universität wien

Diplomarbeit

Entwicklung eines Itempools für den „Mental Arithmetic Test“ (MAT)

Nina Franz

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, im April 2009

Studienkennzahl: 298

Studienrichtung: Psychologie

Betreuer: Ao.Univ.-Prof. Dr. Georg Gittler

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Numerische Fähigkeiten stellen eine wesentliche Dimension für die Messung intelligenten Verhaltens dar, wie schon Thurstone anhand seiner sieben Primärfaktoren postulierte. Nach Thurstone setzt sich intelligentes Verhalten aus den Faktoren „verbal comprehension“, „word fluency“, „number“, „space“, „memory“, „perceptual speed“ und „reasoning“ zusammen. Auch Vernon erfasst in seinem hierarchischen Intelligenz-Modell, neben dem g-Faktor als höchsten Allgemeinheitsgrad der Intelligenz und den „major group factors“ beziehungsweise den „minor group factors“ auch „mathematical abilities“ als einen der spezifischen Faktoren auf der untersten Ebene der Intelligenz (Amelang, Bartussek, Stemmler & Hagemann, 2006).
„Als wichtige Determinanten der Rechenleistung werden derzeit Arbeitsgedächtnis, basales arithmetisches Faktenwissen und die Durchführung von Rechenprozeduren in Betracht gezogen“ (Geary, Brown & Samaranayake, 1991, nach Grube & Barth, 2004, S. 246).

Das dieser Arbeit zugrunde liegende Messinstrument wird als „Mental Arithmetic Test“ (kurz MAT) bezeichnet und ist ein neu entwickeltes Verfahren zur Erfassung numerischer Intelligenz. Zur Lösung der Aufgaben werden hohe Ansprüche an das Gedächtnis gestellt. Es ist grundlegendes mathematisches Wissen von Nöten wie auch logisches Denken. Der Test erfasst somit die Fähigkeitsdimension Kopfrechnen.
Es wird versucht auf kognitive Modelle und Forschungsansätze einzugehen, die für die Aufgabenbewältigung des MAT Relevanz besitzen. So wird vor allem die ACT-R Theorie von Anderson und Lebiere (1998), wie auch deren Verbindungen zum Arbeitsgedächtnismodell von Baddeley (1986) als die zugrundeliegenden kognitiven Modelle erläutert. Da es sich beim MAT um ein nach dem Rasch-Modell konstruiertes Ver-

fahren handelt, werden in Folge die Unterschiede zwischen Klassischer und Probabilistischer Testtheorie thematisiert, wie auch im Anschluss auf das Rasch-Modell und seinen zu prüfenden Forderungen eingegangen wird.

Das vorrangige Interesse dieser Arbeit dient der Evaluierung des Itempools des MAT, wobei die bestehenden Poolaufgaben analysiert und auf Rasch-Homogenität geprüft werden. Ab Kapitel 7 wird auf die Poolaufgaben sowie auf zugrundeliegende Theorien bezüglich der Itemkonstruktion eingegangen. Weiters wird die Stichprobe beschrieben und der Testvorgang erläutert. Es folgt die Prüfung auf Homogenität mit anschließenden Erläuterungen und Interpretation der Ergebnisse.

2 ACT-R Theorie (Anderson & Lebiere, 1998)

In der Arbeit von Arendasy, Sommer und Hergovich (2007) wird angenommen, dass für die mehr oder weniger schwierige Ermittlung von mathematischen Ergebnissen besonders Übungseffekte und Abrufmöglichkeiten von Bedeutung sind:

> Die Leichtigkeit der Ermittlung der lösungsrelevanten Zwischenergebnisse hängt dabei sowohl von der Vertrautheit mit den Rechenoperationen als auch von der Möglichkeit eines Rückgriffs auf lösungsrelevante Ergebnisse im Langzeitgedächtnis ab…[]
> Nach Ashcraft (1995, nach Arendasy, Sommer & Hergovich, 2007, S. 121) und Campbell und Xu (2001, nach Arendasy, Sommer & Hergovich, 2007, S. 121) sind die vier Grundrechnungsarten in unterschiedlichem Ausmaß durch Übung vertraut, wobei Additionen und Subtraktionen im Allgemeinen einfacher fallen als Multiplikationen und Divisionen (Arendasy, Sommer & Hergovich, 2007, S. 121).

Einen entscheidenden Einfluss bei der Bearbeitung mentaler arithmetischer Aufgaben spielt das Gedächtnis, wobei Baddeley erklärt:

> It is often useful to separate out three aspects of any memory system: encoding, the process whereby information is registered; storage, the maintenance of information over time; and retrieval, which refers to the accessing of the information by recognition, recall or implicitly by demonstrating that a relevant task is performed more efficiently as a result of prior experience (Baddeley, 2002, S. 9).

Die meisten der veröffentlichten Studien bezüglich mentaler Rechenaufgaben und der zugrundeliegenden Denkprozesse ähneln sich in ihren Ansätzen deutlich.

Einer der eher jüngeren Erklärungsansätze ist die ACT-R-Theorie („adaptive control of thought-rational“) von Anderson und Lebiere (1998). Es handelt sich hierbei um eine kognitive Theorie über grundlegende Strukturen und Prozesse des Denkens. Anderson schreibt folgendermaßen:

> According to the ACT-R theory, the power of human cognition depends on the amount of knowledge encoded and the effective deployment of the encoded knowledge…[]
> All that there is to intelligence is the simple accrual and tuning of many small units of knowlegde that in total produce complex cognition. The whole

> is no more than the sum of its parts, but it has a lot of parts (Anderson, 1996, S. 355ff).

Die Vorteile der ACT-R-Theorie liegen laut Newell (1990) in der guten Anwendbarkeit auf reelle Probleme und auch der Bezugnahme auf neurowissenschaftliche Daten.

Ihre Ursprünge hat die ACT-Theorie in der „human associative memory (HAM) theory of human memory“ von Anderson und Bower (1973), deren Versuch es war eine Theorie über die Repräsentation von Erinnerungen zu entwickeln (Anderson, 1996). Die HAM-Theorie beschäftigte sich jedoch nur mit dem deklarativen Wissen. Anderson postulierte somit die Unterscheidung zwischen deklarativen und prozeduralem Wissen. Es kam die Idee auf, dass prozedurales Wissen aus sogenannten Produktionsregeln bestehe, was das Produktionssystem ACTE zur Folge hatte. Sieben Jahre wurde fortan mit diesem System gearbeitet, bis es zu einer Weiterentwicklung zur ACT kam (Anderson, 1983), eine Theorie, welche Vermutungen darüber anstellt, wie Produktionsregeln erworben werden. Nach dessen 10jährigen Gebrauch ist das nun aktuelle System die ACT-R (Anderson, 1993b). Dieses System dient als Computersimulationsprogramm welches die Simulation menschlichen Denkens in seiner Umgebung ermöglicht (Anderson, 1996).

2.1 Genereller Aufbau von ACT-R

Anderson (1996) unterscheidet grundsätzlich zwischen deklarativen und prozeduralem Wissen. Das prozedurale Wissen wird durch sogenannte „Produktionsregeln“ dargestellt, das deklarative hingegen besteht aus Einheiten, die „Chunks“ genannt werden. Es ist zu beachten, dass die Bezeichnung „Chunk“ in der ACT-R-Theorie nicht gleichbedeutend ist mit jenen Einheiten, wie sie bei Miller (1956, nach Anderson & Matessa, 1997) verwendet werden (geschätzte Behaltenskapazität des Kurzzeitgedächtnisses von 7±2 Chunks)(Anderson & Matessa, 1997).

Chunks sind schemaartig strukturiert und können graphisch dargestellt werden wie Abbildung 1 zeigt.

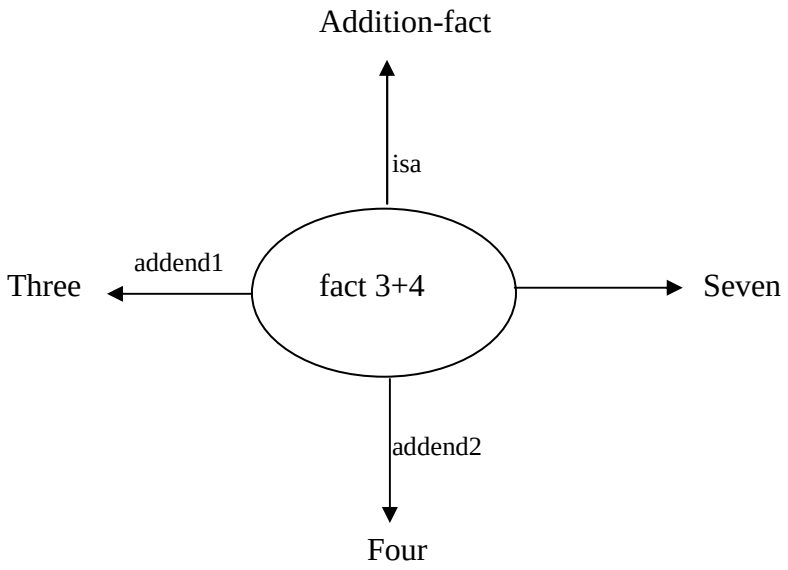

Abbildung 1: graphische Darstellung eines Chunks mit dem Additionsfaktum „3 + 4 = 7“ (Anderson, 1996, S. 356)

Produktionsregeln antworten auf spezifische Ziele und involvieren auch die Neubildung von Teilzielen. Sie beinhalten stets eine Bedingung und eine Aktion:

> IF the goal is to add n1 and n2 in a column
> and n1 + n2 = n3
> THEN set as a subgoal to write n3 in that column (Anderson, 1996, S. 357)

Angenommen es handelt sich um folgende Additionsaufgabe:

$$\begin{array}{r} 531 \\ +248 \\ \hline . \; . \; 9 \end{array}$$

Betrachtet man die Zehnerstelle, würde im Bedingungsteil der Produktionsregel der deklarative Chunk (die Lösung „7“) aufgerufen werden um das gewünschte Teilziel zu erhalten (Anderson, 1996). Wird eine Aufgabe gelöst, stellt diese ein neues Faktum dar, welches wiederum aus dem deklarativen Gedächtnis abgerufen werden kann. Je

öfter ein Chunk benutzt wird, desto besser kann er auch abgerufen werden.

> …, the activation of a chunk is a sum of a base-level activation, reflecting its general usefulness in the past, and an associative activation, reflecting its relevance to the current context. …The activation of a chunk controls both its probability of being retrieved and its speed of retrieval (Anderson, Bothell, Byrne, Douglass, Lebiere & Quin, 2004, S. 1042).

Produktionsregeln können somit nur angewendet werden, wenn das dafür notwendige Wissen im deklarativen Gedächtnis verfügbar ist. Deklaratives und prozedurales Wissen sind also sehr eng miteinander verbunden.

Nach Anderson et. al. (2004) besteht ACT-R aus unterschiedlichen Modulen (siehe Abb. 2) die jeweils für verschiedene Informationsarten zuständig sind. Das visuelle Modul identifiziert Objekte im visuellen Bereich, das manuelle Modul hat kontrollierende Funktion für die Handbewegungen, das deklarative Modul ist für die Wiedergabe von Information aus dem Gedächtnis zuständig und das Zielmodul ist vorhanden um Intentionen aufrecht zu erhalten. Wie aus der Abbildung 2 weiters zu ersehen ist, bestehen nach neurowissenschaftlichen Theorien Verbindungen zu bestimmten kortikalen Regionen, auf die hier jedoch nicht weiter eingegangen wird.

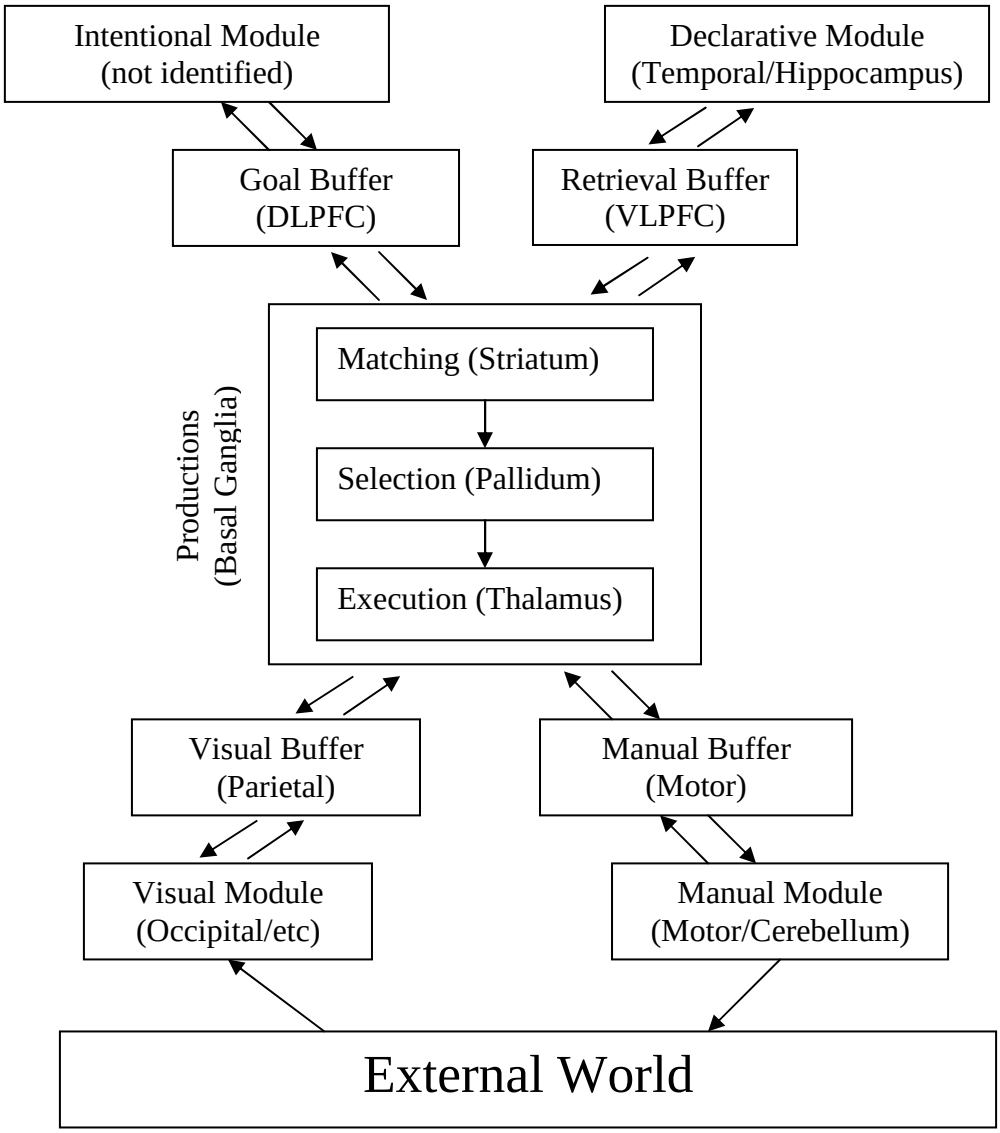

Abbildung 2 : Aufbau der ACT-R-Theorie. DLPFC = dorsolateraler präfrontaler Cortex; VLPFC = ventrolateraler präfrontaler Cortex. (Anderson et. al., 2004, S. 1037)

Die Module enthalten die oben erläuterten Chunks im Zwischenspeicher, die sogenannten „Puffer“, wo die Chunks durch das Produktionssystem erkannt werden. Das zentrale Produktionssystem koordiniert somit das Verhalten durch die Module (Anderson et. al., 2004).

„The production system can detect the patterns that appear in these buffers and decide what to do next to achieve coherent behaviour” (Anderson et. al, 2004, S. 1044).

Der grundlegende Zyklus des zentralen Produktionssystems besteht aus der Zusammenarbeit aller Puffer (Taatgen, Van Rijn, Anderson, 2007).

Eine wichtige Funktion der Produktionsregeln ist die Aktualisierung der Puffer im gesamten ACT-R-System. Das zentrale Produktionssystem ist jedoch nicht sensibel für alle Aktivitäten der Module. Es kann lediglich einen limitierten Informationsanteil verarbeiten, welcher sich im Puffer der Module befindet. Diese Puffer haben Ähnlichkeit zu Baddeley's (1986) „Slave"-Systemen im Arbeitsgedächtnis (Anderson et. al., 2004). Nach Baddeley ist die zentrale Exekutive für den Abruf von benötigten Informationen aus dem Langzeitspeicher zuständig. Die zentrale Exekutive ist ein „Aufmerksamkeitssystem mit begrenzter Kapazität und limitierten Verarbeitungsressourcen". „Bei hoher Anforderung gibt sie Aufgaben an die beiden „Slave"-Systeme weiter" (Gabriel, 2004, S. 20).

2.2 ACT-R im Bezug auf Kopfrechnen (Lebiere, 1999)

Das ACT-R Modell kann, da es ein allgemeines kognitives Modell darstellt, sehr gut auch zur Erklärung von kognitiven Prozessen im numerischen Bereich herangezogen werden.

> Cognitive arithmetic studies the mental representation of numbers and arithmetic facts (counting, addition, subtraction, multiplication, division) and the processes that create, access, and manipulate them. Arithmetic is one of the fundamental cognitive tasks which humans have to master (Lebiere, 1999, S. 5).

Das Lernen mit Zahlen umzugehen bedarf verschiedener Kapazitäten, wie das reine Auswendiglernen und Behalten von Fakten (zum Beispiel das „1*1") und auch stets die Kombination von „Ausprobieren" und „Berechnen".

Es sind in der Regel zwei Klassen an empirischen Phänomenen in der kognitiven Arithmetik zu beobachten. Die Erste bezieht sich auf den Umstand, dass besonders bei Kindern und zu einem bestimmten Teil

auch bei Erwachsenen zwei Lösungsstrategien angewendet werden: der „Abruf aus dem Gedächtnis“ und die „Berechnung“. Einerseits werden die Lösungen zu bestimmten Aufgaben einfach aus dem Gedächtnis abgerufen und andererseits, wenn keine adäquaten Lösungsvorschläge im Gedächtnis zu finden sind, wird die Lösung errechnet. Unter normalen Umständen steigt die kognitive arithmetische Leistung über die Jahre hinweg bis fast zur Gänze mit der Abrufstrategie gearbeitet wird und nicht mehr jeder Schritt iterativ errechnet werden muss.

Das zweite Phänomen involviert die Größe der Aufgaben, den „problem-size effect“. Kinder und Erwachsene benötigen demnach längere Lösungszeiten für Aufgaben, die größere Zahlen enthalten und machen auch mehr Fehler. Betrachtet man dieses Phänomen im Zuge der Berechnungsstrategie erscheint diese Tatsache als einleuchtend, da mehr Rechen- und damit auch mehr Zeitaufwand entsteht wenn man hohe Zahlen miteinander multipliziert oder addiert, was wiederum eine höhere Fehleranfälligkeit bedeutet.
Dieser problem-size-Effekt tritt, wenn auch in einem sehr viel geringerem Ausmaß, auch bei Erwachsenen auf. Es wird angenommen, dass der Grund im verbleibenden Rest an Berechnungsstrategien liegt. Gewisse Studien sprechen jedoch gegen diesen Umstand (Lebiere, 1999). Ein weiteres plausibleres Argument ist, dass kleinere Problemstellungen öfters vorkommen, dadurch häufiger Anwendung finden und daher bei derartigen Aufgaben auch bessere Leistungen erzielt werden.

Im Kindesalter sind im deklarativen Gedächtnis erstmals nur die erlernten Zahlen mit ihren Nachfolgern enthalten mit gewissen einfachen Produktionsregeln. Durch iterative Rechenmethoden können die ersten Berechnungen angestellt werden, die als Chunks dienbar sind und somit theoretisch wieder aus dem deklarativen Gedächtnis abgerufen werden können. Zu Beginn müssen Aufgaben natürlich mit der Berechnungs-strategie bearbeitet werden und sind – auch nach mehrmaligem Rechnen – noch zu schwach im Gedächtnis verankert, sodass die Wahrscheinlich-

keit, dass sie abgerufen werden, nur gering ist. Erst durch häufigere Wiederholungen gewinnen die Chunks an Stärke, sodass sie für die Abrufstrategie verwendet werden können. Chunks erweitern sich dadurch und verschmelzen gegebenenfalls mit anderen Chunks. Zufallsschwankungen spielen dabei eine wesentliche Rolle. Es ist somit auch denkbar, dass falsche Fakten eingespeichert werden und somit fehlerhafte Chunks im deklarativen Gedächtnis vorhanden sind. Aus diesem Grund ist es besonders wichtig, dass inkorrekte Rechenvorgänge so früh wie möglich korrigiert werden.

Viele der jüngsten Entwicklungen der ACT-R-Theorie beschäftigten sich besonders mit mathematischen Problemlösekompetenzen. In seinen Ursprüngen jedoch konzentrierte man sich auf das menschliche Gedächtnis (Anderson & Matessa, 1997).

2.3 Zusammenhänge zwischen ACT-R und Baddeley's Arbeitsgedächtnistheorie

Wie bereits oben kurz erwähnt gibt es gewisse Ähnlichkeiten der ACT-R Theorie mit Baddeley's Arbeitsgedächtnistheorie (1986). Nach Anderson und Lebiere (2003) entsprechen die Zwischenspeicher (Puffer) im ACT-R-Modell den Subsystemen von Baddeley's Arbeitsgedächtnismodell, nämlich der phonologischen Schleife und dem visuell räumlichen Notizblock, die auch „Slave-Systeme" genannt werden.

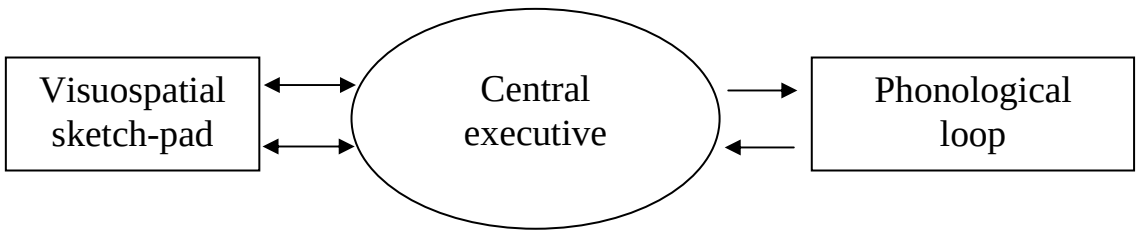

Abbildung 3: das Arbeitsgedächtnismodell von Baddeley und Hitch (1974)

Anderson und Matessa konzentrieren sich in ihrer Publikation von 1997 unter anderem auf Erkenntnisse von Baddeley's Theorie über den zeitabhängigen Effekt des Vergessens und versuchen anhand der ACT-R-Theorie gewisse Schwächen der Theorie aufzufangen. Hier soll jedoch nicht weiter auf methodische Probleme eingegangen werden. Vielmehr wird versucht auf die Zusammenhänge zwischen den beiden Theorien einzugehen.

So wie die Puffer der ACT-R Theorie gewisse Gedächtniselemente speichern und bereit halten, sind auch die beiden Subsysteme von Baddeley und Hitch für die Zwischenspeicherung und Bereitstellung von Information zuständig.

Nach Baddeley (1998) steuert die zentrale Exekutive die bewusste Informationsverarbeitung und koordiniert die Aktionen der phonologischen Schleife und des visuell räumlichen Notizblockes. Die phonologische Schleife ist an der Speicherung und Manipulation von verbaler Information beteiligt, während der visuell räumliche Notizblock für die zeitlich begrenzte Speicherung von visueller und räumlicher Information zuständig ist. Von diesen drei Systemen ist die phonologische Schleife die am meisten untersuchte. Es wird angenommen, dass die phonologische Schleife aus zwei Komponenten besteht, einem temporalen Speicher und einem Prozess der subvokalen Wiederholung.

„Beim Kopfrechnen ist die zentrale Exekutive für die Gesamtkoordination der notwendigen Verarbeitungsprozesse und die phonologische Schleife insbesondere für das Behalten von Aufgabeninformation und Zwischenergebnissen verantwortlich" (vgl. Fürst & Hitch, 2000; Seitz & Schumann-Hengsteler, 2000, nach Grube & Barth, 2004, S. 246).

Kinder benützen am Anfang ihrer Schulkarriere für das Lösen von Rechenaufgaben vorwiegend „verbale Zählstrategien, was die phono-

logische Schleife und die zentrale Exekutive belastet“ (Grube & Barth, 2004, S. 246). Diese Erkenntnis entspricht weitgehend der oben angeführten Rechenstrategie, die laut ACT-R-Theorie überwiegend von Kindern benützt wird, der Berechnungsstrategie. Erst mit zunehmender Übung muss weniger gerechnet und kann mehr aus dem Arbeitsgedächtnis bzw. aus dem deklarativen Gedächtnis abgerufen werden.

Weitere inhaltlich relevante Theorien zu mentalen Rechenprozessen werden aus praktischen Gründen bei der Beschreibung und Itemgeneration des MAT (Kapitel 7), wie auch bei der Interpretation der Ergebnisse selbst angeführt.

3 Klassische Verfahren, die numerische Fähigkeit prüfen

Die numerische Intelligenz wird oft anhand von Untertests in größeren Testbatterien erhoben wie dem Intelligenz-Struktur-Test 2000 R (IST 2000 R) von Amthauer, Brocke, Liepmann und Beauducel (2001), welcher auf dem theoretischen Konzept von Thurstone und Cattell aufbaut (Kubinger, 2006), das Adaptive Intelligenz Diagnostikum (AID 2) von Kubinger und Wurst (2000), welches auf dem Testkonzept von Wechsler und nach dem Rasch Modell konstruiert worden ist oder der Intelligenz-Struktur-Analyse (ISA) von Blum, Didi, Fay, Maichle, Trost, Wahlen und Gittler (1998), wobei bei den gängigen Testbatterien, wie auch den oben genannten, Gedächtnis von mathematischer Fähigkeit getrennt erhoben wird. Die numerischen Intelligenzfunktionen werden z.B. beim ISA durch die beiden Subtests „Praktisches Rechnen“ und „Zahlenreihen fortsetzen“, beim IST 2000 durch die Untertests „Rechenaufgaben“, „Zahlenreihen“ und „Rechenzeichen“ und beim AID 2 durch die Untertests „Angewandtes Rechnen“ und „Unmittelbares Reproduzieren-numerisch“ erfasst.

Laut Manual von z.B. des IST 2000 R soll der Untertest „Rechenaufgaben“ praktisch-rechnerisches Denken bzw. schlussfolgerndes Denken erheben, „Zahlenreihen“ das theoretisch-rechnerische Denken, induktives Denken bzw. die Beweglichkeit und Umstellfähigkeit im Denken ansprechen und der Untertest „Rechenzeichen“ soll ebenfalls schlussfolgerndes Denken erfassen.

Auf der anderen Seite gibt es auch einzelne Verfahren, die, wie auch in den obigen Testbatterien, in der pädagogischen bzw. Förderdiagnostik eingesetzt werden um den jeweiligen Entwicklungs- bzw. Wissensstand der Kinder und Jugendlichen zu erheben, wobei hier nur wenige aus der weitreichenden Palette kurz beschrieben werden:

Der MT-GMS 5 (Mathematiktest – Größen und Maßsysteme für das 5. Schuljahr an Hauptschulen) von Kueffner, H. (1980) ist ein Verfahren um gewisse Fehlertendenzen in den Mathematikaufgaben zu identifizieren (wie zum Beispiel das Finden von Fehlern bei der Verwendung von Längenmaßen) und die Ergebnisse zur Förderung des jeweiligen Schülers heranziehen. Es handelt sich um einen Multiple-Choice-Test mit mathematischen Aufgaben aus dem Alltag welcher als Einzel- oder Gruppentest durchgeführt werden kann

Der DEMAT 1+ (Deutscher Mathematiktest für erste Klassen) von Krajewski, Kuespert, und Schneider (2002), welchen es noch in anderen Klassenstufen gibt, wurde von Lehrplänen der 16 deutschen Bundesländer erstellt und soll die mathematische Kompetenz von Grundschülern erfassen, wobei vorwiegend die Grundrechenarten geprüft werden (Testerklärungen verfügbar unter: http://dbs.univie.ac.at/ suchtitel = PSYNDEXplus-Tests (ab 1945; OvidSP). Copyright (c) 2000-2009 Ovid Technologies, Inc.[Datum des Zugriffs: 07.09.09])..

Der RZD 2-6 (Rechenfertigkeiten- und Zahlenverarbeitungs-Diagnostikum für die 2. bis 6. Klasse) von Jacobs und Petermann (2005) wird vorrangig zur Diagnose von Rechenstörungen eingesetzt und liefert so „Hinweise auf basale Stoerungen [sic] der Informationsverarbeitung“ wie folgende Aufgaben: Mengenschätzen, Regelverständnis und Größenvergleiche (Testerklärungen verfügbar unter: http://dbs.univie.ac.at /suchtitel=PSYNDEXplus-Tests (ab 1945; OvidSP). Copyright (c) 2000-2009 Ovid Technologies, Inc.[Datum des Zugriffs: 07.09.09]).

Der ZRT (Zahlenverarbeitungs- und Rechentests) von Kalbe, Brand und Kessler (2002) ist ein Einzelverfahren zur Erstellung eines Leistungsprofils für Erwachsene mit Hirnschäden. Unter anderem gibt es auch einen Subtest der die Fähigkeit zum Kopfrechnen überprüft. Hierbei wird die Person „instruiert, sich die muendlich [sic] vorgetragenen Rechenaufgaben gut anzuhoeren [sic] und das Ergebnis laut zu

nennen“, „wobei die Zahlen nicht ueber [sic] den Hunderterbereich hinausgehen“ (Testerklärungen verfügbar unter: http://dbs.univie.ac.at/ suchtitel = PSYNDEXplus-Tests (ab 1945; OvidSP). Copyright (c) 2000-2009 Ovid Technologies, Inc.[Datum des Zugriffs: 07.09.09]).

Im Gegensatz zu den anderen genannten Verfahren besteht der VA 3 (Vergleichsarbeiten 3. Schuljahr) von der Arbeitsgemeinschaft für Leistungsmessung in der Schule (1984) aus einem umfassenden Testprozedere, welches eine Schätzung des Leistungsstandes des jeweiligen Schülers ermöglichen soll und als Gruppenverfahren empfohlen wird. Neben vier Diktaten und drei Sonderarbeiten gibt es zwei Mathematikarbeiten, wobei sich eine dieser Arbeiten auf das Kopfrechnen und das schriftliche Rechnen bezieht (Testerklärungen verfügbar unter: http://dbs.univie.ac.at/ suchtitel = PSYNDEXplus-Tests (ab 1945; OvidSP). Copyright (c) 2000-2009 Ovid Technologies, Inc.[Datum des Zugriffs: 07.09.09]).

Der NTA (N-Test Alpha) von Bratfisch und Hagman (2003) ist ein Computerprogramm welches „schnelles und richtiges Kopfrechnen“ erfassen soll (Testerklärungen verfügbar unter: http://dbs.univie.ac.at/ suchtitel = PSYNDEXplus-Tests (ab 1945; OvidSP). Copyright (c) 2000-2009 Ovid Technologies, Inc.[Datum des Zugriffs: 07.09.09]).

Der ZAREKI (Neuropsychologische Testbatterie für Zahlenverarbeitung und Rechnen bei Kindern) von Aster (2001) ist ein neuropsychologisches Verfahren zur Diagnose von Dyskalkulie. Neben Untertests wie das „Anordnen von Zahlen auf einem Zahlenstrahl“ oder das „Zahlen-schreiben“ gibt es auch Kopfrechnen, wobei dem Kind mündlich Aufgaben zum Addieren und Subtrahieren gestellt werden wie zum Beispiel Zwölf plus Sechs oder Vierzehn minus Fünf (Testerklärungen verfügbar unter: http://dbs.univie.ac.at/ suchtitel = PSYNDEXplus-Tests (ab 1945; OvidSP). Copyright (c) 2000-2009 Ovid Technologies, Inc.[Datum des Zugriffs: 07.09.09]).

Die letzten fünf Verfahren beinhalten unter anderem auch einen Untertest „Kopfrechnen“, wobei die Aufgaben entweder mündlich oder schriftlich vorgegeben werden. Diese Aufgaben sind – abgesehen von ihrer niedrigeren Komplexität - vergleichbar mit dem „Mental Arithmetik Test“, bis auf die Tatsache, dass alle genannten Verfahren nach der Klassischen Testtheorie konstruiert wurden. Der Mental Arithmetik Test soll sich hingegen in dieser Arbeit in allen seinen acht Versionen als Rasch-homogen herausstellen.

Das Hauptaugenmerk liegt hier somit auf dem Rasch-Modell und seinen Modelltests, was zur Folge hat, dass dementsprechend im Theorieteil ein grober Überblick über diese Art der Testung in der Probabilistischen Testtheorie gegeben wird.

4 Klassische Testtheorie (KTT) vs. Probabilistische Testtheorie (PTT)

In der Testtheorie werden zwei große Bereiche unterschieden, die Klassische und die Probabilistische Testtheorie. Die PTT trägt international die Bezeichnung „Item-Response-Theorie“ (IRT) (Amelang & Zielinski, 2002). Die ältere der beiden Theorien ist die Klassische Testtheorie (KTT). „Die Klassische Testtheorie ist gegenwärtig die Grundlage der meisten psychologischen Testverfahren“ (Bühner, 2006, S. 25). Der später aufgekommene Ansatz der Probabilistischen Testtheorie versucht gewisse Nachteile der Klassischen Testtheorie zu vermeiden (vgl. Amelang & Zielinski, 2002). Die KTT stellt jedoch für die meisten Verfahren die Konstruktionsgrundlage dar (Amelang und Zielinski, 2002).

> Historisch zurückblickend muss gesagt werden, dass die klassische Testtheorie eine für die Weiterentwicklung der Psychologie entscheidende Rolle gespielt hat. Sie führte die diagnostische Testpsychologie und damit verbunden auch manche Bereiche der Experimentalpsychologie aus dem Stadium intuitiver Deutung von Befunden heraus, sie förderte die Quantifizierung und Objektivierung psychologischer Forschung und unterstützte ein naturwissenschaftliches Selbstverständnis der Psychologie (Fischer, 1974, S.17).

„Der Ausgangspunkt für die Entwicklung der KTT war die Feststellung von Spearman (1910, nach Amelang & Zielinski, 2002), dass messfehlerbehaftete Variablen miteinander niedriger korrelieren müssen, als sie es ohne Fehlerbehaftetheit tun würden“ (Amelang & Zielinski, 2002, S. 34).
Daraus ist laut Bühner (2006, S. 22) zu schließen, dass sich die KTT mit „Messungen und deren Ungenauigkeit“ beschäftigt, die Probabilistische Testtheorie jedoch „direkt die Annahme der Eindimensionalität prüft“.

Bevor wir uns weiter mit der Probabilistischen Testtheorie beschäftigen sollen noch basale Annahmen der KTT erläutert werden:

„Ein grundlegender Begriff, auf welchem die klassische Testtheorie aufbaut, ist der des „true score““ (Fischer, 1974, S.27). Dabei nimmt man an, dass jede Testung fehlerbehaftet ist und daher der „wahre Wert“ vom beobachteten Testwert unterschieden werden muss, da der „Messfehler“ den beobachteten Wert belastet (Fischer, 1974, S.27).

Das bedeutet nun, dass sich jeder beobachtete Wert (X) in einem Test aus dem true score (konstanter wahrer Wert T) und einem Messfehler (E) zusammensetzt (Amelang & Zielinski, 2002).

$$X = T + E$$

Der Messfehler legt sich laut Amelang und Zielinski (2002) somit über den wahren Wert und führt dazu, dass sich der beobachtete Testwert mehr oder weniger vom wahren Wert unterscheidet. „Inhaltlich umfasst das Konzept des Messfehlers die Gesamtheit aller unsystematischen und nicht kontrollierbaren oder vorhersagbaren Einflussgrößen, die auf das Messergebnis einwirken können“ (Amelang & Zielinski, 2002, S. 34).

Rost (1996, S. 35) meint, dass sich der Messfehler durch zwei Eigenschaften kennzeichnen lässt:

> Erstens wird der wahre Wert durch den Messfehler entweder über- oder unterschätzt, was den „Meßfehler [sic] von einem systematischen Fehler“ unterscheidet, der „im Mittel neutral“ ist. D.h. „der Mittelwert oder Erwartungswert der Meßfehlervariable [sic] E ist über eine große Anzahl von Personen 0“:
>
> (I) Erw (E) = 0.
>
> Zweitens korreliert der Messfehler nicht mit dem wahren Wert, was bedeutet, dass nicht systematisch über- oder unterschätzt werden darf (bei hohen bzw. bei niedrigen Werten). Ansonsten würde das einen „systematischen Fehler“ darstellen; d.h.:
>
> (II) Korr (E,T) = 0.
>
> Der Messfehler darf auch nicht mit anderen Variablen korrelieren, „also nicht mit den wahren Werten einer Variable Y“:
>
> (III) Korr (Ex, Ty) = 0,

Weiters darf der Messfehler auch nicht mit anderen Messfehlern Ey korrelieren:

(IV) Korr (Ex, Ey) = 0.

„Diese vier Gleichungen (I) bis (IV) nennt man auch die Axiome der klassischen Testtheorie.“ „Sie wurden von Gulliksen (1950, nach Rost 1996) formuliert und beschreiben nichts anderes als die Eigenschaften eines Meßfehlers [sic]“ (Rost, 1996, S. 35).

Kubinger betont, dass die Verfahren der KTT generell mit Vorsicht zu betrachten sind, da die Daten oft sehr stark von der Stichprobe abhängig sind und sich auch je nach deren Zusammensetzung die Lösungshäufigkeiten deutlich ändern (Kubinger, 2006).

Die Probabilistische Testtheorie vermeidet dieses Problem, wie auch Bühner (2006) bemerkt, da es hier möglich ist von der Stichprobe unabhängige Werte zu ermitteln, was jedoch mit einem sehr viel größeren Aufwand in der Itemkonstruktion einhergeht. In der KTT behilft man sich mit „Gütekriterien für alle Teilstichproben“, jedoch werden diese in der Praxis oft nur mangelhaft zur Verfügung gestellt (Bühner, 2006, S. 32).

Neben der Stichprobenabhängigkeit der KTT ist auch die Erhebung der latenten Dimensionen problematisch bzw. wird derer nicht genügend nachgegangen, denn lt. Fischer (1974, S. 125) muss eine „Testtheorie“, die „eine psychologische Theorie des Testverhaltens sein“ soll, „die Frage zu untersuchen gestatten, wieviele [sic] Grunddimensionen („latent traits“, Faktoren) die Testleistung der Personen bestimmen“. Die Klassische Testtheorie setzt jedoch den Testrohwert, bis auf zufällige Messfehler, mit dem wahren Wert gleich womit an der Problemstellung vorbeigegangen wird, da nicht erhoben oder kontrolliert wird ob auch wirklich nur eine latente Dimension gemessen wird. Die Klassische

Testtheorie nimmt somit Eindimensionalität an ohne diese zu prüfen (Fischer, 1974).

Wie auch bei Stelzl (1979, S. 653) zu ersehen ist „bietet die klassische Testtheorie“ für die „Beurteilung der Homogenität“ „in erster Linie Trennschärfekoeffizienten für die Items und die innere Konsistenz der Skala an“, welche jedoch – wie schon oben erwähnt – stichprobenabhängig bzw. auch von der Schwierigkeit der Items abhängig sind und generell „auf keinen formalisierten Homogenitäts-begriff Bezug nehmen“. „Die probabilistische Testtheorie hat demgegen-über Homogenität begrifflich klar als Eigenschaft des Tests, unabhängig von der Verteilung der Personenparameter, definiert“ (Stelz, 1979, S. 653).

> In der Probabilistischen Testtheorie geht es im Gegensatz zur Klassischen Testtheorie darum, wie Antworten auf Items zustande kommen. Aus diesem Grund werden Antwortmuster untersucht. Die beobachteten Antwortmuster müssen einem bestimmten Modell folgen. Dieses Modell sagt voraus, dass mit steigender Personenfähigkeit die Wahrscheinlichkeit einer Itemlösung zunimmt. Die Lösungswahrscheinlichkeit für ein bestimmtes Item hängt (1) von der Fähigkeit oder Eigenschaftsausprägung einer Person sowie (2) der Schwierigkeit eines Items ab. Diese Beziehung zwischen Personenfähigkeit und Itemlösungswahrscheinlichkeit ist probabilistisch. Das heißt, auch eine Person mit geringer Fähigkeit im Vergleich zur Schwierigkeit eines Items hat eine, wenn auch geringe, Wahrscheinlichkeit, ein solches Item zu lösen (Bühner, M., 2006, S.33).

Auch was die Prüfung der Intervallskaleneigenschaft angeht, bietet die Klassische Testtheorie keine zufriedenstellende Lösung.

> Intervalleigenschaften der Skalen werden stillschweigend vorausgesetzt, da sie die Mindestanforderung für die verwendeten statistischen Größen in der KTT sein müssen. Das RASCH-Modell führt nachweislich zu Rationalskalen (vgl. FISCHER, 1968 a, nach Henning, 1975, S. 11).

Trotz der vielen Kritik ist die KTT – wie schon oben erwähnt - die Grundlage der meisten psychologischen Testverfahren. „Nach Rost (1999, S. 140, nach Bühner, M., 2006, S. 25) basieren 95 Prozent aller Tests auf der Klassischen Testtheorie“. „Ein großer Vorteil der Klass-ischen Testtheorie liegt in ihrer einfachen Anwendbarkeit“ (Henard, 2000, nach Bühner, M., 2006, S.26).

4.1 Überwindung der klassischen Testtheorie

„Der erste systematische Ansatz zur Überwindung der klassischen Testtheorie geht auf Guttman (1950, nach Fischer, 1974, S. 137) zurück."

> Die Anwendbarkeit der klassischen Testtheorie in der Soziologie war von Haus aus in Frage gestellt, sind doch die meisten manifesten Variablen der Soziologie qualitativ. Zwangsläufig wurden deshalb in der Soziologie schon zu einer Zeit, als es in der Psychologie noch keine Alternative zur klassischen Testtheorie gab, allgemeinere Ansätze zu einer Theorie des Messens entwickelt, die auch qualitative Variablen mit einschliesst [sic] (Fischer, 1974, S.137).

Laut Fischer (1974, S. 148) hat sich erstmals Guttman von der Vorstellung befreit, „dass die latente Dimension und die beobachtbaren, qualitativen oder quantitativen Variablen gleichzusetzen" sind. Guttman „geht von der Annahme eines unendlichen Universums von Items aus, welche die gleiche Eigenschaft ξ messen" (Fischer, 1974, S. 138). Er meinte, dass sich die Aufgaben lediglich durch ihre „Schwierigkeit" unterscheiden, „worunter die Häufigkeit positiver Beantwortungen in der Versuchspersonenpopulation zu verstehen ist". Die Versuchspersonen werden dann bzgl. ihrer Antworten „in eine Rangreihe hinsichtlich der zu messenden Einstellung gebracht". Vorraussetzung dafür ist jedoch, dass die positive Beantwortung eines Items i durch die Vp v die positive Beantwortung aller leichteren Items j (j leichter als i) bedingt (Fischer, 1974, S.138).

> Die syntaktische Formulierung des Guttmanschen Messmodells lautet:
> Es gibt ein abzählbar unendliches Universum von Items (i = 1, 2, ...) und eine abzählbar unendliche Population von Vpn (v = 1,2, ...). Jedes Item wird durch einen Parameter σi, die Schwierigkeit des Items, und jede Vp durch einen Parameter ξv, die Einstellung der Vp, charakterisiert. Wenn $\xi v \geq \sigma i$, dann antwortet die v-te Vp auf das i-te Item positiv (mit „ja", „stimmt", und so weiter; oder sie löst die Aufgabe, wenn es sich um einen Leistungstest handelt; symbolisch +) (Fischer, 1974, S.139).

Das Antwortverhalten würde nun folgendermaßen aussehen:

(1) $a_{vi} = \begin{cases} 1 \text{ wenn Vp v mit + antwortet,} \\ 0 \text{ wenn Vp v mit – antwortet (Fischer, 1974, S.139)} \end{cases}$

Die fundamentale Annahme des Modells lautet nach Fischer (1974, S. 139) dann:

(2) $p(+/v, i) = 1$ für $\xi v \geq \sigma i$ und 0 für $\xi v < \sigma i$.

Die Funktion (2) wird als „Itemcharakteristik“ bezeichnet. Je zwei Personen mit $\xi v < \sigma i \leq \xi w$ werden durch das i-te Item allein eindeutig in eine Rangreihe gebracht (Fischer, 1974, S.139).

Die Idee dieses Modells ist, dass „die Lösungswahrscheinlichkeit für einen unteren Bereich der Eigenschaftsausprägung 0 ist und an einer bestimmten Stelle auf 1 springt“, wenn so zu sagen die plötzliche Erkenntnis, wie Rost (1996, S. 104) es beschreibt, eintritt.

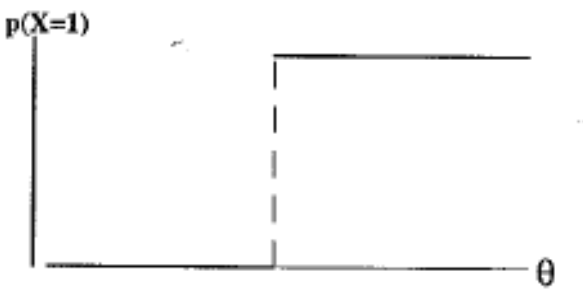

Abbildung 4: Die Itemfunktion einer Guttman-Skala (Rost, 1996, S. 104)

„Haben alle Items diese Form einer Itemcharakteristik und lässt [sic] sich die Stelle dieser ‚plötzlichen’ Einsicht auf derselben latenten

Dimension anordnen, so beschreibt“ die in Abbildung 5 „wiedergegebene Schar von ICC’s die Items eines Tests“ (Rost, 1996, S. 104).

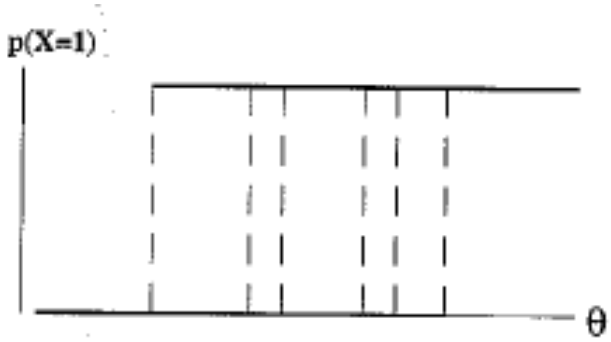

Abbildung 5: Die Itemfunktionen mehrerer Items einer Guttman-Skala (Rost, 1996, S. 104)

Rost (1996, S. 104) bezeichnet dieses Modell, welches als „Guttman-Skala“ bekannt ist, auch als „Alles-oder-Nichts-Itemcharakteristik“, da man eine Aufgabe entweder lösen kann, oder nicht. „Bis zu einem gewissen Fähigkeitsgrad kann man es nicht lösen, darüber hinaus kann man es infolge einer entsprechenden Einsicht oder eines ‚Aha’-Erlebnisses lösen“ (Rost, 1996, S. 104).
Die Guttman-Skala ergibt sich aus diesen beiden obigen Annahmen, nämlich, dass „alle ICC’s stufenförmig von einer 0-Wahrscheinlichkeit zu einer 1-Wahrscheinlichkeit springen und daß [sic] die latente Dimension, auf der diese Sprungstellen angesiedelt sind (repräsentiert durch die X-Achse), für alle Items dieselbe ist“, wobei die zweite Annahme die „Itemhomogenität“ darstellt (Rost, 1996, S. 104).

„Sowohl Itemunterschiede als auch Personenunterschiede“ lassen sich „nur auf Ordinalskalenniveau bestimmen“, was laut Rost (1996, S. 105) die „wichtigste Eigenschaft“ dieses Modells darstellt. Diese Eigenschaft ist in Abbildung 5 zu ersehen, da „sich alle Personen, deren Fähigkeitsausprägungen zwischen den Sprungstellen zweier benachbarter Items liegen, in ihrem Testverhalten nicht voneinander unterscheiden“,

somit kann aus deren Verhalten im Test nicht „auf unterschiedliche Eigenschaftsausprägungen geschlossen werden“ (Rost, 1996, S. 105).

> Es können maximal so viele Eigenschaftsausprägungen von Personen unterschieden werden, wie es Items gibt, plus eins. Bei 6 Items gibt es 7 Bereiche auf dem latenten Kontinuum, die sich aufgrund des Antwortverhaltens unterscheiden lassen [siehe Abb. 5: es gibt 7 Abschnitte in dieser Abbildung].
> Diese 7 Personengruppen entsprechen genau den 7 möglichen Scoregruppen, die es bei 6 Items gibt. Die Gruppe, die am weitesten links liegt, hat kein Item gelöst, die zweite Gruppe hat genau ein Item gelöst usw. Alle Personen, die in dieselbe Scoregruppe fallen, d.h. denselben Personenscore aufweisen, haben auch genau dasselbe Antwortpattern [Antwortmuster] produziert (Rost, 1996, S. 105).

Eine weitere „Eigenschaft des Modells ist die Symmetrie zwischen Item und Personen“. „Sie ist semantisch sofort einsichtig, wenn man sich das Lösen eines Items als „Sieg“ einer Vpn über das Item und das „Nichtlösen-Können“ als „Sieg“ des Items über die Person vorstellt“ (Fischer, 1974, S. 140).
Fischer (1974, S. 140) bemerkt auch weiter, dass „jede Aussage über die Relation zwischen zwei Personenparametern“ „unabhängig von allen anderen Personenparametern“ ist, „da ja nur das Vorhandensein oder Fehlen eines Items zwischen den beiden Parametern entscheidet, ob eine Aussage gemacht werden kann oder nicht“ (Fischer, 1974, S.140).

Man kann nun die Personen nur dahingehend miteinander vergleichen, dass man sie in eine Rangreihe bringt. Um die Differenz quantifizieren bzw. „um die Fähigkeiten der Personen auf einem höheren Skalenniveau als dem der Ordinalskala berechnen zu können“ müsste man „die Schwierigkeiten der Items kennen“ (Rost, 1996, S. 105).
Da die Messung nur ordinal erfolgt enthält die Guttman-Skala laut Fischer (1974, S. 140) „deshalb keinerlei metrische Information, welche in der Vertauschung (Intransivität) von Itemantworten liegen müsste“. „Wären die Antworten der Vpn nicht deterministisch, sondern probabilistisch, dann wäre zu erwarten: Je näher zwei Items mit $\sigma i < \sigma j$ beisammen liegen, desto häufiger wird es vorkommen, dass eine Vp das

schwierigere Item j bejaht (löst) und das leichtere Item i verneint (nicht löst)“ (Fischer, 1974, S.140).

Da es sich jedoch hier nicht um einen probabilistischen Zugang handelt, sondern um einen deterministischen „genügt schon eine einzige Antwort einer Vp, die von der Struktur [2] abweicht, um das Modell zu widerlegen“ (Fischer, 1974, S.141). Guttmans Modell ist laut Fischer (1974, S. 148) somit nicht praxisnah, da er „keinerlei Zufallseinflüsse in seinem Modell vorsah“. „Gerade die Zufallskomponenten in fast jeder Art von Beobachtung, welche wir in den Sozial- und Verhaltenswissenschaften machen, sind aber die Ursache dafür, dass die unseren Theorien zugrundeliegenden Begriffe bis zu einem gewissen Grad vage und unverbindlich sind“ (Fischer, 1974, S. 148).

Erst Lazarsfeld (1950, nach Fischer, 1974, S. 149) hat die, „für die Sozial- und Verhaltenswissenschaften notwendigen Konsequenzen“ gezogen. „Seine Untersuchungen haben ihre wesentlich vertiefte Fortsetzung in Arbeiten von RASCH (1960, 1961, 1966a, nach Fischer, 1974, S. 149) gefunden.“

Die empirische Psychologie prägenden Ideen Lazarsfelds beruhen auf folgenden Grundprinzipien:

1) Trennung von latenter Dimension und beobachtbaren Variablen; diese werden nur als Indikatoren oder Symptome der ersteren aufgefasst;
2) Berücksichtigung des probabilistischen Charakters der Beobachtungen, d.h. Absage an den Determinismus;
3) Prinzip der lokalen stochastischen Unabhängigkeit, d.h. die Kovariation zwischen den beobachtbaren menschlichen Verhaltensweisen – um die Betrachtung auf die Psychologie einzuschränken – wird als Resultat ihrer Abhängigkeit von einer gemeinsamen latenten Variable angesehen, nicht als direkte Abhängigkeit zwischen den Verhaltensweisen selbst.
4) Die Definition und Messung einer latenten Eigenschaft erfordert die wiederholte Beobachtung von Symptomen oder Indikatoren dieser Eigenschaft, nicht notwendigerweise die wiederholte Beobachtung desselben Indikators (Fischer, 1974, S. 149).

4.2 Dichotome Latent-Trait-Modelle

In weiterer Folge wollen wir uns näher mit einer Gruppe der Item-Response-Theorie bzw. Probabilistischen Theorie beschäftigen und zwar mit den dichotomen Latent-Trait-Modellen.

Laut Amelang und Zielinski (2002, S. 71) sind in der psychologischen Diagnostik Latent-Trait-Modelle am häufigsten vertreten.
„Latent-Trait-Modelle gehen davon aus, dass sowohl die Ausprägungen verschiedener Probanden auf den latenten Traits als auch die traitbezogenen Anforderungen der Items an die Person jeweils durch einen Parameter, nämlich einen einzelnen numerischen Kennwert, charakterisiert werden können“ (Amenlang & Zielinski, 2002, S. 71).
Es wird hierzu zwischen „Personenparametern“ und „Itemparametern“ unterschieden. Personenparameter stellen die Fähigkeit bezüglich des „latenten Traits“ dar und Itemparameter die Schwierigkeit bzw. die Anforderung des Items and die Fähigkeit der jeweiligen Person (Amenlang & Zielinski, 2002, S. 71).

Die Zusammenhänge der Items können durch latente Klassen erklärt werden; mittels latent class analysis (LCA) von Lazarsfeld (1950), wobei die Voraussetzung für die Anwendung einer LCA die, weiter unten erläuterte, lokale stochastische Unabhängigkeit darstellt (Reiter, 1996).
„Das Ziel der LCA besteht also darin, die Personen in (latente) Klassen einzuteilen, die dadurch definiert sind, daß [sic] die Personen innerhalb der Klassen dieselben und zwischen den Klassen verschiedene Lösungswahrscheinlichkeiten haben“ (Rost & Georg 1991, S. 56, nach Reiter 1996, S. 33).

Bezüglich der „itemcharakteristischen Funktion“ kann man zwischen „deterministischen“ und „probabilistischen Modellen“ unterscheiden (Amenlang & Zielinski, 2002, S. 72).
„Deterministische Modelle gehen davon aus, dass das Antwortverhalten der Probanden durch die Item- und Personenparameter vollständig bestimmt ist. Probabilistische Modelle hingegen nehmen eine stochastische Beziehung zwischen dem Antwortverhalten des Probanden und den Personen- und Itemparametern an“ (Amenlang & Zielinski, 2002, S. 72).

Das Deterministische Modell entspricht dem Modell von Guttmann (1950), welches als „Vorläufer der später entwickelten probabilistischen Latent-Trait-Modelle angesehen werden kann“ (Amenlang & Zielinski, 2002, S. 72). Die Grundüberlegung dieses Modells ist, dass „für jedes dichotom beantwortete Item“ ein „bestimmter Wert auf der ξ-Skala“ existiert, „ab dem das Item gelöst wird (bzw. dem Item zugestimmt wird)“ (Amenlang & Zielinski, 2002, S. 72). Es muss somit ein striktes System eingehalten werden – beantwortet eine Person ein Item positiv, muss diese auch die davor liegenden Items positiv beantworten. Dieses System entspricht somit einer Ordinalskala und erlaubt keine Aussagen über Distanzen (Amenlang & Zielinski, 2002).

Wie schon im vorigen Kapitel erwähnt ist das Modell von Guttman sehr praxisfern, da jederzeit mit Zufallseinflüssen zu rechnen ist die diese Ordnung mit Leichtigkeit zerstören können.

> Deshalb wurden bereits von Guttman selbst „Reproduzierbarkeits-koeffizienten“ eingeführt, welche davon abhängen, wie viele Rangplatz-vertauschungen vorliegen; sie erlauben eine Beurteilung, ob die Modell-abweichungen noch als tolerierbar angesehen werden können, oder ob die Annahme der Itemhomogenität verworfen werden muss (Amenlang & Zielinski, 2002, S. 73).

Im Gegensatz dazu werden in Probabilistischen Modellen „monoton steigende Funktionen als IC-Funktion angenommen“ (Amenlang & Zielinski, 2002, S. 73).

Hierbei arbeitet man nicht mit kausalen Aussagen, sondern mit Wahrscheinlichkeiten. Jeder Ausprägung einer latenten Variable wird eine Wahrscheinlichkeit zugewiesen, mit der ein bestimmtes Item gelöst wird (Amenlang & Zielinski, 2002).

In weiterer Folge wird in dieser Arbeit näher auf einen speziellen Fall der Probabilistischen Latent-Trait-Modelle eingegangen, dem Rasch-Modell, welchem ein eigenes Kapitel gewidmet wird.

„Neben den genannten dichotomen Latent-Trait-Modellen“ gibt es noch eine „Vielzahl weiterer Modelle“, die „ebenfalls probabilistisch“ sind, „sich aber u. a. durch die Art der manifesten und/oder latenten Variablen und die Art der verwendeten Modellparameter“ unterscheiden und meist eine Weiterentwicklung des in Folge präsentierten dichotomen Rasch-Modells darstellen (Amelang & Zielinski, 2002, S. 85).

4.2.1 Das dichotome logistische Modell nach Rasch

Das dichotome logistische Modell von Rasch ist das am einfachsten aufgebaute unter den probabilistischen Testmodellen (Amelang & Zielinski, 2002). Es gilt, wie der Name schon sagt, nur für dichotome Daten.

Das Modell beinhaltet nur einen Itemparameter, nämlich den Schwierigkeitsparameter σi, und wird deshalb auch 1-PL Modell („Dichotome logistische Testmodell"; „Ein-parametrige logistische Modell") genannt (Kubinger & Jäger, 2003). Im Gegensatz dazu existieren Modelle, die mit mehreren Parametern arbeiten, wie das 2-PL Modell oder das 3-PL Modell von Birnbaum, welche als weitere Parameter den Diskriminationsparameter α_i (2-PL oder „Zwei-parametriges Modell") und den Rateparameter β_i (3-PL oder „Drei-parametriges Modell") berücksichtigen (Kubinger & Jäger, 2003).

Die Wahrscheinlichkeit, dass eine Testperson v ein Item i löst, hängt im 1-PL Modell vom „Personenparameter ξ_v" und vom „Itemparameter σ_i" ab. Es herrscht jedoch kein deterministischer Zusammenhang, wie das bei Guttman der Fall ist, sondern ein probabilistischer, d.h. dass die Lösungswahrscheinlichkeit mit Zunahme der Fähigkeit ξ_v und konstanter Itemschwierigkeit σ_i steigt (Kubinger & Jäger, 2003).

Die Grundgleichung des Rasch-Modells lautet daher:

$$P\ (+/v,i) = \exp\ (\xi v - \sigma i) / [1 + \exp\ (\xi v - \sigma i)].$$

„Das durch diese Gleichung definierte Testmodell wurde 1960 von dem Dänen Georg Rasch erstmals im Detail untersucht und dargestellt und wird seitdem als Rasch-Modell bezeichnet" (Rost, 1996, S. 124).

Fischer und Molenaar (1995) betonen, dass der Gewinn dieser Gleichung von Rasch nicht nur in der Erklärung der gegebenen Antworten liegt, sondern auch in der Möglichkeit der Vorhersage der wahrscheinlichen Verteilung der Antworten von Personen für andere Items.

> The importance of (this equation) is that it enables us not only to explain post hoc how this person responded to this item, but also to predict from a person and item parameter the probability distribution for the answer of this person to other items, and of other persons to this item, without actually observing the response to such person-item pairs (Fischer & Molenaar, 1995, S. 9).

„Die Parameter ξ_v und σ_i sind“, wie bei Fischer (1974, S.209) zu finden, „reelle Zahlen zwischen - ∞ und + ∞“ und wie aus der Formel zu ersehen ist, „hängt die Wahrscheinlichkeit einer richtigen Antwort nur von der Differenz zwischen dem Personen- und dem Aufgabenparameter ab“.
Aus der Formel ist auch weiters zu ersehen (Kubinger, & Jäger, 2003, S. 416), dass die Wahrscheinlichkeit für die Lösung eines Items bei größer werdender Fähigkeit ξ und immer größerer Aufgabenleichtigkeit σ gegen 1.0 strebt. Ist die Fähigkeit gering und das Item schwer, stellt das den umgekehrten Fall dar und die Wahrscheinlichkeit für eine Lösung des Items liegt bei 0.0. Liegt die Aufgabenschwierigkeit und die Fähigkeit der Person auf einer Ebene, also wenn $\xi v = \sigma i$, dann erhält man eine Lösungswahrscheinlichkeit von 0.50 und bietet den meisten Informationswert bezüglich der Fähigkeit der Person.

Laut Fischer (1974, S. 217) erlaubt „das dichotome logistische Modell die Messung von Aufgabenleichtigkeitsparametern auf einer Verhältnisskala, bzw. nach Logarithmieren der Parameter die Messung von Schwierigkeitsunterschieden auf einer absoluten Skala“.
Daraus resultiert, dass Fähigkeitsunterschiede zwischen Personen direkt interpretierbar sind bzw. dass je zwei Items miteinander verglichen werden können ohne dass das Ergebnis des Vergleiches von der Verteilung der Personenparameter in der benützten Stichprobe abhängt (Gittler & Arendasy, 2003).

> Das zweikategorielle Modell von Rasch erlaubt Schlussfolgerungen und Aussagen, welche im Rahmen der klassischen Testtheorie unmöglich oder unsinnig wären; man denke nur an die Rationalskaleneigenschaft der Personen- und der Aufgabenparameter. Allerdings zahlt man für diese Vorteile auch Tribut; dem Modell liegt eine Reihe von Annahmen zugrunde, welche bei seiner Anwendung empirisch getestet werden müssen. Es ist zwar nicht möglich, jede einzelne Annahme für sich und unabhängig von den anderen zu prüfen, doch ihre Gesamtheit, d.h. das Modell als Ganzes, kann einer empirischen Kontrolle unterworfen werden (Fischer, 1974, S. 281).

4.2.1.1 Anforderungen an das RM

Das Rasch-Modell hat unter der Gruppe der Latent-Trait-Modelle gewisse Besonderheiten bezüglich seiner Modelleigenschaften. In diesem Abschnitt wird auf die besonderen Eigenschaften des dichotomen logistischen Modells von Rasch näher eingegangen.

Im Überblick sind dies folgende:

- Lokale stochastische Unabhängigkeit
- Spezifische Itemcharakteristik (IC)
- Stichprobenunabhängigkeit der Parameterschätzungen (spezifische Objektivität der Vergleiche)
- Erschöpfende Statistiken

4.2.1.1.1 Lokale stochastische Unabhängigkeit

Das grundlegende Ziel von Item-Response-Theorien ist ja die Messung von latenten mit Hilfe von manifesten Variablen. Die Items eines Tests sollen „Indikatoren der latenten Variablen ξ“ (Amelang & Zielinski, 2002, S. 69) sein, was auch unter anderem bedeutet, dass die Items untereinander korrelieren müssen, da sie dieselbe Dimension erheben sollen. „Die latente Variable“ sollte demnach „Ursache für die Korrelation“ der Items sein (Amelang & Zielinski, 2002, S. 69). Die

Testitems bezeichnet man daher auch als homogen und sollten das auch über alle eventuellen Subtests sein, da man annimmt, dass jedes Item Indikator für *dieselbe* latente Variable ist oder wie Klauer erläutert:

> The Rasch model predicts that an individual´s ability level is invariant over subtests of the total test, and thus, all subtests measure the same latent trait (Klauer, 1991, S. 213).

Eine wichtige Voraussetzung für die Itemhomogenität ist die lokale stochastische Unabhängigkeit (Bühner, 2006). Diese Annahme der lokalen stochastischen Unabhängigkeit besagt, dass es unbedeutend ist in welcher Reihenfolge die Items gelöst werden bzw. welche Items bereits gelöst worden sind und welche nicht. Die Lösung einer Aufgabe hängt einzig und allein von der Personenfähigkeit und der Aufgabenschwierigkeit ab (Fischer, 1974).

> Diese Verursachung kann überprüft werden, in dem man die latente Dimension auf einem bestimmten Wert (auf einer lokalen Stufe, z. B. ξ_v oder ξ_w) konstant hält. Sind die Items homogen, so muss sich nun die lokale stochastische Unabhängigkeit zeigen, welche darin besteht, dass die Korrelationen zwischen den Items auf diesen Stufen verschwinden. Folglich kann bei Vorliegen der lokalen stochastischen Unabhängigkeit auf Itemhomogenität bezüglich ξ geschlossen werden (Amelang & Zielinski, 2002, S. 69).

Wenn man nun davon ausgeht, dass die Lösungswahrscheinlichkeit nur von der Personenfähigkeit abhängt, ist laut Stelzl (1979, S. 653) „nicht gefordert, daß [sic] die gemessene Fähigkeit im psychologischen Sinn elementar sein müßte [sic]: Sie kann selbst Funktion anderer Fähigkeiten sein“ bzw. wie Gittler (1999) bemerkt muss sich bei der „Modellannahme der Eindimensionalität der Testaufgaben“ die Fähigkeit nicht aus „einer einzigen zugrundeliegenden psychologischen Dimension“ zusammensetzen.“ Laut Gittler (1999) ist meist zwischen „Primärfähigkeiten“ und anderen „Ko-Faktoren“ zu unterscheiden, die ebenfalls die zu messende Dimension beeinflussen. „Es ist sogar möglich, dass Ko-Faktoren kompensatorische Wirkung haben“ und „dennoch dieselbe Fähigkeit (Personenparameter) im“ jeweiligen Test „aufweisen.“

Als Beispiel nennt Stelzl (1979, S. 653) die „Fähigkeit zum Lösen verbaler Denkaufgaben“. Diese setzt sich nun aus der gewichteten

„Summe (a_1 und a_2 Gewichtszahlen) aus Satzverständnis x_1 und logischer Fähigkeit x_2 zusammen, also $x_v = a_1x_{v1} + a_2x_{v2}$." Dieser Fall würde nicht gegen die Homogenität des Tests sprechen.
Das Prinzip der Homogenität wäre laut Stelzl (1979, S. 653) „erst verletzt, wenn ein Teil der Items überwiegend Sprachverständnis, ein Teil überwiegend logischen Denkens prüfen würde, wenn also die Gewichtszahlen von Item zu Item variieren würden (im Extremfall so, dass die einen Items nur von der einen, die anderen Items nur von der anderen Dimension abhängen)."

4.2.1.1.2 Spezifische Itemcharakteristik (IC)

Die Itemcharakteristikkurve zeigt den systematischen Zusammenhang zwischen manifesten Variablen und den zugrundeliegenden latenten Dimensionen, bzw. den Zusammenhang zwischen Personenfähigkeit und Aufgabenschwierigkeit (Nepita, 1991). Das bedeutet nun laut Fischer (1974, S. 154) dass „jedes Item i" „mit der latenten Dimension ξ durch eine eindeutige, aber nicht notwendig umkehrbar eindeutige Funktion fi(ξ) verknüpft" ist.

$$P\ (+/i, \xi) = fi(\xi).$$

„Jede Person mit dem Fähigkeitsgrad ξ hat dieselbe Chance fi(ξ), die Aufgabe i zu lösen" (Fischer, 1974, S. 154). fi(ξ) wird „Item-charakteristik" genannt, wobei nach Fischer zwischen „monotonen" und „nicht monotonen" Itemcharakteristiken unterschieden wird (siehe Abbildung 6).

> Im Bereiche von Intelligenztests findet man meistens monotone Items, bei Konstruktion von Einstellungsfragebogen hingegen oft auch nicht monotone Items, zum Beispiel die Aussage: „Boxen ist nicht schlechter oder besser als irgendein anderer Sport." Sie wird vermutlich am ehesten von solchen Vpn bejaht, die weder eine extrem ablehnende Einstellung zum Boxen haben, noch davon besonders begeistert sind. Die Extreme lehnen diese Aussage, wenn auch aus verschiedenem Grunde, ab (Fischer, 1974, S. 155).

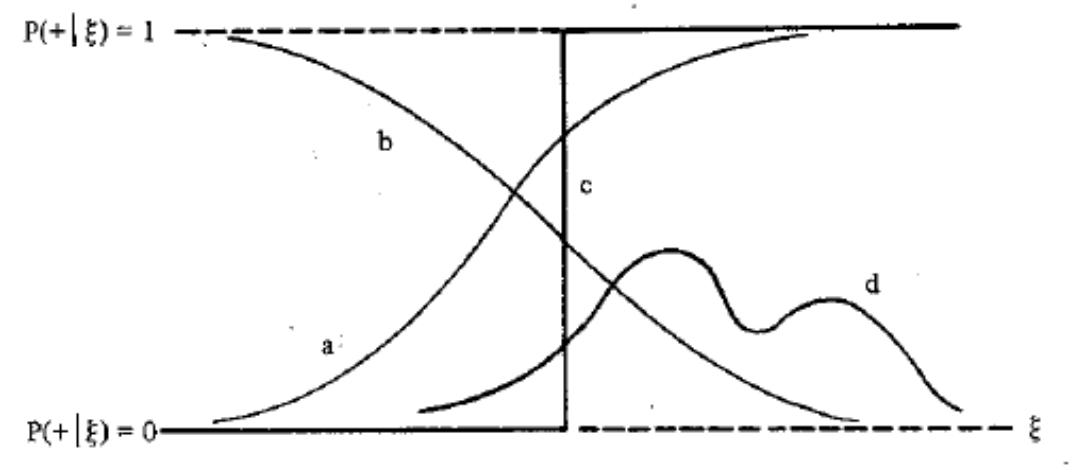

Abbildung 6: „Zwei streng monotone Itemcharakteristiken (a und b), c die Charakteristik eines Guttman-Items und d eines nicht monotonen Items“ (Fischer, 1974, S. 155).

In einem rasch-homogenen Test sollen daher nur Items mit einer monotonen IC sein.

„Daß [sic] das Rasch-Modell nur einen Itemparameter hat, nämlich den Schwierigkeitsparameter, hat zur Folge [wie in Abbildung 7 zu sehen], daß [sic] alle Itemfunktionen den gleichen Anstieg haben und somit parallel bezüglich der X-Achse verschoben sind“ (Rost, 1996, S. 125).

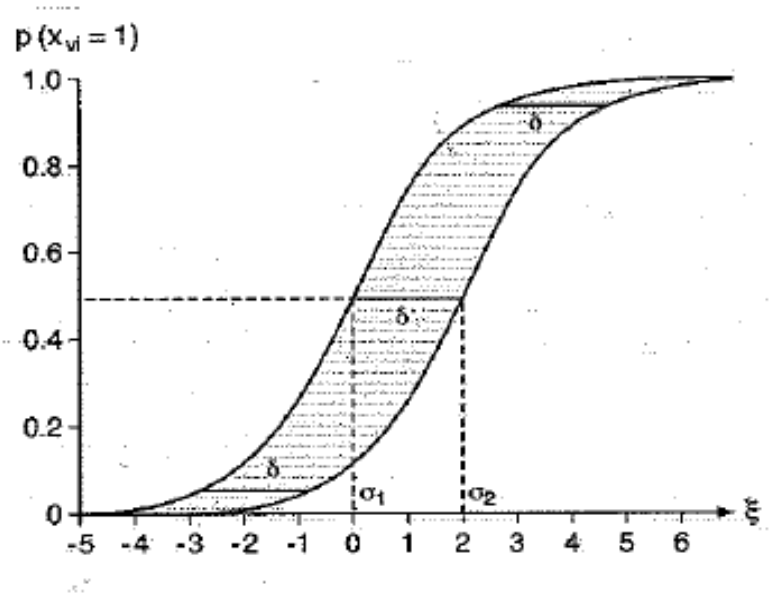

Abbildung 7: zwei parallele IC-Kurven, welche sich nur durch ihre Position auf der ξ-Achse unterscheiden; die Kurven entsprechen zwei rasch-homogenen Items (Amelang & Zielinski, 2002, S. 82).

„Die Parallelität der Itemfunktionen ist ein bedeutsames Merkmal des Rasch-Modells. Es bedeutet, daß [sic] alle Items eines Tests dieselbe Trennschärfe haben, wenn das Rasch-Modell für diesen Test gilt“ (Rost, 1996, S. 125).

4.2.1.1.3 Stichprobenunabhängigkeit der Parameterschätzungen bzw. „spezifische Objektivität der Vergleiche“

Ein großer Vorteil des Rasch-Modells liegt in der Möglichkeit „spezifisch objektive Vergleiche“ anstellen zu können d.h. generalisierbare Aussagen treffen zu können (Gittler & Arendasy, 2003). Das bedeutet nun, dass ein Vergleich „zweier Items (σi – σj) unabhängig davon“ ist welche Personen für die Untersuchungen herangezogen worden sind und auch, dass „Unterschiede zwischen den Personen-parametern (ξv – ξw)“ „unabhängig von den verwendeten Items festgestellt werden“ können (Amelang & Zielinski, 2002, S. 82).

Der Begriff der Stichprobenunabhängigkeit ist laut Rost (1996, S. 127) jedoch „irreführend, weil die Modellparameter des Rasch-Modells nur dann stichprobenunabhängig sind, wenn das Rasch-Modell in der untersuchten Population gilt.“ Die Geltung des Rasch-Modells muss daher erst für die entsprechende bzw. interessierende Population erhoben werden. „Will man“ nun „für einen Test untersuchen, ob das Rasch-Modell gilt, so ist es keineswegs beliebig, welche Personen- und Itemstichprobe man untersucht“ (Rost, 1996, S. 127).

4.2.1.1.4 Erschöpfende Statistiken

Kann Rasch-Homogenität festgestellt werden, gilt automatisch das Prinzip der „erschöpfenden Statistik“ d.h. dass es „bei Modellkonformität“ „für die Schätzung der Parameter gleichgültig“ ist „welche Items von welchen Personen gelöst wurden, entscheidend ist nur die Anzahl“ (Amelang & Zielinski, 2002, S. 77).
Das bedeutet laut Nepita (1991, S. 19), „dass die Anzahl der gelösten Aufgaben genügt, um die gesamte Information über die Leistungsfähigkeit einer Testperson“ zu erhalten.
Es muss somit „nicht für jede Person ein eigener Personenparameter berechnet“ werden, „sondern“ „alle Personen mit demselben Summenscore“ bekommen „auch denselben Personenparameter“ (Rost, 1996, S. 125). „Es ist also nicht nötig [bei Geltung des Rasch-Modells], die Konfiguration der Testantworten als Ganzes zu interpretieren“ (Fischer, 1974, S. 195).

5 Methoden der Parameterschätzung

Wie schon mehrfach erwähnt werden Personen- und Itemparameter unterschieden, wobei es unterschiedliche Methoden der Schätzung gibt. Laut Reiter (1996, S. 14) existieren drei Gruppierungen an Schätzmethoden: Bei der „unbedingten maximum likelihood Methode“ (UML-Methode) werden „Item- und Personenparameter simultan geschätzt“. Die bedingte maximum likelihood Methode (CML-Methode) ist laut Rost (1988, S. 253) „theoretisch befriedigender, aber praktisch oft schwerer zu handhaben“. Dieses Verfahren geht „von der Wahrscheinlichkeit der Daten unter der Bedingung der erschöpfenden Statistiken für die Personenparameter“ aus (Rost, 1988, S. 253). Es werden für die Dauer der Schätzung die Personenparameter außer Acht gelassen. Laut Amelang und Zielinski (2002, S. 77) erfolgt die Schätzung der Itemparameter in der Regel nach der „CML-Methode“, „welche im Unterschied zur unbedingten Maximum-Likelihood-Methode die Konsistenz der Schätzung nicht beeinträchtigt“.

Eine dritte Möglichkeit wäre die „Marginal-Maximum-Likelihood Methode (MML-Methode), bei der die Personenparameter durch Integrieren der Wahrscheinlichkeitsfunktion nach θv berechnet werden“ (Reiter, 1996, S. 15). „Die Marginal-Maximum-Likelihood Methode verwendet die Verteilungsfunktion der Personenparameter, was eine große Fehlerquelle darstellt“ (vgl. MOLENAR 1995b, S. 48f, nach Reiter, 1996, S. 15).

5.1 Modelltests bzw. goodness-of-fit tests

„Da [] das eindimensionale RM die vergleichsweise restriktivsten Modellannahmen beinhaltet, kommt der Frage nach deren Prüfbarkeit große Bedeutung zu: Denn erst dann, wenn die Modellvoraussetzungen durch einen bestimmten Datensatz erfüllt werden, sind die wünschenswerten Modelleigenschaften Implikationen der mathematischen Struktur des Modells“ (Gittler, 1986, S. 386).

„Das einfachste Vorgehen besteht darin, die postulierten Stichprobenunabhängigkeiten zu hinterfragen und die Probandenstichprobe nach einem relevanten Kriterium (z.B. Alter, Geschlecht, Sozialisation, etc., oder nach dem untersuchten Persönlichkeitsmerkmal selbst, []) in zwei oder mehrere Substichproben zu unterteilen und in jeder der Substichproben getrennte Itemparameterschätzungen vorzunehmen“ (Amelang & Zielinski, 2002, S. 79). Mit Hilfe eines grafischen Modelltests kann man, sehr anschaulich in einem Streuungsdiagramm, die jeweiligen zwei Itemparameter miteinander vergleichen. Wie bei Kubinger (2006, S. 87) beschrieben wäre der Idealfall „das Bild einer durch den Ursprung gehenden 45°-Geraden“. In der Praxis ist meist ein mehr oder weniger breiter Punkteschwarm um die Hauptdiagonale zu erkennen. []…„je näher die Itemparameter an der Hauptdiagonalen zu liegen kommen, desto größer ist die Stichprobenunabhängigkeit und desto eindeutiger die Rasch-Homogenität“ (Amelang & Zielinski, 2002, S. 79).
Genügt eine solche Methode nicht gibt es verschiedene numerische Verfahren, „die als globale Modelltests prüfen sollen, ob die Daten insgesamt den Modellannahmen entsprechen, oder ob sich signifikante Abweichungen zeigen“ (Stelzl, 1979 S. 654).

Es wird hier aus ökonomischen Gründen nur auf den Likelihood ratio test von Andersen (1973) näher eingegangen.

5.1.1 Likelihood ratio test von Andersen (1973)

Laut Kubinger und Jäger (2003, S. 418) ist „der bekannteste inferenzstatistische Modelltest“ der sogenannte „(bedingte) Likelihood-Ratio-Test von Andersen“. „Damit wird geprüft, ob die beobachteten Daten durch die in verschiedenen Teilstichproben separat geschätzten Itemparameter wesentlich besser erklärt werden können als durch die entsprechenden Schätzungen in der Stichprobe insgesamt“ (Kubinger & Jäger, 2003, S. 418).

> The proposed goodness of fit test is based on a comparison between difficulties estimated from different score groups an over-all estimates (Andersen, 1973, S. 123).

Praktisch gesehen bedeutet es nun, wenn man die gesamte Stichprobe der Modellprüfung unterzieht, dass k-1 freie Parameter vorhanden sind (k-1, da ein Item für die Normierung vorgesehen ist). Teilt man nun die Stichprobe in zum Beispiel 2 Untergruppen wie Geschlecht (männlich/weiblich) ergibt das eine weniger restriktive Testung mit zwei mal k-1 Freiheitsgraden; d.h. man hat doppelt so viele freie Parameter wie bei der Gesamttestung. Die Idee von Andersen liegt nun darin, dass viele freie Parameter eine bessere Anpassung an das Modell erlauben und somit genauer messen als wenn man die gesamte Stichprobe auf Rasch-Homogenität überprüft (Rost, 2004).

Der Modelltest von Andersen (1973) ermöglicht dadurch im Vergleich zu anderen älteren Modelltests „eine theoretisch fundierte Signifikanz-prüfung für den Gesamttest“ (Stelzl, 1979 S. 654).

> Für die Itemparameter werden sowohl aus den Gesamtdaten wie auch aus Teilstichproben, die nach dem Testrohwert oder auch nach anderen Kriterien gebildet werden können, gesondert die bedingten Maximum-Likelihood-Schätzungen berechnet. Die Übereinstimmung der Parameter aus g Teilstichproben wird folgendermaßen geprüft:

$$\lambda = [L(e)]/[\overset{g}{\Pi} L_r(e)] \qquad \chi^2 = -2 \log \lambda$$

$$r = 1$$

$$df = (g-1)\,(k-1)$$

k = Itemzahl
g = Zahl der Teilstichproben, in die die Gesamtstichprobe unterteilt wurde.
L (e) = bedingte Likelihood der Gesamtdaten bei Verwendung der Parameterschätzungen aus den Gesamtdaten
Lr (e) = bedingte Likelihood der Daten aus der r-ten Teilstichprobe, bei Verwendung der aus dieser Teilstichprobe errechneten Parameter (Stelzl, 1979 S. 655).

Wie bei Andersen (1973, S. 123) zu ersehen ist, ergibt sich nun – basierend auf den Schätzungen innerhalb der Scoregruppe und den Gesamtschätzungen der Itemschwierigkeiten – ein bedingtes Wahrscheinlichkeitsverhältnis, welches chi^2-verteilt ist, wenn das Rasch-Modell gilt.

5.1.1.1 Kritik an Andersen's Modelltest

Das Modell nach Rasch arbeitet im Gegensatz zu anderen Modellen (wie schon oben illustriert) nur mit dem Schwierigkeitsparameter, es kommt somit ohne den Parameter der Diskriminationsstärke aus, was laut Andersen (1973) ein bis dato ungelöstes Problem darstellt, wenn die Diskriminationsstärken der Items nicht ident sind.

> If e.g. the item discriminating powers are substantially different, we get substantially different item parameter estimates from a group of individuals with high raw scores than from a group of individuals with low raw scores (Andersen, 1973, S. 124).

„Das vielfach praktizierte Vorgehen, das Datenmaterial nach dem Testrohwert und einigen nebenher angefallenen Kriterien wie Alter und Geschlecht zu teilen, ist nicht geeignet, Inhomogenitäten gleichsam automatisch anzuzeigen“ (Stelzl, 1979, S. 652). Stelzl hat in ihren Arbeiten festgestellt, dass der Likelihood ratio test von Andersen (1973) nicht dazu geeignet ist Modellabweichungen zu entdecken, „wenn in den Daten Rasch-homogene Subskalen enthalten sind, die unterschiedliche latente Dimensionen erfassen“ (Gittler, 1999, S. 77).

„Basierend auf einem Vorschlag von Wollenberg (1979, nach Gittler, 1999, S. 77) konnte Formann (1981, nach Gittler, 1999, S. 77) jedoch zeigen, dass bei Verwendung von Einzelitems als Teilungskriterium diesem Mangel effizient begegnet werden kann“ (Gittler, 1999, S. 77).

„Denn ohne genauere Vorkenntnis sind nur die einzelnen Items unmittelbar Ausdruck all dessen, was an verschiedenen Eigenschaften durch einen Test gemessen wird, so dass durch eine derartige Modellprüfung alle Inhomogenitäten aufdeckbar sein müssten, die auf das Zusammenspiel mehrere Dimensionen zurückführbar sind“ (Formann, 1981, S. 553f).
Weiters müsste laut Formann (1981, S. 554) anhand eines solchen Vorgehens auch nachvollziehbar sein, „welche Items in dem Sinn als zusammengehörig angesehen werden können, als sie ein und dieselbe latente Dimension erfassen“.

5.1.2 Sensibilität von spezifischen Modelltests in Abhängigkeit der Modellverletzungen

Wie bei Mair und Ledl (2006) schon zu Beginn angemerkt wird, kann die Ursache für eine Modellabweichung verschiedene Formen annehmen indem die, weiter oben angegebenen, Voraussetzungen für das Rasch Modell verletzt sind, wobei nicht eindeutig gesagt werden kann wie genau die Teststatistiken auf bestimmte Modellverletzungen reagieren.

Die Statistik ist laut Glas und Verhelst nicht allein daran interessiert einen einzigen allumfassenden Modelltest zur Verfügung zu stellen, sondern ist bemüht auf die spezifischen Modellverletzungen einzugehen.

> The statistics presented not only support the purpose of a global overall model test, but also provide information with respect to specific model

> violations, such as violation of sufficiency of the sum score, strictly monotone increasing and parallel item response functions, unidimensionality, and differential item functioning (Glas & Verhelst in Fischer & Molenaar, 1995, S. 69).

Glas und Verhelst führen in Fischer und Molenaar (1995, S. 69), unter anderem, je nach Modellverletzung die jeweils sensiblen Modelltests an:

> ...*"tests that focus on the assumptions of sufficiency of the sum score and of strictly monotone increasing and parallel item response functions"*.
>
> - Martin-Löf (1973) T-test
> - Van den Wollenberg (1982) Q_1-test
> - Glas (1988a, 1989) R_1-test
> - Molenaar (1983) U_i-test
> - S_i- and M-test (Verhelst & Eggen, 1989; Verhelst, Glas, & Verstralen, 1994)

In den Studien von Mair und Ledl (2006) geht ebenfalls hervor, dass der Q1-Test von Wollenberg und der R1- bzw. der R2-Test von Glas sensibel für die Verletzung der Annahme der parallelen Item-charakteristikkurven der Items sind. In ihren Arbeiten erwiesen sich jedoch auch der Likelihood ratio test von Andersen, der Fischer-Scheiblechner-Test und der Wald-Test empfindlich für diese Art der Modellverletzung.

> „Not only the estimates obtained from subgroups formed on the basis of background variables should be approximately equal, also estimates obtained at different score levels should, within chance limits, be equal" (Glas & Verhelst in Fischer & Molenaar, 1995, S. 70).
>
> - Andersen's likelihood ratio test (Andersen, 1973b)
> - Fischer-Scheiblechner test (Fischer & Scheiblechner, 1970; Fischer, 1974)

- Wald test

„The assumptions of unidimensionality of the parameter space and of local stochastic independence are the focus of …“ (Glas & Verhelst in Fischer & Molenaar, 1995, S. 70).

- Martin-Löf (1973, 1974) likelihood ratio test
- Van den Wollenberg (1982) Q_1-test
- Glas (1988a, 1989) R_1-test

Laut Mair und Ledl (2006) ist es insbesonders der Q2 von Wollenberg und der R2 von Glas, welche das höchste statistische Vermögen in der Erkennung der Verletzung der Eindimensionalität und der lokalen stochastischen Unabhängigkeit aufweisen.

Beschreibungen bzw. eine vertiefende Einführung der genannten Modelltests sind bei Glas und Verhelst in Fischer und Molenaar (1995) nachzulesen.

6 Itemselektion im Rasch-Modell

Erfüllt ein Datensatz die notwendigen Modellvoraussetzungen nicht, muss nach den zugrundeliegenden Ursachen gesucht werden. Das Problem kann an der Aufgabenstellung selbst (also den einzelnen Items) oder an der untersuchten Stichprobe liegen. Die Stichprobe kann dem Rasch-Modell in dem Sinn nicht genügen als das die Testteilnehmer kein rationales Lösungsverhalten gezeigt haben könnten. Es genügt laut Aussage von Gittler nur ein geringer Prozentsatz an unmotiviert bzw. unseriös arbeitenden Personen für das Zustandekommen von Inhomogenität.

Andererseits könnten Items einzelne Subgruppen bevorzugen bzw. benachteiligen, was ebenfalls den Voraussetzungen eines Rasch-homogenen Verfahrens zuwider laufen würde. Es wäre nun naheliegend diese Aufgaben einfach zu entfernen, sobald sie als signifikant angezeigt werden, jedoch argumentiert Gittler:

> Das routinemäßige Ausscheiden von Items, welche aufgrund „technisch" gegebener Kriterien (z.B. z-Werte; graphische Modelltests) als inhomogen identifiziert wurden, kann – wie [Validierungsstudien] bestätigen – zu artifizieller Modellanpassung führen (Gittler, 1986, S. 386).

Dieser „technisch begründeten Itemselektion" wird die „psychologisch-inhaltlich begründete Selektionsstrategie" (Gittler, 1986, S. 386) gegenübergestellt, welche die Modelltestung von Beginn an und bereits in ihrer Planung begleiten sollte. Dabei sollten vor dem Selektionsprozess, wie vielmehr vor jeglichen Aktionen, inhaltliche Überlegungen angestellt werden („a-priori-Hypothesen über eventuelle „systematische Unterschiede" zwischen Item- und/oder Personengruppen") um „konsistentere und diagnostisch relevantere Ergebnisse" zu erhalten (Gittler, 1986, S. 386).

Im Gegensatz zur technisch begründeten Itemselektion liegen jetzt bereits vor dem ersten Analyseschritt einige „Selektionshypothesen“ vor, die aufgrund inhaltlich-psychologischer Einsichten gewonnen wurden. Sind nun die folgenden globalen Modellkontrollen signifikant, so können diese Hypothesen überprüft werden. Neuerliche Inhomogenitäten sollten zur Generierung anderer, vielleicht präziserer Hypothesen auf inhaltskritischer Basis Anlaß [sic] geben, wobei durchaus technisch gegebene Selektionshinweise mit berücksichtigt werden sollen (Gittler, 1986, S. 393).

7 Der „Mental Arithmetic Test“ (MAT)

Der „Mental Arithmetic Test“ ist ein Verfahren, wessen Aufgabenbewältigung eine Kombination aus Gedächtnis-, und Rechenfähigkeit, wie auch logischem Denken verlangt. Die Probanden müssen drei Zwischenergebnisse im Gedächtnis behalten und zusätzlich nach bestimmten Regeln vorgehen um das Endergebnis berechnen zu können. Diese Form numerische Intelligenz zu erheben bedarf also, neben generellen arithmetischen Fähigkeiten, auch ein hohes Ausmaß an Merkfähigkeit und misst somit die Fähigkeitsdimension „Kopfrechnen“.
„Die Fähigkeitsdimension Kopfrechnen stellt eine kognitive Fertigkeit dar, mit Zahlen und Operatoren schnell und effizient, ohne materielle Hilfsmittel umgehen zu können, um logisch-mathematische und numerische Problemstellungen zu lösen“ (Gabriel, 2004, S. 9).

7.1 Beschreibung und Itemgenerierung des MAT

Alle Poolitems entsprechen der Form aus Tabelle 1. Die rot markierten Zahlen stellen die Zwischenergebnisse dar, die temporär im Gedächtnis zu behalten sind und nicht notiert werden dürfen. Die blau markierte Zahl ist das Endergebnis, welches stets das Resultat nach einer Division oder Multiplikation ist. Das erste und das zweite Zwischenergebnis (in unserem Fall „23“ und „49“) ist immer Ergebnis einer Addition oder Subtraktion. Ist das erste Zwischenergebnis kleiner als das zweite, sind die beiden Teilergebnisse zu addieren (hier müsste man also 23 und 49 addieren, da das erste Zwischenergebnis „23“ kleiner ist als das zweite „49“), was zum nächsten Zwischenergebnis (hier „72“) führt. Anderenfalls, wenn das erste Zwischenergebnis größer als das zweite ist,

müssen die beiden Zahlen subtrahiert werden. Kann die Aufgabe nicht gelöst werden soll die Ziffer „0“ ins Antwortfeld eingegeben werden.

Tabelle 1: Item 5 mit rot markierten Zwischenergebnis und blau markiertem Endergebnis

Pool 5	C	D	E	F	G
-	67	44		23	
	72	/	9	=	8
+	32	17		49	

Es soll bei dieser Art der Vorgabe laut Gabriel (2004) hauptsächlich das Kurzzeitgedächtnis angesprochen werden. Das Abrufen von gespeichertem Wissen aus dem deklarativen Gedächtnis soll vermieden werden, um die eigentliche Rechenleistung der Personen zu erhalten. Umgelegt auf das ACT-R-Modell von Anderson soll möglichst nicht die Abrufstrategie Verwendung finden, sondern vorwiegend die Berechnungsstrategie.

Wie in der Tabelle 1 zu ersehen ist, wird die horizontale Vorgabeform der Rechenaufgaben mit der vertikalen kombiniert. Trbovich und Lefevre (2003) konnten in ihrer Untersuchung zeigen, dass die Teilnehmer die Aufgaben, welche vertikal formatiert waren, schneller lösen konnten als jene, welche horizontal präsentiert wurden. Weiters wurden die horizontal präsentierten Aufgaben auch als wesentlich schwieriger empfunden als die vertikal formatierten. Eine Begründung dürfte darin liegen, dass die vertikale Präsentation in der Regel der Form entspricht, welche in der Grundschule gelehrt wird und weitgehend im Alltag Verwendung findet (Gabriel 2004, S. 59).

Es wurde in der Arbeit von Trbovich und Lefevre (2003) gemäß Hitch (1978) und Hayes (1973) angenommen, dass die vertikale Präsentation der Rechenaufgaben temporäre Repräsentationen im visuell räumlichen Notizblock aktiviert und somit in der vertikalen Bedingung eher der

visuell räumliche Notizblock involviert ist, während in der horizontalen Vorgabe der Aufgaben angenommen wurde, dass die Teilnehmer eine Vielzahl an Lösungsmethoden zeigen würden. Beispielsweise könnte die Aufgabe 79 + 4 auch folgendermaßen gelöst werden: 80 + (4 - 1). Es wird vermutet, dass bei solchen alternativen Lösungsansätzen eher die phonologische Schleife zum Tragen kommt, da versucht werden wird durch subvokale Wiederholung die jeweiligen Zwischenergebnisse im Gedächtnis zu behalten. Aus diesem Grund wird angenommen, dass die horizontale Präsentation der Aufgaben mit größerer Wahrscheinlichkeit den phonologischen Speicher aktivieren wird als die vertikale Präsentation. Trbovich und Lefevre (2003) überprüften diese Vermutung indem sie phonologische bzw. visuelle Gedächtnisaufgaben mit vertikal bzw. horizontal präsentierten Rechenaufgaben kombinierten. Es zeigte sich, dass die Leistung in den Aufgaben schlechter war, welche das phonologische Gedächtnis in Anspruch genommen haben, wenn die Rechenaufgaben horizontal präsentiert wurden und dass die Leistung schlechter war in den Aufgaben, welche das visuelle Gedächtnis beschäftigten, wenn die Rechenaufgaben vertikal präsentiert worden sind. Dieses Ergebnis ist laut Trbovich und Lefevre der Beweis dafür, dass phonologische wie auch visuelle Aspekte des Arbeitsgedächtnisses bei mentalen arithmetischen Problemstellungen angesprochen werden in Abhängigkeit der Präsentationsform (vertikal oder horizontal).
Um die Items des MAT zu lösen werden somit phonologische wie auch visuelle Aspekte des Arbeitsgedächtnisses in Anspruch genommen, wobei jedoch angenommen wird, dass bei dieser Form der Problemstellung die phonologischen Anteile überwiegen bzw. dass auch die vertikal präsentierte Teilaufgabe mittels phonologischer Schleife gelöst wird, da keines der Zwischenergebnisse niedergeschrieben werden darf.

Bis auf Poolitem 16 (siehe Kapitel 8), welches nur einstellige Zahlen enthält, beinhalten die Aufgaben ein- bis dreistellige Zahlen. Trbovich

und Lefevre (2003) konnten zeigen, dass bei den horizontal präsentierten Aufgaben diejenigen Problemstellungen schneller gelöst wurden, bei welchen die zweistellige Zahl vor der einstelligen platziert wurde (zum Beispiel 63 + 4) und nicht umgekehrt (zum Beispiel 4 + 63). Es wird hier angenommen, dass es sich zwischen drei- und zweistelligen Zahlen ähnlich verhält. In den Poolaufgaben wurde demnach stets die größere vor die kleinere Zahl platziert.

Bei der Generierung der Items wurde anhand der obigen Erkenntnisse wie auch nach Gabriel (2004, Kapitel 4) darauf geachtet bestimmte Regelmäßigkeiten einzuhalten.

1. Bei der oberen wie auch bei der unteren horizontalen Präsentation wurde darauf geachtet, dass die größere vor der kleineren Zahl steht.
2. Der Zahlenraum 1 bis 999 sollte eingehalten werden.
3. Es sollten keine gleichen Zahlen in der Angabe vorkommen.
4. Die mittlere horizontale Präsentation sollte entweder eine Divisions- oder eine Multiplikationsaufgabe sein.
5. Der Multiplikator bzw. der Divisor sollte zwischen 2 und 99 liegen.
6. Es sollten keine gleichen Ziffern in der Einerstelle in Zwischenergebnissen und Angabe enthalten sein.
7. Bei keinen Zahlen sollte in der Einerstelle eine Null stehen, mit Ausnahme vom Endergebnis.
8. Das Ergebnis ist immer eine ganze Zahl.
9. Die Zwischenergebnisse sollten nicht in der Angabe enthalten sein.

Die Schwierigkeit der Items variiert laut Gabriel (2004) nach gewissen Kriterien.

> Erstens durch die Anzahl der Ziffern der Zahlen, wobei sich der Zahlenraum auf 1 bis 999 beschränkt. Zweitens stellt bei Addition und Subtraktion die Anzahl der Überträge pro Aufgabe einen Indikator für die Schwierigkeit der Aufgabe dar (Ashcraft et al., 1992; Hitch, 1978; Lemaire et al., 1996; Salthouse & Coon, 1994; nach Gabriel, 2004, S. 60). Drittens ändert bei Multiplikationen bzw. Divisionen die Höhe der Zahl des Multiplikators bzw.

> Divisors die Schwierigkeit dieser Aufgabe. Viertens erhöht sich die Schwierigkeit der Aufgabe mit der Anzahl der im temporären Speicher zu behaltenden Information und fünftens besteht die aufsteigende Schwierigkeitsreihenfolge der vier Grundrechnungsarten nach Mittring (2001, nach Gabriel, 2004, S. 60) in der Reihenfolge Addition, Subtraktion, Multiplikation und Division.

Wie bereits oben erwähnt, konnte Hitch (1978) nachweisen, dass mit der Anzahl der Überträge (carrying) die Schwierigkeit steigt. Er unterscheidet zwischen vier Problemstufen: „NC“ („no carrying“), „CT“ („carrying in the tens“), „CH“ („carrying in the hundreds“) und „CTCH“ („carrying in both tens and hundreds“) (Hitch, 1978, S. 304). Die Ziffer Null wurde nie verwendet. Der Großteil der Teilnehmer rechnete erst mit den Einer-, dann mit den Zehner- und dann mit den Hunderterstellen (UTH = units, tens, hundreds). Nur ein kleinerer Teil der Personen variierte die Strategie je nach Anforderung bezüglich der Überträge. In der Regel wurde die Reihenfolge UTH eingehalten, wenn Überträge vorhanden waren. Waren keine Überträge vorhanden rechneten die Teilnehmer in der Reihenfolge HTU. Die Lösungszeit stieg mit der Anforderung an Überträgen (NC → CT→ CH → CTCH). Diese Studie ist hier auch deshalb von Belang, da mit drei- und zweistelligen Zahlen gearbeitet worden ist, was auch beim MAT der Fall ist. Es ist also anzunehmen, dass auch hier in ähnlicher Weise mit den Zahlen gearbeitet worden ist.

7.2 Vorgabe/Instruktion des MAT

Die Testungen mit dem MAT erfolgten größtenteils computerunterstützt, womit auch die Instruktion schriftlich durchgeführt wurde. Zu Beginn jeder Testung musste ein Probandencode (pbcode = Matrikelnummer), wie auch Alter, Geschlecht und höchste abgeschlossene Schulbildung angegeben werden wie in Abbildung 8 ersichtlich ist. Danach wurde anhand eines fiktiven Items, wie aus Tabelle 2 zu ersehen, der Rechenhergang illustriert und anschließend zwei Beispielitems zum eigenständigen Rechnen vorgegeben. Erst darauffolgend wurde mit der eigentlichen Testung begonnen.

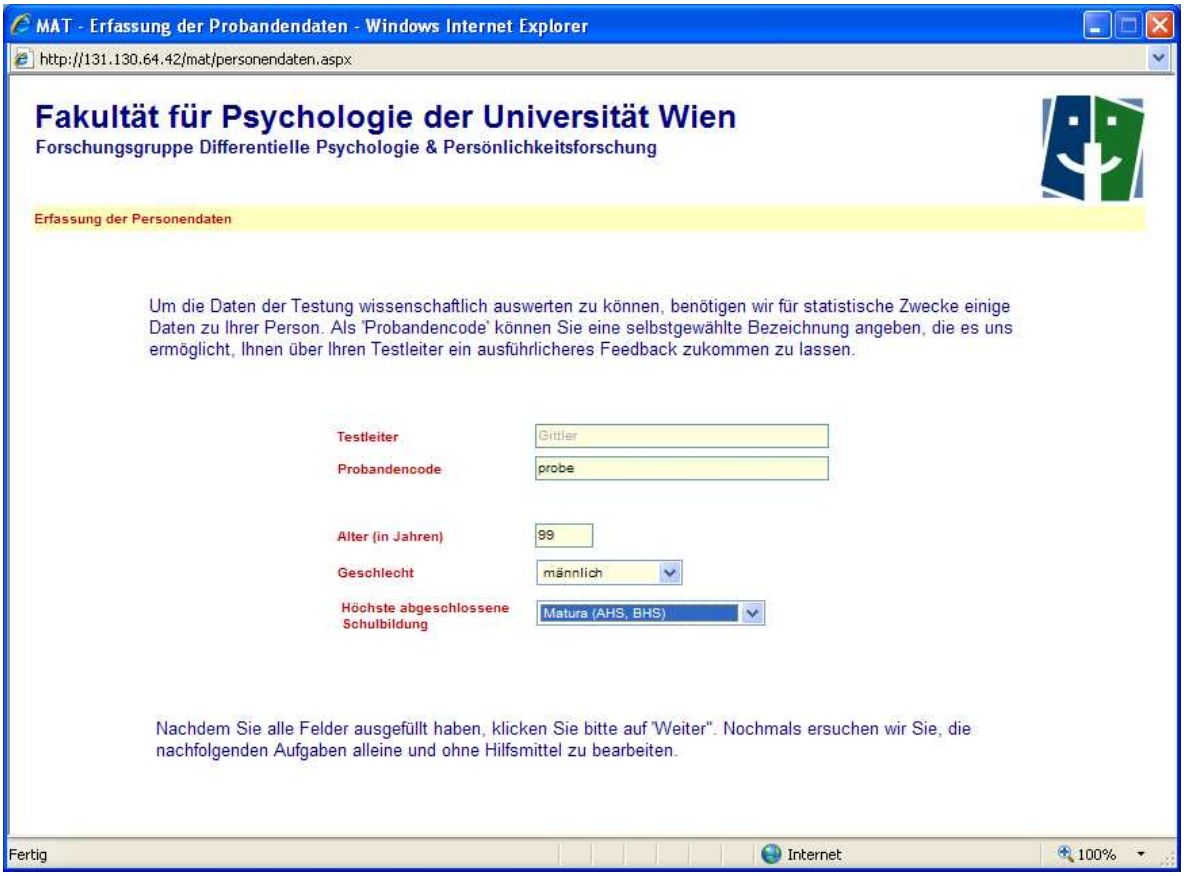

Abbildung 8: verwendeter Rahmen zur Vorgabe des MAT.

Tabelle 2: Beispielitem

	C	D	E	F	G
+	6	2			
			x	9 =	.....
-	5	4			

7.3 die 8 Versionen des MAT

Der Menthal Arithmetic Test besteht aus acht Versionen mit jeweils 23 bzw. 16 Aufgaben. Alle Versionen haben 4 „Linkitems“ gemein (Poolitems: 16, 17, 20, 23), welche hier fett unterlegt sind. Diese sind notwendig, um neben den Einzelanalysen der jeweiligen Versionen auch ein unvollständiges Rasch-Modell zu erheben. Von den insgesamt 109 Poolitems wurden lediglich 95 vorgegeben.

- R1: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, **16**, **17**, 18, 19, **20**, 21, 22, **23**
- R2: 1, 13, 14, 15, **16**, **17**, 18, 19, **20**, 21, 22, **23**, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
- A: **16**, 24, 58, 70, **23**, 80, 95, 46, **20**, 62, 83, 98, **17**, 34, 65, 37
- B: **16**, 32, 59, 73, **23**, 82, 96, 52, **20**, 63, 88, 108, **17**, 42, 71, 38
- C: **16**, 43, 60, 75, **23**, 84, 101, 53, **20**, 64, 89, 109, **17**, 44, 72, 39
- D: **16**, 50, 66, 77, **23**, 85, 26, 55, **20**, 68, 91, 25, **17**, 45, 74, 40
- E: **16**, 54, 67, 78, **23**, 86, 27, 57, **20**, 76, 92, 28, **17**, 47, 90, 41
- F: **16**, 56, 69, 79, **23**, 87, 33, 61, **20**, 81, 94, 31, **17**, 51, 36, 49

Insgesamt wurde mit n 1400 Daten gearbeitet, welche sich wie folgt auf die 8 Versionen verteilen:

Tabelle 3: Aufteilung der 1400 Personen auf die jeweiligen Versionen

VERSION

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	A	202	14,4	14,4	14,4
	B	135	9,6	9,6	24,1
	C	130	9,3	9,3	33,4
	D	130	9,3	9,3	42,6
	E	134	9,6	9,6	52,2
	F	118	8,4	8,4	60,6
	R1	328	23,4	23,4	84,1
	R2	223	15,9	15,9	100,0
	Gesamt	1400	100,0	100,0	

Wie man aus Tabelle 3 ersehen kann haben die meisten Teilnehmer Version R1 vorgelegt bekommen.

8 Empirischer Teil

Die mir zur Verfügung gestellten 1720 Personendaten stammen von Erhebungen aus dem Zeitraum WS 2006/2007 bis WS 2007/2008 im Zuge des Prüfungsbonus für das Fach „Differentielle Psychologie“. Es handelt sich daher um eine Stichprobe, die sich hauptsächlich aus Studenten (im überwiegenden Teil aus dem Fach Psychologie) zusammensetzt. Es erfolgt im Anschluss an die deskriptive eine statistische Auswertung der Daten mit dem Ziel der Überprüfung der Homogenität der Items nach dem Modell von Rasch (1960).

8.1 Zusammensetzung der Stichprobe

Die gesamte Stichprobe (n = 1720) setzt sich prozentuell aus gleich vielen männlichen wie auch weiblichen TestteilnehmerInnen zusammen (Männer: n = 858; Frauen: n = 862; entspricht einer Verteilung von 50:50; siehe Tabelle 4: deskriptive Statistik von Männern und Frauen). Das durchschnittliche Alter liegt bei 23 Jahren wie man in Tabelle 6 und Abbildung 9 ersehen kann.

Tabelle 4: deskriptive Statistik von Männern und Frauen

SEX

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	m	858	49,9	49,9	49,9
	w	862	50,1	50,1	100,0
	Gesamt	1720	100,0	100,0	

Die durchschnittliche Bearbeitungszeit in der Versionen R1 und R2 (lange Version mit 23 Items) betrug 30,05 Minuten (Median 27,57

Minuten) und bei den Versionen A bis F (kurze Version mit 16 Items) 27,79 Minuten (Median 25,91 Minuten) wie in Tabelle 5 ersichtlich.

Tabelle 5: deskriptive Statistik der Bearbeitungszeit in Minuten für die langen und die kurzen Versionen des MAT

Statistiken

TDAUER_M

Gr.1: k=23 - Lange Version	N	Gültig	670
		Fehlend	0
	Mittelwert		30,0533
	Median		27,5750
	Standardabweichung		14,7567
	Varianz		217,7603
	Minimum		1,57
	Maximum		107,25
Gr.2: k=16 Kurze Version	N	Gültig	1050
		Fehlend	0
	Mittelwert		27,7887
	Median		25,9083
	Standardabweichung		14,1879
	Varianz		201,2951
	Minimum		1,65
	Maximum		119,22

Tabelle 6: deskriptive Statistik von Alter und Bildungsgrad

Statistiken

		AGE	EDU
N	Gültig	1720	1720
	Fehlend	0	0
Mittelwert		29,16	3,16
Median		23,00	3,00
Standardabweichung		12,36	1,02
Varianz		152,81	1,05
Minimum		16	0
Maximum		68	6

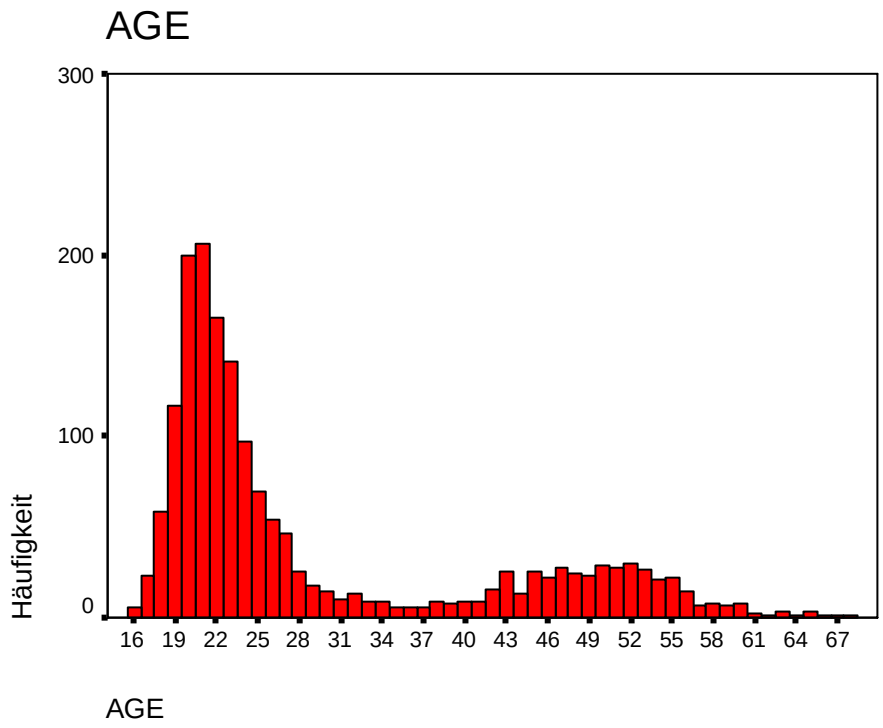

Abbildung 9: Altersverteilung (n = 1720)

63 % der Personen gab Matura (codiert als „3“) als höchste abgeschlossene Ausbildung an wie Tabelle 6 und Abbildung 10 zeigt.

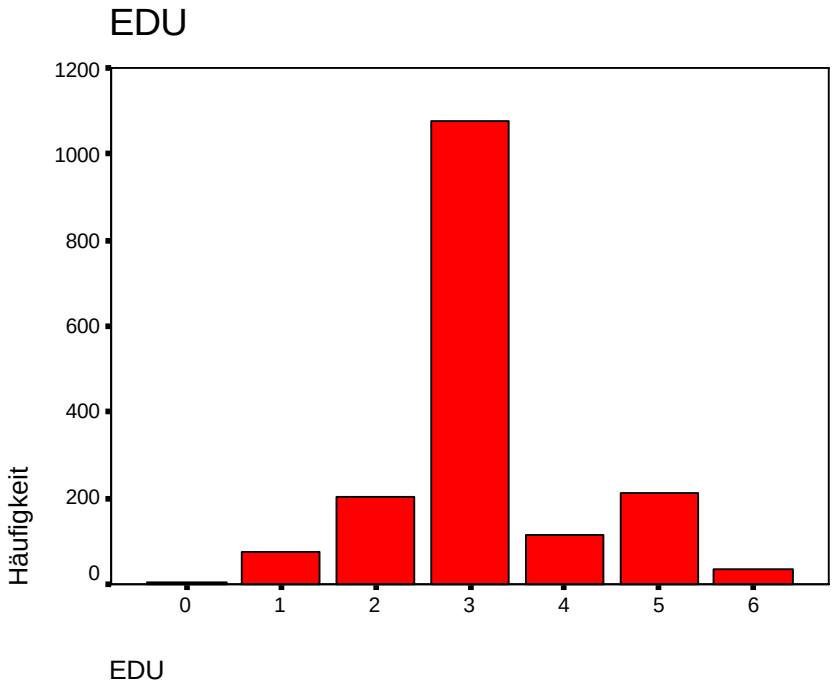

Abbildung 10: Verteilung der Schulbildung (0: Sonderschule, 1: Volksschule/Hauptschule, 2: Mittelschule, 3: Matura, 4: Hochschule, 5: abgeschlossenes Studium, 6: Doktoratsstudium)

8.2 Zustandekommen der nach dem RM untersuchten Personendaten (n=1400)

Bevor mit der Modellprüfung begonnen worden ist, wurden die Daten nach dem Prinzip der „Inhaltlich begründeten Itemselektion“ (Gittler, 1986) aufbereiten.

Die 1720 Personendaten wurden anhand der Verteilung eines Prüfungs-bonus erfasst. Jeder Student hatte somit die Möglichkeit mehrere Personen online am MAT teilnehmen zu lassen. Die Daten sind im SPSS nach Matrikelnummer (pbcode) sortiert d.h. pro Matrikelnummer gibt es mehrere Personeneinträge (vpnr).

Es wurden zunächst die gesamten Daten gruppenweise (anhand der Matrikelnummer) aufgrund der Lösung (av) und Bearbeitungszeit (tdauer_m) auf adäquates Lösungsverhalten hin überprüft. Dabei hat man sich vorwiegend an den durchschnittlichen Lösungszeiten (27,76 Min. für 23 Items; 26,55 Min. für 16 Items) orientiert. Es wurde daher beschlossen alle Personen als unseriös zu betrachten, welche die Aufgaben in weniger als 7 Minuten gelöst und mehr als die Hälfte richtig beantwortet haben. Das würde bedeuten, dass für Bearbeitung pro Item höchstens 20 Sekunden Zeit zur Verfügung stehen würden, was anhand der durchschnittlichen Lösungszeiten wie auch anhand von Testläufen sehr unwahrscheinlich wäre. Haben nun bei einer Matrikelnummer alle beziehungsweise der Großteil der Personen die Aufgaben unglaubwürdig schnell und vorwiegend richtig gelöst, musste angenommen werden, dass die Aufgaben nicht seriös bearbeitet worden sind. Es wurden in Folge alle Daten der jeweiligen Matrikelnummer ausgeschieden, da die Zuverlässigkeit des jeweiligen Testleiters generell in Frage gestellt werden musste. Haben Personen bis auf ein oder zwei Aufgaben keines der Items richtig beantwortet und liegt die Bearbeitungszeit unter 7 Minuten wurden diese Personen ebenfalls ausgeschieden, da vermutet

werden musste, dass es sich hier um unmotivierte Teilnehmer handelt. Weiters wurden alle Personen ausgeschieden, die keines der Items richtig bearbeitet haben, da auch hier kein Informationswert über die jeweilige Person zu erhalten wäre.

Es wurden somit folgende 38 Teilnehmer ausgeschieden:

Tabelle 7: Auflistung der im ersten Durchlauf ausgeschiedenen Personeneinträge (pbcode aus Gründen der Anonymität frei erfunden)

pbcode	vpnr	tdauer_m	av
A	384	5,22	111111111111111111111111
A	807	21,5	1111111111111111
A	383	5,75	111111111111111111111111
A	1049	9,62	1111111001101111
A	382	2,62	111111111111111111111111
A	1210	6,18	1111100001101111
A	381	2,58	111111111111111111111111
A	1377	6,45	1110110010111110
A	1536	7,82	1111111101011001
A	1699	6,73	1110110100111010
B	1268	7,62	1111111111111111
B	903	5,72	1111011111111111
B	1754	5,28	1111011111111011
B	1112	5,27	1111111111111111
B	1432	5,03	1111011111111010
B	1603	7,82	1111100111111111
C	226	20,9	11001000000001000010000
C	750	16,13	0111011101001001
C	225	3,67	11011011001011011110111
C	749	2,97	0001001111010101
C	224	19,25	11101110101011111110111
C	748	3,85	0111011100010101
C	223	4,3	11101101110011011011110
C	747	1,68	0101011111101111
C	746	1,97	0100001111001000
C	745	3,15	0001000110000100
D	1494	1,65	1000000000000000
E	171	3,82	00000000000000000000100
F	830	8,62	1111101110111110
F	829	8,5	1011011111011111
F	828	5,42	1111111111111111
F	827	5,55	1111111111111100
F	825	23,32	1111111101101000
F	826	33,02	1000011111111111
G	99	1,57	11111001001011001111100
H	485	2,12	10000000000000000000000
I	974	3,83	0000000000000000
J	484	1,72	10000000000000000000000

Anhand von Vorerhebungen der Rasch-Homogenität für nunmehr 1682 Daten musste festgestellt werden, dass sich fast in allen Versionen Item 16 bezüglich seiner Itemschwierigkeiten als sehr auffällig herausstellte.
Die Aufgabe 16 ist unter allen vorgegebenen Aufgaben die am leichtesten zu lösende, da die Angabe allein einstellige Zahlenwerte enthält und somit der Gedächtnisaufwand wie auch der Schwierigkeitsgrad wohl nicht all zu hoch ist.

Pool 16

	C	D	E	F	G
+	8	5		13	
		6	*	4 =	24
-	9	2		7	

Item 16 ist in den Versionen A bis F das erste Item und nimmt daher – trotz der vorigen zwei Einführungsitems - den Stellenwert eines warming-up-Items ein. Warming-up-Items sind unter Umständen nicht als gleichwertig mit den übrigen Aufgaben anzusehen, da die Versuchspersonen sich oft erst an die Anforderung der Aufgabe gewöhnen müssen. Poolitem 16 unterscheidet sich somit aufgrund seiner Konstruktion, da lediglich einstellige Zahlen in der Angabe vorhanden sind, und seiner Positionierung grundlegend von den anderen Poolaufgaben. Die Vermutung liegt nahe, dass Item 16 weniger numerische Intelligenz prüft, als vielmehr die Motivation zur Durchführung der Testung. Da man bemüht ist die Anzahl der unseriösen bzw. unmotivierten Bearbeitungen einzuschränken, wurde beschlossen alle Personen auszuscheiden, welche Item 16 nicht gelöst haben. Es wurden somit weitere 282 Personen ausgeschieden, womit nunmehr mit n = 1400 Personen gearbeitet wird.
Es soll hier nicht der Anschein erweckt werden eine „perfekte“ Stichprobe erzeugen zu wollen in der alle unzulänglichen Personen so lange ausgeschieden werden bis Rasch-Homogenität die Folge ist. Da es sich um die erste Poolanalyse der gesamten Daten handelt riskiert man ein Ausscheiden von „guten“ Testteilnehmern um das Ausscheiden von

Items aufgrund einer schlechten Stichprobe zu verhindern (Rost, 2004). Es soll sicher gestellt werden, dass die verbleibenden Personen motiviert sind und rational arbeiten, womit die Voraussetzung für eine Überprüfung nach dem Rasch-Modell erfüllt wäre.

Für die verbleibenden 1400 Daten wird nun die deskriptive Statistik neuerlich aufgeführt.

Das Verhältnis von Frauen und Männern hat sich wie aus Tabelle 8 ersichtlich, nicht verändert.

Tabelle 8: deskriptive Statistik von Männern und Frauen

SEX

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	m	700	50,0	50,0	50,0
	w	700	50,0	50,0	100,0
	Gesamt	1400	100,0	100,0	

Alter und Bildungsgrad haben sich in ihren Werten nicht verändert wie man aus Tabelle 9 ersehen kann.

Tabelle 9: deskriptive Statistik von Alter und Bildungsgrad

Statistiken

		AGE	EDU
N	Gültig	1400	1400
	Fehlend	0	0
Mittelwert		29,22	3,15
Median		23,00	3,00
Standardabweichung		12,48	1,01
Varianz		155,65	1,01
Minimum		16	0
Maximum		68	6

Die durchschnittliche Bearbeitungszeit in der Versionen R1 und R2 (lange Version mit 23 Items) beträgt nunmehr 30,53 Minuten (Median 27,67 Minuten) und bei den Versionen A bis F (kurze Version mit 16

Items) 28,48 Minuten (Median 26,58 Minuten) wie in Tabelle 10 ersichtlich.

Tabelle 10: deskriptive Statistik der Bearbeitungszeit in Minuten für die langen und die kurzen Versionen des MAT

Statistiken

TDAUER_M

Gr.1: k=23 - Lange Version	N	Gültig	551
		Fehlend	0
	Mittelwert		30,5261
	Median		27,6667
	Standardabweichung		14,6636
	Varianz		215,0205
	Minimum		3,83
	Maximum		107,25
Gr.2: k=16 Kurze Version	N	Gültig	849
		Fehlend	0
	Mittelwert		28,4838
	Median		26,5833
	Standardabweichung		14,0203
	Varianz		196,5688
	Minimum		1,78
	Maximum		119,22

8.3 Überprüfung der Geltung des Rasch-Modells

Alle Berechnungen des Rasch-Modells wurden mit der Software LpcM-WiN (Fischer & Ponocny-Seliger, 1998) durchgeführt und für alle Homogenitätsprüfungen wurde der Likelihood-Quotienten-Test von Andersen (1973) verwendet (siehe 5.1.1). Hierfür werden „die Itemparameterschätzungen aus vordefinierten Personenuntergruppen [bezüglich Geschlecht, Alter, etc.] auf statistische Gleichheit geprüft“ (Gittler & Arendasy, 2003). Es wurde ein Signifikanzniveau von 1 % gewählt. Die jeweils untersuchten Versionen des MAT stellen die Forschungshypothese bzw. Nullhypothese H_0 dar (vgl. Fischer, 1974, S. 281).

8.3.1 Kriterienbildung

Die Gültigkeit des Rasch-Modells wurde mittels interner Kriterien (Mean/Median) wie auch externer Kriterien (Geschlecht, Alter, Bearbeitungszeit) geprüft. Da zwei der acht Versionen aus 23 Items (Version R1 und R2 = Gruppe 1) und sechs aus 16 Items (Version A bis F = Gruppe 2) bestehen, war es notwendig die gesamte Stichprobe, aufgrund der unterschiedlichen Durchschnittswerte der Versionen R1/R2 und A-F im Kriterium zeit_gr, zunächst temporär in zwei Gruppen zu splitten um Mittelwertsberechnungen für die jeweilige Gruppe vornehmen zu können.

Für das Kriterium der Bearbeitungszeit ergab sich ein Median von 27,53 Minuten für R1 und R2 und ein Median von 25,90 Minuten für die Versionen A bis F. Es wurde hier somit in „wenig Bearbeitungszeit“ = 0 und „viel Bearbeitungszeit“ = 1 unterteilt.
Das Kriterium des Geschlechts wurde wie folgt dichotomisiert:
w = 1; m = 0.

Das Kriterium des Alters konnte für alle n 1400 Personen in gleicher Weise unterteilt werden, da sich die Mediane (23 Jahre) in den beiden Gruppen 1 und 2 nicht unterschieden haben: < 23 Jahre = 0; > 23 Jahre = 1.

8.3.2 vollständiges Rasch Modell der 8 Versionen

Alle 8 Versionen stellen separat für sich vollständige Rasch-Modelle dar und werden auch als solche im LpcM-WiN gehandhabt.

8.3.2.1 Rasch Modell für Version R1

8.3.2.1.1 Überprüfung nach den internen Kriterien

Zu Beginn wurde nach dem Kriterium „Mittelwert und Median“ der 328 Daten geprüft. Dabei wurden die Daten in eine Gruppe mit niedrigem (n_{1mean} = 144; n_{1med} = 171) und hohem Rohscore (n_{2mean} = 184; n_{1med} = 157) geteilt. Es ergab sich ein signifikanter LRT nach Andersen von 47.12_{mean} bzw. 46.66_{med} ($p<0.01$; df = 21). 21 Freiheitsgrade deshalb, da das Item 16 bei allen Berechnungen ausgenommen wurde, d.h. „Itemanzahl minus 1, minus Item 16“.

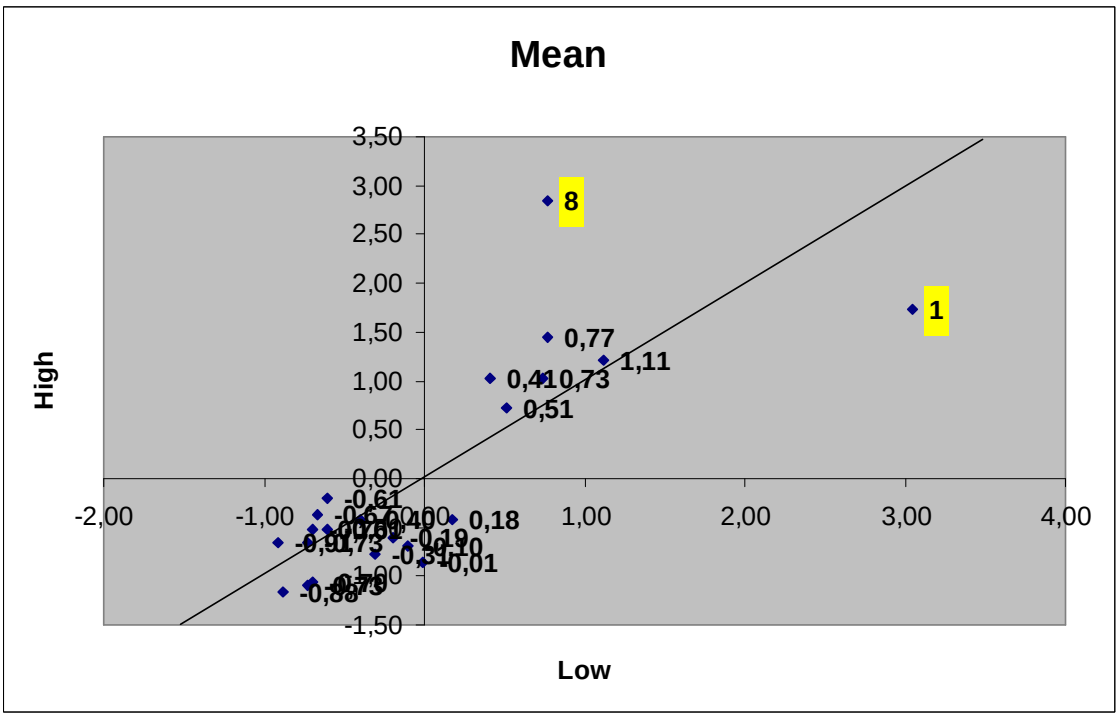

Abbildung 11: graphische Modellkontrolle bezüglich des Kriteriums „Mittelwert“: Gegenüberstellung der Itemleichtigkeiten der Subgruppen „hoher Rohscore“ = „High“ und „niedriger Rohscore“ = „Low“ für Version R1.

Bei der Gegenüberstellung der Itemleichtigkeiten beider Fähigkeitsgruppen (Abbildung 11) präsentierte sich Item 8 als sehr auffällig. Nach dessen Ausscheiden konnte ein Absinken des LRT auf 37.81 ($p<0.01$; df = 20) beim Mittelwert, wodurch knapp Rasch-Homogenität der verbleibenden 21 Items verfehlt worden ist, und auf 35.24 ($p>0.01$; df = 20) beim Median, was Homogenität zur Folge hatte. Da aufgrund der Parameterschätzwerte der beiden Subgruppen Poolitem 2 ebenfalls sehr auffällig war, wurde auch dieses ausselektiert, jedoch ohne Erfolg. Der LRT für den Median ist weiter deutlich gesunken und der LRT für den Mittelwert erwies sich weiterhin als signifikant. Erst durch zusätzliches Ausscheiden von Poolitem 12 folgte Rasch-Homogenität für die verbleibenden 19 Items auf 1Prozent- wie sogar auf 5Prozent-Niveau mit einem nicht signifikanten LRT von 27.59 bzw. 23.50 bezüglich der Kriterien „Mittelwert bzw. Median des Rohscores“ (df = 18) wie in Tabelle 11 zu ersehen ist.

Tabelle 11: Auflistung der Rasch-Modelle für die internen Kriterien Mean und Median für die Version R1 (α-Niveau = 1 %)

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Mean	n_1 = 144	328	-	22	47.12	< 0.01
	n_2 = 184					
Median	n_1 = 171				46.66	< 0.01
	n_2 = 157					
Mean	n_1 = 144		8	21	37.81	< 0.01
	n_2 = 184					
Median	n_1 = 170				35.24	0.02
	n_2 = 158					
Mean	n_1 = 134		8, 2	20	37.61	< 0.01
	n_2 = 194					
Median	n_1 = 164				24.13	0.19
	n_2 = 164					
Mean	n_1 = 122		8, 2, 12	19	27.59	0.07
	n_2 = 206					
Median	n_1 = 190				23.50	0.17
	n_2 = 138					

Warming-Up-Items spielen lediglich in Version R1 und R2 eine Rolle, da in den übrigen Versionen Poolitem 16 den Platz der ersten Aufgabe einnimmt, welches, wie oben bereits erläutert, ausgeschieden wurde. In Version R1 und R2 stellt Poolitem 1 das Warming-Up-Item dar, welches zwar stets etwas abseits der anderen Items seinen Platz einnimmt (wie in Abbildung 11 zu sehen), jedoch nahe der 45°Geraden liegt und somit die jeweiligen Subgruppen nicht benachteiligt bzw. bevorzugt, womit die Rasch-Homogenität der Versionen R1 und R2 nicht gestört ist.

8.3.2.1.2 Überprüfung nach den externen Kriterien

Das externe Teilungskriterium des Geschlechts teilte die 328 Personen in 180 Männer und 148 Frauen und das Kriterium nach Alter in 179 Personen „jünger als 23“ 149 Personen „älter als 23“. Die Modellprüfung

anhand der Kriterien Geschlecht und Alter ergab für alle 22 Items (23 minus Item 16) Rasch-Homogenität.
Das Kriterium der Bearbeitungszeit stellte die „Langsamen“ ($zeit_{n1}$ = 174) und die „Schnellen“ ($zeit_{n2}$ = 154) gegenüber. Bei der Überprüfung der Modellgeltung nach diesem Kriterium stellte sich Poolitem 6 als problematisch dar. Man erhielt zu Beginn einen signifikanten LRT nach Andersen von 40.33 (df = 21) und nach Ausscheiden des Items 6 einen nicht signifikanten LRT von 29.99 (df = 20) bei einem α-Niveau von 0.01 und 0.05 wie in Tabelle 12 zu sehen ist.

Tabelle 12: Auflistung der Rasch-Modelle für die externen Kriterien der Version R1 (α-Niveau = 1 %). n_1 beim Kriterium Zeit (Bearbeitungszeit) stellt die Subgruppe der „Langsameren“ dar; n_2 die der „Schnelleren“.

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Geschlecht	m = 180	328	-	22	26.37	0.19
	w = 148					
Alter	< 23 = 179		-	22	30.87	0.08
	> 23 = 149					
Zeit	n_1 = 174		-	22	40.33	< 0.01
	n_2 = 154					
	n_1 = 174		6	21	29.99	0.07
	n_2 = 154					

8.3.2.2 Rasch Modell für Version R2

8.3.2.2.1 Überprüfung nach den internen Kriterien

Die Modellprüfung nach den Kriterien „Mean“ und „Median“ ergab einen nicht signifikanten LRT von 34.74 für den Mittelwert und einen signifikanten LRT von 46.47 für den Median (df = 21). Wie auch bei Version R1 zeigte sich Poolitem 2 (Position 13) aufgrund der Parameterschätzwerte auffällig. Wie in Tabelle 13 zu ersehen ist erhält man für die

übrigen 21 Aufgaben nach Selektion dieses Items Rasch-Homogenität bezüglich des internen Kriteriums (LRT = 21.55_{mean}; 30.84_{med}; df = 20).

Tabelle 13: Auflistung der Rasch-Modelle für die internen Kriterien Mean und Median für die Version R2 (α-Niveau = 1 %)

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Mean	n_1 = 88	223	-	22	34.74	0.03
	n_2 = 135					
Median	n_1 = 115				46.47	< 0.01
	n_2 = 108					
Mean	n_1 = 83		2	21	21.55	0.37
	n_2 = 140					
Median	n_1 = 133				30.84	0.06
	n_2 = 90					

8.3.2.2.2 Überprüfung nach den externen Kriterien

Die Überprüfung der Modellgeltung anhand der externen Kriterien (sex_gr; age_gr; zeit_gr) zeigte keine Auffälligkeiten und somit durchwegs Rasch-Homogenität für die 22 Items (siehe Tabelle 14).

Tabelle 14: Auflistung der Rasch-Modelle für die externen Kriterien der Version R2 (α-Niveau = 1 %).

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Geschlecht	m = 106	223	-	22	19.00	0.59
	w = 117					
Alter	< 23 = 124		-	22	31.24	0.07
	> 23 = 99					
Zeit	n_1 = 99		-	22	24.66	0.26
	n_2 = 124					

8.3.2.3 Rasch Modell für Version A bis F

Bei der Durchführung des Rasch-Modells für die jeweils vollständigen Daten der Versionen A bis F erhielt man anhand der internen Kriterien „Mittelwert und Median des Rohscores“ wie auch anhand der externen Kriterien (sex_gr, alt_gr, zeit_gr) keine signifikanten LRT-Werte (siehe Tabellen 14 bis 19), was bedeutet, dass die Items der Versionen A bis F bei der Überprüfung der Modellgeltung für das vollständige Rasch-Modell homogen sind und somit eindimensional messen.

Tabelle 15: Auflistung der Rasch-Modelle für die internen und externen Kriterien der **Version A** (α-Niveau = 1 %).

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Mean	n_1 = 96	202	-	15	26.23	0.02
	n_2 = 106					
Median	n_1 = 122				20.53	0.11
	n_2 = 80					
Geschlecht	m = 87		-		24.95	0.04
	w = 115					
Alter	< 23 = 13		-		11.77	0.62
	> 23 = 89					
Zeit	n_1 = 106		-		11.89	0.62
	n_2 = 96					

Tabelle 16: Auflistung der Rasch-Modelle für die internen und externen Kriterien der **Version B** (α-Niveau = 1 %).

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Mean	n_1 = 61	135	-	15	19.65	0.14
	n_2 = 74					
Median	n_1 = 73				22.82	0.06
	n_2 = 62					
Geschlecht	m = 60		-		8.80	0.84
	w = 75					
Alter	< 23 = 78		-		14.83	0.39
	> 23 = 57					
Zeit	n_1 = 63		-		19.47	0.15
	n_2 = 72					

Tabelle 17: Auflistung der Rasch-Modelle für die internen und externen Kriterien der **Version C** (α-Niveau = 1 %).

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Mean	$n_1 = 44$	130	-	15	17.04	0.25
	$n_2 = 86$					
Median	$n_1 = 66$				15.58	0.34
	$n_2 = 64$					
Geschlecht	m = 75		-		16.67	0.27
	w = 55					
Alter	< 23 = 73		-		17.17	0.25
	> 23 = 57					
Zeit	$n_1 = 66$		-		17.97	0.21
	$n_2 = 64$					

Tabelle 18: Auflistung der Rasch-Modelle für die internen und externen Kriterien der **Version D** (α-Niveau = 1 %).

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Mean	$n_1 = 63$	130	-	15	10.21	0.75
	$n_2 = 67$					
Median	$n_1 = 77$				4.15	0.99
	$n_2 = 53$					
Geschlecht	m = 61		-		19.06	0.16
	w = 69					
Alter	< 23 = 63		-		10.20	0.75
	> 23 = 67					
Zeit	$n_1 = 59$		-		15.24	0.36
	$n_2 = 71$					

Tabelle 19: Auflistung der Rasch-Modelle für die internen und externen Kriterien der **Version E** (α-Niveau = 1 %).

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Mean	$n_1 = 63$	134	-	15	15.38	0.35
	$n_2 = 71$					
Median	$n_1 = 75$				14.11	0.44
	$n_2 = 59$					
Geschlecht	m = 64		-		14.38	0.42
	w = 70					
Alter	< 23 = 65		-		12.30	0.58
	> 23 = 69					

Zeit	$n_1 = 61$		-		12.67	0.55
	$n_2 = 73$					

Tabelle 20: Auflistung der Rasch-Modelle für die internen und externen Kriterien der **Version F** (α-Niveau = 1 %).

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Mean	$n_1 = 54$	118	-	15	19.95	0.13
	$n_2 = 64$					
Median	$n_1 = 67$				19.36	0.15
	$n_2 = 51$					
Geschlecht	m = 67		-		22.65	0.07
	w = 51					
Alter	< 23 = 58		-		20.11	0.13
	> 23 = 60					
Zeit	$n_1 = 48$		-		18.68	0.18
	$n_2 = 70$					

8.3.3 Unvollständiges Rasch Modell für alle 1400 Daten

Die Überprüfung der Geltung des Rasch Modells über alle 1400 Daten bzw. über alle 8 Versionen hinweg stellt kein vollständiges System mehr dar, da die Versionen R1 und R2 aus 23 und A bis F aus 16 Aufgaben bestehen. Es werden somit unvollständige Rasch-Modelle gerechnet.

8.3.3.1 Überprüfung nach den internen Kriterien

Bei der Modellprüfung nach den Kriterien „Mean“ und „Median“ wurden die 1400 Daten wiederum in eine Gruppe mit niedrigen (n_{1mean} = 613; n_{1med} = 766) und eine mit hohem Rohscore (n_{2mean} = 787; n_{1med} = 634) geteilt. Es ergab sich ein signifikanter LRT nach Andersen von 158.60_{mean} bzw. $152,15_{med}$ ($p<0.01$; df = 93). 93 Freiheitsgrade deshalb, da 95 Items der 109 verwendet wurden d.h. „95 minus 1, minus Poolitem 16“. Die Gegenüberstellung der Itemleichtigkeiten in Abbildung 12 pro Subgruppe (Low = Gruppe mit niedrigem Rohscore und High = Gruppe

mit hohem Rohscore) zeigt, dass Poolitem 38 und 52 auffällig sind und sich deutlich vom Punkteschwarm abspalten. Auch Item 43 liegt etwas abseits, wenn auch schon etwas höher auf der 45°-Geraden. Zunächst wird sich jedoch auf die Items 38 und 52 konzentriert, da diese auch anhand der Itemparameter deutlich auffällig werden.

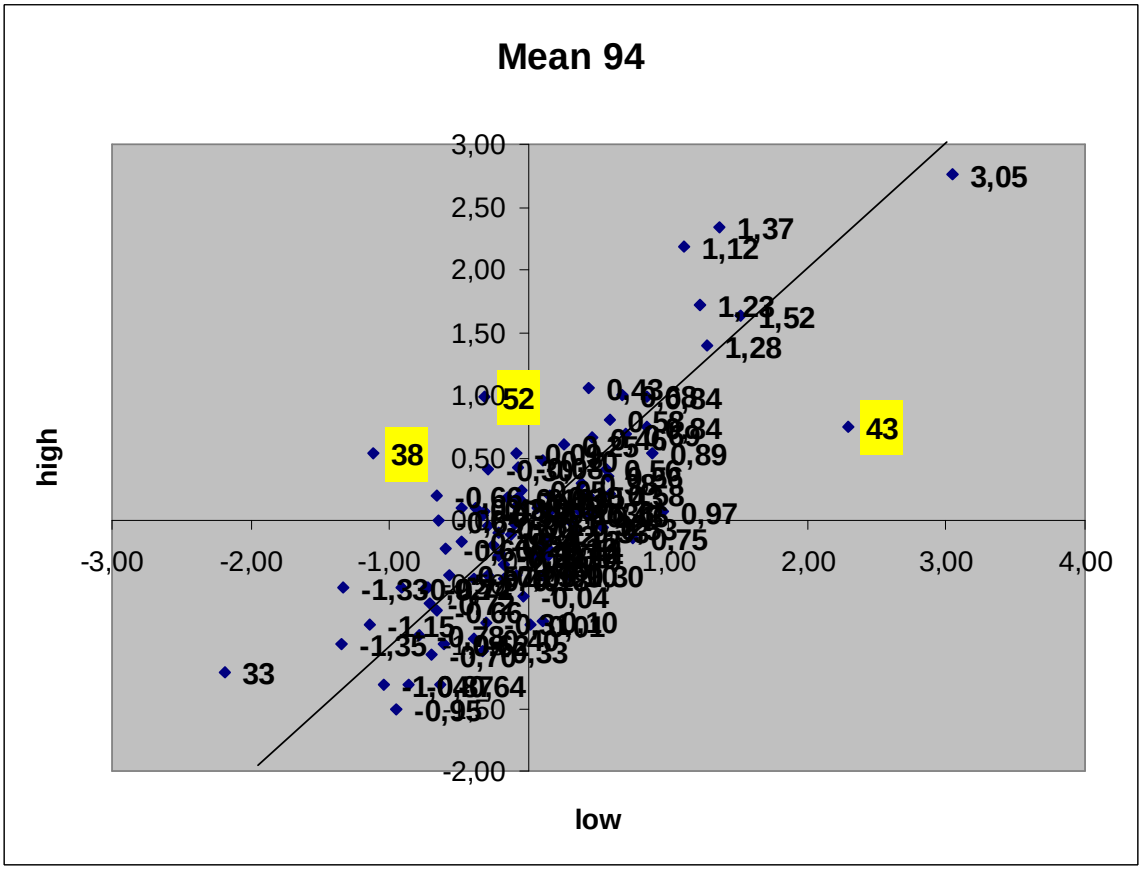

Abbildung 12: graphische Modellkontrolle: Gegenüberstellung der Itemleichtigkeiten der Subgruppen „hoher Rohscore“ (High) und „niedriger Rohscore“ (Low)

Das jeweils alleinige Ausscheiden von Poolitem 38 und 52 hat keine deutliche Auswirkung auf den LRT und auch wenn beide Items zusammen ausgeschieden werden verändert sich der LRT nur geringfügig. Erst durch zusätzliches Ausscheiden von Poolitem 43 fällt der LRT für das Kriterium „Median“ auf 137,51 (bzgl. des Kriteriums „Mean“ verändert sich nichts), wodurch bei weitem jedoch noch keine Homogenität in Sicht ist. Da die Items 2, 8 und 12 aus Version R1 und R2 bezüglich des internen Kriteriums in den vollständigen Rasch-Modellen auffällig geworden sind, wird angenommen, dass auch im

unvollständigen Rasch-Modell, welches alle acht Versionen enthält, diese Items ein Problem darstellen könnten. Ungeachtet der Items 38, 52 und 43 werden die Items 2, 8 und 12 ausgeschieden, wodurch sofort Rasch-Homogenität bezüglich des Kriteriums „Median“ folgt. Der LRT für das Kriterium „Mean“ sinkt auf 133,76 (df = 90; p < 0.01). Da sich durch zusätzliches Ausscheiden der beiden vorigen Poolitems 38 und 52, der LRT für Mean nur wenig senkt, wird ein weiteres, laut Schätzung der Itemleichtigkeiten auffälliges, Item (Poolitem 59) ausgeschieden, was wieder eine deutliche Senkung des LRT für Mean (124,97 bei df = 87) zur Folge hat (wie auch ein erneutes Absinken des ohnehin schon homogenen Wertes für den Median), jedoch noch keine Rasch-Homogenität nach sich zieht. Die Items 38, 52 und 59 befinden sich in Version B und noch ein weiteres Poolitem dieser Version des MAT wird auffällig; Item 108, durch wessen Ausscheiden jedoch keine Homogenität folgt. Erst durch das zusätzliche Ausscheiden von Poolitem 43 (das schon in Abbildung 11 auffällig war) und 56 kann auch für das Kriterium „Mean“ mit einem LRT von 116,59 (df = 84) knapp Rasch-Homogenität auf einem α-Niveau von 1 % erreicht werden, wobei zu beachten ist, das der LRT für den Median (wie in Tabelle 21 deutlich zu sehen) schon nach alleinigem Ausscheiden der Items 2, 8 und 12 nicht mehr signifikant gewesen ist.

Tabelle 21: Auflistung der Rasch-Modelle für die internen Kriterien Mean und Median für das unvollständige Rasch Modell für alle Versionen (n = 1400, α = 1 %)

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Mean	n_1 = 613	1400	-	94	158.60	< 0.01
	n_2 = 787					
Median	n_1 = 766				152.15	< 0.01
	n_2 = 634					
Mean	n_1 = 598		2, 8, 12	91	133.76	< 0.01
	n_2 = 802					
Median	n_1 = 798				112.87	0.05
	n_2 = 602					
Mean	n_1 = 594		2, 8, 12, 38, 52, 59	88	124.97	< 0.01
	n_2 = 806					

Median	n_1 = 803				102.43	0.12
	n_2 = 597					
Mean	n_1 = 588		2, 8, 12, 38, 52, 59, 108	87	121.43	< 0.01
	n_2 = 812					
Median	n_1 = 799				97.97	0.18
	n_2 = 601					
Mean	n_1 = 606		2, 8, 12, 38, 52, 59, 108, 43	86	119.89	< 0.01
	n_2 = 794					
Median	n_1 = 814				93.96	0.24
	n_2 = 586					
Mean	n_1 = 605		2, 8, 12, 38, 52, 59, 108, 43, 56	85	116.59	0.01
	n_2 = 795					
Median	n_1 = 810				88.31	0.35
	n_2 = 590					

8.3.3.2 Überprüfung nach den externen Kriterien

Die Überprüfung der Modellgeltung anhand der externen Kriterien (sex_gr; age_gr; zeit_gr) zeigte keine Auffälligkeiten wie in Tabelle 22 zu ersehen ist, was für eine eindimensionale Messung bezüglich der angegebenen Subgruppen spricht.

Tabelle 22: Auflistung der Rasch-Modelle für die externen Kriterien für das unvollständige Rasch-Modell aller Versionen (α-Niveau = 1 %).

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Geschlecht	m = 700	1400	-	94	121.78	0.02
	w = 700					
Alter	<23=753		-	94	108.25	0.13
	>23=647					
Zeit	n_1 = 701		-	94	118.92	0.04
	n_2 = 699					

8.3.3.3 Best Test Design (Wright & Stone, 1979)

Betrachtet man die Ergebnisse des unvollständigen Rasch Modells aller 1400 Daten sind es neun Poolitems, welche auffällig werden. Neben den

Items 2, 8 und 12, welche schon in Version R1 bzw. R2 auffällig geworden sind, kommen noch sechs weitere hinzu, die jedoch ausschließlich für das interne Kriterium „Mean“ ausschlaggebend sind, da für das Kriterium „Median“ bereits Homogenität erreicht worden ist. Fast alle diese Items stammen aus der Version B (38, 52, 59, 108), 43 aus der Version C und 56 aus Version F. Es ist auffällig, dass die Version B, sobald sie mittels vollständigen Rasch-Modells geprüft wird, bezüglich aller internen und externen Kriterien Rasch-Homogenität vorweist und somit kein einziges dieser im unvollständigen Rasch-Modell auffälligen Items ein Problem darstellt.
Version B wird deshalb einer genaueren Analyse mittels „Best Test Design“ nach Wright und Stone (1979) unterzogen. Es wird somit ermittelt, ob die Stichprobe der Version B den Anforderungen im Rasch-Modell genügt.
Man beginnt bei der Analyse nach Wright und Stone (1979) damit, alle Personen auszuscheiden, welche entweder alles richtig oder alles falsch beantwortet haben, da diese Probanden keinerlei Informationsgewinn liefern. Generell gilt, dass eine fähigere Person stets eine höhere Erfolgswahrscheinlichkeit haben sollte und dass sich jede Person leichter mit einfachen, denn mit schweren Items tun sollte. Ist die Differenz zwischen Item und Person nicht groß, dann trägt ein Item mehr zur Messung einer Person bei, als wenn Item und Person weit voneinander entfernt liegen. Je größer die Differenz zwischen Item und Person, desto höher die Anzahl an Items, die gebraucht werden, um eine präzise Messung zu gewährleisten und desto weniger effizient ist ein jedes Item.

„In general, the test length necessary to maintain a specified level of measurement precision is inversely proportional to the relative efficiency of the items used” (Wright & Stone, 1979, S. 75).

Die T-Werte zeigen, verglichen mit dem kritischen z-Wert (1%, zweiseitig) 2,58, dass sieben Versuchspersonen (979, 1027, 1102, 991, 1019, 1096 und 1142) signifikant vom Modell abweichen. Die übrigen

Teilnehmer, welche Version B vorgelegt bekommen haben, entsprechen den Bedingungen des Rasch-Modells. Diese sieben Teilnehmer werden nun aus den Berechnungen ausgeschlossen und mit n = 1393 Personen weiter gerechnet. Es wird somit ein neues write-out-file erstellt um zu sehen, ob sich die Modelltests bezüglich der internen Kriterien verändern (siehe Tabelle 23)

Tabelle 23: neue Auflistung der Rasch-Modelle für die internen Kriterien Mean und Median für das unvollständige Rasch Modell für alle Versionen (n = 1393, α = 1 %)

Kriterium	Subgr.	n	ausgeschiedene Poolitems	k	LRT (χ^2)	p
Mean	n_1 = 610	1393	-	94	157.19	< 0.01
	n_2 = 783					
Median	n_1 = 763				151.44	< 0.01
	n_2 = 630					
Mean	n_1 = 595		2, 8, 12	91	131.86	< 0.01
	n_2 = 798					
Median	n_1 = 795				112.48	0.05
	n_2 = 598					
Mean	n_1 = 599		2, 8, 12, 38, 52	89	130.79	< 0.01
	n_2 = 794					
Median	n_1 = 804				101.94	0.15
	n_2 = 589					
Mean	n_1 = 591		2, 8, 12, 38, 52, 59	88	122.95	< 0.01
	n_2 = 802					
Median	n_1 = 800				102.14	0.13
	n_2 = 593					
Mean	n_1 = 585		2, 8, 12, 38, 52, 59, 108	87	119.62	< 0.01
	n_2 = 808					
Median	n_1 = 796				97.99	0.18
	n_2 = 597					
Mean	n_1 = 603		2, 8, 12, 38, 52, 59, 108, 43	86	118.16	0.01
	n_2 = 790					
Median	n_1 = 811				94.03	0.24
	n_2 = 582					

Wie man aus Tabelle 23 ersehen kann, gibt es zunächst nur minimale Veränderungen in den Daten. Die LRT-Werte für Mean sind durchgehend niedriger als vorher, was zur Folge hat, dass nur acht statt neun Items ausgeschieden werden müssen um Rasch-Homogenität bezüglich der

internen Kriterien zu erhalten. Neben den Items 2, 8 und 12 stammen die folgenden auffälligen Items, bis auf Item 43 (aus Version C), zwar noch immer aus der Version B, jedoch muss nun angenommen werden, dass der Grund der Inhomogenität im unvollständigen Rasch-Modell (n = 1393) nicht in der Stichprobe, sondern in den Poolitems selbst zu suchen ist. Es folgt somit eine Analyse der einzelnen auffällig gewordenen Items.

8.4 Itemanalyse

Es wird nun nach inhaltlichen Erklärungen für die Auffälligkeit der einzelnen Poolitems gesucht. Die Prüfung erfolgt nach den oben genannten Kriterien (Kapitel 7.1) für die Itemkonstruktion. Es werden anhand der Itemparameter die Schwierigkeiten für die Subgruppen erfasst um eventuell, basierend auf der zugrundeliegenden Literatur, Erklärungen für die fehlende Homogenität zwischen den jeweiligen Subgruppen zu finden. Weiters wird auf systematisch gehäufte und fehlerhafte Antwortalternativen geachtet um mögliche Anhaltspunkte für die entstandene Inhomogenität zu erhalten.

Pool 8

	C	D	E	F	G
-	78	55		23	
	98	/	14	=	7
+	53	22		75	

Poolitem 8 wird lediglich in Version R1 bezüglich des Kriteriums „Mean“ und „Median“ auffällig. Wie die Itemleichtigkeiten für Mean und Median zeigen, fällt Item 8 den Personen mit hohem Rohscore deutlich leichter (σ = -2,85) als jenen mit niedrigem Rohscore (σ = -0,77).

Tabelle 24: Häufigkeiten der Antwortalternativen für Poolitem 8.

P8

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0	20	1,4	3,6	3,6
	3	9	,6	1,6	5,3
	4	9	,6	1,6	6,9
	5	3	,2	,5	7,4
	6	15	1,1	2,7	10,2
	7	472	33,7	85,7	95,8
	8	10	,7	1,8	97,6
	9	2	,1	,4	98,0
	10	1	,1	,2	98,2
	12	2	,1	,4	98,5
	14	1	,1	,2	98,7
	37	1	,1	,2	98,9
	77	1	,1	,2	99,1
	151	1	,1	,2	99,3
	718	1	,1	,2	99,5
	756	1	,1	,2	99,6
	824	1	,1	,2	99,8
	1400	1	,1	,2	100,0
	Gesamt	551	39,4	100,0	
Fehlend	System	849	60,6		
Gesamt		1400	100,0		

Tabelle 24 zeigt, dass 472 Personen (85,7 %) die richtige Antwort „7“ angegeben wurde und lediglich 15 Mal (2,7 %) die Antwort „6“ bzw. auch einige Male die Antwort „8“, „3“ und „4“, wobei es sich um einfache Rechenfehler handeln dürfte.

Das Item enthält keinerlei Überträge und nur zweistellige Zahlenangaben, sodass die Anforderungen an das Kurzzeitgedächtnis offenbar nicht besonders hoch sind. Weiters sind die Zwischenlösungen „23“ und „75“ in „gebrochener Form“ (d.h. dass die Zwischenlösungen nicht als ganze Zahlen in der Angabe enthalten sind, sondern deren einzelne Ziffern verstreut in den Zahlen der Angabe zu finden sind, wie es bei Item 8 bei der ersten Teillösung „23“ der Fall ist … Ziffer „2“ in „22“ und Ziffer „3“ in „53“) in der Angabe enthalten, was eine Gedächtnishilfe darstellen kann.

Pool 2

	C	D	E	F	G
+	35	23		58	
	16	*	5	=	80
-	65	23		42	

Poolitem 2 enthält keinerlei Überträge und lediglich zweistellige Zahlenangaben, sodass die Anforderungen an das Kurzzeitgedächtnis wiederum wohl nicht sehr hoch sind.

Was bei diesem Item auffällig ist, dass die Zahl 23 gleich zwei Mal in der Angabe vorkommt, wie auch gleiche Ziffern in den Einerstellen der Angabe vorhanden sind, womit die aufgestellten Konstruktionskriterien verletzt werden.
Poolitem 2 wird in Version R1 und R2 bezüglich des Kriteriums „Mean“ und „Median“ auffällig. Wie man anhand der Itemleichtigkeiten für Mean und Median in der Version R1 und R2 erkennen kann, fällt Item 2 den Personen mit hohem Rohscore deutlich schwieriger (σ_{R1} = 0,85; σ_{R2} = 0,74) als jenen mit niedrigem Rohscore (σ_{R1} = 0,01; σ_{R2} = -0,47). Tabelle 25 zeigt, dass neben der richtigen Antwort „80“ (73 %), 34 Mal (6 %) die Antwort „90“ gegeben wurde und auch 15 Mal die Antwort „500“.

Tabelle 25: Häufigkeiten der Antwortalternativen für Poolitem 2.

P2

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0	1	,1	,2	,2
	4	1	,1	,2	,4
	5	1	,1	,2	,5
	15	1	,1	,2	,7
	16	4	,3	,7	1,5
	20	2	,1	,4	1,8
	25	2	,1	,4	2,2
	30	9	,6	1,6	3,8
	55	1	,1	,2	4,0
	60	4	,3	,7	4,7
	65	7	,5	1,3	6,0
	70	6	,4	1,1	7,1
	75	11	,8	2,0	9,1
	80	402	28,7	73,0	82,0
	85	4	,3	,7	82,8
	90	34	2,4	6,2	88,9
	100	1	,1	,2	89,1
	106	1	,1	,2	89,3
	110	1	,1	,2	89,5
	115	1	,1	,2	89,7
	125	1	,1	,2	89,8
	130	9	,6	1,6	91,5
	135	1	,1	,2	91,7
	145	1	,1	,2	91,8
	150	2	,1	,4	92,2
	165	1	,1	,2	92,4
	170	1	,1	,2	92,6
	180	8	,6	1,5	94,0
	189	1	,1	,2	94,2
	210	1	,1	,2	94,4
	230	3	,2	,5	94,9
	280	1	,1	,2	95,1
	450	1	,1	,2	95,3
	455	1	,1	,2	95,5
	500	15	1,1	2,7	98,2
	530	2	,1	,4	98,5
	720	1	,1	,2	98,7
	730	6	,4	1,1	99,8
	1500	1	,1	,2	100,0
	Gesamt	551	39,4	100,0	
Fehlend	System	849	60,6		
Gesamt		1400	100,0		

Die Anordnung der Zahlen in der Angabe erscheint in diesem Beispiel markant und eventuell irreführend. Beide horizontale Aufgaben ergeben die Zahl „8“ in ihrer Einerstelle (siehe Tabelle 26). Eventuell ist das oftmalige Auftreten der Antwort „90“ darin begründet, dass statt des Zwischenergebnisses „16“ sich die Zahl „18“ quasi aufdrängt (aufgrund des oftmaligen Auftretens der Zahl „8“). 18 mal 5 ergibt die Antwort „90“. Es ist jedoch nicht auszuschließen, dass es sich hierbei erneut um einen einfachen Rechenfehler handelt, der sich eventuell schon in den vorhergehenden Zwischenergebnissen ereignet hat.

Tabelle 26: Darstellung der Teilaufgabe von Poolitem 2

35		23
	16	
65		23

Das Auftreten der Antwort „500“ (15 mal) dürfte allein daher zustande gekommen sein, dass die beiden Zwischenergebnisse „58“ und „42“ nicht subtrahiert, sondern fälschlicherweise addiert worden sind, was das Ergebnis „500“ zur Folge hat.

Poolitem 2 ist eine einfach zu lösende Aufgabe, verleitet jedoch zu falschen Lösungsansätzen, sodass es als Poolitem für den MAT nicht geeignet erscheint.

Pool 12

	C	D	E	F	G
+	212	57		269	
	112	/	8	=	14
-	189	32		157	

Poolitem 12 wird lediglich in Version R1 bezüglich des Kriteriums „Mean“ und „Median“ auffällig und fällt den Personen mit niedrigem Rohscore

leichter (σ = 0,10) als jenen mit hohem Rohscore (σ = 0,70).

Item 12 enthält in keinem einzigen Rechenschritt Überträge, besteht jedoch teils aus dreistelligen Zahlenangaben, wodurch der Schwierigkeitsgrad laut Gabriel (2004, S. 60) steigt.

In der Angabe ist die zweite Zwischenlösung „157“ gebrochen in der Angabe enthalten, wie auch die dritte Zwischenlösung „112“, was eine Gedächtnisstütze darstellen könnte. Weiters sind die Ziffern in den Einerstellen in Zwischenergebnissen und Angabe gleich, was den Konstruktionskriterien zuwider läuft.
Im letzten Rechenschritt (Division durch 8) ist das Teilergebnis der Division („32“) in der Angabe enthalten („112:8 … 8 geht in 11 1 Mal, bleibt 3; 8 geht in 32 4 mal“ …was zum Endergebnis „14“ führt), was wiederum eine Gedächtnisstütze darstellt, vorausgesetzt der Teilnehmer stellt die Rechenschritte in angeführter Art und Weise an.

Tabelle 27 zeigt, dass neben der richtigen Antwort „14“ (72 %) deutlich oft die falsche Antwort „6“ (37 Mal = 7 %) bzw. „13“ (20 Mal = 3,6 %) auftritt. Eine mögliche Erklärung für das Zustandekommen der Antwort „6“ könnte das irrtümliche Addieren der unteren horizontalen Teilaufgabe sein (wobei diese Vorgehensweise schwieriger ist, da nun Überträge zu berücksichtigen wären), was das Zwischenergebnis „221“ zur Folge hätte (welches wiederum gebrochen in der Angabe zu finden wäre). Das mittlere Zwischenergebnis wäre somit „48“ mit dem Endergebnis „6“. Das Ergebnis „13“ scheint die Folge einfacher Rechenfehler zu sein. Item 12 dürfte somit, aus nicht klar ersichtlichem Grund, zu falschen Lösungsansätzen verleiten und eignet sich deshalb und unter oben erwähnten Aspekten kaum als Poolitem.

Tabelle 27: Häufigkeiten der Antwortalternativen für Poolitem 12.

P12

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0	24	1,7	4,4	4,4
	1	2	,1	,4	4,7
	2	2	,1	,4	5,1
	4	7	,5	1,3	6,4
	5	4	,3	,7	7,1
	6	37	2,6	6,7	13,8
	8	1	,1	,2	14,0
	9	1	,1	,2	14,2
	10	1	,1	,2	14,3
	12	9	,6	1,6	16,0
	13	20	1,4	3,6	19,6
	14	395	28,2	71,7	91,3
	15	9	,6	1,6	92,9
	16	8	,6	1,5	94,4
	18	1	,1	,2	94,6
	19	1	,1	,2	94,7
	20	2	,1	,4	95,1
	26	2	,1	,4	95,5
	29	1	,1	,2	95,6
	39	5	,4	,9	96,6
	50	1	,1	,2	96,7
	51	1	,1	,2	96,9
	53	2	,1	,4	97,3
	55	1	,1	,2	97,5
	64	1	,1	,2	97,6
	96	3	,2	,5	98,2
	112	2	,1	,4	98,5
	126	1	,1	,2	98,7
	144	1	,1	,2	98,9
	258	1	,1	,2	99,1
	728	1	,1	,2	99,3
	816	1	,1	,2	99,5
	832	1	,1	,2	99,6
	896	2	,1	,4	100,0
	Gesamt	551	39,4	100,0	
Fehlend	System	849	60,6		
Gesamt		1400	100,0		

Pool 6

	C	D	E	F	G
+	23	14		37	
	14	*	12	=	168
-	47	24		23	

Poolitem 6 wird auffällig, wenn die Stichprobe der Version R1 anhand der Bearbeitungszeit geteilt wird. Wie an den Itemleichtigkeiten zu erkennen ist, fällt dieses Item den Personen mit weniger Bearbeitungszeit (also den Schnelleren) deutlich leichter (σ = 0,24) als jenen mit mehr Bearbeitungszeit (σ = 1,06).

Alle Zwischenergebnisse sind ausnahmslos in der Angabe enthalten („23“ und „14“ sogar als ganze Zahlen und „37“ in gebrochener Form). Es ist anzunehmen, dass die Teilnehmer mit weniger Bearbeitungszeit diese Hilfen eher in Anspruch nehmen und dadurch schneller sind.

Die Ausgabe der Antworthäufigkeiten für Poolitem 6 zeigt keine besonderen Auffälligkeiten. Neben der richtigen Antwort „168“ (68 %), kommt die Antwort „164“ 16 Mal (3 %) und die Antwort „720“ 15 Mal vor. „720“ ist wohl dadurch zustande gekommen, dass die beiden Zwischenergebnisse „37“ und „23“ nicht subtrahiert, sondern fälschlicherweise addiert wurden. Das Zustandekommen der Antwort „164“ dürfte Folge eines einfachen Rechenfehlers sein.

Die folgenden Poolitems werden im unvollständigen Rasch-Modell (n 1393) auffällig:

38

	C	D	E	F	G
-	349	217		132	
	946	/	11	=	86
+	461	353		814	

Poolitem 38 fällt den Personen mit hohem Rohscore deutlich leichter (σ = 0,56) als jenen mit niedrigem Rohscore (σ = 1,03).
In der Angabe sind einzelne Ziffern der Zwischenlösungen enthalten - wenn auch in verstreuter Form - die als Gedächtnishilfe nützlich sein können. Um das erste Zwischenergebnis besser im Gedächtnis halten zu können müsste man sich lediglich die Zahlenreihenfolge „1 2 3“ in etwas veränderter Reihenfolge merken. Es sind kaum Überträge enthalten, jedoch besteht die Angabe ausschließlich aus dreistelligen Zahlen, wodurch der Schwierigkeitsgrad steigt, besonders wenn die Hilfen nicht in Anspruch genommen werden.

Die Häufigkeitsausgabe der alternativen Lösungsangaben zeigt keine Auffälligkeiten.

52	C	D	E	F	G
-	265	127		138	
	294	/	21	=	14
+	81	75		156	

In der Angabe des Poolitems 52 sind, bis auf einige einzelne Ziffern, kaum Hilfen enthalten, wie auch in jedem Rechenschritt Überträge zu beachten sind, was den Schwierigkeitsgrad und die Gedächtnisanforderungen ansteigen lässt. Der letzte Rechenschritt „Division durch 21“ könnte eventuell durch die „81“ in der Angabe Erleichterung erfahren wenn die Rechenschritte in folgender Weise angestellt werden: „21 in 29 … 1 Mal enthalten; bleibt 8; 21 in 84 … 4 Mal enthalten“. Das erste Zwischenergebnis wäre ebenfalls leicht durch die Angabe „127“ zu merken, da, um „138“ zu erhalten lediglich jeweils eine Eins zur Zahl 2 und 7 hinzugezählt werden muss.
Item 52 fällt den Personen mit hohem Rohscore deutlich leichter (σ = -1,01) als jenen mit niedrigem Rohscore (σ = 0,23).

Die Häufigkeitsausgabe der alternativen Lösungsangaben zeigt keine Auffälligkeiten.

59		C	D	E	F	G
	+	98	71		169	
		14	*	11	=	154
	-	932	777		155	

Item 59 enthält deutlich oft die Ziffer „7“, verletzt die aufgestellten Konstruktionskriterien jedoch nicht. In den ersten beiden Rechenschritten (die beiden Horizontalen) sind Überträge zu beachten, wodurch besonders die zweite horizontale Aufgabenstellung die Gedächtniskapazitäten mehr beansprucht, da auch keine augenscheinlichen Hilfen in der Angabe zu finden sind. Die beiden letzten Rechenschritte sind aufgrund der günstigen Zahlenkombination und der fehlenden Überträge relativ einfach zu bewältigen.
Poolitem 59 fällt den Personen mit hohem Rohscore leichter (σ = -0,02) als jenen mit niedrigem Rohscore (σ = 0,56).

Die Häufigkeitsausgabe der alternativen Lösungsangaben zeigt keine Auffälligkeiten.

108		C	D	E	F	G
	+	363	118		481	
		156	/	12	=	13
	-	619	294		325	

Poolitem 108 enthält in der Angabe einzelne Ziffern der Zwischenlösungen in verstreuter Form, wie auch die Zahlenfolgen der Zwischenlösungen günstig fallen (z.B. für „325“ … „3 + 2 = 5“ oder für „156“ … „1 + 5 = 6“), was eine gewisse Gedächtnisstütze darstellen kann. Die Angabe besteht jedoch ausschließlich aus dreistelligen Zahlen wie auch in jedem Rechenschritt Überträge zu beachten sind, was die

Anforderungen an das Kurzzeitgedächtnis erhöht und es zu einem eher schwieriger zu lösendem Item macht.
Item 108 fällt den Personen mit hohem Rohscore leichter (σ = -0,56) als jenen mit niedrigem Rohscore (σ = -0,01).

Die Häufigkeitsausgabe der alternativen Lösungsangaben zeigt keine Auffälligkeiten.

43	C	D	E	F	G
+	38	23		61	
	42	/	7	=	6
-	63	44		19	

Item 43 fällt den Personen mit niedrigem Rohscore deutlich leichter (σ =
-2,31) als jenen mit hohem Rohscore (σ = -0,76) und dürfte aufgrund der lediglich zweistelligen Angabe und der in der Angabe enthaltenen Hilfen, bezogen auf die Itemparameter und im Vergleich zu den übrigen auffällig gewordenen Items gesehen, ein einfacher zu lösendes Item darstellen. Die Aufgabe enthält zwar in jedem Rechenschritt Überträge, jedoch sind diese relativ einfach zu lösen wie zum Beispiel der dritte Rechenschritt (vertikale Aufgabenstellung) zeigt. Die Problemstellung ist ökonomischer zu bewerkstelligen betrachtet man die Aufgabenstellung ganzheitlicher indem man statt „61 – 19“ eventuell „60 – 20“ und anschließend „2“ hinzurechnet.

Die Häufigkeitsausgabe der alternativen Lösungsangaben zeigt keine Auffälligkeiten.

8.4.1 Interpretation zur Itemanalyse

Es wird nun versucht, anhand der obigen Itemanalyse zu erläutern, warum die auffällig gewordenen Items die diversen Subgruppen in unterschiedlicher Weise bevorzugen bzw. benachteiligen und somit keine

Stichprobenunabhängigkeit mehr gegeben ist. Zu Beginn soll eine globalere Betrachtungsweise der analysierten Items über die 8 Versionen des MAT hinweg erfolgen. Es wird erst auf die vollständigen Rasch-Modelle (einzelne Betrachtung der acht Versionen) und dann auf das unvollständige Rasch-Modell (alle acht Versionen zusammen) eingegangen. Anschließend werden Vermutungen über die Reaktions-richtung der Subgruppen der internen und externen Teilungskriterien angestellt.

In den Versionen R1 und R2 gibt es, bis auf Item 6, nur bezüglich des internen Kriteriums Auffälligkeiten. Allein Poolitem 2 wird in beiden Versionen (R1 und R2) auffällig, die Items 8, 12, 6 nur in Version R1. Einziger Unterschied zwischen den beiden Versionen ist, dass die Itemvorgabe in R1 in aufsteigender Itemnummerierung erfolgte und in R2 diese Anordnung nicht eingehalten wurde. Diese drei Items werden demnach nur dann auffällig, wenn sie eher zu Beginn gestellt werden, was auf einen gewissen Reihenfolge-, wenn nicht sogar Übungseffekt hindeuten könnte. Da es offensichtlich nicht unbedeutend ist, an welcher Stelle die Items 8, 12 und 6 vorgegeben werden, verletzen diese Aufgaben, nebst der spezifischen Objektivität auch das Prinzip der lokalen stochastischen Unabhängigkeit und sind somit aus zweifachem Grund nicht für den Itempool eines Rasch-homogenen Verfahrens geeignet.
Hingegen sind die Versionen A bis F mit sehr großer Wahrscheinlichkeit Rasch-homogen.

Werden die Homogenitätsprüfungen mit dem unvollständigen Rasch-Modell durchgeführt, ist festzustellen, dass alle Auffälligkeiten ausschließlich bezüglich des internen Kriteriums auftreten. Wie schon weiter oben illustriert, sind alle auffälligen Poolitems, bis auf Item 43 (Version C) und 56 (in Version F), in Version B zu finden (Items 38, 52, 59, 108), obwohl die Version B (wie auch C und F), als vollständiges Modell betrachtet, keine Inhomogenitäten zeigt.

Die Modellprüfung anhand des Kriteriums „Median“ ergab schon nach Selektion der Items 2, 8 und 12 Rasch-Homogenität. Um auch bezüglich des internen Kriteriums „Mean“ Rasch-Homogenität zu erreichen mussten alle neun Items ausgeschieden werden.

8.4.1.1 Auffälligkeiten anhand des internen Kriteriums

Folgende Poolitems werden bezüglich des internen Kriteriums auffällig:
8, 2 und 12 in der Version R1 bzw. 2 auch in Version R2;
38, 52, 59, 108 und 43 im unvollständigen Rasch-Modell.

Anhand der Itemparameter der auffällig gewordenen Poolitems wird vermutet, dass die Teilnehmer mit niedrigem Rohscore eventuellen Hilfen in den Angaben und anderen Erleichterungen, wie günstigen Zahlenkombinationen und wenig Überträgen, mehr Beachtung schenken als Personen mit hohem Rohscore. Diese Annahme wird deshalb getroffen, da bei allen auffällig gewordenen Items, bei jenen eindeutige Hilfen in den Angaben zu finden und wenig bis keine Überträge enthalten waren, die Personen mit niedrigem Rohscore überraschenderweise deutlich niedrigere Itemschwierigkeiten aufgewiesen haben als die an sich fähigeren Personen. Werden die Aufgaben in Summe schwieriger durch dreistellige Angaben und Überträge, verschwindet dieser Vorteil, auch wenn eventuell Hilfen in den Angaben enthalten sind. Lediglich Poolitem 8 zeigt, dass es Personen mit niedrigem Rohscore, trotz relativ vieler Hilfen in den Angaben, schwieriger fällt als den fähigeren Teilnehmern.

Item 2 und 12 enthalten Hilfen und keinerlei Überträge, sodass diese Aufgabe, nun wie erwartet, den Personen mit niedrigem Rohscore leichter fällt.
Die Items 38, 52, 59 und 108 hingegen fallen den fähigeren Teilnehmern leichter, da weniger eindeutige Hilfen zu finden sind, Überträge zu

beachten sind und es sich fast ausschließlich um dreistellige Zahlenangaben handelt. Poolitem 43 fällt hingegen erwartungsgemäß den Personen mit niedrigem Rohscore leichter, da zwar Überträge vorhanden sind, dieser Umstand aber dadurch aufgehoben wird, dass sehr viele Hilfen und andere Erleichterungen in Anspruch genommen werden können.

Fast alle angegebenen Items enthalten Hilfen, womit darauf geschlossen wird, dass bei jenen Aufgaben möglicherweise andere – nicht beabsichtigte - Lösungsstrategien als bei den übrigen Items verwendet werden und dadurch jene Aufgaben eventuell nicht dasselbe messen, wie der Rest des Itempools. Neben Kopfrechenfähigkeit werden somit wohl auch andere Fähigkeiten abgeprüft, wodurch die Eindimensionalität dieser Items gefährdet erscheint. Weiters dürfte keine Stichprobenunabhängigkeit vorliegen, da die angegebenen Items die jeweiligen Subgruppen nicht in gleicher Weise bevorzugen bzw. benachteiligen. Demzufolge eignen sich jene Items vermutlich nicht für den Pool des MAT.

Items 2 und 12 verletzen die aufgestellten Konstruktionskriterien und verleiten, wie auch Item 6, zu falschen Lösungsansätzen, sodass auch aus diesem Grund keine Eignung für den Pool besteht.

8.4.1.2 Auffälligkeiten anhand der externen Kriterien

Allein Item 6 wird bezüglich des Zeitkriteriums in Version R1 auffällig. Offenbar verwenden die Teilnehmer mit kurzer Bearbeitungszeit die Hilfen in den Angaben effizienter, als Personen mit höheren Bearbeitungszeiten. Da dieses Item deutlich den Konstruktionskriterien widerspricht und die jeweiligen Subgruppen möglicherweise nicht dieselben Lösungsstrategien anwenden und dadurch diese Aufgabe

eventuell nicht dasselbe Konstrukt misst wie die übrigen Aufgaben, eignet sich Item 6 nicht für den Pool eines Rasch-homogenen Verfahrens.

Ansonsten gibt es keine Auffälligkeiten bezüglich der externen Kriterien, was wohl darauf zurückzuführen ist, dass, laut Formann (2003), das interne Kriterium eine höhere Sensitivität aufweist als das externe.

9 Zusammenfassung

Hauptaufgabe dieser Arbeit bestand in der Analyse der Personendaten wie der in den acht Versionen vorgegebenen Poolitems des MAT. Es sollte eine rein technische Itemselektion (Gittler, 1986) vermieden werden, indem versucht wurde, die auffällig gewordenen Items anhand inhaltlicher Überlegungen und entsprechender Literatur zu analysieren um somit eine eventuelle Selektion inhaltlich zu begründen.

Nach anfänglichen 1720 Personendaten wurde nach Selektion von augenscheinlich unseriös bzw. unmotiviert arbeitenden Testteilnehmern (u. a. Ausscheiden der Teilnehmer, die Item 16 nicht gelöst haben) nunmehr mit 1400 Daten weitergearbeitet. Im Laufe der statistischen Erhebungen und der Analyse der Version B, mittels „Best Test Design“ nach Wright und Stone (1979), wurden weitere sieben Personen ausgeschieden und die Rasch-Modelle letztendlich mit 1393 Personendaten weitergeführt.

Neben dem schon in den Vorerhebungen ausgeschiedenen Item 16, wurden neun weitere Items als ungeeignet für den Itempool des MAT erachtet. Vier Items (2, 8, 12 und 6) betreffen die vollständigen Modelle R1 und R2, wobei Items 2, 8 und 12 bezüglich des internen Kriteriums der Version R1 auffällig werden und lediglich Item 2 auch in Version R2 auffällig wird. Allein Item 6 wird in Version R1 bezüglich des externen Kriteriums der Bearbeitungszeit auffällig. Fünf Items (38, 52, 59, 108 und 43) betreffen das unvollständige Rasch-Modell für alle 1393 Personendaten, also alle acht Versionen zusammen.

Selektiert man die genannten Items, erhält man Rasch-Homogenität für die verbleibenden 85 Poolitems des MAT. Die Items der bereinigten

Versionen messen also mit sehr großer Wahrscheinlichkeit eindimensional und weisen somit hohe Kontentvalidität auf. Dementsprechend ist daraus zu schließen, dass innere Konsistenz herrscht und alle Items die gleiche Trennschärfe aufweisen (Rost, 2004) und somit kein sogenanntes ‚Attenuationparadoxon' mehr zu befürchten ist (Gulliksen, 1945). Damit wurden die, laut Arendasy, Sommer und Hergovich (2007, S. 122), „zwei wichtigsten Vorteile bei Geltung des RM", nämlich die „Personenhomogenität, d.h. die statistische Äquivalenz der Itemschwierigkeitsschätzungen in unterschiedlichen Teilstichproben einer Population und [] die Itemhomogenität der generierten Testaufgaben" mit großer Wahrscheinlichkeit erfüllt.

Bei der weiteren Generierung von Poolaufgaben, nach der Art des „Mental Arithmetic Test", wird empfohlen neben den aufgestellten Konstruktionskriterien ebenfalls darauf zu achten, dass die Zwischenergebnisse nicht nur in ihrer vollständigen Form (als ganze Zahlen) in der Angabe vermieden werden, sondern auch nicht in gebrochener Weise dort zu finden sind, wie es bei den meisten auffällig gewordenen Items zu beobachten war.

10 Literaturverzeichnis

Amelang, M. & Zielinski, W. (2002). *Psychologische Diagnostik und Interventionen*. 3. Auflage. Berlin [u.a.]: Springer.

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

Amthauer, R., Brocke, B., Liepmann, D. & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)*. Göttingen: Hogrefe.

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

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1993b). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1996). A simple theory of complex cognition. *American Psychologist*, 51(4), 355-365.

Anderson, J. R. & Bower, G. H. (1973). *Human associative memory*. Washington, DC: Winston and Sons.

Anderson, J. R. & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review*, 104(4), 728-748.

Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Anderson, J. R. & Lebiere, C. (2003). The Newell Test for a theory of Cognition. Behavioral and brain sciences, 26.

Anderson, J. R., Bothell, D., Byrne, M., Douglass, D., Lebiere, C., & Quin, Y. (2004). An integrated theory of mind. *Psychological Review*, 111, 1036-1060.

Arbeitsgemeinschaft für Leistungsmessung in der Schule (Org).SO (o.J.). (1984). Vergleichsarbeiten 3. Schuljahr (VA 3). Diktate. Mathematikarbeiten. Sonderarbeiten. Essen: Tellus.

Arendasy, M. E., Sommer, M. & Hergovich, A. (2007). Psychometrische Technologie. *Diagnostica*, 53(3), 119-130.

Aster, M.v. (2001). *Testverfahren zur Dyskalkulie. Zareki.* Frankfurt am Main: Swets.

Baddeley, A. D. (1986). *Working memory*. Oxford, England: Oxford University Press.

Baddeley, A. D. & Hitch, G.J. (1974). Working memory. In G. A. Bower (ed.), *Recent Advances in Learning and Motivation*, Vol. 8 (pp. 47-89). New York: Academic Press.

Baddeley, A. D. (1998). Working memory. *Life Sciences*, 321, 167-173.

Blum, F., Didi, H. J., Fay, E., Maichle, U., Trost, G., Wahlen, J. H. & Gittler, G. (1998). *Intelligenz Struktur Analyse (ISA). Ein Test zur Messung der Intelligenz*. Frankfurt: Swets & Zeitlinger B. V., Swets Test Services.

Bratfisch, O; Hagman, E.SO. (2003). *N-Test Alpha. Schnelles und richtiges Kopfrechnen [Computerprogramm]*. Moedling: Schuhfried.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson Studium.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests: Grundlagen und Anwendungen*. Bern; Wien [u.a.]: Huber.

Fischer, G. H., & Molenaar, I. W., Eds. (1995). Rasch Models: *Foundations, Recent Developments, and Applications*. NY [u.a.]: Springer.

Fischer, G., & Scheiblechner, H. (1970). Algorithmen und programme für das probabilistische testmodell von rasch [algorithms and programs raschs probabilistic test model]. *Psychologische Beiträge*, 12, 23-51.

Formann, A. K. (1981). Über die Verwendung von Items als Teilungskriterium für Modellkontrollen im Modell von Rasch. *Zeitschrift für experimentelle und angewandte Psychologie*, 28(4), 541-560.

Formann, A. K. (2003). *Testtheorie und Testkonstruktion: Skriptum zur Vorlesung*. Neuauflage 2003. Universität Wien.

Gabriel, M. (2004). *Entwicklung einer Rasch-homogenen und LLTM-kalibrierten Skala zur Erfassung der Fähigkeitsdimension Kopfrechnen*. Unveröffentlichte Diplomarbeit: Universität Wien.

Gittler, G. (1986). Inhaltliche Aspekte bei der Itemselektion nach dem Modell von Rasch. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 33(3), 386-412.

Gittler, G. (1999). Sind Raumvorstellung und Reasoning separierbare Fähigkeitsdimensionen? Dimensionalitätsanalysen zweier Rasch-skalierter Tests: 3 DW und WMT. *Diagnostica*, 45(2), 69-81.

Gittler, G. & Arendasy, M. (2003). Psychometrische Effekte von Farbgebung. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 24(1), 23-31.

Glas, C. (1988). The derivation of some tests for the rasch model from the multinominal distribution. *Psychometrika*, 53, 525-546.

Glas, C. (1989). *Contributions to estimating and testing rasch models (doctoral thesis).* Enschede: University of Twente.

Glas, C. & Verhelst, N. (1995). Testing the rasch model. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (p. 69-96). New York: Springer.

Grube, D. & Barth, U. (2004). Rechenleistung bei Grundschülern. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology*, 18(3/4), 245–248.

Gulliksen, H. (1945). The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika*, 10(2), 79-91.

Guttman, L. (1950). The basis of scalogram analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, J.A. Clausen (eds.), *Studies in social psychology in world war II*, Vol. IV.Princeton/N.J.: Princeton Univ. Press.

Hayes, J. R. (1973). On the function of visual imagery in elementary mathematics. In W. Chase (Ed.), *Visual information processing* (pp. 177-214). New York: Academic Press.

Henning, H. J. (1975). *Skalenanalyse und Rasch-Modell.*

Hitch, G. J. (1978). The role of short-term working memory in mental arithmetic. *Cognitive Psychology,* 10, 203-323.

Jacobs, C. & Petermann, F. (2005). *RZD 2-6. Rechenfertigkeiten- und Zahlenverarbeitungs-Diagnostikum fuer die 2. bis 6. Klasse.* Goettingen: Hogrefe.

Kalbe, E., Brand, M. & Kessler, J. (2002). *ZRT. Zahlenverarbeitungs- und Rechentest.* Goettingen: Beltz.

Klauer, C. K. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika*, 56(2), 213-228.

Krajewski, K., Kuespert, P. & Schneider, W. (2002). *DEMAT 1+. Deutscher Mathematiktest fuer erste Klassen.* Goettingen: Beltz.

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

Kubinger, K. D. (2006). *Psychologische Diagnostik*: *Theorie und Praxis psychologischen Diagnostizierens*. Göttingen: Hogrefe.

Kubinger, K. D. & Wurst, E. (2000). *Adaptives Intelligenz Diagnostikum – Version 2.1 (AID 2)*. Göttingen: Beltz.

Kueffner, H.SO (o.J.). *Mathematik-Test Größen und Maßsysteme für das 5. Schuljahr an Hauptschulen (MT-GMS 5).* Fehlerorientierte Tests: Konzept und Bewährungskontrolle. Weinheim: Beltz (Originaltitel: Methoden der Konstruktion und Analyse fehlerorientierter Tests. Ein

Beitrag zur Weiterentwicklung der Lerndiagnostik in Schule und Unterricht. Inaugural-Dissertation. Hagen: Fernuniversität-Gesamthochschule 1980).

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, J.A. Clausen (eds.), *Studies in social psychology in world war II*, Vol. IV (pp. 362-412). Princeton/N.J.: Princeton Univ. Press.

Lebiere, C. (1999). The dynamics of cognition: An ACT-R model of cognitive arithmetic. *Kognitionswissenschaft*, 8, 5-19.

Mair, P., & Ledl, T. (2006). Monte Carlo simulations for Rasch model tests. *Memorias del XX Foro Nacional de Estadistica*, 83-94.

Martin-Löf, P. (1973). Statistika Modeller. Anteckningar fran seminarier lasaret 1969-1970, utarbetade av Rolf Sundberg. Obetydligt ändrat nytryck, October 1973. Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistisk vid Stockholms Universitet.

Martin-Löf, P. (1974). *The Notion of Redundancy and Its Use as a Quantitative Measure of the Discrepancy between a Statistical Hypothesis and a Set of Observational Data.* Scand J Statist 1: 3-18. University of Stockholm.

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika*, 48, 49-72.

Nepita, B. (1991). *Dimensionalitätsanalyse aller Rasch-homogenen Intelligenztests zur Frage der Dimensionalität der Intelligenz*. Unveröffentlichte Diplomarbeit: Universität Wien.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

(Reiter, C.) Schreiner, C. (1996). *Raschmodell, Latent-Class-Analysis und Mischverteilungsraschmodell: Versuch einer Rasch-Skalierung des Anstrengungsvermeidungstests mit polytomen Antwortvariablen*. Unveröffentlichte Diplomarbeit: Universität Wien.

Rost, J. (1988). *Quantitative und qualitative probabilistische Testtheorie*. Bern: Huber.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion*. Bern: Huber.

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

Stelzl, I. (1979). Ist der Modelltest des Rasch-Modells geeignet, Homogenitätshypothesen zu prüfen? Ein Bericht über Simulationsstudien mit inhomogenen Daten. *Zeitschrift für experimentelle und angewandte Psychologie*, 26(4), 652-672.

Taatgen, N. A., Van Rijn, H., Anderson, J. R. (2007). An integrated theory of prospective time interval estimation: the role of cognition, attention, and learning. *Psychological Review*, 114(3), 577-598.

Trbovich, P. L. & Lefevre, J. (2003). Phonological and visual working memory in mental addition. *Memory & Cognition*, 31(5), 738-745.

Verhelst, N.D. & Eggen T.J.H.M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek* (PPON-rapport, nr. 4). Arnhem: CITO.

Verhelst N.D., Glas C.A.W. & Verstralen H.H.F.M. (1994). *OPLM*: *One Parameter Logistic Model. Computer program and manual.* Arnhem: CITO.

Wollenberg, A. L. van den (1982). Two new test statistics for the Rasch model. *Psychometrika*, 47, 123-140.

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Testerklärungen aus Kapitel 3: verfügbar unter: http://dbs.univie.ac.at/ suchtitel = PSYNDEXplus-Tests (ab 1945; OvidSP). Copyright (c) 2000-2009 Ovid Technologies, Inc.[Datum des Zugriffs: 07.09.09]).

11 Anhang

11.1 Bearbeitungs- und Lösungshäufigkeiten der vorgegebenen Poolitems

Poolnummer	1	2	3	4	5	6	7	8	9	10	11	12	13
bearbeitet	551	551	551	551	551	551	551	551	551	551	551	551	551
gelöst	531	402	335	345	473	372	391	472	375	352	484	395	471
gelöst in %	96,4	73,0	60,8	62,6	85,8	67,5	71,0	85,7	68,1	63,9	87,8	71,7	85,5

Poolnummer	14	15	16	17	18	19	20	21	22	23	24	25	26
bearbeitet	551	551	1393	1393	551	551	1393	551	551	1393	202	130	130
gelöst	384	341	1393	1015	384	484	927	446	379	1082	125	89	83
gelöst in %	69,7	61,9	100,0	72,9	69,7	87,8	66,5	80,9	68,8	77,7	61,9	68,5	63,8

Poolnummer	27	28	31	32	33	34	36	37	38	39	40	41	42
bearbeitet	134	134	118	128	118	202	118	202	128	130	130	134	128
gelöst	97	81	57	77	90	109	48	115	74	80	61	79	93
gelöst in %	72,4	60,4	48,3	60,2	76,3	54,0	40,7	56,9	57,8	61,5	46,9	59,0	72,7

Poolnummer	43	44	45	46	47	49	50	51	52	53	54	55	56
bearbeitet	130	130	130	202	134	118	130	118	128	130	134	130	118
gelöst	111	101	87	149	95	79	99	79	86	87	75	57	75
gelöst in %	85,4	77,7	66,9	73,8	70,9	66,9	76,2	66,9	67,2	66,9	56,0	43,8	63,6

Poolnummer	57	58	59	60	61	62	63	64	65	66	67	68	69
bearbeitet	134	202	128	130	118	202	128	130	202	130	134	130	118
gelöst	89	137	74	76	64	103	66	65	98	86	82	82	82
gelöst in %	66,4	67,8	57,8	58,5	54,2	51,0	51,6	50,0	48,5	66,2	61,2	63,1	69,5

Poolnummer	70	71	72	73	74	75	76	77	78	79	80	81	82
bearbeitet	202	128	130	128	130	130	134	130	134	118	202	118	128
gelöst	121	65	60	84	59	89	89	76	99	61	124	65	74
gelöst in %	59,9	50,8	46,2	65,6	45,4	68,5	66,4	58,5	73,9	51,7	61,4	55,1	57,8

Poolnummer	83	84	85	86	87	88	89	90	91	92	94	95	96
bearbeitet	202	130	130	134	118	128	130	134	130	134	118	202	128
gelöst	122	81	80	98	78	63	84	65	64	72	63	123	89
gelöst in %	60,4	62,3	61,5	73,1	66,1	49,2	64,6	48,5	49,2	53,7	53,4	60,9	69,5

Poolnummer	98	101	108	109
bearbeitet	202	130	128	130
gelöst	130	83	86	99
gelöst in %	64,4	63,8	67,2	76,2

Zusammenfassung

Das dieser Arbeit zugrunde liegende Messinstrument bezeichnet sich als „Mental Arithmetic Test“ (kurz MAT) und ist ein neu entwickeltes Verfahren zur Erfassung numerischer Intelligenz. Zur Lösung seiner Aufgaben ist das Kurzzeitgedächtnis gefordert, grundlegendes mathematisches Wissen von Nöten wie auch logisches Denken und erfasst somit die Fähigkeitsdimension Kopfrechnen.
Das vorrangige Interesse dieser Arbeit dient der Entwicklung eines Itempools für den MAT, wobei die bestehenden Poolaufgaben analysiert und auf Rasch-Homogenität geprüft werden.
Es soll eine rein technische Itemselektion (Gittler, 1986) vermieden werden indem versucht wird die auffällig gewordenen Items anhand der aufgestellten Konstruktionskriterien und mit Hilfe entsprechender Literatur zu analysieren um somit eine eventuelle Selektion inhaltlich zu begründen.

Nach anfänglichen 1720 Personendaten wird nach Selektion von augenscheinlich unseriös bzw. unmotiviert arbeitenden Testteilnehmern (u. a. das Ausscheiden aller Teilnehmer, welche Item 16 nicht gelöst haben) nunmehr mit 1400 Daten weitergearbeitet. Im Laufe der statistischen Erhebungen und der Analyse der Version B mittels „Best Test Design“ nach Wright und Stone (1979) werden weitere sieben Personen ausgeschieden und die Rasch-Modelle letztendlich mit 1393 Personendaten weitergeführt.

Neben dem schon in den Vorerhebungen ausgeschiedenen Item 16, werden neun weitere Items (2, 6, 8, 12, 38, 52, 59, 108, 43) als ungeeignet für den Itempool des MAT erachtet. Selektiert man die genannten Items erhält man Rasch-Homogenität für die verbleibenden 85 Poolitems des MAT. Die Items der bereinigten Versionen messen mit sehr großer Wahrscheinlichkeit eindimensional und sind somit inhaltlich valide.

Abstract

The testing instrument which is used here is called „Mental Arithmetic Test“ (MAT) and is a new generated method for testing numeric abilities. For solving its items there is a need on good short-time memory, basic mathematical knowledge and reasoning to fulfil the dimension of mental arithmetic.
The major task is generating an item pool for the MAT by analysing the existing items to examine its homogeneity by the Rasch model. The notable items should be analysed according to an appropriate Literature and the construction criteria in order to avoid a technical item-selection (Gittler, 1986) and to support and give reasons for a selection.

At the beginning all participants (n 1720) were extracted that were assumed to be dubious or unmotivated (among others the extraction of all persons that fail item 16) and as a result it was continued with 1400 participants. During the statistical analysis by “Best test design” by Wright und Stone (1979) it was necessary to extract further seven participants and so at last it was resumed with 1393 persons.

In addition to item 16, which was because of its construction already determined as an unqualified item at the beginning of analysis, it was required to rule out further nine items (2, 6, 8, 12, 38, 52, 59, 108, 43) to get homogeneity by Rasch for the remaining 85 pool items of the MAT.
The revised items of the versions are very probably considered as one-dimensional and therefore are valid contently.

Lebenslauf

Nina FRANZ
Dörfleserstraße 36/1/1/1
2242 Prottes (NÖ)

geboren am: 9. März 1983 in Mistelbach
Staatsbürgerschaft: Österreich

Ausbildung

10/1997 – 06/2002	Handelsakademie Gänserndorf (NÖ)
10/2002 – 2009	Psychologiestudium an der Hauptuniversität Wien Schwerpunkte: klinische Psychologie und Differentielle Psychologie

Beruflicher Werdegang und weitere Ausbildungen

1999 - 2009	Anstellung bei Fa. Ing. Franz KEG in Prottes (NÖ) als Bürohilfskraft
11/2006	Ausbildung zum Taxilenker im Wifi Wien
2007	Erlangung der Führerscheinklassen C, D, E, F

Praktika

1999 Exploration &	Ferialjob in der Buchhaltungsabteilung der OMV Austria Production GmbH Gänserndorf
2007	6-Wochen-Praktikum im Ambulatorium Märzstraße für die Behandlung und Betreuung behinderter Kinder und Jugendlicher
2008	sechsstündiges Seminar zum Thema „Demonstrationen zur Veränderungsmessung mittels LLTM“
1997 – 2007	ehrenamtliche Tätigkeit im Landespensionisten und -pflegeheim Gänserndorf „Barbaraheim“
2009	ehrenamtliche Tätigkeit in der Behindertenhilfe „Geh mit uns“ in 2201 Kapellerfeld

Sprachkenntnisse

Englisch und Französisch in Wort und Schrift
Italienisch Grundkenntnisse