeln und erste psychometrische Untersuchungen*. Vortrag am 47. Kongress der Deutschen Gesellschaft für Psychologie, Bremen, September 2010.

Powers, D. A. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin, 100*, 67-77. Retrieved from http://ovidsp.tx.ovid.com/sp-3.8.1a/ovidweb.cgi

R Core Team (2014). R: A language and environment for statistical computing (Version 3.0.3) [Computer Software]. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Kopenhagen: Danish Institute for Educational Research.

Raymond, M. R., Neustel, S. & Anderson, D. (2009). Same-Form Retest Effects on Credentialing Examinations. *Educational Measurement: Issues and Practice, 28*(2), 19-27. doi:10.1111/j.1745-3992.2009.00144.x

Reeve, C. L. & Bonaccio, S. (2008). Does test anxiety induce measurement bias in cognitive ability tests? *Intelligence, 36*(6), 526-538. doi:10.1016/j.intell. 2007.11.003

Reeve, C. L. & Lam, H. (2005). The Psychometric Paradox of Practice Effects Due to Retesting: Measurement Invariance and Stable Ability Estimates in the Face of Observed Score Changes. *Intelligence, 33*, 535-549. doi:10.1016/ j.intell.2005.05.003

Reeve, C. L. & Lam, H. (2007). The Relation between Practice Effects, Test-Taker Characteristics and Degree of g-Saturation. *International Journal of Testing, 7*(2), 225-242. doi:10.1080/15305050701193595

Roediger, H. L., III & Karpicke, J. D. (2006). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science, 1*(3), 181-210. doi:10.1111/j.1745-6916.2006.00012.x

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2. Aufl.). Bern: Huber.

Ruch, W. W., Stang, S. W., McKillip, R. H. & Dye, D. A. (1994). *Employee Aptitude Survey. Technical Manual* (2nd ed.). Glendale, California: Psychological Services, Inc. Retrieved from http://stagecp.psionline.com/pdfs/ EASTechMan_3-3-94.pdf

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online, 8*(2), 23-74. Retrieved from http://www.dgps.de/fachgruppen/methoden/mpr-online/issue 20/art2/mpr130_13.pdf

Schnipke, D. L. & Scrams, D. J. (1997). Modeling Item Response Times with a Two-State Mixture Model: A New Method of Measuring Speededness. *Journal of Educational Measurement, 34*(3), 213-232. doi:10.1111/j.1745-3984.1997.tb00516.x

Schott, F. & Wieberg, H.-J. W. (1984). Regelgeleitete Itemkonstruktion. Ein Verfahren zur Definition von Itemuniversa und deren kontentvalider Abbildung in Itemmengen für Tests und Treatments. *Diagnostica, 30*(1), 47-66.

Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly, 50*(3), 345-362. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/3-2008/03_Sonnleitner.pdf.

Spearman, C. (1927). *The abilities of man: their nature and measurement*. London: Macmillan.

Stelzl, I. (1982). *Fehler und Fallen der Statistik - für Psychologen, Pädagogen und Sozialwissenschaftler*. Bern: Huber.

te Nijenhuis, J., van Vianen, A. E. M. & van der Flier, H. (2007). Score gains on *g*-loaded tests: No *g*. *Intelligence, 35*(3), 283-300. doi:10.1016/j.intell.2006.07.006

Vandenberg, R. J. & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods, 3*(1), 4-70. doi:10.1177/109442810031002

Weiß, R. H. (1987). *Wortschatztest (WS) und Zahlenfolgentest (ZF). Ergänzungstests zum Grundintelligenztest CFT 20. Handanweisung*. Göttingen: Hogrefe.

Anhang

A. Organisatorische Schreiben

Landesschulrat für Niederösterreich
Rennbahnstraße 29
3109 St. Pölten

Herrn
Gottfried Berndl
Per E-Mail: gottfried.berndl@univie.ac.at

Sachbearbeiterin:
Mag. Christina Unterberger

t: +43 2742 280 5370
f: +43 2742 280 1111
e: christina.unterberger@lsr-noe.gv.at

Beilage(n): 0

Präs.-420/466-2009

Datum: 14.12.2009

Betrifft:
Genehmigung der Durchführung einer empirischen Untersuchung

Der Landesschulrat für Niederösterreich genehmigt die Durchführung der vorgelegten empirischen Untersuchung zum Thema „Logisch-schlussfolgerndes Denken – Erprobung eines neuen Tests mit Zahlenreihen" durch Herrn Gottfried Berndl.

Die Untersuchung darf in dem vorliegenden Umfang antragsgemäß an folgenden Schulen in Niederösterreich durchgeführt werden:

Allgemeinbildende höhere Schulen:	Gymnasium Englische Fräulein der Vereinigung von Ordensschulen Österreichs, St. Pölten Bundesgymnasium und Bundesrealgymnasium Lilienfeld Bundesgymnasium und Bundesrealgymnasium Wolkersdorf Bundesgymnasium und Bundesrealgymnasium Tulln Bundesgymnasium und Bundesrealgymnasium Schwechat
Berufsbildende mittlere und höhere Schule:	Bundeshandelsakademie und Bundeshandelsschule St. Pölten

Parteienverkehr
Dienstag 8-12 Uhr

http://www.lsr-noe.gv.at
office@lsr-noe.gv.at
DVR:0064394

Amtsstunden
Mo.-Fr. 8-16 Uhr

Auf die Einhaltung der Datenschutzbestimmungen darf hingewiesen werden, außerdem ist vor Beginn der Erhebungen das Einverständnis der Eltern bzw. Erziehungsberechtigten und die Zustimmung der jeweiligen Direktion einzuholen. Die an dieser Untersuchung teilnehmenden SchülerInnen sind vor Beginn der Erhebung ausdrücklich auf die Freiwilligkeit ihrer Teilnahme hinzuweisen, außerdem ist deren Anonymität jedenfalls zu wahren. Es ist darauf zu achten, dass die Durchführung der Untersuchung jeweils längstens eine Unterrichtsstunde in Anspruch nimmt.

Für den Amtsführenden Präsidenten
Dr. F r e u d e n s p r u n g
Wirkl. Hofrat

Abbildung 16. Bescheid des Landesschulrates für Niederösterreich.

Gottfried Berndl
[redacted]
✉ [redacted]
Tel. [redacted]

Liebe Eltern,

Im Rahmen meiner Diplomarbeit an der Universität Wien führe ich derzeit eine Untersuchung an der Schule Ihres Kindes durch. Dabei geht es um die Erprobung eines neuen Tests, der zur Messung des logisch-schlussfolgernden Denkens dient.

Dazu möchte ich den Schülern und Schülerinnen zu zwei Zeitpunkten unterschiedliche Formen des sogenannten Zahlenreihentests zur Bearbeitung am Computer vorgeben. Jede Aufgabe des Tests besteht aus einer Zahlenreihe, in welcher eine Zahl fehlt. Aufgabe der Schüler und Schülerinnen ist es, die Regel, nach welcher die Zahlenreihe aufgebaut ist, zu finden, um die gesuchte Zahl zu bestimmen.

Beispiel: **1 2 4 7 11 (?) 22**

In diesem Beispiel wird die Regel „Addition mit einer schrittweise um eins ansteigenden Zahl" angewendet. Die gesuchte Zahl ist somit 16.

Für Ihr Kind kann die Testung eine gute Lernerfahrung für künftige Eignungstests darstellen. Zudem biete ich Ihrem Sohn / Ihrer Tochter eine individuelle Rückmeldung der Testergebnisse an, welche er / sie mittels eines selbst gewählten Passwortes über eine Internetseite abrufen kann.

Die Daten, die ich über ihr Kind sammle, werden selbstverständlich streng vertraulich behandelt und auch nicht an die Klassenlehrer bzw. die Schule weitergegeben. Ihr Kind wird zu keinem Zeitpunkt nach seinem Namen gefragt.

Ich bitte Sie um Ihr Einverständnis zur Teilnahme Ihres Kindes an der Untersuchung. Für Rückfragen stehe ich Ihnen unter der oben angegebenen E-Mail-Adresse bzw. Telefonnummer gerne zur Verfügung. Bitte füllen Sie in jedem Fall die beigefügte Einverständniserklärung aus.

Ich verbleibe mit Dank für Ihr Interesse und hoffe auf eine gute Kooperation!

Gottfried Berndl

✂ ----------

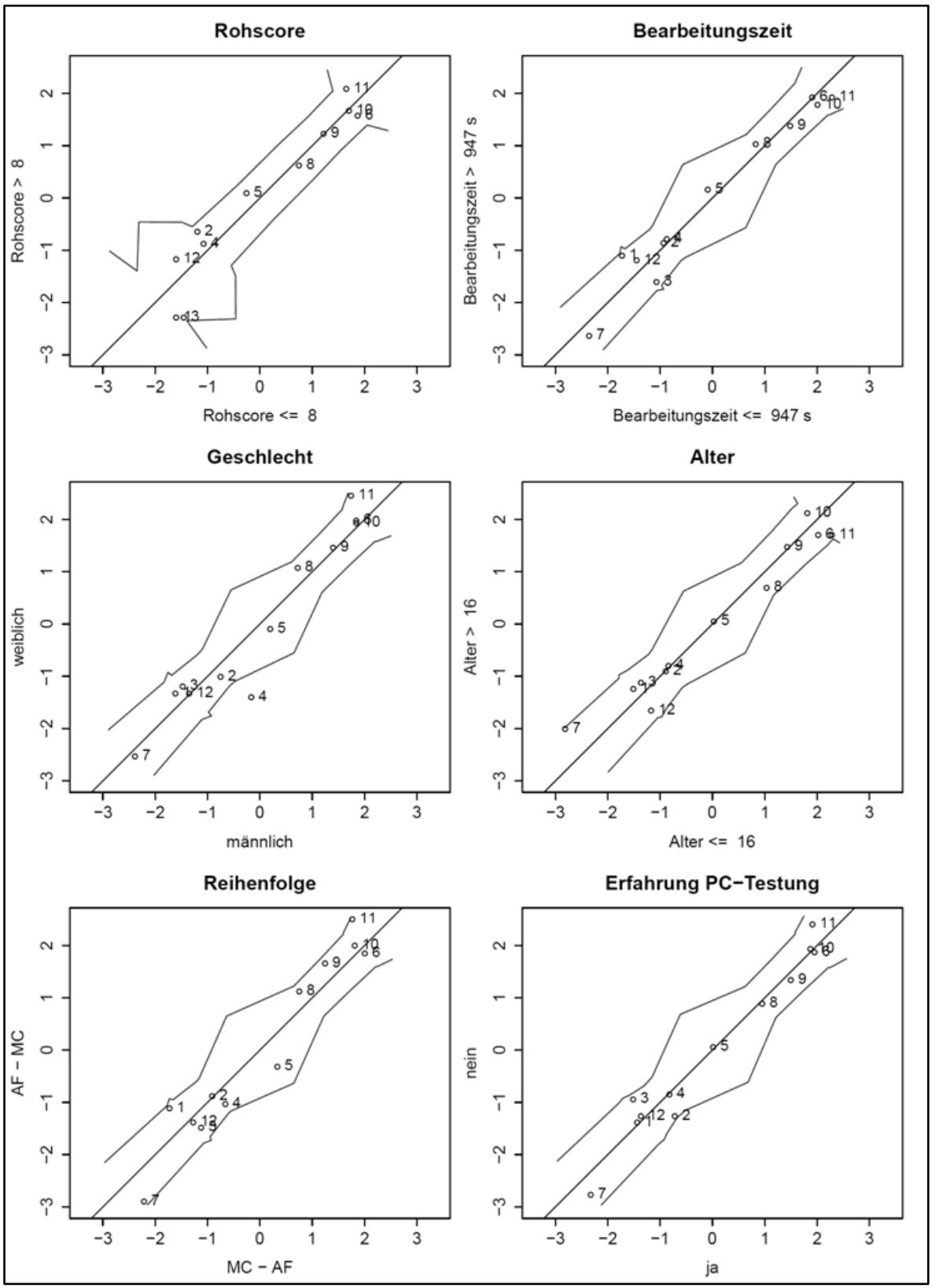
Ich bin mit der Teilnahme meines Sohnes / meiner Tochter _____
(bitte Namen eintragen) an der Untersuchung zum Zahlenreihentest einverstanden

❒ ja ❒ nein (Zutreffendes bitte ankreuzen)

_____
(Unterschrift)

Abbildung 17. Elternbrief.

B. Ergänzende Ergebnisdarstellungen



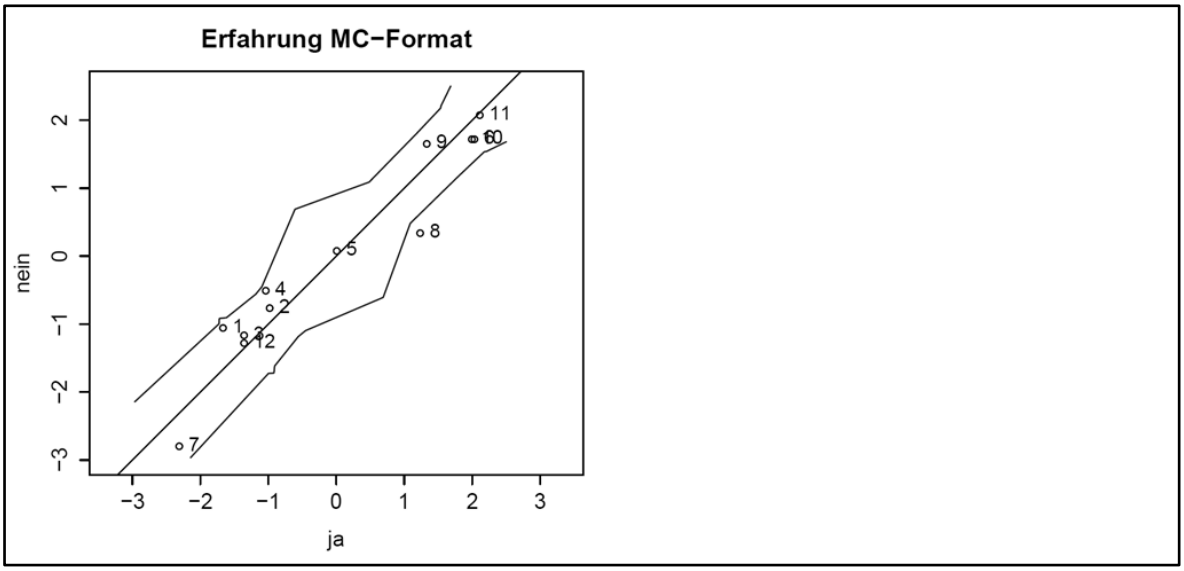


Abbildung 18. Grafische Modelltests zu den Rasch-Modell-Analysen zu Testzeitpunkt eins. Für das Teilungskriterium Rohscore wird Item 7 nicht dargestellt, da es in der Gruppe der Leistungsstärkeren von allen gelöst wurde. Die eingezeichneten Konfidenzbänder wurden für $\alpha = 0{,}05$ ermittelt.

Tabelle 11

Geschätzte Item-Schwierigkeitsparameter $\hat{\sigma}_i$, Standardschätzfehler und Konfidenzintervalle ($\alpha = 0{,}05$) aus den Rasch-Modell-Analysen zu Testzeitpunkt eins

	$\hat{\sigma}_i$	Std.schätzfehler	KI - untere Grenze	KI - obere Grenze
I1	-1,42	0,21	-1,83	-1,02
I2	-0,90	0,18	-1,26	-0,55
I3	-1,29	0,20	-1,68	-0,90
I4	-0,84	0,18	-1,18	-0,49
I5	0,02	0,15	-0,27	0,32
I6	1,91	0,16	1,60	2,22
I7	-2,47	0,30	-3,05	-1,89
I8	0,92	0,14	0,64	1,20
I9	1,43	0,15	1,14	1,72
I10	1,89	0,16	1,58	2,19
I11	2,09	0,16	1,77	2,40
I12	-1,33	0,20	-1,73	-0,94

Tabelle 12

Geschätzte Personenparameter $\hat{\xi}_v$ *und Standardschätzfehler aus den Rasch-Modell-Analysen zu Testzeitpunkt eins*

Rohscore	$\hat{\xi}_v$	Std.schätzfehler
2	-2,27	0,86
3	-1,62	0,77
4	-1,07	0,73
5	-0,55	0,72
6	-0,03	0,72
7	0,50	0,73
8	1,05	0,75
9	1,63	0,78
10	2,30	0,86
11	3,23	1,10
12	4,21	NA

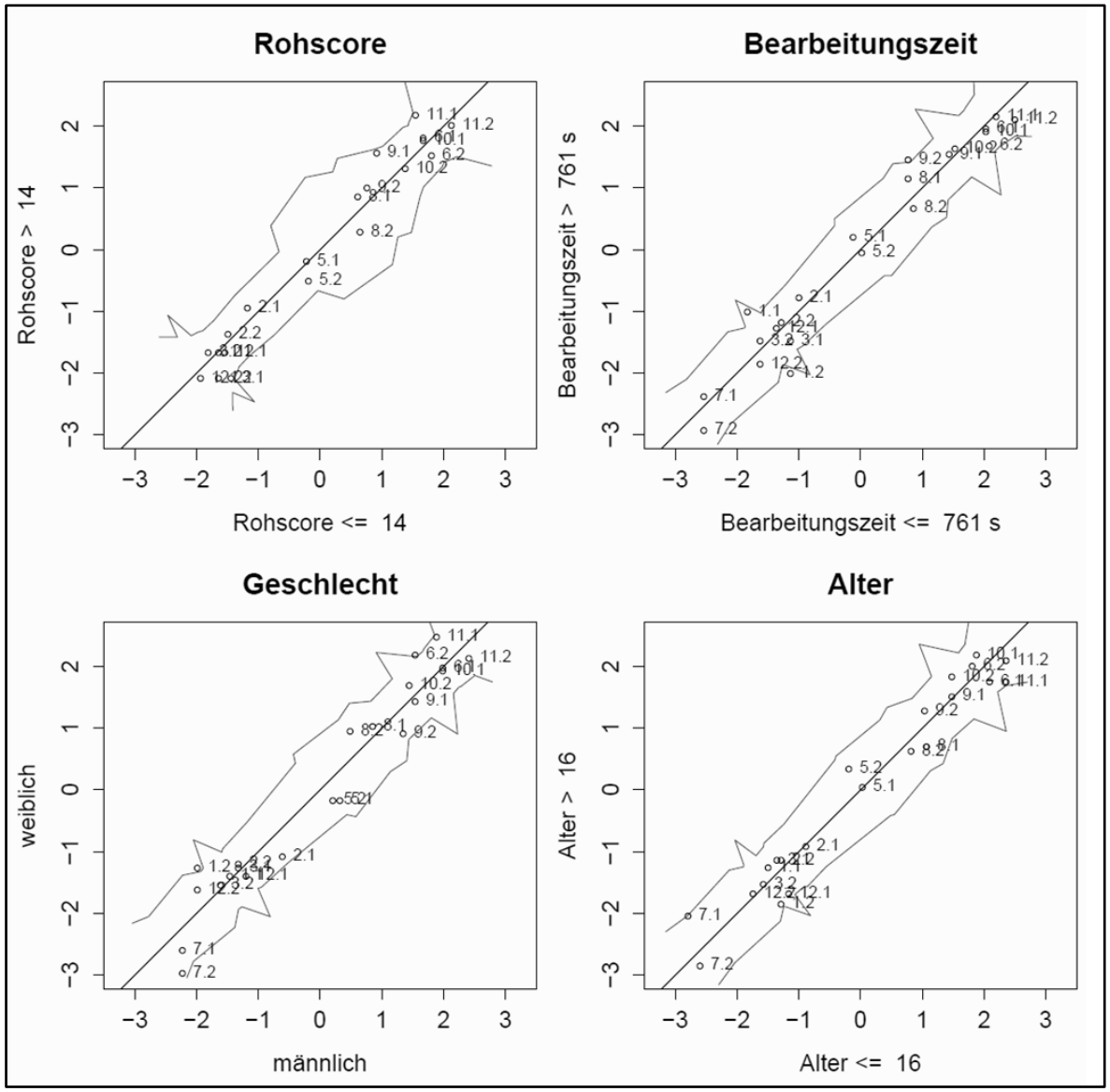


Abbildung 19. Grafische Modelltests zu den Rasch-Modell-Analysen nach Ausschluss von Item 4 zu beiden Testzeitpunkten. Für das Teilungskriterium Rohscore wird Item 7 nicht dargestellt, da es in der Gruppe der Leistungsstärkeren von allen gelöst wurde. Die eingezeichneten Konfidenzbänder wurden für $\alpha = 0{,}05$ ermittelt.

Tabelle 13

Geschätzte Item-Schwierigkeitsparameter $\hat{\sigma}_{it}$, Standardschätzfehler und Konfidenzintervalle ($\alpha = 0{,}05$) aus den Rasch-Modell-Analysen für itemspezifische Veränderung nach Ausschluss von Item 4

	$\hat{\sigma}_{it}$	Std.schätzfehler	KI - untere Grenze	KI - obere Grenze
I1.1	-1,42	0,21	-1,83	-1,00
I2.1	-0,90	0,18	-1,26	-0,54
I3.1	-1,28	0,20	-1,68	-0,89
I5.1	0,03	0,15	-0,28	0,33
I6.1	1,98	0,16	1,67	2,29
I7.1	-2,47	0,31	-3,07	-1,87
I8.1	0,95	0,15	0,66	1,24
I9.1	1,48	0,15	1,18	1,78
I10.1	1,95	0,16	1,64	2,27
I11.1	2,16	0,16	1,84	2,48
I12.1	-1,33	0,21	-1,73	-0,92
I1.2	-1,46	0,21	-1,88	-1,04
I2.2	-1,24	0,20	-1,63	-0,85
I3.2	-1,56	0,22	-2,00	-1,13
I5.2	-0,02	0,16	-0,33	0,28
I6.2	1,85	0,16	1,55	2,16
I7.2	-2,69	0,34	-3,35	-2,03
I8.2	0,75	0,15	0,46	1,04
I9.2	1,10	0,15	0,81	1,39
I10.2	1,57	0,15	1,27	1,87
I11.2	2,27	0,17	1,94	2,60
I12.2	-1,72	0,23	-2,18	-1,27

Tabelle 14

Strukturmatrizen der LLTM-Modelle 1-3

Modell 1

Modell 2

Modell 3

	σ_1	σ_2	σ_3	σ_5	σ_6	σ_7	σ_8	σ_9	σ_{10}	σ_{11}	σ_{12}	λ	γ
I1.1	1	0	0	0	0	0	0	0	0	0	0	0	0
I2.1	0	1	0	0	0	0	0	0	0	0	0	0	0
I3.1	0	0	1	0	0	0	0	0	0	0	0	0	0
I5.1	0	0	0	1	0	0	0	0	0	0	0	0	0
I6.1	0	0	0	0	1	0	0	0	0	0	0	0	0
I7.1	0	0	0	0	0	1	0	0	0	0	0	0	0
I8.1	0	0	0	0	0	0	1	0	0	0	0	0	0
I9.1	0	0	0	0	0	0	0	1	0	0	0	0	0
I10.1	0	0	0	0	0	0	0	0	1	0	0	0	0
I11.1	0	0	0	0	0	0	0	0	0	1	0	0	0
I12.1	0	0	0	0	0	0	0	0	0	0	1	0	0
I1.2	1	0	0	0	0	0	0	0	0	0	0	1	1
I2.2	0	1	0	0	0	0	0	0	0	0	0	1	1
I3.2	0	0	1	0	0	0	0	0	0	0	0	1	1
I5.2	0	0	0	1	0	0	0	0	0	0	0	1	1
I6.2	0	0	0	0	1	0	0	0	0	0	0	1	1
I7.2	0	0	0	0	0	1	0	0	0	0	0	1	0
I8.2	0	0	0	0	0	0	1	0	0	0	0	1	0
I9.2	0	0	0	0	0	0	0	1	0	0	0	1	0
I10.2	0	0	0	0	0	0	0	0	1	0	0	1	0
I11.2	0	0	0	0	0	0	0	0	0	1	0	1	0
I12.2	0	0	0	0	0	0	0	0	0	0	1	1	0

Anmerkungen. Basisparameter: σ_1 - σ_3 und σ_5 - σ_{12}…Ausgangsschwierigkeiten der Items, λ…allgemeiner Wiederholungsparameter bzw. Retest-Effekt, γ…für das MC-Format spezifischer Retest-Effekt.

Tabelle 15

Gegenüberstellung der Schätzungen der Itemparameter aus dem Rasch-Modell für itemspezifische Veränderung ohne Item 4 (Modell 0) und LLTM-Modell 2

	Modell 0	Modell 2
I1.1	-1,42	-1,44
I2.1	-0,90	-1,07
I3.1	-1,28	-1,42
I5.1	0,03	0,00
I6.1	1,98	1,92
I7.1	-2,47	-2,58
I8.1	0,95	0,85
I9.1	1,48	1,29
I10.1	1,95	1,76
I11.1	2,16	2,21
I12.1	-1,33	-1,52
I1.2	-1,46	-1,65
I2.2	-1,24	-1,27
I3.2	-1,56	-1,62
I5.2	-0,02	-0,20
I6.2	1,85	1,71
I7.2	-2,69	-2,78
I8.2	0,75	0,65
I9.2	1,10	1,08
I10.2	1,57	1,55
I11.2	2,27	2,01
I12.2	-1,72	-1,72

Anmerkungen. Die Itemparameter aus Modell 2 ergeben sich aus der Modellgleichung von Modell 2: $\sigma_i + q_t\lambda$, wobei σ_i die Ausgangsschwierigkeit der Items darstellt, λ den Wiederholungsparameter und q_t zu Zeitpunkt eins gleich null und zu Zeitpunkt zwei gleich eins ist. Die Itemparameter zu Zeitpunkt eins (I1.1 bis I12.1) für Modell 2 entsprechen somit gleichzeitig den Basisparametern σ_1 - σ_3 und σ_5 - σ_{12}. Die Darstellung erfolgt wiederum in Item-Schwierigkeitsparametern.

Lebenslauf

Persönliche Angaben

Gottfried Berndl
geboren am 5. März 1981 in St.Pölten
österreichischer Staatsbürger
gottfried.berndl@gmx.net

Ausbildung

2002 – aktuell	Diplomstudium Psychologie, Universität Wien Schwerpunkt: Psychologische Diagnostik
2001 – 2002	Studienzweig Betriebswirtschaft, WU Wien
1995 – 2000	Höhere Technische Lehr- und Versuchsanstalt St.Pölten, Abteilung Elektrotechnik Reifeprüfung mit Auszeichnung am 5. Juni 2000
1987 – 1995	Volks- und Hauptschule, Ober-Grafendorf

Fachbezogene Berufserfahrung

08/2013	psychologische Betreuung von Kindern bei einer Lernwoche (bei Mag. Flickinger und Dr. Koch)
01/2011 – 06/2011	Praktikum und ehrenamtliche Tätigkeiten an der neuroonkologischen Ambulanz, Universitätsklinik für Kinder- und Jugendheilkunde, AKH Wien (bei Mag. Pletschko und Dr. Leiss)
10/2009 – 06/2013	Studienassistent am Arbeitsbereich Psychologische Diagnostik
04/2009 – 07/2009	Praktikum am Arbeitsbereich Psychologische Diagnostik, Fakultät für Psychologie (bei Mag. Poinstingl)

Spezielle Weiterbildungen

seit 2010	zertifiziert zum AID 2 Testleiter