universität wien

D I S S E R T A T I O N

Titel der Dissertation

„Die Anwendung des dynamischen Testmodells von Kempf
auf unterschiedliche Datensätze“

Verfasserin

Mag. rer. nat. Marlis Posch

angestrebter akademischer Grad

Doktorin der Naturwissenschaften (Dr. rer. nat.)

Wien, 2008

Studienkennzahl lt. Studienblatt:	A 091 298
Dissertationsgebiet lt. Studienblatt:	Psychologie
Betreuer:	Univ.-Prof. Dr. Anton Formann MSc

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Inhaltsverzeichnis

1 Vorwort und Danksagung **5**

2 Einleitung **7**

3 Dynamisches Lernen **12**

3.1 Personenspezifische Lernmodelle 13

3.2 Operations- und itemspezifische Lernmodelle 18

3.3 Reaktionskontingente Lernmodelle 21

3.3.1 Markov-Modelle 22

3.3.2 Modelle aus der mathematischen Lerntheorie 26

4 Das dynamische Testmodell von Kempf **35**

4.1 Modelldarstellung 36

4.2 Schätzung der Item- und Transferparameter des Kempf-Modells 38

4.3 Schätzung der Personenparameter des Kempf-Modells 43

4.4 Goodness-of-Fit-Statistiken 45

4.4.1 Modellgeltungstest für das Kempf-Modell 45

4.4.2 Reduktion zum Rasch-Modell 46

5 Programm zur Schätzung der Modellparameter **48**

5.1 Technische Angaben 48

5.2 Schätzung der Rasch-Modell Parameter 49

5.3 Struktur des Programms 50

5.4 Graphische Benutzeroberfläche 52

5.4.1 Java-Programm 53

5.4.2 Leitfaden für Benutzer/innen 54

5.5 Ausgabe 60

6 Anwendung des dynamischen Testmodells **62**
6.1 Simulation von Daten 62
6.1.1 Simulationsprogramm 63
6.1.2 Automatisierung der Parameterschätzung für Simulationsreihen und Übertrag in SPSS 64
6.1.3 Ergebnisse einer Simulationsreihe mit 8 Items 66
6.1.4 Ergebnisse einer Simulationsreihe mit 20 Items 88
6.2 Mathematiksubtest der PISA-Studie 111
6.3 Zufallsauswahl aus Items von Bahrick & Hall 115
6.4 3DW-Daten von Gittler 118
6.5 SPM-Daten von Schmöger 122
6.5.1 SPM Subtest C, Erwachsene 123
6.5.2 SPM Subtest C, Kinder 126
6.5.3 SPM Subtest E, Kinder 128
6.6 WMT-Daten von Weber 131
6.7 WMT-Daten von Formann, Waldherr & Piswanger 134
6.8 Water-Level Tasks von Formann 137

7 Diskussion und Kritik **141**

8 Zusammenfassung **145**
Literaturverzeichnis 149
Lebenslauf 155

1 Vorwort und Danksagung

Diese Dissertation entstand aus Interesse an einem Modell, das 1974 entwickelt wurde, jedoch dann so gut wie in der Versenkung verschwand. In der Fachliteratur finden sich nicht viele Verweise darauf, es wurde bis jetzt nicht wieder aufgegriffen. Im Rahmen der vorliegenden Dissertation wurde das Thema „dynamisches Testmodell Kempf“ wieder aufgerollt, die dahinter stehende Theorie behandelt und vor allem die Anwendung mittels eines PC-Programms aktualisiert und getestet. Das Modell bzw. das Programm soll mit aktueller Software und auf aktuelle Datensätze angewendet werden. In dieser Arbeit wurde versucht, die Originalschreibweise von Formeln so gut es geht beizubehalten, in manchen Fällen war es auf Verständnisgründen jedoch wichtig, eine andere Schreibweise anzunehmen.

Das Verfassen dieser Dissertation war ein langer, aber auch interessanter Prozess, auf dem mich viele Menschen begleitet und unterstützt haben. Auf diesem Weg möchte ich für die fachliche und menschliche Unterstützung und Hilfe Danke sagen.

Mein Dank geht an den Hauptbetreuer meiner Dissertation Prof. Anton Formann, der mir das Thema zugedacht und nahe gebracht hat. Er hat an der Erweiterung des Fortran-Programms und der Verbesserung und Anpassung der Programmstruktur maßgeblich mitgewirkt.

Mein Ansprechpartner in allen Fragen zu neuen Programmen war mein Freund Mag. Fritz Wottawa. Er stand mir auch bei der Erstellung der graphischen Benutzeroberfläche und bei der Automatisierung der Simulationen mit Rat und Tat zur Seite. Ohne ihn hätte sich die Fertigstellung der Arbeit um lange Zeit verzögert.

Ohne die Hilfe von vielen liebe Kollegen hätte ich keine Daten zur Verfügung gehabt, um das adaptierte Fortran-Programm auch anwenden zu können. Mein Betreuer Prof. Formann stellte mir Daten zu den Water-Level Tasks aus seinem Forschungspraktikum II zur Verfügung; mein Zweitbetreuer Prof. Georg Gittler überließ mir netterweise Testdaten und ein paar Hintergrundinformationen zum 3DW; Dr. Karin Waldherr gab mir Testdaten zur revidierten Fassung des WMT; Mag. Michaela Schmöger stellte mir SPM Daten aus diversen

Forschungspraktika zur Verfügung; Prof. Harry Bahrrick gab mir per E-Mail die Zusicherung, dass ich den Datensatz von Bahrrick & Hall ohne weiteres verwenden darf; Univ. Doz. Ivo Ponocny ließ mir Ergebnisse des Mathematik-Tests der PISA-Studie zukommen; last but not least überließ mit Dr. Michael Weber einen WMT-Datensatz, lieferte mir aber auch in vielen spannenden Diskussionen wertvolles Feedback für meine Arbeit.

Auch meinen Eltern sage ich hiermit Danke für ihre liebe Unterstützung, die sie mir im Laufe der Jahre gegeben haben.

2 Einleitung

> „In vielen Fällen erfordert die Lösung neuer Problemstellungen einen Transfer von vorangegangenen Lösungsprozessen derart, dass Denkoperationen, die bisher in anderem Zusammenhang aufgetreten sind und geübt wurden, nun in neuen Konstellationen ablaufen.“ (Spada, 1976, S.22)

Dann tritt Lernen auf. Es weist also auch eine dynamische Komponente auf, die auf bisher gemachten Erfahrungen beruht. Die vorliegende Arbeit soll daher zunächst dynamisches Testen und Lerntheorie im Allgemeinen behandeln. Anschließend sollen verschiedene Arten dynamischer Lernmodelle nach thematischer Einordnung vorgestellt und verglichen werden. Diese Lernmodelle beziehen sich auf das Lernen während einer Testung bzw. Testbearbeitung. Auf Lernen zwischen zwei Testungen wird in diesem Rahmen nicht näher eingegangen.

Das zentrale Thema der vorliegenden Arbeit bildet das dynamische Testmodell von Kempf (1974). Eigenheiten und Modelldarstellungen werden eingehend behandelt. Bezug nehmend auf ein Originalprogramm von Kempf & Mach (1975) wird ein adaptiertes PC-Programm, von der Verfasserin kurz „DynTest“ genannt, zur Parameterschätzung vorgestellt.

Veränderungsmessung, und somit auch die Messung von Lernen war in der Geschichte der Psychologie oft und zu verschiedenen Zeiten Gegenstand von Untersuchungen und Theorien. In der folgenden Einleitung findet sich ein kurzer Abriss der Geschichte des dynamischen Testens und von mathematisch-psychologischen Modellen.

Ursprünglich wurde die Veränderung der Fähigkeit eines Individuums als Störgröße angesehen, da

> „z.B. das Konstrukt der Intelligenz einschließlich seiner Ausdifferenzierung in unterschiedliche Teilfähigkeiten als relativ zeit- und situationsinvariant definierte Eigenschaft verstanden“

wird (Guthke & Wiedl, 1996, S. 4). Auch die Methodik zur Erfassung dieser dynamischen Komponente war nicht ausreichend, sogar bei der Testkonstruktion und -durchführung wur-

de darauf geachtet, dass die Testaufgaben möglichst wenig störenden Einflüssen unterliegen, um eine möglichst hohe Reliabilität zu gewährleisten. Mit der Entwicklung der dynamischen Testdiagnostik, also der Diagnostik der intraindividuellen Variabilität wurde auch der Bedarf nach Modellen, die diese Veränderungen ausdrücken, immer stärker. Speziell Lerntests wurden in diesem Zusammenhang untersucht und entwickelt. Zusammenfassend kann die dynamische Testdiagnostik durch das Folgende definiert werden.

> „Dynamische Testdiagnostik ist ein Sammelbegriff für testdiagnostische Ansätze, die über die gezielte Evozierung und Erfassung der intraindiviuellen Variabilität im Testprozess entweder auf eine validere Erfassung des aktuellen Standes eines psychischen Merkmales und/oder seiner Veränderbarkeit abzielen.“ (Guthke & Wiedl, 1996, S. 8)

Lernpotential und Lernfähigkeit gehören zu diesem Feld der Veränderung. Lerntests nehmen einen besonderen Stellenwert im Bereich der dynamischen Testdiagnostik ein. Obwohl bis zu den Anfängen des 20. Jahrhunderts dynamisches Testen auf „herkömmliche“ Leistungs- und Intelligenztests beschränkt war, wurde die Idee von Lerntests verbreiteter, z.B. Buckingham (1921) sagt:

> „Theoretisch würde daraus folgen, dass die Messung des aktuellen Fortschritts repräsentativen Lernens den besten Intelligenztest darstellen würde ... Die meisten der jetzt gebräuchlichen Tests sind nicht Tests zur Erfassung der Lernfähigkeit („capacity to learn“), sondern dessen, was schon gelernt wurde.“ (Buckingham, 1921, S. 211, in Guthke & Wiedl, 1996, S. 18)

Es bildete sich ein Konzept der Lernfähigkeit in Abgrenzung zum Konzept der Intelligenz heraus, das auch diagnostisch genutzt wurde, anfangs vor allem für retardierte Kinder. Es wurden Trainingseffekte und die psychische Entwicklung von Kindern untersucht (vgl. Wygotski, 1934). Auch Kern (1930) beschäftigte sich mit dem Effekt von Übung und fasste die dynamischen Komponenten innerhalb jedes Individuums so zusammen:

> „Wir besitzen keine ausreichende Bürgschaft, dass die von der Prüfung als gut begabt befundene Prüflinge nach mehrfacher Wiederholung nicht starke Leistungsabfälle aufweisen und sind erst recht nicht gegen die Überraschung gesichert, dass Prüflinge, die wir auf Grund des Prüfungsausfalles als schlecht begabt zensieren, sich mit einem Male als hervorragend befähigt erweisen.“ (Kern, 1930, S. 464, in Guthke & Wiedl, 1996)

In späteren Jahren stand das so genannte „coaching“ im Mittelpunkt der Forschung, das sich auf die Unterweisungen während und vor der Testung bezieht, da durch solches Feedback die Testperformance wesentlich verändert werden kann (siehe Wiseman, 1954). Verschiedene Coachingmethoden und -intensitäten wurden verglichen, leistungsverändernde Maßnahmen wurden inventarisiert. Besonders einschneidend für die dynamische Testdiagnostik war die Formulierung der Axiome von Zubin (1950), z.B

1. „dass jedes Individuum zunächst als eigenes Universum zu betrachten sei, das erst nach tieferer Erforschung mit anderen in Gruppen zusammengefasst werden dürfe,
2. dass es für jedes Individuum und jedes Merkmal ein spezifisches Performanzniveau gäbe, zu dem der beobachtete Testwert eine Stichprobe darstelle, und
3. dass jedes Individuum und jedes Merkmal auch durch einen Grad an Variabilität mit einem je spezifischen Muster („Spielbreite“) gekennzeichnet sei.“ (Guthke & Wiedl, 1996, S. 38)

Merkmale und Fähigkeiten einer Person können sich also im Zustand von Fluktuation befinden.

Die Hinwendung zum Individuum fand auch bei den statistischen Methoden statt, auch experimentelle Einzelfallanalysen wurden durchgeführt. Anstatt in der Vergangenheit erworbenes Wissen abzufragen, schlugen Psychologen wie Jensen (1961) vor, die direkte Lernfähigkeit mittels Aufgaben zum unmittelbaren Behalten, seriellen Lernen und Paarassoziationslernen zu erheben. Rohwer (1971) definierte die Lernfähigkeit als „die Fähigkeit zu Erwerb, Behalten und Produzieren neuer Informationen“ (S. 192, in Guthke & Wiedl, 1996, S. 45). Hier waren die „Lerntaktiken“ die Analyseeinheiten.

Severson (1976) und seine Mitarbeiter konzentrierten sich im Rahmen der „*Lernprozessdiagnostik*“ auf die Einflüsse von Arten der Aufgabenpräsentation, verschiedenen Instruktionsformen und Arten der Verstärkung bei der Vermittlung von Lesefertigkeiten. Ab den 70er Jahren des 20. Jahrhunderts wurde generell die Lernfähigkeit unter verschiedenen Interventionen während Lang- (d.h. Test-Training-Test) und Kurzzeitlerntests (d.h. *eine* Testung) verstärkt Gegenstand der Forschung. In diesem Sinne wurden auch die Untersuchungs- und Trainingsprozeduren standardisiert und verschiedene Validitätsaspekte empirisch kontrolliert (vgl. Guthke, 1972). Andere Konzepte dynamischer Untersuchungsverfahren waren

beispielsweise das *Learning Potential Assessment Device* von Feuerstein et al. (1979) und die „*Lernpotentialdiagnostik*“ (*Learning Potential Assessment*) von Budoff et al. (1971), bei der in Personen eingeteilt wurde, die ihre hohe Leistung beibehalten konnten („*high scorer*“), Personen, die ihre Leistung deutlich steigerten („*gainer*“) und Personen, die niedrige Leistung nicht verbessern konnten („*non-gainer*“).

Neuere dynamische Lerntests sind auf die Messung von Behalten und Transfer von Lernerfahrungen ausgerichtet. Einsatzgebiete sind vor allem die Förderdiagnostik, Berufseignungsdiagnostik und Rehabilitation (siehe Guthke & Wiedl, 1996). Dynamische Lerntests umfassen die bereits erwähnten zwei Hauptarten.

Langzeit-Lerntests Sie bestehen aus 3 Phasen, der ersten Testphase, einer Pädagogisierungs- bzw. Trainigsphase und einer Posttestphase. Für diese Art von Lerntest wird die Veränderungsmessung zur Gewinnung statistischer Messzahlen herangezogen, da der Lerngewinn oder -verlust durch den Prä- und Posttestvergleich gemessen werden kann. In der vorliegenden Arbeit wird nicht näher auf Prä- / Posttestveränderungen eingegangen.

Kurzzeit-Lerntests Bei dieser Art von Tests wird die Pädagogisierungsphase direkt in den Testprozess miteinbezogen. Es kommt zu Feedbacks und Lösungshinweisen während der Testung, somit muss nur eine einzige Testung vorgenommen werden. Die statistische Auswertung gestaltet sich bei Kurzzeit-Lerntests etwas schwieriger. Eine Möglichkeit ist es, die „Empfänglichkeit für Hilfen“, oder die Latenzzeit bis zum Auftreten des ersten Fehlers zu messen (vgl. Guthke & Wiedl, 1996). Eine andere Möglichkeit bezieht sich auf dynamische Test- bzw. Lernmodelle, die den Lerngewinn innerhalb eines Tests von Item zu Item messen und im folgenden genauer unter die Lupe genommen werden sollen.

Die Geschichte der mathematischen Modelle ist mindestens eben so lang. Bereits im Jahre 1837 forderte Hebart eine mathematische Formulierung psychologischer Theorien und unternahm den Versuch, die Gesetze der Newton’schen Mechanik auf die Psychologie zu übertragen.

Weber und Fechner gelang es erstmals, mit dem Weber-Fechner Gesetz den Zusammenhang zwischen physischen und mentalen oder psychischen Prozessen mathematisch auszudrücken. Die ersten Jahrzehnte des 20. Jahrhunderts brachten einen Aufschwung der experimentellen Psychologie und der mathematischen Statistik mit sich. Mathematische Modelle

wurden immer alltäglicher angewendet, erste Höhepunkte wurden mit der klassischen Testtheorie und der Faktorenanalyse erreicht.

> „Diese erste Phase der Entwicklung war dadurch gekennzeichnet, dass die jeweiligen Modellannahmen in erster Linie aus Gründen der mathematischen Einfachheit gewählt wurden und sich nur sekundär an den Erfordernissen des psychologischen Forschungsgegenstandes orientierten. Zugleich waren diese Modelle auf eine universelle Anwendbarkeit hin ausgelegt und wurden infolgedessen häufig als bloße Methoden missverstanden.“ (Kempf, 1974, S.14)

Ende des zweiten Weltkrieges wurde Kritik an der klassischen Testtheorie laut. Erkenntniskritische und wissenschaftstheoretische Überlegungen wurden häufiger geäußert, Namen wie Guttman, Lazarsfeld, Rasch und Fischer prägten die Forschungslandschaft. Dennoch wurden die Modellannahmen nur selten reflektiert und mit den inhaltlichen Theorien in Einklang gebracht. Kempf (1974) definiert daher die wichtigsten Aufgaben innerhalb des psychologischen Forschungsprozesses folgendermaßen:

- „die Präzisierung psychologischer Konzepte,
- die Herstellung einer eindeutigen Zuordnung zwischen inhaltlich-psychologischen Theorien und den Methoden ihrer Überprüfung.“ (S. 16)

Die zweite Anforderung von Kempf (1974) betrifft die mathematische Handhabung des Modells. Essenziell sei

- „die einwandfreie Bestimmbarkeit der Modellparameter,
- die einwandfreie Vergleichbarkeit der Modellparameter (sofern die zu formalisierende Theorie Aussagen über Relationen zwischen Modellparametern trifft),
- die einwandfreie Prüfbarkeit der Modellstruktur.“ (S. 17)

Ein solches Modell, das inhaltlich zu psychologischen Theorien passt, formulierte er schließlich selbst. Es sollte als Grundidee für die dynamischen Testmodelle gelten.

3 Dynamisches Lernen

In der Item-Response-Theorie gibt es eine zentrale Annahme - die lokale stochastische Unabhängigkeit.

> „Betrachtet man die Durchführung eines Tests als ein Experiment ..., so stellt die Beobachtung des Verhaltens mehrerer Personen bei verschiedenen Items ... eine Messwiederholung dar. Da alle Itemantworten von denselben Personen stammen, und durch die zu messende Personeneigenschaft bedingt sind, werden keine unabhängigen Beobachtungen realisiert.
>
> Hält man die zu messende Personeneigenschaft jedoch konstant, z.B. indem man nur eine Person betrachtet oder nur Personen mit derselben Ausprägung der latenten Variable, so müssen die Items experimentell unabhängig bearbeitet werden.
>
> Diese spezielle Art von Unabhängigkeit nennt man lokale stochastische Unabhängigkeit.“ (Rost, 2004, S. 69)

Die Antwortvariablen sollen also bedingt unabhängig voneinander sein. Dadurch setzt sich die Wahrscheinlichkeit eines ganzen Antwortmusters aus dem Produkt der Wahrscheinlichkeiten der Antworten auf die Items zusammen. Bezüglich dieser zentralen Annahme ergeben sich aber Schwierigkeiten, wenn es zu Veränderungen der latenten Personenfähigkeit innerhalb einer Testung kommt. Diese Veränderungen stellen eine direkte Verletzung der lokalen stochastischen Unabhängigkeit dar. In dynamischen Test- bzw. Lernmodellen soll genau diese Veränderung der Personenfähigkeit während eines Tests untersucht und gemessen werden. Es gibt jedoch verschiedene Arten von dynamischen Testmodellen: Eine Art bezieht sich nur auf die Personenfähigkeit und fällt somit unter die Kategorie *personenspezifische Lernmodelle*, eine andere Art auf die Items oder zugrunde liegende Operationen - *operations- und itemspezifische Lernmodelle* und eine letzte Art hängt personen- und itemunspezifisch ausschließlich von vorangegangenen Antwortmustern ab und wird unter dem Begriff *reaktionskontingente Lernmodelle* subsumiert.

3.1 Personenspezifische Lernmodelle

Der Ansatz des personenspezifischen Lernens geht auf das Konzept der Lerntests (siehe Abschnitt 2) zurück. Während der Bearbeitung von Lerntests wird versucht, die Veränderung der Personenfähigkeit als Indikator für die individuelle Lernfähigkeit zu messen. Der Lerngewinn soll also *personenspezifisch* gemessen werden. Die Lernvorgänge sind personenspezifisch, insofern sie von der Anzahl der pro Person bearbeiteten Items abhängen, jedoch nicht von den Reaktionen der Personen. Ein besonderes Beispiel für personenspezifisches Lernen bieten Klauer & Sydow (1992), die sich besonders mit der Entwicklung eines probabilistischen Modells für Kurzzeitlerntests beschäftigten. Klauer & Sydow meinen, das Konzept eines Lerntests beruhe auf der Ansicht, dass

1. „das Testverhalten beim Lerntest durch *zwei* Faktoren, Fähigkeitsstatus und Lernfähigkeit, beschrieben sei,
2. dass es bedeutende interindividuelle Unterschiede in der Lernfähigkeit gebe,
3. dass die Lernfähigkeit außerdem diagnostische Informationen liefere, die in dem Status nicht berücksichtigt sei, das heißt, dass Lernfähigkeit und Status relativ unabhängig, zumindest aber nicht perfekt korreliert seien, und
4. dass die Lernfähigkeit für die Prognose zukünftiger Leistungen möglicherweise wichtiger sei als der Status, zumindest aber einen zusätzlichen Beitrag liefere.“ (S. 175)

Nachtests im Vergleich zu Vortests, oder Kurzzeittests im Vergleich mit herkömmlichen Tests zeigen oft einen leichten Gewinn an Vorhersagegüte, besonders für den unteren Leistungsbereich (Guthke, 1990; Flammer & Schmid, 1982). Das könnte für die Berücksichtigung einer Lernfähigkeit sprechen. Die Indizes für Fähigkeitsstatus und Lernfähigkeit werden dabei allerdings vermengt. Klauer & Sydow entwickelten daher ein eigenes Modell, das zwischen den beiden Faktoren differenzieren kann. Durch dieses Modell können unverzerrte Schätzwerte für die Varianz der Fähigkeitsstatus- und der Lernfähigkeitsvariablen und deren Korrelationen gewonnen werden, was bei anderen Modellen zu einem erheblichen Problem geführt hatte (Zimmermann & Williams, 1982a, 1982b).

Die Wahrscheinlichkeit, dass ein Proband mit Fähigkeitsparameter ξ die i-te Aufgabe löst, wird nun durch $f_i(\xi)$ beschrieben.

$$P(X_i = 1|\xi) = f_i(\xi). \tag{3.1}$$

Als Itemcharakteristik f_i nehmen Klauer & Sydow als Ausgangsannahme die logistische Funktion des Rasch-Modells an.

$$f_i(\xi) = \frac{exp(\xi - \sigma_i)}{1 + exp(\xi - \sigma_i)} \tag{3.2}$$

Zusätzlich zum Rasch-Modell postulieren sie auch einen Lernzuwachsparameter ψ. Jede Aufgabe i des Lerntests kann zu einem Lernzuwachs ψ im Fähigkeitswert ξ der Person führen. Dieser ist abhängig von der Person, der Aufgabe und der erhaltenen Hilfestellung. Für den Lerngewinn wird ein Item also entweder von der Person spontan gelöst, oder sie erhält eine Hilfestellung.

Außerdem treffen Klauer & Sydow die vereinfachte Annahme, dass dieser Lernzuwachs ψ zwar personenspezifisch, aber gleich groß für alle Aufgaben sei. Die Wahrscheinlichkeit für eine richtige Antwort des Probanden ist nun

$$P(X_i = 1|\xi, \psi) = f_i(\xi + w_i\psi), \tag{3.3}$$

wobei w_i die bis zur i-ten Aufgabe akkumulierten $i-1$ Lerngelegenheiten zusammenfasst. Bei Item 1 ist $w_1 = 0$, bei Item 10 ist $w = 9$ usw.

Die Statusvariable ξ und die Lernfähigkeitsvariable ψ seien weiters in der Population bivariat normalverteilt mit unbekannten Varianzen σ^2_ξ, σ^2_ψ und unbekannter Korrelation $\rho_{\xi,\psi}$. Die Populationsmittelwerte seien Null. Auf Grund dieser Annahmen wird ein neues Antwortmuster $x = (x_1, ..., x_j)^t$ festgelegt. Die Wahrscheinlichkeit für dieses Antwortmuster ist dann

$$P(X = x|\xi, \psi) = \mu(\xi, \psi)h(x)exp(X\xi + Y\psi) \tag{3.4}$$

wobei

$$\mu(\xi, \psi) = \prod_{j=1}^{J}(1 + exp(\xi + w_j\psi - \sigma_j))^{-1},$$

$$h(x) = exp(-\sum_{j=1}^{J} x_j\sigma_j),$$

$$X = \sum_{j=1}^{J} x_j \text{ und } Y = \sum_{j=1}^{J} w_j x_j.$$

Es ergeben sich erschöpfende Statistiken für ξ und ψ, nämlich einmal der Testwert X und einmal der gewichtete Score Y.

Zur eindeutigen Definition des Modells sind jedoch noch Normierungen von Nöten. Analog zu faktorenanalytischen Modellen kann man Status und Lernfähigkeit als Faktoren auffassen, die laut Klauer & Sydow in das i-te Item mit den Ladungen $v_i = 1$ für den Fähigkeitsstatus und mit w_i für die Lernfähigkeit miteingehen.

$$P(X_i = 1|\xi, \psi) = f_i(v_i\xi + w_i\psi) \tag{3.5}$$

Die notwendigen Lineartransformationen der Ladungsvektoren $v = (v_1, ...v_J)^t$ und $w = (w_1, ...w_J)^t$ umfassen die Gleichsetzunge der Länge

$$|v| = |w|$$

und die Orthogonalität der beiden Ladungsvektoren

$$v^t w = 0.$$

Die transformierten Werte für w_i ergeben sich bei Konstanthaltung der Ladungen des Statusfaktors ξ $v_i = 1$ mit

$$w_i = \sqrt{\frac{12}{(J-1)(J+1)}} \left(i - \frac{J+1}{2} \right).$$

Dies wäre die Gestaltung eines einfachen Modells für dichotom codierte Daten. Es kann ebenfalls für mehrkategorielle Daten ausgedehnt werden, z.B. „mit Hilfe gelöst“, „ohne Hilfe gelöst“, etc..

> „Das Modell postuliert einen linearen Zuwachs in der aktuellen Fähigkeit als Funktion der Zeit beziehungsweise der Lerngelegenheiten. Damit fällt es in die große Klasse der so genannten *straight-line growth*-Modelle, deren Eigenschaften zum Beispiel von Rogosa and Willett (1985) detailliert analysiert wurden.“ (Klauer & Sydow, 1992, S.179)

Klauer & Sydow (1992) analysierten in der Folge mit Hilfe dieses Modells Kurzzeitlerntests auf Lernprozesse und erhöhte Anpassungsgüte. Dazu konstruierten sie zusätzliche vereinfachte Modelle. Im ersten wird die Annahme getroffen, dass es keine bedeutenden interindividuellen Unterschiede in der Lernfähigkeit gibt. Die Varianz σ_ψ^2 wird also im Vorhinein gleich Null gesetzt. Es wird des weiteren durch ein zweites Modell geprüft, ob es nötig ist, Status- und Lernfähigkeit als nicht redundante Faktoren anzusehen, also ob ein Verlust der Anpassungsgüte auftritt, wenn der Absolutwert der Korrelation $\rho_{\xi,\psi}$ gleich 1 gesetzt wird.

> „Bei diesen Analysen liegen auf der Seite der Daten die Häufigkeiten vor, mit denen einzelne Antwortmuster in der untersuchten Stichprobe auftreten. Diese werden mit den Wahrscheinlichkeiten verglichen, die das jeweilige Modell für das Auftreten des Antwortmusters in der Population vorhersagt. Man erkennt, dass die Modellprüfung und -vergleiche hier wie anderswo auf Aussagen über die Population von Probanden beruhen und keineswegs prüfen, ob das Testverhalten der untersuchten Individuen dem Modell folgt.“ (Klauer, 1988, in Klauer & Sydow, 1992)

Zur Parameterschätzung wird folgendermaßen vorgegangen. Die Wahrscheinlichkeit eines Antwortvektors $x = (x_1, ..., x_J)^t$ bei J Aufgaben ergibt sich durch

$$P(X = x) = \int \prod_{j=1}^{J} f_j^{x_j}(\xi + w_j\psi)(1 - f_j(\xi + w_j\psi))^{1-x_j} dN(\xi, \psi). \tag{3.6}$$

$N(\xi, \delta)$ bezeichnet die bivariate Normalverteilung. Die Aufgabenparameter, die Varianzen σ_ξ^2 und σ_ψ^2 sowie die Kovarianz $\sigma_{\xi,\psi}$ bedingen die Wahrscheinlichkeit. Die Likelihood der Daten berechnet sich aus

$$L = \prod_{x \in \Omega} P(X = x)^{n_x}, \tag{3.7}$$

wobei Ω die Menge der in der Stichprobe tatsächlich auftretenden verschiedenen Antwortmuster und n_x die Häufigkeit des Antwortvektors x bezeichnet. Für die Parameterschätzung werden wie üblich die Maxima der Funktion gesucht. Zur Berechnung der $P(X = x)$ schlagen Klauer & Sydow (1992) ein Verfahren vor, das aus der bivariaten Normalverteilung der Variablen ξ und ψ seinen Nutzen zieht. Mit geeigneten Koeffizienten a, b und c als Linearkombination zweier unabhängig normalverteilter Variablen u und v mit der Varianz 1 können

die Parameter wie folgt dargestellt werden

$$\xi = au \text{ und } \psi = bu + cv$$

bei den Varianzen von

$$\sigma_{\xi}^2 = a^2 \text{und } \sigma_{\psi}^2 = b^2 + c^2$$

und der Kovarianz von

$$\sigma_{\xi,\psi} = ab.$$

Als Normierungen werden $a > 0$ und $c \geq 0$ festgelegt. Nach Ersetzen der Parameter ξ und ψ in der Formel für $P(X = x)$ kann man nun über die univariate Standardabweichung integrieren. Für die numerische Auswertung des Integrals kann das Gauß-Hermite Verfahren angewendet werden, welches sehr effizient ist (siehe Bock & Aitkin, 1981).

Auch für die Formulierung der vereinfachten Analysemodelle für die Kurzzeitlerntests ist die Reparametrisierung äußerst nützlich, es können einfache lineare Hypothesen über die neuen Parameter a, b und c gebildet werden. $\sigma_{\psi}^2 = 0$ entspricht nun den Hypothesen $b = 0$ und $c = 0$. $\rho_{\xi,\psi} = 1$ kann durch $c = 1$ ersetzt werden. Zur Maximierung der Likelihoodfunktion wenden Klauer & Sydow (1992) ein konjugiertes Gradientenverfahren an.

Zunächst sollte die Anpassungsgüte der Kurzzeitlerntests überprüft werden, also ob das Basismodell die Häufigkeiten der einzeln beobachteten Antwortmuster zufrieden stellend beschreibt. Durch die große Anzahl an denkbar möglichen Antwortmustern müssen jedoch mehrere Antwortmuster zusammengefasst werden.

> „Da für die Lerntests die gemeinsame Verteilung des Testwerts X und der suffizienten Statistik Y für die Lernfähigkeit besonders interessant ist, teilen wir jede Rohwertgruppe entlang der Terzile der Verteilung von Y innerhalb der Rohwertgruppe noch einmal in drei etwa gleich große Gruppen mit den kleinsten, mittleren und größten Werten von Y." (Klauer & Sydow, 1992, S. 181)

Es wird ein χ^2-Test mit der Statistik

$$G^2 = -2\sum_{x,y} n_{x,y} log\left(\frac{m_{x,y}}{n_{x,y}}\right) \tag{3.8}$$

durchgeführt, wobei $n_{x,y}$ die beobachteten Häufigkeiten des jeweiligen Y-Bereiches bei gegebenem Testwert und $m_{x,y}$ die Modellvorhersagen für die Zellen aufgrund der geschätzten

Parameter darstellt. Die Anzahl der Freiheitsgrade ist um 1 geringer als die Differenz aus der Anzahl der Zellen und der Anzahl der Modellparameter. Dieser G^2-Wert ist allerdings nicht ganz minimiert, da die Modellparameter nicht aus den aggregierten Daten, sondern anhand der Likelihoodfunktion geschätzt wurden.

Des weiteren können auch noch die vereinfachten Modelle mit $b = 0$ und $c = 0$ bzw. nur $c = 0$ mit dem ursprünglichen Basismodell verglichen werden. Das Maximum L_1 für das Basismodell und das Maximum L_2 für die zusätzlichen Modelle müssen hierfür berechnet werden. Es erfolgt ein weiterer χ^2-Test durch die Statistik

$$G^2 = -2log\left(\frac{L_2}{L_1}\right) \tag{3.9}$$

mit df gleich der Anzahl der Parameter, die im vereinfachten Modell gleich Null gesetzt werden.

3.2 Operations- und itemspezifische Lernmodelle

Im Gegensatz zu personenspezifischem Lernen ist auch ein operationsspezifisches Lernen durch das Üben einzelner Operationen denkbar. Dadurch würde die Operationsschwierigkeit verringert. Spada (1976) geht auf das linear logistische Denkmodell von Scandura (1973) im Hinblick auf operationsspezifisches Lernen ein. Ausgehend von der Grundstruktur des linearen logistischen Denkmodells von Scandura, das sich wiederum auf das linear logistische Testmodell (LLTM) von Fischer (1972)

$$p_{vi} = \frac{exp(\xi_v - \sum_{j=1}^{m} f_{ij}\eta_j + c)}{1 + exp(\dots)} \tag{3.10}$$

bezieht, wird eine Erweiterung des Modells vorgenommen, um operationsspezifisches Lernen berücksichtigen zu können. Die Konstanz der Operationsparameter wird vorausgesetzt.

η_j im Intervall $[0, +\infty]$ bezeichne einen Operationsschwierigkeitsparameter, der noch nicht durch operationsspezifisches Lernen verändert wurde. τ_{ij} bezeichne den Effekt des Übens von Operation j auf die Verringerung der Operationsschwierigkeit bis zum Zeitpunkt der Bearbeitung von Aufgabe i. Die Differenz $(\eta_j - \tau_{ij})$ wäre dann die verbleibende Schwierigkeit der Operation zu diesem Zeitpunkt. Die Operationsparameter η_j sind genauso wie die operationsspezifischen Lernparameter τ_{ij} für alle Versuchspersonen gleich. Man kann nun

diese Annahmen über die Formulierung von Nebenbedingungen der Aufgabenparameter so formulieren

$$\sigma_i = \sum_{j=1}^{m} f_{ij}(\eta_j - \tau_{ij}) + c \text{ für alle } i = 1,2,\ldots,k. \tag{3.11}$$

Da die Anzahl der Operations- und Lernparameter jedoch in dieser Darstellung zu groß ist, können die Parameter nicht geschätzt werden. Zur Reduzierung der Parameter müssen die Annahmen über die τ_{ij} restringiert werden.

Der Effekt der Übung hängt von der Form und Häufigkeit der Übung ab. Wenn vorausgesetzt wird, dass die Häufigkeit der Übung bekannt ist, kann diese auf folgende Weise aus der Aufgabenstrukturmatrix erschlossen werden

$$h_{ij} = \sum_{u=1}^{i-1} f_{uj}, \tag{3.12}$$

wobei f_{uj} die Häufigkeit von Operation j bei der Aufgabe u und h_{ij} die Häufigkeit einer Übung von Operation j durch Aufgabenbearbeitung bis zum Zeitpunkt der Vorgabe von Aufgabe i darstellen.

> „Die erste Übung einer Operation (während der Bearbeitung von Aufgaben) hat - so vermute ich - den größten Effekt auf die Abnahme der Operationsschwierigkeit. Mit zunehmender Übungshäufigkeit wird wahrscheinlich der zusätzliche Effekt jeder weiteren Übung immer geringer, bis schließlich eine weitere Abnahme der Operationsschwierigkeit durch Übung aufgrund von Sättigungseffekten nicht mehr erreicht werden kann.“ (Spada, 1976, S. 148)

Die Restriktionen zur Parameterreduktion sind

$$\tau_{ij} = h^*_{ij}\beta_j \quad \text{mit } 0 \leq \beta_j \leq \eta_j \tag{3.13}$$

und

$$h_{ij}^* = f(h_{ij}) \quad \text{mit } f(h_{ij}) \begin{cases} = 0 \text{ für } h_{ij} = 0 \\ > f(h_{ij}-1) \text{ für alle } h_{ij} = 1,2,\ldots \\ \longrightarrow 1 \text{ für } h_{ij} \longrightarrow \infty \end{cases} \tag{3.14}$$

$$\text{und } [f(h_{ij}+1) - f(h_{ij})] < [f(h_{ij} - f(h_{ij}-1)]$$

$$\text{für alle } h_{ij} = 1,2,\ldots.$$

Die Lernparameter τ_{ij} sind also abhängig von der Übung der einzelnen Operationen und von den Parametern β_j. Letztere werden mit transformierten Werten, anstatt mit den Übungshäufigkeiten multipliziert. h_{ij}^* ist auf das Intervall [0,1] beschränkt und strebt für ein wachsendes h_{ij} asymptotisch gegen 1. Daher markiert β_1 die maximale durch Übung erzielbare Verringerung der Operationsschwierigkeit j. D.h. h_{ij}^* legt fest, welcher Anteil des erzielbaren Übungsgewinns nach einer bestimmten Operationsanzahl erreicht worden ist. Es ist auch möglich, dass überhaupt kein Lernen durch Übung stattfindet, nämlich dann, wenn die Differenz $(\eta_j - \beta_j) = \eta_j$ ist.

Die Transformation der Übungshäufigkeiten findet mittels der Funktion f statt, die streng monoton, verzögert wachsend und auf das Intervall $[0,1]$ beschränkt ist. Nach Art des Rasch-Modells wird sie in logistischer Form angeschrieben.

$$h_{ij}^* = \frac{h_{ij} b}{1 + h_{ij} b} \quad \text{mit } h_{ij} = 1,2,\ldots \tag{3.15}$$

$$\text{und } b > 0.$$

Der Faktor b gewichtet in dieser Gleichung die Übungshäufigkeiten und legt fest, wie rasch h_{ij}^* gegen 1 bzw. $h_{ij}^* \beta_j$ gegen den maximal erreichbaren Übungsgewinn β_j strebt. Faktor b ist gleich groß für alle Operationen, die zum Lösen von homogenen Aufgaben benötigt werden. Er ist kein zu schätzender Parameter, sondern hypothetisch festzulegen. Die Aufgabenstruktur $((f_{ij}))$ und die transformierten Übungshäufigkeiten h_{ij}^* sollen somit für die Schätzung der Parameter bekannt sein.

Die Aufgabenparameter können mit

$$\sigma_i = \sum_{j=1}^{m} f_{ij}(\eta_j - h_{ij}^* \beta_j) + c = \sum_{j=1}^{m} (f_{ij}\eta_j - f_{ij} h_{ij}^* \beta_j) + c. \tag{3.16}$$

verallgemeinert werden. Das erweiterte operationsspezifische linear logistische Denkmodell von Spada (1976) kann nun in folgender Weise angeschrieben werden:

$$p_{vi} = \frac{exp(\xi_v - \sum_{j=1}^{m} f_{ij}(\eta_j - h^*_{ij}\beta_j) + c}{1 + exp(\dots)}. \tag{3.17}$$

Die Anzahl der Parameter beträgt $2m$, da jede Operation durch zwei Parameter - die Anfangsschwierigkeit und den maximalen Übungsgewinn - festgelegt wird. Die Wahrscheinlichkeit einer korrekten Itemlösung wird zurückgeführt auf einen Personen- und einen Itemparameter, wobei der Itemparameter als über die Zeit variabler Operationsparameter charakterisiert wird. Die Veränderung ist abhängig von Art und Umfang der erfolgten Übung. Bezogen auf operationsspezifisches Lernen hängt also die Aufgabenschwierigkeit sowohl von der Struktur der Aufgabe an sich ab, als auch von der Struktur der zuvor vorgelegten Items und der Stelle, an der das Item vorgelegt wird.

> „Der Übungstransfer ist operationsspezifisch in dem Sinne, dass die Verringerung der Schwierigkeit einer Aufgabe ausschließlich auf jene zu ihrer Lösung benötigten Operationen zurückgeht, die durch Übung bei vorangegangenen Aufgaben leichter geworden sind. Operationsspezifisches Lernen hat somit im Allgemeinen eine unterschiedliche Verringerung der Schwierigkeit einzelner Aufgaben zur Folge.“ (Spada, 1976, S. 152)

Die Aufgabenstrukturmatrix zur Schätzung der Parameter enthält die Aufgabenstruktur $((f_{ij}))$ und die mit diesen Häufigkeiten multiplizierten transformierten Übungshäufigkeiten h^*_{ij}. Die Schätzung der Parameter wird ohne Berücksichtigung der dargestellten Nebenbedingungen über die Operationsparameter η_j und die Lernparameter β_j durchgeführt, die festlegen, in welchem Wertebereich die Parameter liegen sollen. Für eine möglichst gezielte Modellgeltungskontrolle sollte die Aufgabensequenz in verschiedenen Personenstichproben systematisch variieren.

3.3 Reaktionskontingente Lernmodelle

Die Maßzahl bei reaktionskontingenten Lern- oder Testmodellen bezieht sich auf den Lerngewinn oder -verlust in Abhängigkeit davon, ob eine Person ein Item tatsächlich gelöst hat oder nicht. Das Lernen findet reaktionskontingent statt, d.h. personen- und itemunspezifisch

abhängig vom bisher gezeigten Verhalten in einem Test. Ein Lerneffekt fällt also anders aus, je nachdem ob Items vorher gelöst wurden oder nicht.

> „Generell sind beide Richtungen denkbar, nämlich dass man nur dann lernt, wenn man ein Item gelöst hat, weil man ein 'reinforcement' (dt. Verstärkung) aufgrund der gelungenen Lösung erhält. Es ist aber auch denkbar, dass man einen Lerneffekt nur bei nicht-gelösten Aufgaben erzielt, denn nur bei solchen gibt es noch etwas zu lernen, z.B. durch die nachträgliche Mitteilung des korrekten Lösungsweges." (Rost, 2004, S. 291)

Innerhalb der Gruppe der reaktionskontingenten Lernmodelle kann wiederum eine weitere Differenzierung in 2 verschiedene Subgruppen getroffen werden. Zum einen gibt es Modelle aus der Gruppe der *Markov-Modelle*, zum anderen Modelle, die aus der *mathematischen Lerntheorie* resultieren.

3.3.1 Markov-Modelle

Gemischte und latente Markov-Modelle bieten eine Möglichkeit, dynamische Prozesse darzustellen. Für gemischte latente Markov-Modelle gilt die lokale stochastische Unabhängigkeit der manifesten Variable bei Konstanthaltung der latenten Variable nicht, sondern eine spezielle Art der lokalen stochastischen Abhängigkeit.

Markov-Modelle im Allgemeinen setzen sich aus einer Verschmelzung von so genannten Markov-Ketten nach Andrei Andrejewitsch Markov und Mischverteilungsmodellen zusammen (Langeheine & Van de Pol, 1990).

> „Markov models are aimed at modeling the transition probabilities between two or more different states at consecutive time points. That means, a person may be in a state A at time point $t-1$ and moves to another state, say B, at time point t." (Rost, 2002, p. 55)

Markov-Ketten an sich sind probabilistische Modelle, die die Übergangswahrscheinlichkeit von Zustand A zum Zeitpunkt $t-1$ zum Zustand B zum Zeitpunkt t mittels eines Transitionsparameters $\tau_{A,t-1,B,t}$ beschreiben. Markov-Modelle gelten für diskrete oder kategorielle Daten. Die manifesten Daten sind Häufigkeiten von beobachteten Mustern von Variablen zu drei oder mehr Zeitpunkten. Weiters können manifeste von latenten Markov-Modellen

unterschieden werden. Das Auftreten von Messfehlern spielte eine große Rolle für die Entwicklung dieser Unterscheidung. Während in manifesten Markov-Modellen keine Messfehler auftreten, ist dies bei latenten Modellen nicht der Fall. Neben den Übergangsparametern wurde dafür ein zweiter Parametertyp eingeführt: die Wahrscheinlichkeit, einen Indikator für den Zustand A zu beobachten, während Zustand A wirklich gegeben ist. Die Wahrscheinlichkeiten $\rho_{I(A)|A}$ sind bedingte Wahrscheinlichkeiten, die die Stärke des Zusammenhangs zwischen dem Indikator $I(A)$ und dem indizierten Zustand A angeben. Bei manifesten Markov-Modellen sind alle ρ-Parameter gleich 1.

Zum ersten Zeitpunkt, $t = 1$, wird die Verteilung der manifesten diskreten Variablen X von einer latenten Verteilung mit den Parametern δ_A und δ_B festgelegt. Diese beschreiben die Wahrscheinlichkeiten, dass sich die Person in Zustand A oder B befindet. Es werden also die bedingten Wahrscheinlichkeiten der manifesten Variable X durch $\rho_{x|A}$ und $\rho_{x|B}$ den Zuständen A und B zugeteilt. Indikatoren für die Zustände A und B sind hierbei die Kategorien von X. Die Überprüfung der Modellgültigkeit kann mittels χ^2-Statistik nach Pearson oder L^2-Likelihood-Ratio erfolgen. (Langeheine & Van de Pol, 1990)

Zur Illustration des Konzepts von latenten Markov-Modellen dient ein sehr einfaches Beispiel. Es soll nur eine einzige manifeste Variable X mit lediglich zwei Kategorien ($x = 0$: „Item nicht lösen“ und $x = 1$: „Item lösen“) und nur zwei (gleich bleibende) latente Zustände A („die Person ist hoch motiviert“) und B (“die Person ist gar nicht motiviert“) für jeden Zeitpunkt geben. Die Verteilung der manifesten Variablen über die Zeit würde dann mit

$$\begin{aligned} t = 1 : & p(x_1) = \delta_A \cdot \rho_{x_1=1|A} + \delta_B \cdot \rho_{x_1=1|B} \\ & \text{und } p(x_1 = 0) = 1 - p(x_1 = 1). \end{aligned} \tag{3.18}$$

beschrieben werden. Die latenten Zustände definieren also etwas Ähnliches wie eine Latent Class-Struktur. Die Antwortwahrscheinlichkeit wird durch die Summe des Produktes eines „Klassengrößenparameters“ δ und einer bedingten Antwortwahrscheinlichkeit ρ festgelegt.

> „In contrast to ordinary latent class analysis, a person does not have to stay in a particular class, but moves with probability $\tau_{A,1,B,1}$ from class A to class B, and with probability $\tau_{B,1,A,1}$ in the opposite direction.“ (Rost, 2002)

Die Wahrscheinlichkeit, in einer Klasse zu bleiben, wird demnach durch die Komplementärwahrscheinlichkeiten dazu festgelegt. Die Antwortwahrscheinlichkeit zum zweiten Zeit-

punkt ist dann

$$\begin{aligned} t = 2 : p(x_2 = 1) = \delta_A \cdot (\tau_{A,1,A,2} \cdot \rho_{x_2=1|A} + \tau_{A,1,B,2} \cdot \rho_{x_2=1|B}) \\ + \delta_B \cdot (\tau_{B,1,A,2} \cdot \rho_{x_2=1|A} + \tau_{B,1,B,2} \cdot \rho_{x_2=1|B}) \\ \text{und } p(x_2 = 0) = 1 - p(x_2 = 1). \end{aligned} \tag{3.19}$$

Die Antwortwahrscheinlichkeit in dem zweiten Glied dieser Markov-Kette hängt also von vier Übergangswahrscheinlichkeiten τ ab, von vier additiven Elementen, die sich auf die Kombinationen der zwei Zustände *AA*, *AB*, *BA*, und *BB* beziehen. Die Anzahl der latenten Klassen, in diesem Beispiel vier Klassen, steigt exponentiell mit der Anzahl der Zeitpunkte. Bei einem dritten Zeitpunkt wären $2^3 = 8$ Klassen notwendig usw. Durch die Tatsache, dass jede Person den Zustand von Zeitpunkt zu Zeitpunkt wechseln kann, muss eigentlich das gesamte Muster der Zustände über die Zeit als Kategorie der latenten Klassenvariable angesehen werden.

Gemischte latente Markov-Modelle, also „Mixed-Markov Models“, sind dann latente Markov-Modelle mit verschiedenen Modellparametern δ, ρ und τ in verschiedenen Subpopulationen. Diese „Mischvariablen“ müssen nicht unbedingt latent sein, sie können auch beobachtet, also manifest sein. Der Unterschied zwischen Mixed-Markov Modellen und latenten Markov-Modellen wird von Langeheine & Van de Pol (1990) so beschrieben:

> „Mixed Markov Modelle postulieren eine bestimmte Anzahl von Klassen, von denen jede durch einen eigenen Markov Prozess auf dem manifesten Niveau gekennzeichnet ist. Aber die Zugehörigkeit zu einer Klasse bleibt für ein Individuum konstant über die Zeit. In Latent Markov Modellen kann ein Individuum dagegen von Zeitpunkt zu Zeitpunkt von einer in eine andere Klasse wechseln. Diese latenten Übergangswahrscheinlichkeiten gelten allerdings für die gesamte Stichprobe.“ (S. 93)

Als Spezialfall eines allgemeinen Mixed-Markov Modells kann etwa das „Mover-Stayer Modell“ von Blumen, Kogan & McCarthy angesehen werden (1955). Ihnen fiel auf, dass das manifeste Markov-Modell zu viel Veränderung nach vielen Übergängen vorhersagt. Sie schlugen daher ein Modell vor, in dem es nur zwei Klassen gibt - die „mover“, die einer gewöhnlichen Markov-Kette entsprechen, und die „stayer“, die mit einer Wahrscheinlichkeit von 1 in derselben Kategorie bleiben. Dieses Modell wäre ein 2-Klassen Mixed-Markov Modell. Nach Schwierigkeiten bei der Parameterschätzung und einer Überschätzung der

„stayer“ wurden neue Parameterschätzer von Goodman (1961) und Morgan et al. (1983) vorgestellt. Langeheine & Van de Pol (1990) betrachteten das „Mover-Stayer Modell“ daher nur als Spezialfall des Mixed-Markov Modells.

Ein weiteres von Langeheine & Van de Pol (1990) betrachtetes Spezialmodell ist das „Black & White Modell“ von Converse (1964, 1970). Dieses Modell geht von der Annahme aus, dass sich eine Stichprobe in zwei extrem verschiedene Untergruppen aufteilen lässt, wobei der eine Teil ein perfekt stabiles Antwortverhalten über die Zeit zeigt, für den zweiten Teil hingegen nur der Zufall gilt. Es ist also notwendig, die δs und τs der zweiten Klasse auf Gleichwahrscheinlichkeit zu fixieren. Auch mit diesem Modell gab es Schwierigkeiten, hier lagen diese in der nicht zufrieden stellenden Modellgeltung. Converse (1964) schlug daher ein Erweiterung des Modells um eine dritte Klasse von Personen vor, die er als „true changers“, also nicht zufällige Wechsler von einer Kategorie zur nächsten bezeichnet. Diese Annahme führt wiederum zu den latenten Markov Modellen (Langeheine & Van de Pol, 1990). Van de Pol, Langeheine und de Jong (1996) entwickelten eine eigene Software für latente Markov-Modelle.

Latente Markov-Modelle könnten natürlich auch als eigener Typ von Item-Response-Modellen angesehen werden. Die Zeitpunkte der Markov-Modelle sind dann die aufeinander folgenden Antworten in einem Test, die Parameter ρ die bedingten Itemwahrscheinlichkeiten. $\rho_{x_i|A}$ ist die Wahrscheinlichkeit, das Item i zu lösen, wenn die Person sich im Zustand A befindet, $\rho_{x_1|B}$ die selbe Wahrscheinlichkeit, wenn sich die Person in Zustand B befindet. Die Zustände im Markov-Modell definieren die latente diskrete Variable. Die dynamischen Komponenten während der Testung liegen auf der Hand.

> „Different from most IRT models, it is *not* assumed that the latent variable stays constant during test administration. In the contrary, the latent Markov model parameterizes the *change* of the latent state during test performance. Latent Markov models as IRT models focus on latent change as a *qualitative* event, i.e. moving from one state to another.“ (Rost, 2002)

Der Parameter $\tau_{B,i-1,A,i}$ legt die Wahrscheinlichkeit fest, sich in Zustand A zu befinden während man Item i löst, nachdem man zum Zeitpunkt der Bearbeitung von Item $i-1$ in Zustand B war. Jede Person kann nach jedem Item ihren Zustand ändern. Nach (3.18) hat jedes Item zwei Schwierigkeitsparameter, nämlich die Lösungswahrscheinlichkeiten für die zwei Zustände. Wie bereits oben erwähnt, hängt die latente Variable für ein Item i vom ganzen vorangegangenen *Zustandsmuster* ab, und nicht nur von ausschließlich Zustand A oder B

(siehe (3.19)). Lokale stochastische Unabhängigkeit ist also *nicht* gegeben. Inhaltlich interpretiert könnte ein Wechsel der Zustände nach bestimmten Items z.B. ein Dazulerneffekt, ein Sinken der Konzentrationsfähigkeit oder ein Wechsel von bestimmten kognitiven Strategien sein.

Latente Markov-Modelle könnten auch für „Mastery/Non-Mastery"-Lernmodelle (nach Macready und Dayton, 1980) herangezogen werden.

3.3.2 Modelle aus der mathematischen Lerntheorie

Die zweite Art dynamischer Lernmodelle entwickelte sich aus der Verallgemeinerung von mathematischen Lernmodellen.

Zwei Modelle wurden im Rahmen der mathematischen Lerntheorie besonders bekannt.

1. Das *Modell von Verhelst & Glas (1993)*. Ausgehend von einer Idee von Fischer (u.a. 1983) wird ein Item als eine Sammlung von virtuellen Items angesehen, von denen jeder Versuchsperson eines aufgrund der Antworten auf die vorangegangenen Items vorgelegt werden soll. In diesem Modell wird das Rasch-Modell kombiniert mit dem missing-data Konzept und mit linearen Restriktionen der Parameter, sodass eigentlich ein LLTM mit inkomplettem Design entsteht. Der entstehende Transfer oder Lerneffekt hängt nicht von der ursprünglichen Fähigkeit ξ der Person ab. Jede Veränderung der Personenfähigkeit kann in eine Veränderung der Itemschwierigkeit umgewandelt werden. Die Itemschwierigkeit hängt somit von einem intrinsischen Parameter und einer dynamischen Komponente ab, die sowohl von der Reihenfolge der Items als auch von der spezifischen Lerneffektanfälligkeit des Items abhängt.

2. Das *Modell von Kempf (1974)*. In diesem Fall hängen die Parameter von den partiellen Antwortmustern beispielsweise bis zum Item I_{i-1} ab und beeinflussen die Lösungswahrscheinlichkeit von Item I_i. Dieses Modell wird ab Abschnitt 4 genauer behandelt.

Anfang der 1960er stellte die mathematische Lerntheorie ein wichtiges Forschungsgebiet der Psychologie dar und lieferte einen Ausgangspunkt für formale Lernmodelle (siehe Verhelst & Glas, 1995). Zur Verdeutlichung der Theorie stelle man sich ein klassisches T-Labyrinth Lernexperiment vor: Ein Tier wird in ein T-Labyrinth gesetzt und muss sich zwischen rechtem und linkem Gang entscheiden. Wenn es den einen Gang wählt, bekommt es Futter als Belohnung, beim anderen Gang nichts.

> „In a simple learning model, it is assumed that (a) learning (i.e.m a change in the tendency to choose the alley which yields the food reinforcer) occurs only on reinforced trials; (b) the 'inherent' difficulty of the situation is constant, and (c) there are no initial differences between the animals in the tendency to choose the reinforced alley.“ (Verhelst & Glas, 1995, S.198)

Dieses Experiment ist rein subjektkontrolliert, der Ausgang hängt nur vom Verhalten des Versuchstiers ab und wird nicht vom Versuchsleiter beeinflusst. Wenn die Durchgänge mit Testitems gleichgesetzt werden, dann wäre $\sigma_i = \sigma$, da die Itemschwierigkeit konstant ist, genauso wie die anfängliche Fähigkeit des Versuchsobjektes ξ. Diese Annahme der Invariabilität der Schwierigkeits- und Fähigkeitsparameter ist typisch für die Lernmodelle, die zwischen 1955 und 1970 entwickelt wurden. Sie muss jedoch auch gleichzeitig bedeuten, dass Experimente nur unter konstanten Bedingungen durchgeführt werden können. Diese Homogenitätsannahme kam durch einen Mangel an Werkzeugen zum Erfassen individueller Unterschiede zu Stande. Formal kann das Modell durch das „Ein-Operator Beta Modell“ von Luce (1959) ausgedrückt werden. Die Wahrscheinlichkeit eines Erfolges in Durchgang i nach j Erfolgen vorangegangener Durchgänge ist gegeben durch

$$P(X_i = 1|\nu, R_i = j) = \frac{\nu\alpha^j}{1+\nu\alpha^j}, \tag{3.20}$$

wobei $\nu = exp(\xi - \sigma)$ und $\alpha = exp(\delta)$ ist. Wenn Lernen nach einem unverstärkten Durchgang auftritt, so wird das ausgedrückt durch

$$P(X_i = 1|\nu, R_j = j) = \frac{\nu\alpha_1^j\alpha_2^{i-j-1}}{1+\nu\alpha_1^j\alpha_2^{i-j-1}}, \tag{3.21}$$

mit $\alpha_1 = exp(\delta)$ und $\alpha_2 = exp(\varepsilon)$, das wiederum dem „Zwei-Operatoren Modell“ von Luce (1959) entspricht. Als Beispiel kann die logistische Variante des „Ein-Durchgang Perseverationsmodells“ von Sternberg (1959) dienen. Im oben genannten T-Labyrinth-Experiment wurde eine höhere Autokorrelation im Antwortmuster Xzwischen den Durchgängen gefunden, als erwartet. Das läßt darauf schließen, dass vorangegangene Antworten oder Verhaltensweisen tendenziell wiederholt werden. Im Modell von Sternberg, das sich auf diese Annahme stützt, wird die Wahl einer nicht-verstärkten Antwort bzw. Verhaltensweise als Erfolg definiert.

$$p_i = (1-b)a^{i-1}p_{i-1} + bX_{i-1}, \text{ für } i \geq 2, 0 < a, b < 1, \tag{3.22}$$

$p_i = P(X_i = 1)$ und a ist ein Parameter für die Lernrate und b ein Perseverationsparameter, der die Tendenz zur Wiederholung der vorherigen Antwort angibt. Nach Sternberg (1963, in Verhelst & Glas, 1995) sieht die logistische Entsprechung des Modells so aus:

$$logit(p_i) = \xi + (i-1)\psi + \delta X_{i-1}, \text{ für } i \geq 2, \tag{3.23}$$

wobei $\xi = logit(p_1)$ als konstant behandelt wird. ψ ist hierbei ein Parameter für die Lernrate, δ ein Perseverationsparameter. Das logistische Modell ist flexibler als (3.22), aufgrund der Restriktionen, denen der Perseverationsparameter b unterliegt. Alternierende, also nicht perseverierende Antworten müssten in dem Fall durch ein anderes Modell ausgedrückt werden. Ein positives δ im logistischen Modell (3.23) hingegen bedeutet eine Tendenz zur Perseveration, ein negatives δ eine Tendenz, verschiedene Alternativen zu wählen.

Das logistische Modell verletzt jedoch die Forderung nach lokaler stochastischer Unabhängigkeit. Angenommen ein Einstellungsfragebogen soll mittels Rasch-Modell überprüft werden und man hat den Verdacht, dass eine Tendenz zu wechselnden Antworten besteht, kann man (3.23) anpassen. Man setzt $\psi = 0$ und lässt Variationen in den Leichtigkeitsparametern β_i und der latenten Variable ξ zu. Man nimmt $2k-1$ virtuelle Items an, $(i,0),(i,1)$ für $i > 1$ und $(1,1) \equiv (1,0)$. Diese Itempaare sind geordnet, der jeweilige zweite Teil der Paare entspricht der vorhergehenden Antwort.

Die lokale stochastische Unabhängigkeit kann mittels Likelihood-Ratio Test überprüft werden. Man vergleicht zu diesem Zweck das Modell mit einem restringierten Modell, bei dem $\delta = 0$ gesetzt wurde und somit genau dem Rasch-Modell entspricht.

3.3.2.1 Das Modell von Verhelst & Glas (1993)

Die Kontrolle über die Veränderung von Verhalten hängt in der mathematischen Lerntheorie von zwei Klassen von Ereignissen ab,

> „one is the behavior of the responding subject itself; the other comprises all events that occur independently of the subject's behavior, but which are assumed to change that behavior. Models that only allow for the former class are called 'subject controlled'; if only external control is allowed, the model is 'experimenter controlled'; and models where both kinds of control are allowed are labelled 'mixed models'. . . . In the sequel ist will be assumed, that all controlling events can be binary coded, that the subject control can be modelled through the

> correctness of the responses on past items, and that experimenter control expresses iteself at the level of the item.“ (Verhelst & Glas, 1995, S. 190)

Für das Modell sei nun X der Vektor der Antwortvariablen (0 = nicht korrekt, 1 = korrekt) und Z der binäre Vektor, der ausdrückt, ob eine Verstärkung nach dem Item stattgefunden hat oder nicht. Verstärkung bedeutet, dass die Person nach der Bearbeitung eines Items über die richtige Lösung informiert wird. Z sei unabhängig von X.

Der partielle Antwortvektor $X^i(i > 1)$ wird definiert als

$$X^i = (X_1, \ldots, X_{i-1}), \tag{3.24}$$

der partielle Verstärkungsvektor $Z^i(i > 1)$ als

$$Z^i = (Z_1, \ldots, Z_{i-1}). \tag{3.25}$$

Die allgemeinste Form des Modells von Verhelst & Glas wird folgendermaßen angeschrieben:

$$P(X_i = 1|\xi, x^i, z^i) = \frac{exp[\xi - \sigma_i + f_i(x^i) + g_i(z^i)]}{1 + exp[\xi - \sigma_i + f_i(x^i) + g_i(z^i)]}. \tag{3.26}$$

ξ ist die latente Variable, σ_i der Schwierigkeitsparameter von Item I_i, x^i und z^i sind die Realisationen von X^i und Z^i, $f_i(.)$ und $g_i(.)$ reellwertige Funktionen. Da diese Funktionen diskret und finit sind, kann man deren Werte bereits als Parameter ansehen. Dieses Modell stellt auch gleichzeitig die Verallgemeinerung zu (3.23) mit $\beta_i = 0, g_i(Z^i) = (i-1)\psi$ und $f_i(X^i) = \delta X_{i-1}$ dar.

Das generalisierte Modell kann jedoch nicht identifiziert werden, da die Anzahl der Parameter die Anzahl der möglichen Antwortmuster bei weitem übersteigt. Daher müssen dem verallgemeinerten Modell Restriktionen auferlegt werden. Eine häufige Restriktion der mathematischen Lerntheorie wird auch hier eingesetzt, die Funktionen f_i und g_i sollen in ihren Werten symmetrisch sein. Das führt zu Modellen mit vertauschbaren Operatoren. Da die Werte dichotom sind, bedeutet das auch, dass der Gültigkeitsbereich von f_i und g_i auf die Summe der Elemente der Vektoren x^i und z^i restringiert wird. Die Variablen R_i und S_i werden definiert als

$$R_i = \begin{cases} \sum_{j=1}^{i-1} X_j, & (i > 1), \\ 0, & (i = 1), \end{cases} \tag{3.27}$$

und

$$S_i = \begin{cases} \sum_{j=1}^{i-1} Z_j, & (i > 1), \\ 0, & (i = 1), \end{cases} \tag{3.28}$$

mit den Realisationen r_i und s_i, und der bereits erwähnten Annahme der Symmetrie der Funktionen g_i und f_i, was schließlich zu dem Modell

$$P(X_i = 1|\xi, r_i, s_i) = \frac{exp[\xi - \sigma_i + \delta_i(r_i) + \gamma_i(s_i)]}{1 + exp[\xi - \sigma_i + \delta_i(r_i) + \gamma_i(s_i)]} \tag{3.29}$$

führt. $\delta_i(0)$ und $\psi_i(0)$ sind definiert als 0 für alle i. Wenn alle δ und ψ gleich 0 sind, heißt das, dass überhaupt kein Transfer stattfindet und das Modell mit dem herkömmlichen Rasch-Modell gleichzusetzen ist. Wenn alle δ gleich 0 sind und wenigstens ein ψ nicht, resultiert daraus ein versuchsleiterkontrolliertes Modell. Wenn alle ψ gleich 0 sind und wenigstens ein δ nicht, ist das Modell subjektkontrolliert, in allen anderen Fällen gemischt. Da in diesem symmetrischen Modell jedoch kein Vergessen auftreten kann, ist diese Herangehensweise zwar elegant, aber eher unrealistisch, daher sollte die Forderung nach Symmetrie zumindest teilweise fallen gelassen werden. Mithilfe des missing-data Konzeptes können die vorangegangenen Modelle an das herkömmliche Rasch-Modell angepasst werden.

Angenommen es gibt ein reales Item I_i, das mit einer Anzahl von virtuellen Items $(i, j), j = 0, \ldots, i-1$ zusammenhängt. Das virtuelle Item (i, j) wird nun allen Versuchspersonen vorgelegt, die genau j korrekte Antworten auf die $i-1$ vorangegangenen realen Items gegeben haben. Das Antwortmuster X hängt mit einem Designvektor $D(X)$ zusammen. Dessen Elemente $D(X)_{ij}$ für $i = 1, \ldots, k$ und $j = 0, \ldots, i-1$ sind definiert durch

$$D(X)_{ij} = \begin{cases} 1 & \text{wenn } R_i = j, \\ 0 & \text{andernfalls.} \end{cases} \tag{3.30}$$

Aus dem Antwortmuster X wird das Antwortmuster $Y(X)$ mit den Elementen $Y(X)_{ij}$ für $i = 1, \ldots, k$ und $j = 0, \ldots, i-1$, die definiert sind durch

$$Y(X)_{ij} = \begin{cases} 1 & \text{wenn } D(X)_{ij} = 1 \text{ und } X_i = 1, \\ 0 & \text{wenn } D(X)_{ij} = 1 \text{ und } X_i = 0, \\ c & \text{wenn } D(X)_{ij} = 0, \end{cases} \tag{3.31}$$

wobei c eine beliebige Konstante $\neq 0$ oder 1 ist. $Y(X)$ und $D(X)$ sind eindeutige Transformationen von X. In diesem Modell kann die Wahrscheinlichkeit eines beobachteten Antwortmusters x durch folgende Gleichung beschrieben werden

$$\begin{aligned} P(x|\xi;\varepsilon) &= P(x_1|\xi;\varepsilon)\prod_{i>1}P(x_i|x^i,\xi;\varepsilon) \\ &= \frac{exp[\sum_{i=1}\sum_{j=0}^{i-1}y(x)_{ij}d(x)_{ij}(\xi+\varepsilon_{ij})]}{\prod_{i=1}\prod_{j=0}^{i-1}[1+exp(\xi+\varepsilon_{ij})]^{d(x)_{ij}}}, \end{aligned} \tag{3.32}$$

wobei ε einen $k(k+1)/2$ großen Vektor mit den Elementen ε_{ij} für $i = 1,\ldots,k$ und $j = 0,\ldots,i-1$ darstellt. Die Elemente ε_{ij} wiederum bestehen aus $\delta_i(j) - \sigma_i$. Die Einzelantwort x_i wird ersetzt durch $y(x)_{ij}$, die Variable $d(x)_{ij}$, die mit den Itemvorgaben zu tun hat, wird immer durch die vorherigen Antworten x^i bestimmt.

Die Modelldarstellung (3.32) ist äquivalent zum ursprünglichen Modell (3.26), sie ist eine Verallgemeinerung der Likelihood-Funktion des Rasch-Modells mit unvollständigen Designs. Für versuchsleiterkontrollierte oder gemischte Modelle ist eine ähnliche Verallgemeinerung denkbar und möglich, nur wäre beim versuchsleiterkontrollierten Modell der Designvektor von Z abhängig anstatt von X, und die Probleme bei der Parameterschätzung wären ein wenig anders.

Das erste Problem bei der Parameterschätzung des subjektkontrollierten Modells ist die Identifizierbarkeit. Da $\xi - \sigma_i + \delta_i(j)$ den selben Wert hat wie $\xi^* - \sigma_i^* + \delta_i^*(j)$, mit $\sigma_i^* = \sigma_i - c - d_i$, $\delta_i^*(j) = \delta_i(j) - d_i$ und $\xi^* = \xi - c$ für ein beliebiges c und d_i für $i = 1,\ldots,k$, müssen den Parametern $k+1$ Restriktionen auferlegt werden, um das Modell identifizierbar zu machen.

Diese Probleme beim so genannten „multi-stage-testing“ wurden von Glas (1988) untersucht. Multi-stage-testing bezeichnet eine Festlegung der Reihenfolge der Tests durch die Reihenfolge der Testscores der Versuchspersonen (Verhelst & Glas, S. 194). Das subjektkontrollierte Modell und seine virtuellen Items können als multi-stage-testing Design mit jeweils nur einem Item pro Test angesehen werden, wobei der nächste Test (also Item) vom Summenscore R_i abhängt, der bei den vorangegangenen Tests erreicht wurde.

> „The main result of Glas is the conclusion that, in the case of a multi-stage design, the CML estimation equations have no unique solution, while MML generally does yield consistent estimates.“ (Verhelst & Glas, 1995, S. 194f)

In (3.32) erweist sich die CML-Schätzung als ausgesprochen schwierig. Das Test-Design und der Summenscore sind erschöpfende Statistiken für ξ. Die CML-Methode schätzt die Likelihood-Funktion abhängig von sowohl dem Summenscore, als auch dem Design. Der Antwortvektor $Y(X)$, bedingt durch Summenscore und Design, ist jedoch vollständig bestimmt, was dazu führt, dass die Likelihood gleich 1 ist und somit nicht benutzt werden kann.

Bei versuchsleiterkontrollierten Modellen ist dies völlig anders. Der Designvektor ist unabhängig von den Antworten der Versuchspersonen, es gibt einen Summenscore r und k virtuelle Items, somit $\binom{k}{r}$ verschiedene Antwortmuster und die CML-Methode kann angewendet werden.

Im Fall der subjektkontrollierten Modelle jedoch muss die MML-Schätzmethode herangezogen werden. ξ wird als Zufallsvariable mit einer Wahrscheinlichkeitsdichtefunktion $g(\xi;\varphi)$ angesehen, wobei φ einen Parametervektor darstellt. Die Wahrscheinlichkeit eines beobachteten Antwortmusters A, mit der Auftretenshäufigkeit n_A, ist festgelegt durch

$$P(x;\varepsilon,\varphi) = \int_{-\infty}^{+\infty} P(x|\xi;\varepsilon)g(\xi;\varphi)d\xi. \tag{3.33}$$

Für alle möglichen Antwortmuster A unterliegt die Anzahl der n_A einer parametrischen multinomialen Verteilung mit Index $n = \sum_A n_A$ und den Parametern $P(x;\varepsilon,\varphi)$ für alle binären k-Vektoren A. Der Logarithmus der Likelihood-Funktion ist

$$\ln L(\varepsilon,\varphi;\{A\}) = \sum_A n_A \ln P(A) = \sum_v \ln \int P(A_v|\xi;\varepsilon)g(\xi;\varphi)d\xi, \tag{3.34}$$

wobei $\{A\}$ die Daten bezeichnet und A_v das Antwortmuster der Person S_v. Die simultane Maximierung dieser Funktion unter der Berücksichtigung von φ und ε liefert die MML-Schätzer der Parameter.

Für die MML-Schätzung ist es notwendig, eine Annahme über die Verteilung der latenten Variable in der Population (die Normalverteilung ist hierbei am häufigsten) und über die Ziehung der Stichprobe zu treffen. Prinzipiell könnte die Verteilung jedoch auch aus den Daten geschätzt werden, diese Herangehensweise nennt man auch nicht parametrische-MML, oder semi-parametrische-MML. Ohne Restriktionen der Dichtefunktion gibt es $2k-1$ freie Parameter, d.h. eine notwendige Bedingung für eine korrekte und eindeutige MML-Schätzung ist $k \geq 3$. Mithilfe der Modellgleichungen (3.32) und (3.34) und einigen zusätzlichen linearen Restriktionen für die Parameter ξ können mehrere interessante Spezialfälle untersucht wer-

den. Angenommen, es gibt einen m-dimensionalen Vektor η, wobei $m < k(k+1)/2$. Dann sei $\eta = B\varepsilon$ und B eine konstante Matrix mit dem Rang m. Die Dimension von η soll kleiner sein als die Anzahl der virtuellen Items. Daraus resultiert ein LLTM. Man kann daraus z.B. die folgenden Modelle mit der Normierung $\delta_0 = 0$ identifizieren (vlg. Verhelst & Glas, 1995).

- Das Ausmaß an Lernen hängt davon ab, wie viele vorangegangene Items man vorher gelöst hat. Das bedingt die Restriktion

$$\varepsilon_{ij} = \delta_j - \sigma_i. \tag{3.35}$$

- Durch weitere Restriktionen kann man weiters annehmen, dass dieses Ausmaß nach jedem erfolgreichen Bearbeiten konstant ist.

$$\varepsilon_{ij} = j\delta - \sigma_i \tag{3.36}$$

  Das Modell (3.20) ist ein Spezialfall dieses Modells.

- Man kann auch ein Zwei-Operatoren-Modell formulieren. Die Veränderung der latenten Fähigkeit kann auch noch abhängig von vorangegangenen Fehlern sein. Eine stark verallgemeinerte Version davon ist dann

$$\varepsilon_{ij} = \delta_j - \sigma_i + \rho_{i-j-1}, \tag{3.37}$$

  wobei $\rho_0 = 0$ ist. Das Modell (3.37) ist dann eine Reparametrisierung von (3.32).

- Man kann (3.37) noch weiter spezifizieren, wenn man annimmt, dass das Ausmaß des Lernens unabhängig vom Item ist, mit $\delta_j = l\delta$ und $\rho_j = j\rho$ für $j \neq 0$.

- Für letzteres Modell kann man weiters annehmen, dass eine falsche Antwort genau den gegenteiligen Effekt wie eine richtige Antwort hat, wenn man also $\delta = -\rho$ setzt.

- Ein Modell, in dem das Ausmaß an Lernen, unabhängig von den vorhergehenden richtig beantworteten Items, immer das gleiche ist, also $\delta = \rho$, ist jedoch nicht identifizierbar, wenn jeder Person jedes Item in der gleichen Reihenfolge vorgegeben wird. Man könnte aber die Verstärkung der Personen variabel gestalten und den Test in z.B. 2 verschiedenen Reihenfolgen zwei gleichwertigen Stichproben vorgeben.

- In den vorangegangenen Modellen ist die Fähigkeit zu lernen jeweils unabhängig von der Stelle eines Items in einer Itemfolge. In manchen Fällen ist dies aber nicht realistisch, z.B. wenn Lerneffekte von Gewöhnungs- oder Ermüdungseffekten gemindert werden. Daher kann man eine Grenze für das Ausmaß des Lernens konstruieren, in dem man dem Modell (3.36) die zusätzliche Restriktion

$$\varepsilon_{ij} = i^{-c} j\delta - \sigma_i, \text{ für } c > 2 \tag{3.38}$$

  auferlegt.

4 Das dynamische Testmodell von Kempf

Ein Überblick über verschiedene dynamische Test- und Lernmodelle wurde bereits in den vergangenen Abschnitten gegeben. In Abschnitt 3.3.2 wurden die Modelle aus der mathematischen Lerntheorie erwähnt, zu denen das dynamische Testmodell von Kempf für dichotome Items (1974) zählt. Dieses soll nun in seiner Modelldarstellung und Anwendung genauer vorgestellt werden.

In Kapitel 2 wurde erwähnt, dass es besondere Probleme bereitete, die Lernkomponente in Kurzzeitlerntests testtheoretisch zu erfassen. Das dynamische Testmodell von Kempf bietet eine weitere Möglichkeit, Lernen während der Bearbeitung eines Tests (insbesondere eines psychologischen Leistungs- oder Intelligenztests) zu quantifizieren. Dies wird durch die Einführung eines so genannten „Transferparameters“ erreicht. Auch die Verstärkung oder Abschwächung von Einstellungen in Einstellungs- oder Befindlichkeitsfragebögen könnte damit gemessen werden.

Kempf (1974) fand den Ansatzpunkt für die Anwendung seines dynamischen Modells mit separierbaren Parametern in der Aggressionsforschung. Eine Person, die ihre Aggression ausgedrückt hat, wird in Zukunft eher weniger Aggression verspüren. Diese Annahme kann nur getroffen werden, wenn das Prinzip der lokalen stochastischen Unabhängigkeit der Items (in diesem Fall Aggressionsprovokationen) fallengelassen wird, da die Aggression der Gegenwart offenbar von vorangegangenen Aggressionsäußerungen abhängt. Anstelle der lokalen stochastischen Unabhängigkeit tritt das Prinzip der lokalen seriellen Abhängigkeit, das weit weniger restriktiv ist.

Der Itemscore (a_{vi}) kann als formaler Ausdruck der dynamischen Komponente statt als

$$p\{(a_{vi})\} = \prod_{i=1}^{k} p\{a_{vi}\} \tag{4.1}$$

als

$$p\{(a_{vi})\} = \prod_{i=1}^{k} p\{a_{vi}|s_{vi}\} \tag{4.2}$$

angeschrieben werden, wobei s_{vi} den partiellen Antwortvektor $(a_{v1},\ldots,a_{vi-1})$ bezeichnet und anstelle der Itemcharakteristika $f_i(\xi)$ die bedingten Itemcharakteristiken

$$f_{i.s_{vi}}(\xi) = p\{a_{vi} = 1|(a_{v1},\ldots,a_{vi-1}) = s_{vi}\} \tag{4.3}$$

verwendet werden. Die bedingte Verteilung des Itemscores a_{vi} wird in Abhängigkeit der Antworten auf vorangegangene Items so definiert

$$p\{a_{vi}|s_{vi}\} = [f_{i.s_{vi}}(\xi_v)]^{a_{vi}}\,[1 - f_{i.s_{vi}}(\xi_v)]^{1-a_{vi}}. \tag{4.4}$$

Dabei soll jede einzelne Funktion mit der latenten Dimension ξ streng monoton wachsen. Lokale stochastische Unabhängigkeit tritt also nur dann auf, wenn alle $f_{i.s_{vi}}(\xi)$ für ein festes i gleich sind.

4.1 Modelldarstellung

Der spezielle Modellansatz des Kempf-Modells besagt, dass die bedingte Itemcharakteristikfunktion $f_{i.s_{vi}}(\xi)$ von der Anzahl der korrekt beantworteten vorangegangenen Items abhängt:

$$r_{vi} = \begin{cases} 0 & \text{für } i = 1 \\ \sum_{j=1}^{i-1} a_{vj} & \text{für } i = 2,3,\ldots,k. \end{cases} \tag{4.5}$$

Die Itemcharakteristika können für alle partiellen Antwortvektoren s_{vi} mit dem gleichen Summenscore r_{vi} gleichgesetzt werden. Alle partiellen Antwortvektoren mit gleichem Summenscore haben also äquivalenten Einfluss auf die Wahrscheinlichkeit der richtigen Beantwortung eines Items i

$$f_{i.s_{vi}}(\xi) = f_{i.r_{vi}}(\xi). \tag{4.6}$$

Die Modellstruktur lehnt sich an die BTL-Darstellungsform (siehe Bradley & Terry, 1952, Luce, 1959) des Rasch-Modells an (siehe (4.10)), indem dieser ursprünglichen Form noch der Lern- oder Transferparameter ψ hinzugefügt wird. Die Modellgleichung sieht dann so

aus

$$f_{i.r_{vi}}(\xi) = \frac{\xi_v + \psi_{r_{vi}}}{\xi_v + \sigma_i}, \tag{4.7}$$

mit der Nebenbedingung, dass $\psi_{r_{vi}} < \sigma_i$ sei. σ_i bezeichnet den Itemschwierigkeitsparameter. Der Transferparameter $\psi_{r_{vi}}$ ist abhängig von den r_{vi} vorangegangenen korrekt beantworteten Items. Es ist jedoch dabei nicht wichtig, *welche* vorangegangenen Items gelöst wurden, sondern nur *wie viele*. Er gibt damit also nicht an, bei welchem bestimmten Item ein Lerneffekt wie groß ist, sondern er bedeutet „wenn $i-1$ Items in der Vergangenheit gelöst, bzw. mit 1 beantwortet wurden, ist der Transfereffekt so und so groß“. Er wirkt sich auf die Lösungswahrscheinlichkeit eines Items aus, denn je größer der Transferparameter, umso größer ist die bedingte Lösungswahrscheinlichkeit eines Items i. Kempf (1974, S. 38) beschreibt die Interpretation der Transferparameter auf folgende Weise.

- „Steht der numerische Wert der Transfer-Parameter $\psi_{r_{vi}}$ in einem monoton wachsenden Zusammenhang mit r_{vi}, so kann der Transfer daher als 'Lerngewinn' interpretiert werden.
- Ist die Abhängigkeit der Transfer-Parameter $\psi_{r_{vi}}$ von r_{vi} dagegen monoton fallend, so ist der Transfer als 'Reaktionshemmung' zu interpretieren. (In diesem Sinne kann z.B. die Katharsis als eine Reaktionshemmung für Aggression verstanden werden.)
- Ist der Zusammenhang zwischen $\psi_{r_{vi}}$ und r_{vi} nicht monoton, so sprechen wir von einer 'Fluktuation', welche durch gleichzeitig stattfindende Lern- und Hemmungsprozesse erklärt werden kann, die mit unterschiedlicher Beschleunigung ablaufen.“

Die Intervallskaleneigenschaft der Modellparameter verlangt zur eindeutigen Festlegung noch eine Skalennormierung, nämlich

$$\min(\psi_{r_{vi}}) = 0 \text{ für } r_{vi} = 0, \ldots, k-1 \tag{4.8}$$

und

$$\prod_{i=1}^{k} \sigma_i = 1. \tag{4.9}$$

Die Verwandtschaft mit dem Rasch-Modell sieht man durch folgende Restriktion. Wenn alle $\psi_{r_{vi}} = 0$ gesetzt werden, ist das dynamische Testmodell äquivalent dem herkömmlichen

Rasch-Modell in seiner BTL-Modell-Darstellung

$$\frac{\xi_v}{\xi_v + \sigma_i}. \tag{4.10}$$

Das Modell von Kempf stellt also eine Generalisierung des Rasch-Modells dar. Auch im dynamischen Testmodell

- „ist die Anzahl der gelösten Aufgaben a_{vo} eine erschöpfende Statistik für den Personenparameter v,
- können Vergleiche von Personen (oder Items) in spezifisch objektiver Weise ausgeführt werden und
- existieren CML-Schätzfunktionen für die Strukturparameter.“ (Kempf, 1974, S. 38)

4.2 Schätzung der Item- und Transferparameter des Kempf-Modells

Die Separierbarkeit der Modellparameter ist also gegeben, sodass die Item- und Transferparameter mit Hilfe einer bedingten Maximum-Likelihoodmethode aus der Likelihoodfunktion

$$L = p\{((a_{vi}))|(a_{vo})\} = \prod_{v=1}^{n} p\{(a_{vi})|a_{vo}\} \tag{4.11}$$

geschätzt werden können, wobei $((a_{vi}))$ die Antwortmatrix von n Personen auf k Items und $(a_{vo}) = \sum_{i=1}^{k} a_{vi}$ den Rohscorevektor der Personen darstellt. Ein Rohscore von $a_{vo} = 0$ oder $a_{vo} = k$ ergibt eine Wahrscheinlichkeit von 1, fließt somit nicht in die bedingte Likelihood ein und liefert keine Information. Diese bedingte Likelihood der Antwortmatrix (a_{vi}) kann für die Antworten von n Personen mit $0 < a_{vo} < k$ Rohscores auch so angeschrieben werden

$$L = \prod_{v=1}^{n} p\{(a_{vi})|a_{vo}\} = \prod_{v=1}^{n} \frac{p\{(a_{vi})\}}{p\{a_{vo}\}}. \tag{4.12}$$

Wenn man nun (4.6) und (4.7) in (4.3) und (4.4) einsetzt, so ergibt sich daraus folgende

Likelihoodfunktion für $p\{a_{vi}\}$

$$\begin{aligned} L = p\{(a_{vi})\} &= \prod_{i=1}^{k} p\{a_{vi}|r_{vi}\} \\ &= \prod_{i=1}^{k} \frac{(\xi_v + \psi_{r_{vi}})^{a_{vi}} (\sigma_i - \psi_{r_{vi}})^{1-a_{vi}}}{\xi_v + \sigma_i} \\ &= \prod_{r=0}^{a_{vo}-1} (\xi_v + \psi_r) \prod_{i=1}^{k} \frac{(\sigma_i - \psi_{r_{vi}})^{1-a_{vi}}}{\xi_v + \sigma_i}. \end{aligned} \tag{4.13}$$

$p\{a_{vo}\}$ besteht aus der Summe aller Wahrscheinlichkeiten $p\{(a^*_{vi})\}$ aller möglichen Antwortvektoren (a^*_{vi}), die mit dem Rohscore a_{vo} kompatibel sind (s. Kempf & Hampapa, 1975, S.13)

$$\begin{aligned} p\{a_{vo}\} &= \sum_{(a^*_{vi}|a_{vo})} p\{(a^*_{vi})\} \\ &= \prod_{r=0}^{a_{vo}-1} (\xi_v + \psi_r) \sum_{(a^*_{vi}|a_{vo})} \prod_{i=1}^{k} \frac{(\sigma_i - \psi_{r^*_{vj}})^{1-a^*_{vi}}}{\xi_v + \sigma_i}, \end{aligned} \tag{4.14}$$

wobei $r^*_{vi} = \sum_{j=1}^{i-1} a^*_{vi}$ für $i = 2,3,\ldots,k$ und $r^*_{vi} = 0$ für $i = 1$ darstellen. Durch Einsetzen von (4.13) und(4.14) in (4.12) ergibt sich nun die bedingte Likelihoodfunktion von

$$L = \prod_{v=1}^{n} \frac{\prod_{i=1}^{k} (\sigma_i - \psi_{r_{vi}})^{1-a_{vi}}}{\sum_{(a^*_{vi})|a_{vo}} \prod_{i=1}^{k} (\sigma_i - \psi_{r^*_{vi}})^{1-a^*_{vi}}}. \tag{4.15}$$

Des weiteren bezeichnen Kempf & Hampapa (1975) n_{ri} als Anzahl der Personen, die eine falsche Antwort auf Item i nach $r_{vi} = r$ richtigen Antworten auf die vorangegangenen Items $j = 1,2,\ldots,i-1$ gegeben haben. Für $i = 1,2,\ldots,k$ und $r = 0,1,\ldots,i-1$ tritt der Ausdruck $(\sigma_i - \psi_r)$ n_{ri}-mal im Zähler von (4.15) auf. $N_{k,s}$ sei außerdem die Anzahl von Personen, die s falsche Antworten auf k Items gegeben haben, so dass $a_{vo} = k - s$ ist. Da

$$G(k,s) = \sum_{(a^*_{vi})|k-s} \prod_{i=1}^{k} (\sigma_i - \psi_{r^*_{vi}})^{1-a^*_{vi}} \tag{4.16}$$

N_{k-s}-mal im Nenner von (4.15) vorkommt, kann (4.15) zu

$$L = \frac{\prod_{i=1}^{k} \prod_{r=0}^{i-1} (\sigma_i - \psi_r)^{n_{ri}}}{\prod_{s=1}^{k-1} G(k,s)^{N_{k-s}}} \tag{4.17}$$

vereinfacht werden. Die Schätzgleichungen

$$\begin{aligned} &\sum_{r=0}^{\alpha-1} \frac{r_{r\alpha}}{\sigma_\alpha - \psi_r} - \sum_{s=1}^{k-1} N_{k-s} \frac{\partial G(k,s)/\partial \sigma_\alpha}{G(k,s)} = 0 \text{ für } \alpha = 1,\ldots,k \\ \text{und } &\sum_{i=\beta+1}^{k} \frac{n_{\beta i}}{\psi_\beta - \sigma_i} - \sum_{s=1}^{k-1} N_{k-s} \frac{\partial G(k,s)/\partial \psi_\beta}{G(k,s)} = 0 \text{ für } \beta = 0,\ldots,k-1 \end{aligned} \tag{4.18}$$

müssen unter der Nebenbedingung $\psi_\beta < \sigma_\alpha$ für alle α und β gelöst werden. Für die Lösung der Schätzgleichungen müssen noch die Eigenschaften der so genannten G-Funktionen (4.16) spezifiziert werden. $G(k,s)$ ist die Summe der Elemente von Produkten von s Faktoren $(\sigma_i - \psi_{r^*})$. σ_{i_j} sei der Itemparameter im j-ten Faktor, a_{vi} sei gleich 0 für die $j-1$ vorangegangenen Items und gleich 1 für $r_{vi_j} = i_j - j$ der Items bei $i < i_j$. Ein Itemparameter σ_i soll weiters auch nur einmal pro Produkt auftreten, die Produkte selbst werden über alle möglichen Kombinationen von Itemparametern summiert. Dann kann i_j nicht größer werden als $k-s+j$ und die G-Funktionen $G(k,s)$ können so angeschrieben werden

$$G(k,s) = \sum_{i_1=1}^{k-s+1} \sum_{i_2=i_1+1}^{k-s+2} \cdots \sum_{i_s=1_{s-1}+1}^{k} \prod_{j=1}^{s} (\sigma_{i_j} - \psi_{i_j-1}). \tag{4.19}$$

Kempf et al. (1975, S. 15) setzen (4.19) gleich mit

$$\hat{G}(k,s) = \sum_{m=0}^{s} \delta_m(k-s)\gamma_{s-m}(k) \cdot (-1)^m. \tag{4.20}$$

Diese Form beinhaltet die so genannten Delta- und Gamma-Funktionen. $\gamma_{s-m}(k)$ stellt wie im Rasch-Modell die Summe aller möglichen Produkte von $s-m$ Itemparametern $\sigma_1,\ldots,\sigma_k$ ohne Wiederholungen dar.

$$\gamma_{s-m}(k) = \begin{cases} \sum_{i_1=1}^{k-s+m+1} \sum_{i_2=i_1+1}^{k-s+m+2} \cdots \sum_{i_{s-m}=i_{s-m-1}+1}^{k} \prod_{t=1}^{s-m} \sigma_{i_t} \text{ für } m = 0,1,\ldots,s-1 \\ 1 \text{ für } m = s. \end{cases} \tag{4.21}$$

$\delta_m(k-s)$ kennzeichnet die Summe aller möglichen Produkte von m Transferparametern aus dem Set $\psi_0,\ldots,\psi_{k-s}$ mit Wiederholungen

$$\delta_m(k-s) = \begin{cases} 1 \text{ für } m = 0 \\ \sum_{j_1=0}^{k-s} \sum_{j_2=j_1}^{k-s} \cdots \sum_{j_m=j_{m-1}}^{k-s} \prod_{t=1}^{m} \psi_{j_t} \text{ für } m = 1,2,\ldots,s. \end{cases} \tag{4.22}$$

Die Delta-Funktionen können rekursiv über die Formel

$$\sum_{\eta=0}^{m} \psi_{k+1-s}^{m-\eta} \delta_\eta(k-s) = \delta_m(k+1-s) \tag{4.23}$$

gewonnen werden. Jede Delta-Funktion wird durch eine Delta-Funktion mit einem Parameter weniger erklärt. Die erste partielle Ableitung der Delta-Funktionen entspricht

$$\partial \delta_m(k-s)/\partial \psi_r = \sum_{\eta=0}^{m-1} \psi_r^\eta \, \delta_{m-\eta-1}(k-s). \tag{4.24}$$

Dies gilt für alle $r = 0,\ldots,k-s$ und $m > 0$. Für die Fälle $m = 0$ und $r > k-s$ hängen die Delta-Funktionen nicht von ψ_r ab und $\partial \delta_m(k-s)/\partial \psi_r = 0$. Nun kann mit Hilfe der Delta- und Gamma-Funktionen die erste partielle Ableitung der G-Funktionen gebildet werden

$$\partial G(k-s)/\partial \psi_r = \begin{cases} \sum_{m-1}^{s} \gamma_{s-m}(k)(\sum_{j=0}^{m-1} \psi_r^j \delta_{m-1-j}(k-s))(-1)^m \text{ für } r = 0,1,\ldots,k-s \\ 0 \text{ für } r > k-s, \end{cases} \tag{4.25}$$

und weil $\partial \gamma_{s-m}(k)/\partial \sigma_i = \gamma^{(i)}_{s-m-1}(k)$ für $m < s$ und $\partial \gamma_{s-m}(k)/\partial \sigma_i = 0$

$$\partial G(k-s)/\partial \sigma_i = \sum_{m=0}^{s-1} \delta_m(k-s)\gamma^{(i)}_{s-m-1}(k)(-1)^m. \tag{4.26}$$

Ähnlich wie bei der bedingten Maximum-Likelihoodschätzung im Rasch Modell beschreibt $\gamma^{(i)}_{s-m-1}$ die elementaren symmetrischen Funktionen $s-m-1$ter Ordnung der Parameter $\sigma_1,\ldots,\sigma_k$. Schließlich und endlich haben die bedingten Schätzgleichung die folgenden Formen

$$\begin{aligned} &\partial \ln(L)/\partial \sigma_\alpha = \\ &\sum_{r=0}^{\alpha-1} \frac{n_{r\alpha}}{\sigma_\alpha - \psi_r} - \sum_{s=1}^{k-1} N_{k-s} \frac{\sum_{m=0}^{s-1} \delta_m(k-s)\gamma^{(\alpha)}_{s-m-1}(k)(-1)^m}{\sum_{m=0}^{s} \delta_m(k-s)\delta_{s-m}(k)(-1)^m} = 0 \end{aligned} \tag{4.27}$$

für $\alpha = 1,\ldots,k$ und

$$\partial \ln(L)/\partial \psi_\beta = \sum_{i=\beta+1}^{k} \frac{n_{\beta i}}{\psi_\beta - \sigma_i} - \sum_{\substack{s=1 \\ s\leq k-\beta}}^{k-1} N_{k-s} \frac{\sum_{m=1}^{s} \gamma_{s-m}(k)(\sum_{j=0}^{m-1} \psi_\beta^j \delta_{m-1-j}(k-s))(-1)^m}{\sum_{m=0}^{s} \delta_m(k-s)\delta_{s-m}(k)(-1)^m} = 0 \tag{4.28}$$

für $\beta = 0,\ldots,c_{\max}$, wobei $c_{\max}$ den größten beobachteten Rohscore $a_{vo} < k$ bezeichnet.

Das Problem, das sich jedoch aus diesen Gleichungen ergibt, betrifft die Nebenbedingung $\psi_r < \sigma_i$ $(r = 0,\ldots,k; i = 1,\ldots,k)$. Kempf & Hampapa (1975) lösen das Problem, in dem sie lineare Parametertransformationen $\psi_r \rightarrow \psi_r^*$ und $\sigma_i \rightarrow \sigma_i^*$ durchführen, so dass gilt $0 < \psi_r^* < 1 \leq \sigma_i^*$ $r = 0,\ldots,k-1; i = 1,\ldots,k)$ und Hilfsparameter $\phi_r = \ln(\psi_r^*/(1-\psi_r^*))$ und $\eta_i = \ln(\sigma_i^*) - 1$ einführen. Die Lösungen der Schätzgleichungen (4.27) und (4.28) können dann mit

$$\psi_r^* = \exp(\phi_r)/(1+\exp(\phi_r)) \tag{4.29}$$

und

$$\sigma_i^* = 1 + \exp(\eta_i) \tag{4.30}$$

aus den Lösungen von

$$\partial ln(L)/\partial \eta_\alpha = \partial ln(L)/\partial \sigma_\alpha \; (\partial \sigma_\alpha/\partial \eta_\alpha) = 0 \tag{4.31}$$

und

$$\partial ln(L)/\partial \phi_\beta = \partial ln(L)/\partial \psi_\beta \; (\partial \psi_\beta/\partial \phi_\beta) = 0 \tag{4.32}$$

berechnet werden. Als Nebenbedingung für die Hilfsparameter wird

$$\text{MIN}(\psi_r^*) = 1 - \text{MAX}(\psi_r^*) = \text{MIN}(\sigma_i^*) - 1 \tag{4.33}$$

gesetzt.

Die Schätzung der Item- und Transferparameter erfolgt im Ganzen drei mal. Einmal für den Gesamtdatensatz und jeweils einmal für zwei Untergruppen, die für den Modellgeltungstest des Kempf-Modells notwendig sind (siehe Abschnitt 4.4.1). Die Schätzgleichungen werden mittels Gradientenmethode (Fischer & Formann, 1972) iterativ gelöst. Der Vorteil der Gradientenmethode liegt darin, dass nur die ersten partiellen Ableitungen der Likelihood-Funktion benötigt werden. Sie gelangt auch zum absoluten Maximum, egal welcher Startwert

bei der ersten Iteration festgelegt wurde. Es muss nur noch die Genauigkeit der Schätzungen überprüft werden. Dafür wird im Programm ein Genauigkeitstest durchgeführt. Er funktioniert derart, dass die Werte der G-Funktionen nicht nur durch (4.20), sondern auch durch das rekursive System

$$\hat{G}(j,1) = \sum_{i=1}^{j}(\sigma_i - \psi_{i-1}) \text{ für} j = 1,k$$
$$\hat{G}(j,j) = \prod_{i=1}^{j}(\sigma_i - \psi_0) \text{ für} j = 1,k \tag{4.34}$$
$$\text{und}$$
$$\hat{G}((j+1),s) = \hat{G}(j,s) + (\sigma_{j+1} - \psi_{j+1-s})\hat{G}(j,(j-1)) \text{ für} j = 2,k-1; s = 2,j$$

berechnet und $G(k-s)$ und $\hat{G}(k-s)$ mit einander verglichen werden. Im Programm wird das Verhältnis $G(k,s)/\hat{G}(k,s)$ ausgerechnet und das kleinste und größte davon ausgegeben, wenn es eine Abweichung $G(k,s)/\hat{G}(k,s) \neq 1$ gibt. Wenn die Berechnungen genau genug sind, sollten die beiden Verhältnisse nicht wesentlich voneinander abweichen. Ist die Ungenauigkeit zu groß, wird die Prozedur abgebrochen.

4.3 Schätzung der Personenparameter des Kempf-Modells

Kempf (1977) schlug selbst eine Möglichkeit vor, wie zusätzlich zu den Item- und Transferparametern auch die Fähigkeiten von Personen verglichen werden können. Die Likelihood der Datenmatrix $((a_{vi}))$ hängt von den Itemschwierigkeiten lediglich über die Itemrandmatrix $((n_{ri}))$ ab, so dass die bedingte Wahrscheinlichkeit

$$p\{((a_{vi}))|((n_{ri}))\} = \frac{\prod_{v=1}^{n}\prod_{r=0}^{a_{vo}-1}(\xi_v + \psi_r)}{\sum_{((a^*_{vi}))|((n_{ri}))}\prod_{v=1}^{n}\prod_{r=0}^{a^*_{vo}-1}(\xi_v + \psi_r)} \tag{4.35}$$

die Itemparameter nicht mehr beinhaltet. a_{vo} ist eine erschöpfende Statistik für den Personenparameter ξ_v, daher müssen die geschätzten Fähigkeitsparameter für Personen mit demselben Rohscore gleich sein. Setzt man die Parameterschätzer $\hat{\xi}_v = \hat{\xi}_u$ für $a_{vo} = u$ in die Gleichung

(4.35) ein, ergibt sich daraus

$$\frac{\prod_{u=1}^{k-1}\prod_{r=0}^{u-1}(\hat{\xi}_u+\psi_r)^{N_u}}{\sum_{((a_{vi}^*))|((n_{ri}))}\prod_{u=1}^{k-1}\prod_{r=0}^{u=1}(\hat{\xi}_u+\psi_r)^{N_u^*}}. \tag{4.36}$$

Die Parameterschätzer verringern sich, da die Häufigkeit der Rohscores durch die Itemrandsummen festgelegt werden, so dass $N_u = N_u^*$ für $u = 1,\ldots,k-1$, und alle möglichen Antwortmatrizen $((a_{vi}^*))$, die kompatibel mit der Itemrandmatrix $((n_{ri}))$ sind.

> "Since the conditional likelihood $p\{((a_{vi}))|(n_{ri}))\}$ cannot be used as a basis for Parameter estimation, however, such comparisons have no practical relevance, but only interpretative meaning." (Kempf, 1977, p. 313)

Für die Arbeit an der Aktualisierung des Fortran-Programms (siehe Abschnitt 5) zur Schätzung der Parameter des Kempf-Modells wurde auch versucht, die Personenparameter ξ_v zu schätzen. Mittels des Newton-Raphson-Verfahrens könnte man iterativ die Nullstellen der Funktion über die Nullstellen der Tangenten der Funktion berechnen. Man benötigt dafür die logarithmierte Likelihoodfunktion (4.13)

$$f(v) = \sum_{r=0}^{a_{vo}-1}\frac{1}{\xi_v+\psi_r} - \sum_{i=1}^{k}\frac{1}{\xi_v+\sigma_i}, \tag{4.37}$$

und deren erste Ableitung

$$f'(v) = \sum_{r=0}^{a_{vo}-1}\left(-\frac{1}{(\xi_v+\psi_r)^2}\right) + \sum_{i=1}^{k}\frac{1}{(\xi_v+\sigma_i)^2}. \tag{4.38}$$

Dieses Verfahren konvergiert meist sehr rasch, es können jedoch Probleme bei Auffinden der Nullstellen auftreten. In das vorliegende Programm konnte die Schätzung der Personenparameter daher nicht eingebaut werden. Die resultierenden Personenparameter waren teilweise nur unsinnige, viel zu große oder zu kleine Zahlen. Es konnte deswegen auch keine Normierung gefunden werden, mit der negative Personenparameter vermieden werden konnten.

4.4 Goodness-of-Fit-Statistiken

Kempf & Mach (1975) sehen, wie bereits in Abschnitt 4.2 erwähnt, einen Modellgeltungstest für das Kempf-Modell vor. Er gibt an, ob und wie gut das Modell auf die Daten passt. Man kann jedoch noch einen weiteren Test durchführen, der angibt, ob die Transferparameter vernachlässigbar sind und somit das Rasch-Modell angenommen werden kann.

4.4.1 Modellgeltungstest für das Kempf-Modell

Für den Goodness-of-Fit-Test des Modells wird die Gesamtstichprobe (im Programm die Antwort-, bzw. A-Matrix) zunächst durch einen festgelegten Trennwert c_1 in zwei Untergruppen aufgeteilt. Die erste Subgruppe besteht aus Personen mit niedrigem Rohscore $a_{vo} \leq c_1$, die zweite aus Personen mit hohem Rohscore $a_{vo} > c_1$. Diejenigen Items, die von allen Personen positiv oder negativ beantwortet wurden, werden ausgeschieden. Der Cut-off Punkt c_1 kann entweder von dem/der Benutzer/in selbst festgelegt werden, oder er wird vom Programm automatisch so festgelegt, dass die zwei Subgruppen ansatzweise die selbe Größe haben und möglichst wenig Items eliminiert werden müssen (siehe Abschnitt 5.3). Es wird ein Likelihood-Ratio Test, basierend auf einer Approximation an die χ^2-Verteilung durchgeführt. Das Grundprinzip des Ratio-Test ergibt sich aus

$$p\{(a_{vi})|a_{vo} = k-s\} = \frac{\prod_{i=1}^{k}(\sigma_i - \psi_{r_{vi}})^{1-a_{vi}}}{\sum_{m=0}^{s}\delta_m(k-s)\gamma_{s-m}(k)\cdot(-1)^m}. \tag{4.39}$$

Die Verteilung der Antworten jeder Person unter der Bedingung vom Rohscore a_{v0} ist unabhängig vom Personenfähigkeitsparameter und hängt nur von den Item- und Transferparametern ab. Diese können für jede Subgruppe G_v geschätzt werden, in dem man das Produkt von (4.39) über alle Personen der Subgruppe als bedingte Likelihood $L_v = \prod_{v\in G_v} p\{(a_{vi})|a_{v0}\}$ festlegt (siehe Kempf & Hampapa, 1975, S. 24).

Angenommen $G_1,\ldots,G_M$ stellen M disjunkte Subgruppen von Personen dar, so werden restringierte CML Schätzer der Hilfsparameter definiert.

$$\begin{matrix}\hat{\phi}_0^{(v)},\ldots,\hat{\phi}_{c_v}^{(v)} \\ \hat{\eta}_1^{(v)},\ldots,\hat{\eta}_k^{(v)},\end{matrix} \tag{4.40}$$

wobei c_v der größte Rohscore $a_{vo} < k$ in Subgruppe G_v ist.

Wenn das Modell gilt, sollten immer die selben Parameter $\phi_r^{(v)} = \phi_r$ und $\eta_i^{(v)} = \eta_i$ ge-

schätzt werden, egal welche Subgruppe untersucht wird. Durch den Vergleich dieser restringierten CML-Schätzer mit den unrestringierten $(\hat{\phi}_0,\ldots,\hat{\phi}_{c_{max}};\hat{\eta}_1,\ldots,\hat{\eta}_k)$ kann der Modell-Fit berechnet werden. Die bedingte Likelihood-Ratio kann somit so angegeben werden

$$\lambda = \frac{L(\hat{\phi}_0,\ldots,\hat{\phi}_{c_{max}};\hat{\eta}_1,\ldots,\hat{\eta}_k)}{\prod_{v=1}^{M} L^{(v)}(\hat{\phi}_0^{(v)},\ldots,\hat{\phi}_{c_v}^{(v)};\hat{\eta}_1^{(v)},\ldots,\hat{\eta}_k^{(v)})}, \tag{4.41}$$

wobei λ immer ≤ 1 sein muss. Wenn das Modell gilt, weichen die restringierten CML-Schätzer nur gering von den allgemeinen Schätzern ab und λ nähert sich somit 1. Wenn λ weitaus kleiner als 1 ist, wird das Modell verworfen.

Kempf et al. (1975) nehmen an, dass

> "From a theorem by Andersen (1971), it follows that the distribution $-2ln(\lambda)$ converges for $n \to \infty$ to a χ^2-distribution with
>
> $$df = (k-1)(M-1) + \sum_{v=1}^{M} c_v - c_{max} \tag{4.42}$$
>
> degrees of freeedom."(S. 25)

Das Modell wird mit dem asymptotischen Signifikanzniveau α verworfen, wenn $-2\ln(\lambda)$ größer ist als das $(1-\alpha)$te Perzentil der χ^2-Verteilung mit $df = k-1+c_1$ Freiheitsgraden.

4.4.2 Reduktion zum Rasch-Modell

Da das Modell von Kempf und das Rasch-Modell nur dann exakt äquivalent sind, wenn alle Transferparameter $\psi = 0$ sind (siehe Abschnitt 4.1), wurde eine zweite Goodness-of-Fit-Statistik eingeführt. Sie soll zeigen, ob sich die Likelihood des Kempf-Modells signifikant von der des Rasch-Modells unterscheidet. Wenn dies nicht der Fall ist, ist der Effekt der Transferparameter vernachlässigbar und das Modell kann zu einem „simplen“ Rasch-Modell reduziert werden.

Die Likelihood-Ratio wird mit

$$-2(L_{Rasch} - L_{Kempf}) \tag{4.43}$$

mit $df = c_{max}$ Freiheitsgraden gebildet. Unter der Nullhypothese $\psi_0 = \psi_1 = \cdots = \psi_{c_{max}} = 0$,

sind die Likelihoods der beiden Modelle gleich. Diese Hypothese wird verworfen, wenn der χ^2-Wert über dem entsprechenden kritischen Wert liegt, und somit die Transfereffekte nicht vernachlässigbar sind.

5 Programm zur Schätzung der Modellparameter

Kempf & Hampapa bzw. Kempf & Mach (1975) entwickelten gemeinsam ein Fortran-Programm zur Schätzung der Modellparameter sowie zur Durchführung eines Modellgeltungstests. Das Originalprogramm war für Lochkarten konzipiert, dementsprechend musste es für heutige PCs adaptiert werden. Das neu adaptierte Programm wurde von der Verfasserin auf den Namen „DynTest“ getauft.

Folgende Änderungen wurden vorgenommen:

- Die Personenanzahl wurde von 450 auf 1 000 000 Personen heraufgesetzt,
- die Itemanzahl wurde von maximal 20 auf maximal 100 Items heraufgesetzt,
- die Ausgabe des Programms wurde verändert,
- die Berechnung der Gamma-Funktionen wurden verändert, um die Itemanzahl bei zumindest gleicher Schätzgenauigkeit erhöhen zu können,
- die Schätzung der Item- und Personenparameter des Rasch-Modells wurde eingefügt,
- dadurch wurde ein neuer Modellgeltungstest möglich gemacht, und
- schließlich wurde zur Erhöhung der Benutzerfreundlichkeit eine Benutzeroberfläche erstellt.

5.1 Technische Angaben

Alle Berechnungen und Simulationen wurden auf einem HP PC mit Intel(R)Core(TM)2 Duo CPU, E4500 @2.20 GHz, 2.19 GHz und 988 MB RAM durchgeführt.

Die Arbeitsschritte an der Modernisierung bzw. Änderung am Fortran-Programm sowie die Erstellung des ausführbaren Programms von DynTest erfolgten mit Hilfe des Freeware Fortran-Compilers Plato3, Version 3.50 von Silverfrost.

Das Java-Programm wurde mit der Entwicklungsumgebung Eclipse, Version 3.3.1.1 erstellt.

Die Programme zur Datensimulation bzw. automatischen Parameterschätzung und Übertragung in SPSS wurde mit Visual Studio 2008 von Microsoft in C# erstellt.

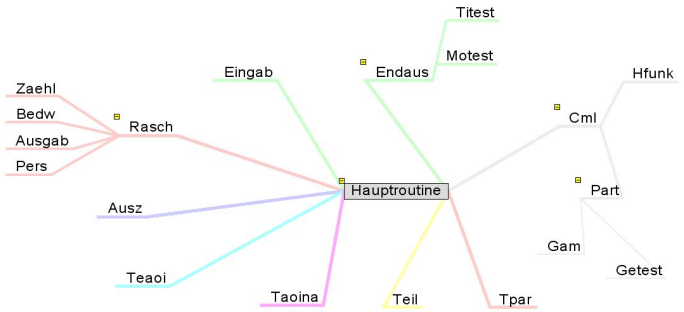
Zur Erstellung der Diagramme und Graphiken sowie zur Berechnung der Statistiken für die Parameterschätzungen wurde SPSS 15.0 verwendet.

5.2 Schätzung der Rasch-Modell Parameter

Im ursprünglichen Programm von Kempf et al. (1974) kommt die Schätzung der Rasch-Modell-Parameter nicht vor, daher wurden vier zusätzliche Subroutinen in das Programm mit aufgenommen. Sie stammen aus dem Fortran-Programm von Formann (in Fischer, 1974). In diesen Subroutinen werden die Itemleichtigkeitsparameter des Rasch-Modells für den ausgewählten Datensatz mittels CML-Schätzung berechnet und in einheitsnormierter, produktnormierter und logarithmierter Form ausgegeben. Zusätzlich werden auch noch die Itemschwierigkeitsparameter mit $\sigma_i = \frac{1}{\varepsilon_i}$ angegeben, um sie direkt mit den Ergebnissen des Itemparameterschätzung des Kempf-Modells vergleichen zu können. Die Personenparameter und die Likelihood des Rasch-Modells werden ebenfalls berechnet. Letztere wird anschließend für den Likelihood-Ratio Test zum Vergleich des Rasch-Modells mit dem Kempf-Modell benötigt.

5.3 Struktur des Programms

Abbildung 5.1: Programmstruktur



Das Fortran-Programm besteht aus der Hauptroutine und 19 Subroutinen (siehe Abbildung 5.1), die folgenden Funktionen haben.

Hauptroutine In ihr werden die Parameter aus der Datei test.ini eingelesen, also etwa die Anzahl der Items und der Personen, der Name des Datensatzes etc.. [1] Des Weiteren können in der Hauptroutine ebenfalls etwaige Anfangswerte für die Parameterschätzungen eingelesen werden. Alle Subroutinen zur Schätzung der Parameter und zur Berechnung der Goodness-of-Fit Tests werden von hier aufgerufen. Der Modellgeltungstest zum Vergleich des Kempf-Modells mit dem Rasch-Modell wird ebenfalls in der Hauptroutine durchgeführt.

Subroutine *Eingab* Hier wird der Datensatz eingelesen. Die Daten werden auf Gültigkeit untersucht und die Versuchspersonen mit nicht gültigen (nicht 0/1 codierten) Daten werden eliminiert.

Subroutine *Rasch* Diese Subroutine stammt ursprünglich aus einem Fortran-Programm von Formann (in: Fischer, 1974). Hier werden die Subroutinen *Zaehl*, *Bedw*, *Pers* und *Ausgab* zur Schätzung der Parameter für das Rasch-Modell aufgerufen.

Subroutine *Zaehl* In dieser Subroutine werden die Antwortvektoren zur Schätzung der Parameter des Rasch-Modells eingelesen und ausgezählt.

[1] Die test.ini Datei wird zuvor durch die Eingaben in die Benutzeroberfläche befüllt (siehe Abschnitt 5.4.1).

Subroutine *Bedw* Hier werden die Itemparameter für das Rasch-Modell mittels CML-Schätzung berechnet. Bei Überschreiten der angegebenen Rechengenauigkeit wird die Schätzung für das Rasch-Modell abgebrochen. In diese Subroutine wurde die neue Schätzung der Gamma-Funktionen nicht eingebaut, da hier weder wegen Itemanzahlen größer als 20, noch wegen zu großer Ungenauigkeiten Probleme auftraten. Die Berechnung würde so nur noch mehr Zeit beanspruchen.

Subroutine *Pers* berechnet die Personenparameter für das Rasch-Modell. Diese werden unlogarithmiert und logarithmiert ausgegeben.

Subroutine *Ausgab* dient lediglich der Ausgabe der Itemparameter des Rasch-Modells. Diese werden einheitsnormiert, produktnormiert, logarithmiert sowie zum direkten Vergleich an die Modelldarstellung des dynamischen Testmodells von Kempf angepasst ausgegeben.

Subroutine *Ausz* In dieser Subroutine werden die Item-Randsummen A_{oi} und die Häufigkeiten der Rohscores der Personen $N_{a_{vo}}$ berechnet sowie die Anzahl der Personen N_{ri}, die ein Item i falsch beantwortet haben, nachdem r_{vi} richtig beantwortet wurden. Personen mit $a_{vo=0}$ oder $a_{vo=k}$, also Personen, die alle oder kein Item gelöst haben, werden hier ebenfalls ausgesondert.

Subroutine *Teaoi* Hier werden Items mit $a_{oi} = 0$, also Items, die nie gelöst worden sind, ausgeschieden.

Subroutine *Taoina* Diese Subroutine löscht alle Items mit $a_{oi} = N$, also Items die immer gelöst wurden. Auch die darauf folgenden Items werden eliminiert, wenn das Item nicht das erste oder letzte im Test war. Des Weiteren wird getestet, ob die Anzahl an verbleibenden Items größer als 3 ist. Wenn nicht, wird die Analyse abgebrochen.

Subroutine *Teil* Hier wird der Trennwert c_1 zur Aufteilung der Stichprobe berechnet. Des Weiteren wird abermals kontrolliert, ob innerhalb der zwei Subgruppen Items mit $a_{oi} = N$ vorkommen. Diese und die folgenden werden eliminiert.

Subroutine *Tpar* Diese Subroutine bildet Datenmatrizen für die beiden Subgruppen.

Subroutine *Endaus* Hier werden die Subroutinen *Titest*, *Ausz*, *Cml* und *Motest* aufgerufen, um die Parameterschätzer für die beiden Subgruppen und den Modellgeltungstest

zu berechnen. Auch die Anzahl der Versuchspersonen pro Subgruppe wird hier ausgegeben.

Subroutine *Titest* *Titest* ist ebenfalls für die Berechnung der Datenmatrizen für die Subgruppen zuständig.

Subroutine *Cml* *Cml* ruft die Subroutinen *Part*, *Hfunk* und *Kempfpers* auf, um die CML-Schätzer der Item-, Transfer- und Personenparameter mittels Gradientenmethode und Methode der „Regula Falsi“ zu berechnen. Die Prozedur wird abgebrochen, wenn das Kriterium nicht erreicht wurde oder die Rechenungenauigkeit zu groß ist. Die Subroutine normiert und transformiert die Parameterschätzer und gibt sie in transformierter, Mitte-normierter und Null-Eins-normierter Form aus.

Subroutine *Hfunk* *Hfunk* berechnet die Werte von $\ln(L)$ in der Richtung des Gradienten.

Subroutine *Part* Hier werden die Delta-, Gamma- und G-Funktionen, die logarithmierte Likelihood sowie deren erste partielle Ableitungen berechnet. *Part* ruft die Subroutinen *Gam* zur Berechnung der Gamma-Funktionen und *Getest* zum Genauigkeitstest auf.

Subroutine *Gam* In dieser Subroutine werden die Gamma-Funktionen berechnet.

Subroutine *Getest* In dieser Subroutine wird der Rechengenauigkeitstest durchgeführt.

Subroutine *Motest* Die Subroutine *Motest* führt schließlich den Modellgeltungstest für das Kempfmodell durch.

5.4 Graphische Benutzeroberfläche

Zur leichteren Handhabung des Programms für den/die Benutzer/in wurde ein Java-Programm mit graphischer Benutzeroberfläche auf das Fortran-Programm aufgesetzt. Der Aufbau und die Elemente des Java-Programms sollen kurz erläutert werden, der folgende Abschnitt enthält außerdem einen Leitfaden für Benutzer. [2]

[2] Im Folgenden wird das Wort „Benutzer“ für Benutzer und Benutzerinnen gebraucht, um die Übersichtlichkeit nicht zu beeinträchtigen.

5.4.1 Java-Programm

Um die Bedienung des adaptierten Fortran-Programms an heutige Gewohnheiten anzupassen, ohne allerdings das eigentliche Programm zu verändern, wurde von der Autorin eine graphische Benutzeroberfläche (Graphical User Interface, GUI) in Java geschrieben, die das Fortran-Programm zur Schätzung der Parameter des Kempf-Modells aufruft und die Ergebnisse in einer Datei speichert.

Um auf die zukünftige Verwendbarkeit und die Kompatibilität mit möglichst vielen Betriebssystemen zu achten, fiel die Wahl der Bibliothek für die graphischen Elemente auf Swing. Diese weist im Gegensatz zu der möglichen Alternative AWT eine betriebssystemübergreifende Gestaltung der eingesetzten Elemente auf.

Das eingesetzte Layout ist ein 15x2 Grid-Layout, d.h. das grundlegende Layout-Element ist eine Tabelle mit 15 Zeilen und zwei Spalten. Eine Besonderheit des Grid-Layouts ist die dynamische Anpassung der Elemente an die Größe des Fensters. Das bedeutet, wenn der Benutzer die Fenstergröße der GUI verändert, passen sich die Elemente, wie z.B. Buttons, Textfelder, Beschriftungsfelder, etc. an. Die Anzahl der Zeilen (15) richtet sich nach den maximal zu wählenden Parametern. Wo es möglich war, wurden die jeweiligen Elemente mit voreingestellten Standardwerten befüllt. Ob sämtliche Felder zu sehen sind oder nicht, richtet sich nach dem Wert von „Anfangsschätzungen einlesen“, da dieser Parameter darüber entscheidet, ob Startwerte für die Kempf-Modell-Parameter aus zwei externen Dateien eingelesen werden oder nicht.

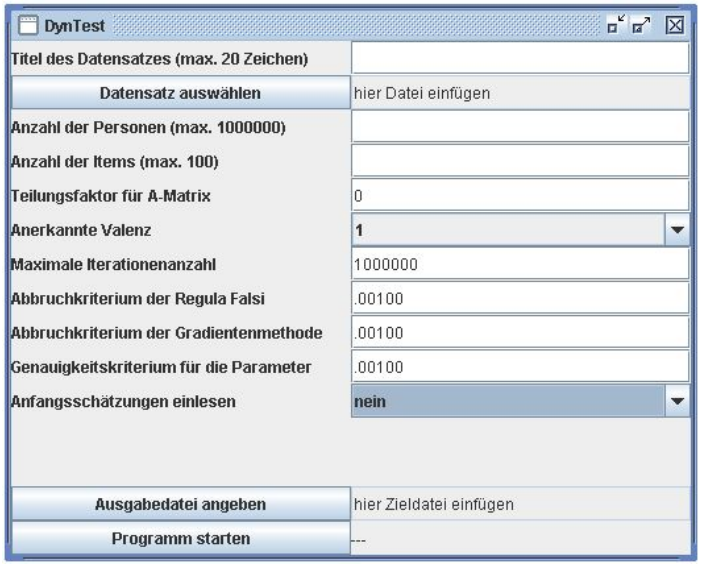
Für die von dem Benutzer frei wählbaren Parameter (der Titel des Datensatzes, die Anzahl der Personen und der Items, der Teilungsfaktor für die A-Matrix, die Maximale Anzahl der Iterationen, das Abbruchkriterien der Regula Falsi und der Gradientenmethode sowie das Genauigkeitskriterium für die Parameter) wurden Textfelder als Eingabemittel gewählt. Für die Parameter, die aus einem vorgegebenen Set gewählt werden können (die anerkannte Valenz und das Einlesen der Anfangsschätzungen), kommen so genannte DropDownListen ohne Möglichkeit der freien Eingabe zum Einsatz. Um die Auswahl der zwei bzw. vier zu verwendenden Dateinamen (der Datensatz, die Anfangsschätzungen der Item- und Transfer-Parameter sowie die Ausgabedatei) möglichst benutzerfreundlich zu gestalten, gibt es die Möglichkeit die jeweiligen Dateinamen mit Hilfe einer Instanz der Klasse *FileChooser* auszuwählen. Deren Namen wird im Anschluss an eine erfolgte Auswahl in ein Bezeichnungsfeld (Label) eingetragen.

Bei Betätigung des Buttons „Programm starten“ werden die erfolgten Eingaben auf Vollständigkeit und Gültigkeit geprüft. Bei positivem Ergebnis der Prüfung werden die Eingaben in die Datei test.ini gespeichert und das Fortran-Programm in einer eigenen Shell gestartet, wobei die Ausgabe auf die in der GUI ausgewählte Ausgabedatei umgeleitet wird.

5.4.2 Leitfaden für Benutzer/innen

Die GUI besteht aus insgesamt 15 Eingabezeilen, von denen jedoch in der Standard-Einstellung lediglich 13 sichtbar sind (siehe Abbildung 5.2).

Abbildung 5.2: Standard-Ansicht der GUI



Der Benutzer kann folgende Parameter eingeben:

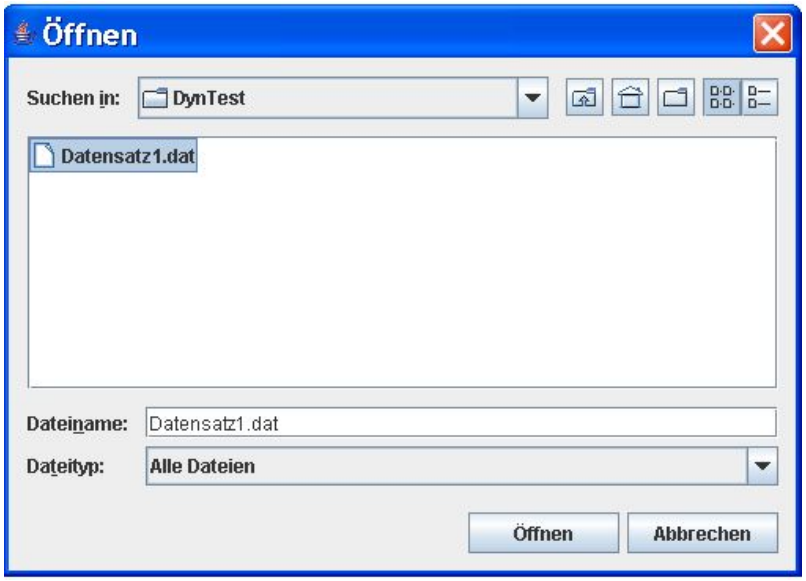
1. Zeile: Den Titel des Datensatzes mit maximal 20 Zeichen. Z.B. WMT_18_Items etc.

2. Zeile: Die Datei des Datensatzes, für den man die Parameter schätzen möchte. Der Datensatz selbst soll eine .dat-Datei sein und darf ab der ersten Spalte nur Nullen und Einsen enthalten. Eine Zeile steht für eine Person. Zwischen den Nullen bzw. Einsen darf sich kein Leer- oder Sonderzeichen befinden (siehe Abbildung 5.3).

Abbildung 5.3: Datensatz

Als Erleichterung für den Benutzer erscheint bei Klicken des Buttons „Datensatz auswählen“ ein neues Fenster, in dem er die Datei aus den vorhandenen Verzeichnissen auswählen kann (siehe Abbildung 5.4). Nachdem die Datei des Datensatzes im Feld „Dateiname“ steht, fügt man sie durch Klicken des Buttons „Öffnen“ der GUI hinzu. Um bei der händischen Eingabe keine Fehler zu machen, kann man den Datensatz nur auf diese Weise auswählen.

Abbildung 5.4: Datensatz auswählen



3. Zeile: Die Anzahl der Personen des Datensatzes. Dieses Feld *muss* händisch ausgefüllt werden. Die Maximalanzahl der Personen beträgt 1 000 000.

4. Zeile: Die Anzahl der Items des Datensatzes. Dieses Feld *muss* man ebenfalls händisch ausfüllen. Die Maximalanzahl der Items beträgt 100.

5. Zeile: Den Teilungsfaktor für die A-Matrix (siehe Abschnitt 4.4.1). Dieser Teilungsfaktor wird später in der Ausgabe als „Vorgegebene Konstante“ bezeichnet, bei der die Antwortmatrix geteilt wird. Als Standard-Wert ist in der GUI „0“ eingestellt, das bedeutet, dass der Teilungsfaktor für jeden Datensatz neu errechnet wird. Wird hier ein Wert > als 0 eingegeben, wird dieser neue Wert als Teilungsfaktor für den ersten Teilungsversuch herangezogen.

6. Zeile: Die anerkannte Valenz, d.h. welcher Wert aus 0 oder 1 als „richtig beantwortet“ gilt. Als Default-Wert ist hier 1 eingestellt, durch das Auswahlmenü ist es jedoch möglich, den Wert auf 0 zu ändern, wenn man dies für eine spezielle Fragestellung - etwa eines Einstellungsfragebogens - benötigt.

7. Zeile: Die maximale Iterationenanzahl bei der Schätzung der Parameter. Als Standard-Wert ist hier eine Maximalanzahl von 1 000 000 Iterationen eingestellt.

8. Zeile: Das Abbruchkriterium der Regula-Falsi (siehe Abschnitt 4.2). Die Schätzung der Modellparameter erfolgt u.a. mithilfe der Regula Falsi. Das Kriterium soll möglichst genau erreicht werden. Der Standard-Wert wurde hier auf 0.001 gesetzt, er kann jedoch auf 5 Nachkommastellen genau verändert werden.

9. Zeile: Das Abbruchkriterium der Gradientenmethode (siehe Abschnitt 4.2). Auch bei dieser Methode soll das Kriterium möglichst genau erreicht werden, als Abbruchswert wurde hier 0.001 verwendet. Dieser Wert kann ebenfalls auf 5 Nachkommastellen beliebig genau verändert werden.

10. Zeile: Das Genauigkeitskriterium für die Parameter. Als Standard-Wert wurde hier 0.001 eingegeben, damit die Parameter möglichst genau geschätzt werden. Auch dieser Wert kann bis zu 5 Nachkommastellen verändert werden.

11. Zeile: Das Einlesen von Anfangsschätzungen für die Parameter. Hier kann der Benutzer aus einem Menü auswählen, ob für die Parameterschätzung des Modells Anfangsschätzungen für die Item- und Transferparameter eingelesen werden sollen oder nicht. Das

Einlesen von Anfangsschätzungen ist nicht notwendig, daher steht die Standardeinstellung auf „nein“. Falls man jedoch Startwerte für die Parameterschätzung festlegen möchte, kann man dies durch Auswahl aus dem Menü ändern. Der Vorgang des Auswählens ist der gleiche, wie für die Datei des Datensatzes.

12. Zeile: Die Auswahl der Datei für die Anfangsschätzungen der Itemparameter Sigma. Nur wenn ausgewählt wurde, dass Anfangsschätzungen für die Parameter eingelesen werden sollen, wird diese Zeile sichtbar (siehe Abbildung 5.5). So wie beim Auswählen des Datensatzes, kann nun wieder aus den eigenen Verzeichnissen eine Datei für die Anfangswerte der Sigma-Parameter ausgewählt werden (siehe Abbildung 5.6). In der angelegten Datei (etwa einer .txt-Datei) muss jeder Anfangsschätzwert als Kommazahl in eine neue Zeile geschrieben werden (siehe Abbildung 5.7), die Datei wird Zeile für Zeile eingelesen.

Abbildung 5.5: GUI mit Einlesen der Anfangsschätzungen für die Parameter

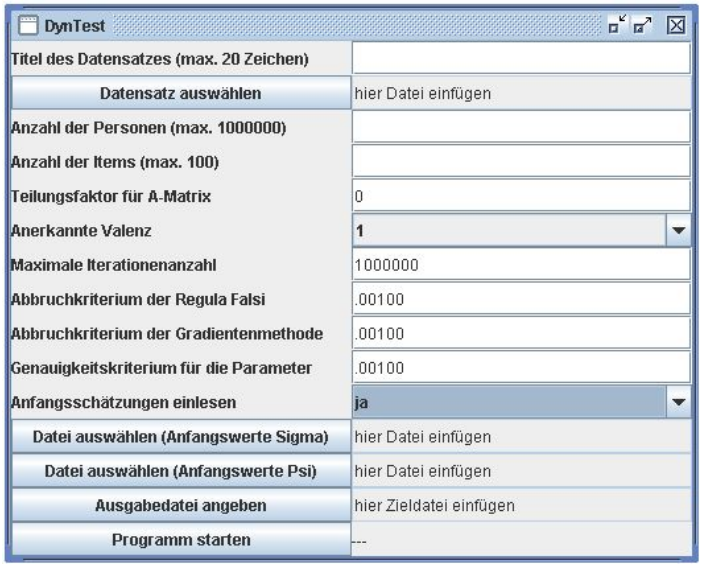


Abbildung 5.6: Auswählen der Anfangsschätzwerte für Sigma und Psi

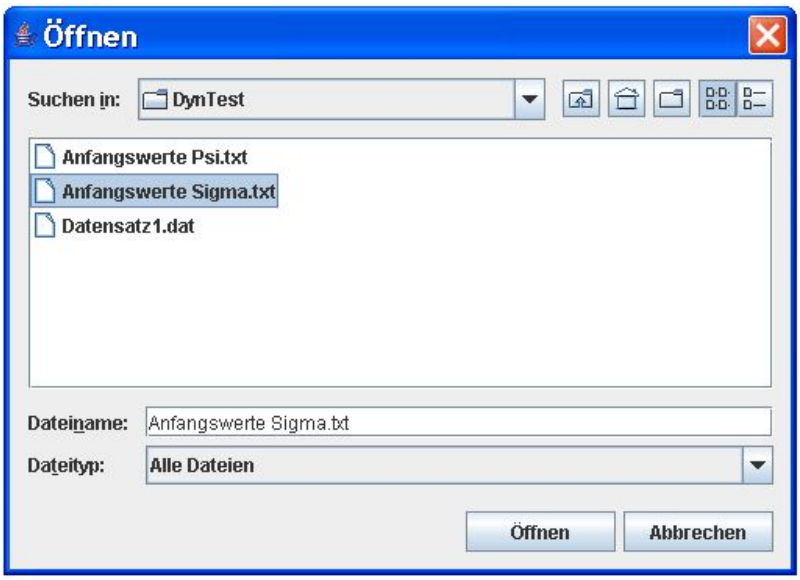


Abbildung 5.7: Datei für Anfangsschätzwerte

13. Zeile: Die Auswahl der Datei für die Anfangsschätzungen der Transferparameter Psi. Die Handhabung erfolgt genauso wie bei der Auswahl der Anfangswerte für die Sigma-Parameter.

14. Zeile: Das Anlegen der Ausgabedatei. Das Programm erstellt die angegebene Zieldatei jedes Mal neu. Man legt eine Ausgabedatei an, indem man aus den eigenen Verzeichnissen

einen Ordner auswählt, in den die Ausgabe gespeichert werden soll, und im Feld „Dateiname“ einen \ und den gewünschten Ausgabenamen eingibt, z.B. C:\Desktop\output1.txt oder Z:\DynTest\Ausgabe1.txt (siehe Abbildung 5.8).

Abbildung 5.8: Ausgabe anlegen

15. Zeile: Der Button „Programm starten“. Bei Anklicken dieses Buttons wird das Fortran-Programm (siehe Abschnitt 5.3) gestartet. Nach Klicken des Buttons wird zusätzlich auf Fehler bei der Eingabe geprüft. Falls eine Eingabe fehlerhaft erfolgt ist (z.B. Buchstaben statt Zahlen eingegeben wurden), wird dies neben dem Button vermerkt (siehe Abbildung 5.9). Wenn keine Fehler aufgetreten sind, startet die Berechnung. Die Ausgabe-Datei wird nach Beenden der Schätzungen automatisch geöffnet (siehe Abbildung 5.10) und in dem Verzeichnis, das ausgewählt wurde, gespeichert. Neben dem „Start“-Button erscheint außerdem nach Abschluss der Schätzungen die Meldung „Eingaben korrekt“. Dieser Vorgang kann jedoch abhängig von der Größe des Datensatzes einige Minuten bis Stunden dauern.

Abbildung 5.9: Fehlerhafte Eingabe

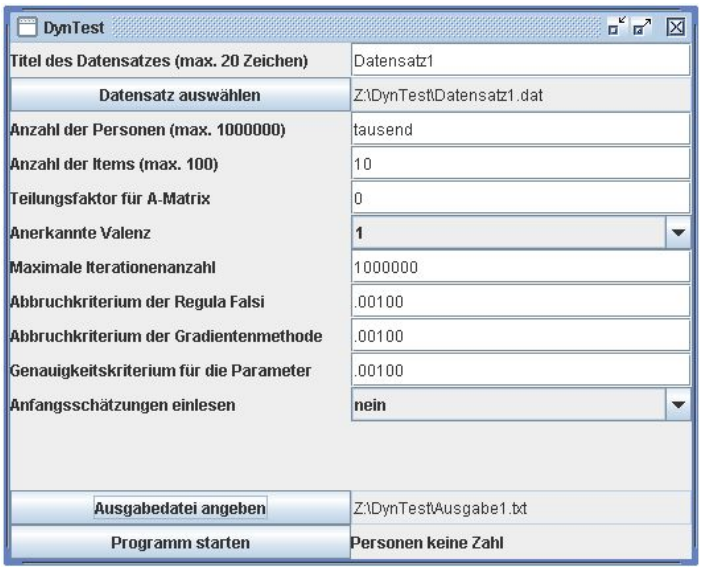
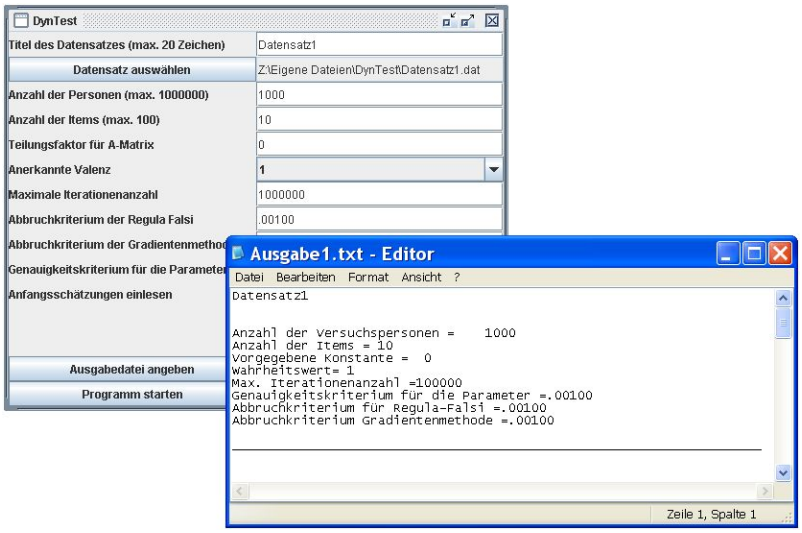


Abbildung 5.10: Automatische Ausgabe



5.5 Ausgabe

Die Ausgabe erfolgt in Form einer .txt-Datei. An oberster Stelle der Ausgabe finden sich die Inhalte der Eingabezeilen aus der GUI. Es werden der Titel des Datensatzes, die Anzahl der Personen, die Anzahl der Items, das Teilungskriterium (also der Wert, der für die Teilung

der Stichprobe in zwei Subgruppen verantwortlich ist), der Wahrheitswert (die Valenz), die Maximale Iterationenanzahl, das Genauigkeitskriterium für die Parameter, das Abbruchkriterium für die Regula-Falsi und das Abbruchkriterium für die Gradientenmethode ausgegeben.

Im Folgenden kann der Benutzer sehen, wie viele Personen wegen ungültiger Werte im Datenfile (also alles andere als „0“ und „1“) gelöscht wurden. [3]

Als nächstes sieht man die Item und Personenparameter für das Rasch-Modell sowie dessen logarithmierte Likelihood. Zusätzlich dazu ist aufgeführt, wie viele Iterationen die Schätzung der Parameter benötigt hat und auf welches Item einheitsnormiert wurde. Die Itemleichtigkeitsparameter des Rasch-Modells sind in einheits-, produktnormierter und logarithmierter Form sowie nach der Schreibweise der Itemparameter des Kempf-Modells angegeben. Die Personenparameter des Rasch-Modells sind sowohl normal, als auch in logarithmierter Form ausgegeben.

Für die Schätzung der Parameter des Kempf-Modells ist angegeben, wie viele Personen, die entweder alle oder kein Item richtig beantwortet hatten, und wie viele Items, die nie oder immer gelöst wurden, ausgeschieden wurden. Die verbliebenen Itemnummern sind zur Überprüfung aufgeführt.

Als nächstes sieht man die Parameterschätzung des Kempf-Modells für die Gesamtstichprobe, mit Angabe der benötigten Iterationen sowie der logarithmierten Likelihood. Die Ausgabe der Parameter umfasst zum Ersten die so genannten „transformierten“ Hilfsparameter η und ϕ (siehe Abschnitt 4.2), zum Zweiten die „Mitte-normierten“ Itemschwierigkeitsparameter σ und Transferparameter ψ (diese sind normiert nach (4.33)) und zum Dritten die „Null-Eins-normierte“ Form (d.h. die Summe der ψ ist 0, das Produkt der σ ist 1).

Anschließend sind dieselben Kennwerte und Parameter für die beiden Teilstichproben gesondert angegeben. Zusätzlich wird hierfür die Stichprobengröße für die erste und die zweite Untergruppe ausgegeben.

Als letztes sind die logarithmierten Likelihoods für die beiden Modellgeltungstests (siehe Abschnitt 4.4) in der Ausgabe zu sehen. D.h. für die Modellgeltung des Kempf-Modells sind die Likelihood für die Gesamtstichprobe des Kempf-Modells sowie die beiden Likelihoods der Teilstichproben und deren Likelihoodquotienten mit den Freiheitsgraden ausgegeben. Für die Überprüfung der Modellgeltung des Rasch-Modells sind die Gesamtlikelihoods des Kempf- und des Rasch-Modells sowie deren Quotienten und die Freiheitsgrade aufgeführt.

[3] Im Idealfall sollte keine Person aus diesem Grund ausgeschieden werden, wenn der Datensatz korrekt eingegeben wurde.

6 Anwendung des dynamischen Testmodells

Kempf (1974) äußerte selbst Kritik an seinem Fortran-Programm und bemängelte die Handhabbarkeit, die Ungenauigkeit sowie die Interpretierbarkeit der Parameter des Modells. Um die Möglichkeiten und Grenzen des Modells bzw. des Programms auszuloten, wurden verschiedene simulierte und reale Datensätze herangezogen. Im Folgenden sollen die Ergebnisse dieser Anwendung dargestellt und diskutiert werden.

Es ist jedoch anzumerken, dass die verwendeten Datensätze teilweise sehr problematisch in ihrer Anwendung auf das dynamische Testmodell von Kempf sind. Auch wurden die Tests von Personen ohne Verstärkung oder Rückmeldung bearbeitet, sodass ein positiver Transfer ausschließlich durch das Einarbeiten in die Materie, und nicht über positives (oder negatives) Feedback stattfinden konnte.

6.1 Simulation von Daten

Die Simulation von Daten ist wichtig, um im ersten Schritt die Parameterschätzung des Programms genauer unter die Lupe zu nehmen. Damit können die Möglichkeiten und Grenzen sowie die Genauigkeit der Schätzungen untersucht werden. In der vorliegenden Arbeit werden die Ergebnisse zweier Simulationsreihen mit 8 bzw. 20 Items vorgestellt. Für erstere wurden jeweils 100 Datensätze mit 100, 500, 1000, 5000 und 100000 Personen erzeugt, für zweitere jeweils 100 Datensätze mit 500, 1000 und 5000 Personen. Um diese Simulationen möglichst effektiv und zeitsparend zu generieren und zu berechnen, wurden zwei C#-Programme mit GUI geschrieben. Ein weiteres C#-Programm mit GUI wurde erstellt, um die Ergebnisse der Parameterschätzungen möglichst schnell und vor allem fehlerfrei in SPSS zu übertragen.

Aus den jeweils 100 Datensätzen pro Personen- und Itemanzahl wurde für die Mitte- und

Null-Eins-normierten Item- und Transferparameter jeweils der Mittelwert gebildet. Man erhält so die Durchschnittsschätzungen von vier Arten von Parameterschätzern. Diese sind dann gut miteinander vergleichbar. Weiters wurden die simulierten Datensätze jeweils zweimal durch gerechnet. Einmal mit den alten Gamma-Funktionen wie im Originalprogramm nach Kempf und ein zweites mal mit den neuen Gamma-Funktionen, die eine größere Itemanzahl zulassen. Im Folgenden wird gezeigt, dass sich die Genauigkeit der beiden Methoden so gut wie gar nicht unterscheidet. Der Vorteil der neuen Methode liegt jedoch wie bereits angeführt in der Möglichkeit, mehr als 20 Items schätzen zu können. Allgemein wurden mit der neuen Methode weniger Iterationen bis zum Erreichen des Genauigkeitskriteriums gebraucht, jedoch mehr Zeit als mit der alten. Im Sinne der Einheitlichkeit sind in der vorliegenden Arbeit nur die Durchschnittswerte der neuen Gamma-Funktionsschätzung angeführt, da diese auch zur Schätzung der echten Datensätze herangezogen wurde.

6.1.1 Simulationsprogramm

Das Simulationsprogramm für die Generierung von Kempf-Modell konformen Datensätzen wurde in C# geschrieben. Zunächst wird die vom Benutzer festgelegte Anzahl von Simulationsdateien in einem Unterordner mit der jeweiligen ausgewählten Personenanzahl (100, 500, 1000, 5000 oder 100 000) erstellt (siehe Abbildung 6.1). Danach werden die Item-, Transfer und Personenparameter statisch mit deren Werten befüllt. Für die Daten in jeder Simulationsdatei wird ein zweidimensionales Datenfeld mit der Größen n und k angelegt. Im Anschluss daran wird die Wahrscheinlichkeit p für die Daten des Kempf-Modells nach der Formel

$$p = \frac{\xi_i + \psi_r}{\xi_i + \sigma_j} \tag{6.1}$$

berechnet. i geht von Person 1 bis n, j geht von Item 1 bis k. r ist die Anzahl der richtig gelösten (also mit 1 kodierten) Items pro Person. Die binären Daten werden mit Hilfe der Funktion Bernoulli auf folgende Weise erstellt. Es werden gleichverteilte Zufallszahlen generiert. Wenn die Zufallszahl kleiner oder gleich der Wahrscheinlichkeit p aus Modellgleichung (6.1) ist, wird der Dateneintrag auf 1 gesetzt, wenn sie größer als p ist auf 0. Diese Daten werden nun in die jeweilige Simulationsdatei geschrieben.

Abbildung 6.1 zeigt die GUI, in der man auswählen kann, wie viele Zeilen - in diesem Fall Personen - eine Datei haben soll und wie viele Dateien erzeugt werden sollen. Außerdem kann man den Pfad angeben, in dem die erzeugten Dateien gespeichert werden sollen.

Als Kontrolle werden nach fertiger Simulation im leeren Rechteck die gewählten Personenparameter angezeigt. Nach drücken des „Simulieren“-Buttons werden die Dateien in fortlaufender Nummerierung im gewünschten Verzeichnis gespeichert.

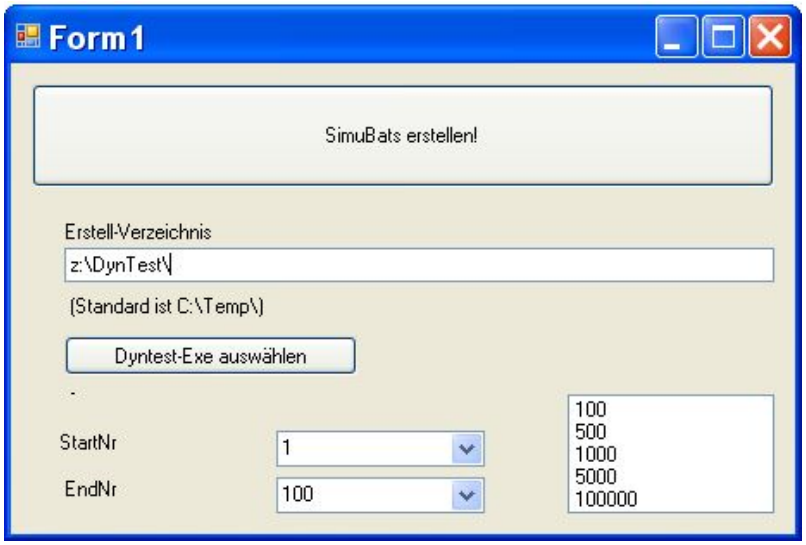
Abbildung 6.1: GUI für die Datensimulation

6.1.2 Automatisierung der Parameterschätzung für Simulationsreihen und Übertrag in SPSS

Damit die DynTest.exe nicht für jede Datei einzeln händisch gestartet werden musste, kam folgendes C#-Programm zum Einsatz. Zunächst wird eine Stapelverarbeitungsdatei simu.bat angelegt. Diese dient der automatisierten Erstellung von .ini-Dateien und der automatisierten Parameterschätzung von allen Simulationsdateien. Jeder erstellten Simulationsdatei wird eine .ini-Datei, die die jeweiligen Parameter wie Dateinamen, Personen- und Itemanzahl sowie Genauigkeitskriterien (siehe Abschnitt 5.4.2) enthält, beigefügt. Für die Parameterschätzung benötigt die DynTest.exe die Datei test.ini (siehe Abschnitt 5.4.1). In diese test.ini wird von der .bat-Datei die jeweilige .ini Datei pro Simulation hinein kopiert, um die Schätzungen für verschiedene Dateien möglich zu machen. Damit wird jede Simulationsdatei geschätzt und die Ausgabe in eine eigene .txt-Datei gespeichert.

Abbildung 6.2 zeigt die GUI, die zu diesem Zweck geschrieben wurde. Man wählt das Verzeichnis, in das die .bat-Datei gespeichert werden soll sowie den Pfad der .exe-Datei aus und gibt an, die Parameter welcher Dateien geschätzt werden sollen. Durch Klicken des „Simu Bats erstellen!“-Buttons wird die .bat Datei im gewünschten Verzeichnis erstellt und muss im Anschluss daran nur noch gestartet werden.

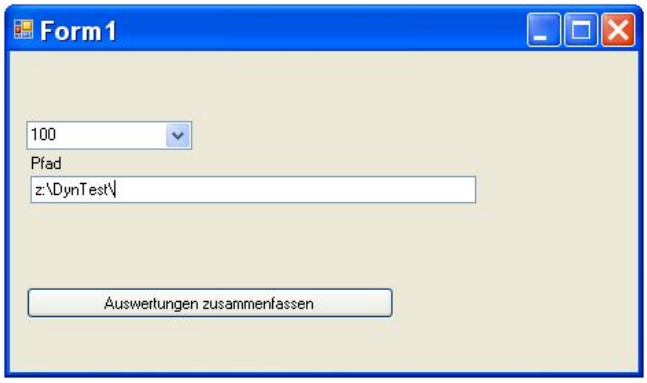
Abbildung 6.2: GUI für die Generierung der .bat-Datei



Um die geschätzten Mitte- und Null-Eins-normierten Item- und Transferparameter für die 100 Dateien pro Personenanzahl in SPSS transformieren zu können, wurde ein weiteres C#-Programm geschrieben. Durch dieses wird eine .sps Syntaxdatei für jede Personenanzahl angelegt. Die geschätzten Parameter werden dann aus den Ausgabedateien eingelesen, geparst und in SPSS-Syntax transformiert. Die Syntax muss dann noch händisch ausgeführt werden, um ein mit allen Parameterschätzern gefülltes .sav-File zu erhalten.

In Abbildung 6.3 wird die GUI für die Erstellung der SPSS-Syntax dargestellt. Man wählt aus, wie viele Dateien zum Syntax zusammengefasst werden sollen und in welchen Verzeichnis dies geschehen soll. Nach Klicken des Buttons „Auswertung zusammenfassen“ wird im gewünschten Verzeichnis für jede Personenanzahl, für die die Datensätze simuliert wurden, eine .sps-Datei erzeugt.

Abbildung 6.3: GUI für die Generierung der SPSS-Syntax



6.1.3 Ergebnisse einer Simulationsreihe mit 8 Items

Die zur Simulation verwendeten Modellparameter wurden willkürlich ohne bestimmte Normierungen von der Autorin festgelegt, um möglichst reale Parameter zu simulieren. Es wurden je 100 Datensätze für 100, 500, 1000, 5000 und 100000 Personen generiert und geschätzt. Tabelle 6.1 zeigt die acht Item- und Transferparameter sowie die sieben Personenparameter.

Tabelle 6.1: Ausgangsparameter der Simulation bei 8 Items

Item	Itemschwierigkeit	Transfer	Personenfähigkeit
1	1.6	0	1
2	2.1	0.5	1.5
3	1.5	0.2	2
4	1.7	0.1	3
5	1.2	0.4	3.5
6	1.5	0.7	4
7	1.4	0.6	4.5
8	1.8	0.4	

Die Verteilung der Personenparameter wurde ebenfalls willkürlich von der Autorin festgelegt. In Tabelle 6.2 sind die Werte der Personenparameter und die dazugehörigen Prozentanteile in der Stichprobe dargestellt.

Tabelle 6.2: Verteilung der Personenparameter

Personenfähigkeit	Prozent
1	1%
1.5	4%
2	15%
3	30%
3.5	15%
4	15%
4.5	20%

6.1.3.1 Simulationen mit 100 Personen und 8 Items

Wie bereits erwähnt, wird hier der Durchschnitt der Parameterschätzer, die mittels der neuen Gamma-Funktionen geschätzt wurden, angegeben. In Tabelle 6.3 finden sich die Item- und Transferparameter des Kempf-Modells für 100 Personen und acht Items in jeweils Mitte-normierter und Null-Eins-normierter Form. [1] Die Schätzdauer betrug mit den neuen Gamma-Funktionen rund vier Minuten, im Vergleich dazu betrug die durchschnittliche Zeit mit den alten Gamma-Funktionen unter eine Minute.

Tabelle 6.3: Geschätzte Parameter des Kempf-Modells bei 100 Personen und 8 Items

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.7893	0.4559	1.3063	-0.1121
2	2.0794	0.7540	1.6380	0.2312
3	1.5628	0.4511	1.0917	-0.0891
4	1.6206	0.3371	1.1550	-0.2001
5	1.1975	0.5259	0.7294	0.0024
6	1.3978	0.7279	0.9302	0.2323
7	1.2927	0.6300	0.8352	0.1159
8	1.5032	0.4351	1.0566	-0.1395

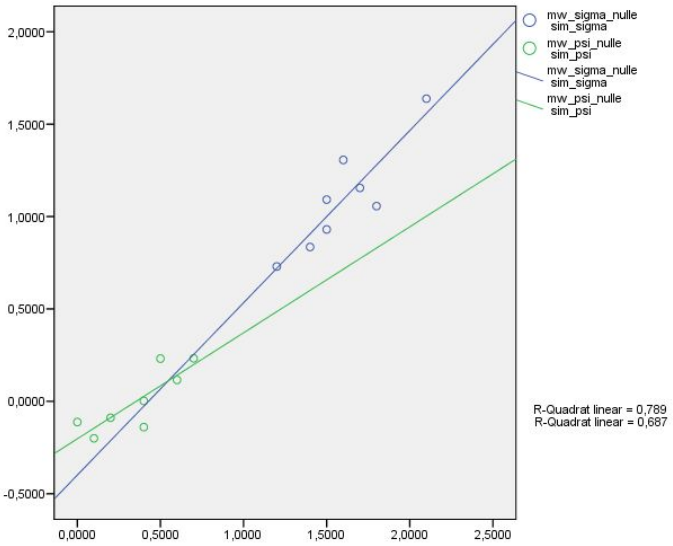
Die geschätzten Parameter weichen noch etwas von den ursprünglich simulierten Parameter ab. [2] In Abbildung 6.4 kann man diese Abweichungen deutlich erkennen. In SPSS wurde

[1] Es ist zu beachten, dass die Itemparameter zwar pro Item gelten, die Transferparameter aber pro (partiellem) Rohscore. Man kann deshalb nicht den ersten Transferparameter als zugehörig zu Item 1 interpretieren, wie dies bei den Itemparametern der Fall wäre. Der erste Transferparameter ist (siehe Abschnitt 4.1) also nicht der Lernparameter bei Item 1, sondern wenn vorher *null* Items, der zweite wenn vorher *ein* Item, egal welches, gelöst wurde usw.. Aus Gründen der Übersichtlichkeit werden für die Tabelle der Kempf-Modell-Parameter aber trotzdem Itemnummern angegeben.

[2] Natürlich können nicht 1 : 1 die selben Parameter herauskommen, da die geschätzten Parameter ja besonders

hier ein sog. „überlagertes Streudiagramm“ mit den durchschnittlichen Null-Eins-normierten Item- („mw_sigma_nulle“) und Transferparametern („mw_psi_nulle“) [3] und den Simulationsparametern („sim_sigma“ und „sim_psi“) erstellt sowie für beide Parameterpaare eine Regressionsgerade durch den Punkteschwarm gelegt. [4] Rechts neben dem Diagramm ist das Bestimmtheitsmaß R^2 für beide Regressionen angegeben. Das Bestimmtheitsmaß der Itemparameter liegt demzufolge bei 0.789, das der Transferparameter bei 0.687, was auf eine mittlere bis hohe Übereinstimmung hinweist. Die Korrelation bzw. das Bestimmtheitsmaß der Transferparameter liegen niedriger als die der Itemparameter, d.h. die Transferparameter werden ungenauer wiedergegeben.

Abbildung 6.4: Parameterschätzung bei 100 Personen und 8 Items



Dass die Unterschiede zwischen dem „Mitte-normierten“ und „Null-Eins-normierten“ Parametern (siehe Abschnitt 5.3) ebenfalls auf einer Lineartransformation beruhen, zeigt die folgende Abbildung 6.5. Es werden wieder im Rahmen eines überlagerten Streudiagramms die Regressionsgeraden zwischen den durchschnittlichen Mitte-normierten und Null-Eins-normierten Item- und Transferparametern durch den Punkteschwarm gelegt. Beide Normie-

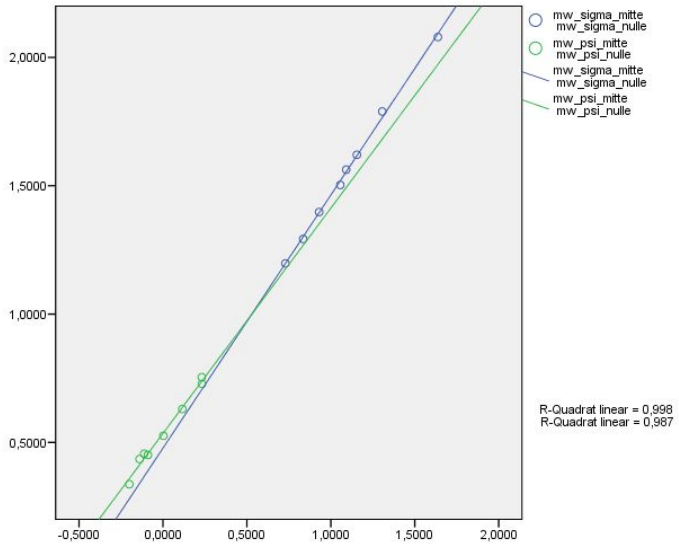
normiert sind, jedoch ist diese Normierung nur eine Lineartransformation und der Zusammenhang müsste was die Größenordnung angeht dennoch erkennbar sein.

[3] Im folgenden beziehen sich die Begriffe „mw_sigma_nulle“ und „mw_psi_nulle“ immer auf die mit den neuen Gamma-Funktionen geschätzten Parameter.

[4] Da der Leser höchstwahrscheinlich mit der Null-Eins-Normierung am meisten vertraut ist, werden die Parameter für die Streudiagramme und die Standardabweichungen bzw. Varianzen ausschließlich in dieser Form angegeben. Die nach (4.33) Mitte-normierten Parameter sind lediglich anders skaliert (siehe in Folge auch Abbildung 6.5 u.a.).

rungen hängen mit den Bestimmtheitsmaßen von 0.998 und 0.987 nahezu perfekt zusammen. Ungenauigkeiten können durch etwaige Rundungsfehler entstehen.

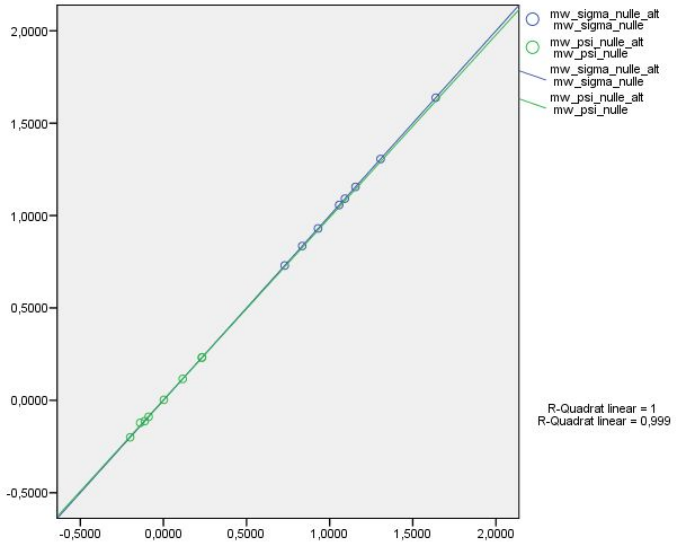
Abbildung 6.5: Parametertransformation bei 100 Personen und 8 Items



Es wurden, wie bereits weiter oben erwähnt, alle Datensätze einmal mit den originalen und einmal mit den neuen Gamma-Funktionen gerechnet. Abbildung 6.6 zeigt die beiden Schätzmethoden (Null-Eins-normiert) in einem überlagerten Streudiagramm mit eingezeichneten Regressionsgeraden. Die Schätzungen der alten Gamma-Funktionen werden mit „mw_sigma_nulle_alt“, die der neuen mit „mw_sigma_nulle“ bezeichnet, das selbe gilt auch für die Transferparameter Ψ. [5] Die Itemparameter hängen mit einem Bestimmtheitsmaß von 1 perfekt, die Transferparameter mit 0.999 fast perfekt zusammen.

[5] Diese Schreibweise wird auch bei den folgenden Simulationen beibehalten.

Abbildung 6.6: Vergleich alte vs. neue Gamma-Funktionen bei 100 Personen und 8 Items



Um einen Richtwert für die Genauigkeit der Schätzungen (mit den neuen Gamma-Funktionen) zu erhalten, wurden über die 100 Datensätze hinweg die Standardabweichungen und die Varianzen für jeden der Null-Eins-normierten acht Itemparameter gebildet. Die Ergebnisse sind in Tabelle 6.4 abzulesen. Die Streuungen für die Schätzer mit den alten Gamma-Funktionen sind mit diesen ident.

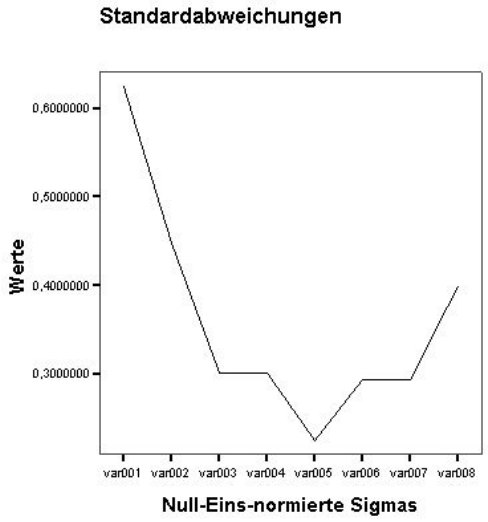
Tabelle 6.4: Statistiken der Null-Eins-normierten Schwierigkeitsparameter bei 100 Personen und 8 Items

Nr.	Standardabweichung	Varianz
1	0.6242	0.3896
2	0.4474	0.2001
3	0.3011	0.0907
4	0.3009	0.0906
5	0.2244	0.0503
6	0.2933	0.0860
7	0.2925	0.0856
8	0.3991	0.1593

Um eine genauere Vorstellung der Streuung der Schätzungen zu erhalten, wurde ein Liniendiagramm für die Standardabweichungen der Itemparameter erstellt. Dieses ist in Abbildung 6.7 zu sehen. Auf der X-Achse erfolgt die Einteilung in die acht Itemparameter, auf der Y-Achse sind die Werte der Standardabweichungen aufgetragen. Nach Abbildung 6.7 ist die

Streuung der Itemparameter beim ersten Parameter am höchsten, sinkt dann bis zum fünften ab und steigt schließlich bis zum letzten wiederum etwas an.

Abbildung 6.7: Standardabweichungen der Null-Eins-normierten Itemparameter bei 100 Personen und 8 Items



Für die acht (Null-Eins-normierten) Transferparameter wurden ebenfalls die Standardabweichungen und Varianzen berechnet und in Tabelle 6.5 dargestellt. [6]

Tabelle 6.5: Statistiken der Null-Eins-normierten Transferparameter bei 100 Personen und 8 Items

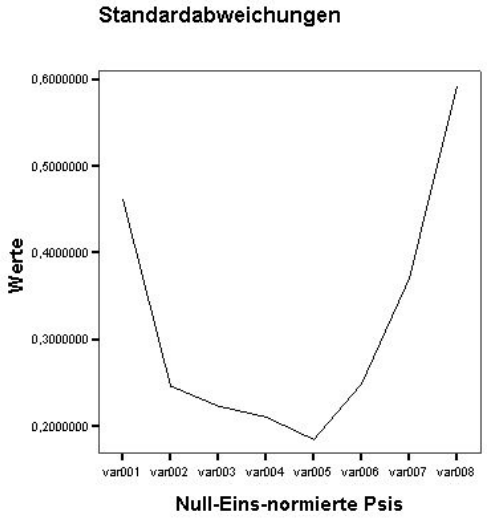
Nr.	Standardabweichung	Varianz
1	0.4616	0.2131
2	0.2460	0.0605
3	0.2228	0.0496
4	0.2102	0.0442
5	0.1843	0.0340
6	0.2484	0.0617
7	0.3703	0.1371
8	0.5926	0.3512

Es wurde auch ein Liniendiagramm für die Standardabweichungen der Transferparameter erstellt, das in Abbildung 6.8 zu sehen ist. Die Einteilung der X-Achse erfolgt nach den acht

[6] Achtung, die Nummer in der Tabelle bezieht sich auf den Transferparameter, nicht auf das Item. Man beachte, dass ja die Transferparameter vom partiellen Rohscore und nicht vom Item an sich abhängen.

Transferparametern, auf der Y-Achse sind wieder die Werte der Standardabweichungen eingetragen. Auch bei den Transferparametern ist die Streuung beim ersten groß, sinkt bis zum fünften weitgehend ab und steigt dann wieder steil an. Die Streuung beim letzten Parameter ist größer als beim ersten. [7]

Abbildung 6.8: Standardabweichungen der Null-Eins-normierten Transferparameter bei 100 Personen und 8 Items



6.1.3.2 Simulationen mit 500 Personen und 8 Items

Dieselben Item- und Transferparameter und das selbe Verhältnis der Personenparameter wie oben wurden verwendet, um 100 Datensätze mit 500 Personen zu simulieren. Für die Schätzung mit den neuen Gamma-Funktionen wurden in etwa sechs Minuten pro Datensatz benötigt, mit den alten eine. Tabelle 6.6 zeigt den Durchschnitt der Ergebnisse der Schätzungen für die Kempf-Modell-Parameter. Sie werden wiederum in Mitte- und Null-Eins-normierter Form angegeben und wurden mit den neuen Gamma-Funktionen berechnet.

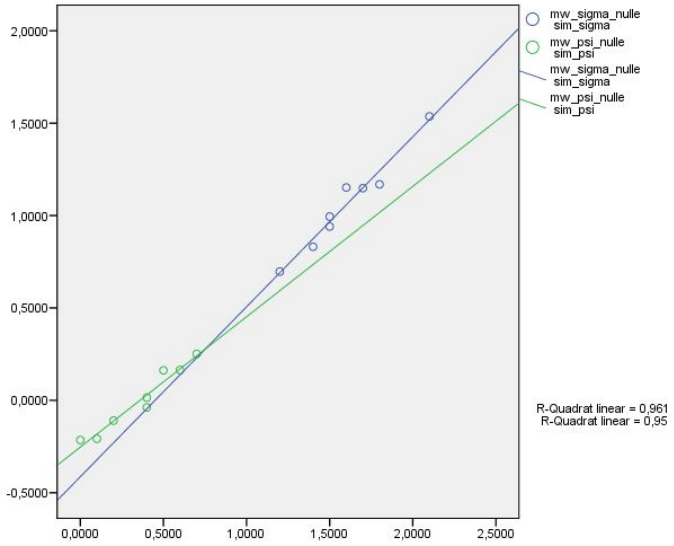
[7] Dieser Umstand ist mit unvorteilhaften N_{ri} (also der Anzahl an Personen, die Item i falsch beantworten, nachdem sie r richtig beantwortet hatten) verbunden, ein Problem das genauer in Abschnitt 6.1.4 erläutert wird. Bei acht Items ist dies jedoch nicht häufig der Fall und nicht so auffällig wie bei 20 Items. Die Streudiagramme, die die Genauigkeit der geschätzten Parameter anzeigen sind eher generell ungenau und weisen nicht nur einen besonderen Ausreißer auf.

Tabelle 6.6: Geschätzte Parameter des Kempf-Modells bei 500 Personen und 8 Items

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.6304	0.3595	1.1517	-0.2145
2	1.9714	0.6922	1.5364	0.1624
3	1.4617	0.4335	0.9942	-0.1098
4	1.5893	0.3340	1.1482	-0.2078
5	1.1642	0.5236	0.6967	0.0146
6	1.3833	0.7403	0.9404	0.2514
7	1.2731	0.6546	0.8312	0.1640
8	1.5770	0.4615	1.1689	-0.0383

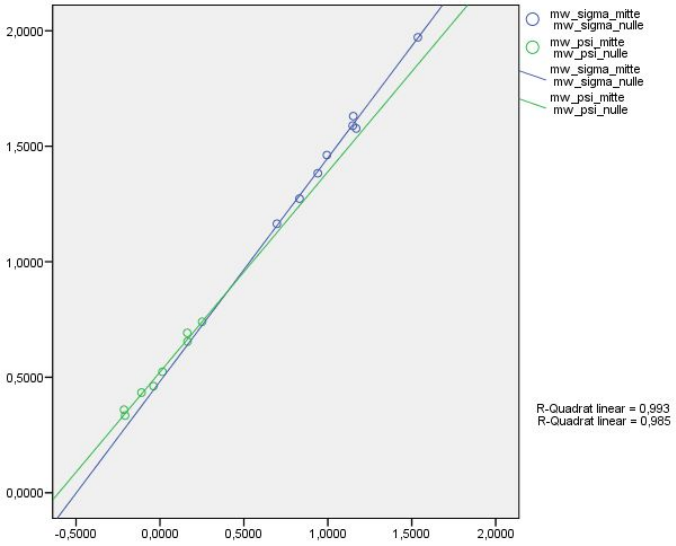
Diese Parameter liegen näher an den ursprünglich simulierten als bei 100 Personen. Abbildung 6.9 zeigt ein überlagertes Streudiagramm der simulierten und dem Durchschnittswert der geschätzten Parameter. Es sind auch die beiden Regressionsgeraden eingezeichnet und das Bestimmtheitsmaß angegeben. Für die Itemparameter liegt das Bestimmtheitsmaß nun bei 0.961, für die Transferparameter bei 0.95. Die Transferparameter wurden also immer noch leicht ungenauer geschätzt.

Abbildung 6.9: Parameterschätzung bei 500 Personen und 8 Items



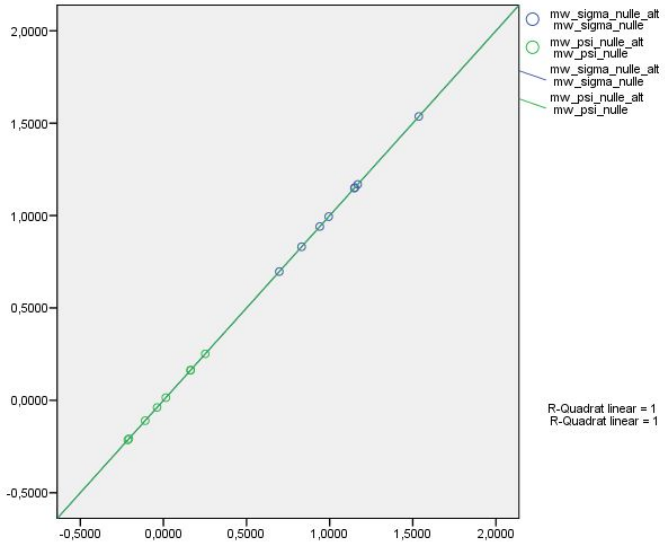
Der Zusammenhang der beiden beiden Normierungsarten wird in Abbildung 6.10 durch ein überlagertes Streudiagramm verdeutlicht. Die Bestimmtheitsmaße liegen bei 0.993 für die Item- und bei 0.985 für die Transferparameter.

Abbildung 6.10: Parametertransformation bei 500 Personen und 8 Items



Der Zusammenhang zwischen den Schätzungen mit den alten und den neuen Gamma-Funktionen ist bei 500 Personen perfekt. Abbildung 6.11 zeigt ein Bestimmtheitsmaß von 1 für Item- und Transferparameter.

Abbildung 6.11: Vergleich alte vs. neue Gamma-Funktionen bei 500 Personen und 8 Items



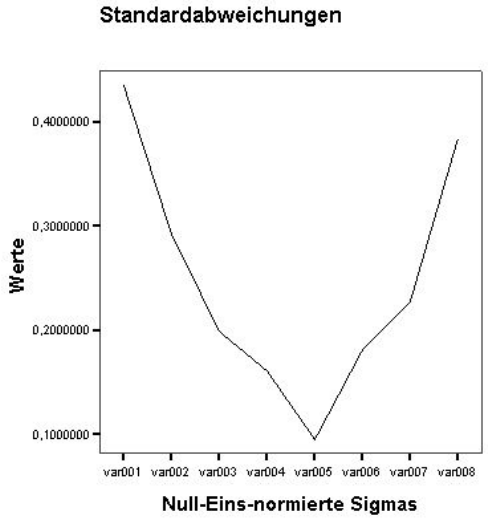
Für die Null-Eins-normierten Schätzungen der Itemparameter wurden auch hier Streuungsmaße berechnet. Standardabweichungen und Varianzen pro Itemparameter sind in Tabelle 6.7 dargestellt.

Tabelle 6.7: Statistiken der Null-Eins-normierten Itemparameter bei 500 Personen und 8 Items

Nr.	Standardabweichung	Varianz
1	0.4346	0.1889
2	0.2918	0.0851
3	0.1986	0.0394
4	0.1607	0.0258
5	0.0948	0.0090
6	0.1810	0.0328
7	0.2275	0.0518
8	0.3843	0.1477

Es wurde wiederum ein Liniendiagramm der Standardabweichungen erstellt (siehe Abbildung 6.12). Auf der X-Achse sind die Itemparameternummern, auf der Y-Achse die Werte aufgetragen. Man kann abermals eine hohe Streuung zu Anfang und am Ende erkennen, bei Itemparameter fünf ist wieder die wenigste Streuung in den Schätzungen vorhanden.

Abbildung 6.12: Standardabweichungen der Null-Eins-normierten Itemparameter bei 500 Personen und 8 Items



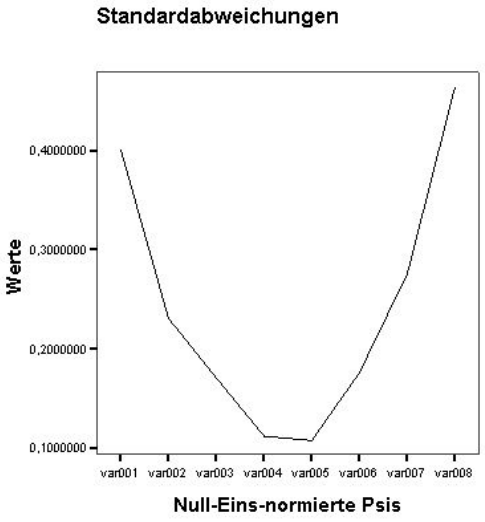
Die selbe Prozedur wurde auch für die Transferparameter bei 500 Personen durchgeführt, wie Tabelle 6.8 zeigt.

Tabelle 6.8: Statistiken der Null-Eins-normierten Transferparameter bei 500 Personen und 8 Items

Nr.	Standardabweichung	Varianz
1	0.4004	0.1603
2	0.2303	0.0531
3	0.1704	0.0290
4	0.1117	0.0125
5	0.1073	0.0115
6	0.1760	0.0310
7	0.2756	0.0759
8	0.4641	0.2154

Die graphische Anschauung der Standardabweichungen zeigt Abbildung 6.13. Die Streuung ist beim ersten Parameter hoch, fällt bis zum vierten und fünften ab und steigt bis zum letzten wieder stark an.

Abbildung 6.13: Standardabweichungen der Null-Eins-normierten Transferparameter bei 500 Personen und 8 Items



6.1.3.3 Simulationen mit 1000 Personen und 8 Items

Auch für 1000 Personen wiederholt sich das selbe Spiel mit den simulierten Parametern und der Bildung des Durchschnitts der Parameterschätzer für die 100 Datensätze. Es wurden durchschnittlich ungefähr sechs Minuten pro Datensatz für die Schätzung mit den neuen Gamma-Funktionen gebraucht, mit den alten wieder eine Minute. In Tabelle 6.9 sind die

Durchschnittswerte der geschätzten Item- und Transferparameter des Kempf-Modells abzulesen.

Tabelle 6.9: Geschätzte Parameter des Kempf-Modells bei 1000 Personen und 8 Items

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.6051	0.3065	1.1072	-0.2658
2	1.9604	0.6635	1.4983	0.1284
3	1.4601	0.4078	0.9743	-0.1372
4	1.6094	0.3190	1.1414	-0.2222
5	1.1786	0.5334	0.6943	0.0158
6	1.3900	0.7520	0.9319	0.2583
7	1.3077	0.6640	0.8518	0.1739
8	1.6096	0.5457	1.1794	0.0601

Abbildung 6.14 zeigt, dass nun bei 1000 Personen die simulierten und die geschätzten Parameter noch stärker zusammenhängen. Im überlagerten Streudiagramm mit Regressionsgeraden ist diesmal ein Bestimmtheitsmaß für die Itemparameter von 0.983 und von 0.991 für die Transferparameter zu sehen.

Abbildung 6.14: Parameterschätzung bei 1000 Personen und 8 Items

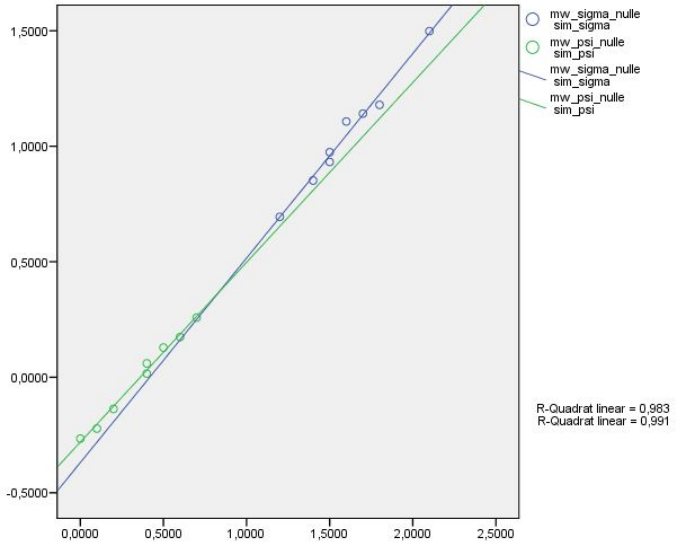
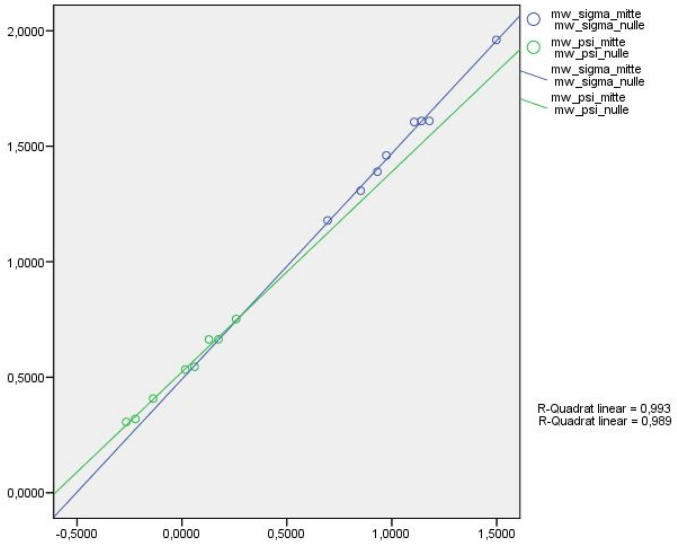


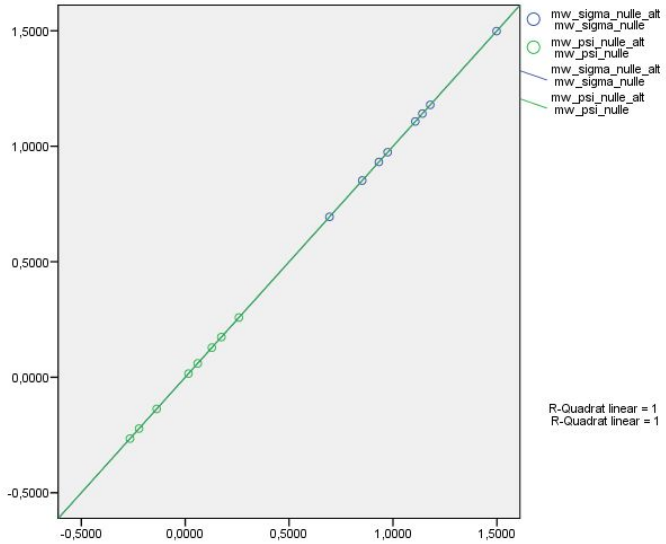
Abbildung 6.15 zeigt ein überlagertes Streudiagramm für die beiden Normierungen. Die beiden Bestimtheitsmaße liegen bei 1000 Personen nun bei 0.993 und 0.989.

Abbildung 6.15: Parametertransformation bei 1000 Personen und 8 Items



Die alte und neue Methode der Gamma-Funktionsschätzung ist auch bei 1000 Personen wieder ident. Abbildung 6.16 zeigt Bestimmtheitsmaße zwischen beiden Varianten von jeweils 1.

Abbildung 6.16: Vergleich alte vs. neue Gamma-Funktionen bei 1000 Personen und 8 Items



Die Streuungsmaße, also Standardabweichung und Varianz für die Null-Eins-normierten Schätzungen der Itemparameter werden in Tabelle 6.10 aufgeführt.

Tabelle 6.10: Statistiken der Null-Eins-normierten Itemparameter bei 1000 Personen und 8 Items

Nr.	Standardabweichung	Varianz
1	0.3715	0.1380
2	0.2215	0.0490
3	0.1661	0.0276
4	0.1013	0.0103
5	0.0613	0.0038
6	0.1608	0.0258
7	0.2180	0.0475
8	0.3371	0.1136

Abbildung 6.17 zeigt die Standardabweichungen für die acht Itemparameter. Ähnlich wie weiter oben bildet das Liniendiagramm ein “U“ mit der wenigsten Streuung beim fünften Item.

Abbildung 6.17: Standardabweichungen der Null-Eins-normierten Itemparameter bei 1000 Personen und 8 Items

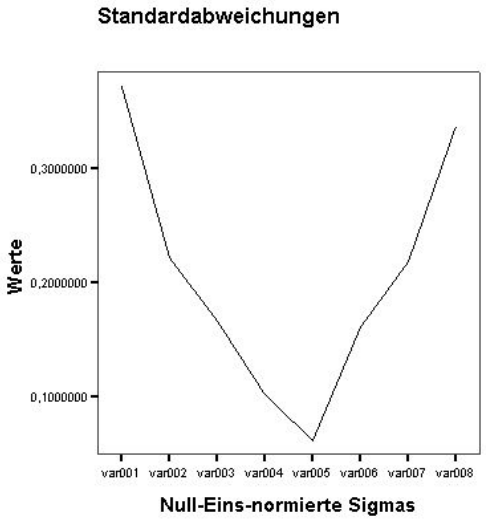


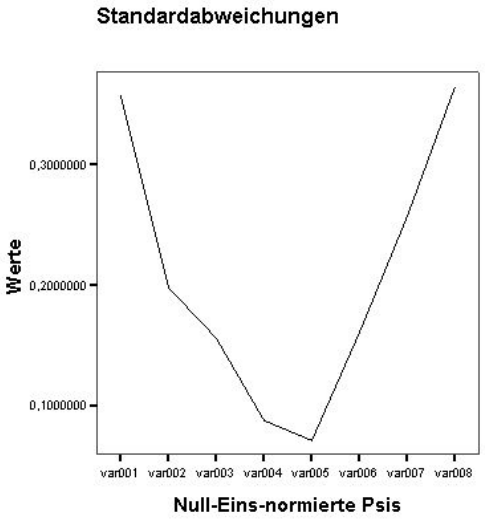
Tabelle 6.11 enthält die Standardabweichungen und Varianzen für die acht Transferparameter.

Tabelle 6.11: Statistiken der Null-Eins-normierten Transferparameter bei 1000 Personen und 8 Items

Nr.	Standardabweichung	Varianz
1	0.3568	0.1273
2	0.1979	0.0391
3	0.1558	0.0243
4	0.0875	0.0077
5	0.0709	0.0050
6	0.1608	0.0259
7	0.2572	0.0661
8	0.3647	0.1330

Die Standardabweichungen werden in Abbildung 6.18 graphisch dargestellt. Der erste und letzte Transferparameter weisen wiederum die größten Streuungen auf, der vierte und fünfte die kleinsten.

Abbildung 6.18: Standardabweichungen der Null-Eins-normierten Transferparameter bei 1000 Personen und 8 Items



6.1.3.4 Simulationen mit 5000 Personen und 8 Items

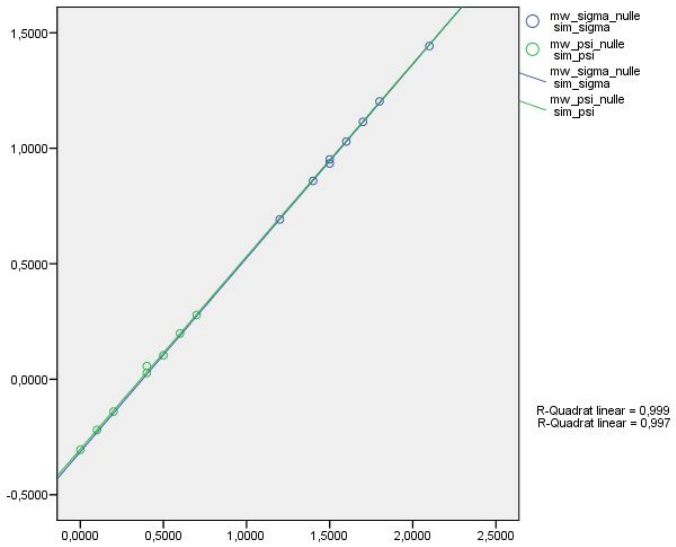
Nach altbewährter Manier wurden mit den gleichen Parametern 100 Datensätze mit 5000 Personen simuliert. Die Schätzdauer mit den neuen Gamma-Funktionen betrug etwa sieben Minuten, mit den alten Gamma-Funktionen wieder eine Minute. Der Durchschnitt der Schätzungen mit den neuen Gamma-Funktionen ist in Tabelle 6.12 aufgeführt.

Tabelle 6.12: Geschätzte Parameter des Kempf-Modells bei 5000 Personen und 8 Items

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.5602	0.2329	1.0292	-0.3052
2	1.9638	0.6309	1.4426	0.1038
3	1.4536	0.3842	0.9342	-0.1401
4	1.6272	0.3005	1.1142	-0.2195
5	1.2023	0.5398	0.6918	0.0272
6	1.4537	0.7822	0.9516	0.2784
7	1.3567	0.6997	0.8587	0.1984
8	1.6929	0.5496	1.2030	0.0571

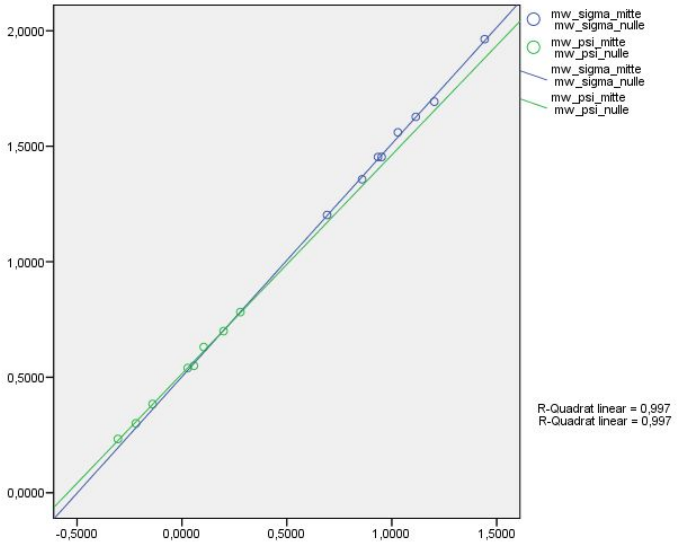
Der Durchschnitt der geschätzten Parameter liegt bei 5000 Personen nun schon der nahe an den gewählten simulierten Werten. Abbildung 6.19 zeigt das überlagerte Streudiagramm für die Null-Eins-normierten und simulierten Item- und Transferparameter mit beiden Regressionsgeraden und Bestimmtheitsmaßen. Die Itemparameter hängen nun mit einem Bestimmtheitsmaß von 0.999, die Transferparameter mit 0.997 fast perfekt mit den simulierten Parametern zusammen.

Abbildung 6.19: Parameterschätzung bei 5000 Personen und 8 Items



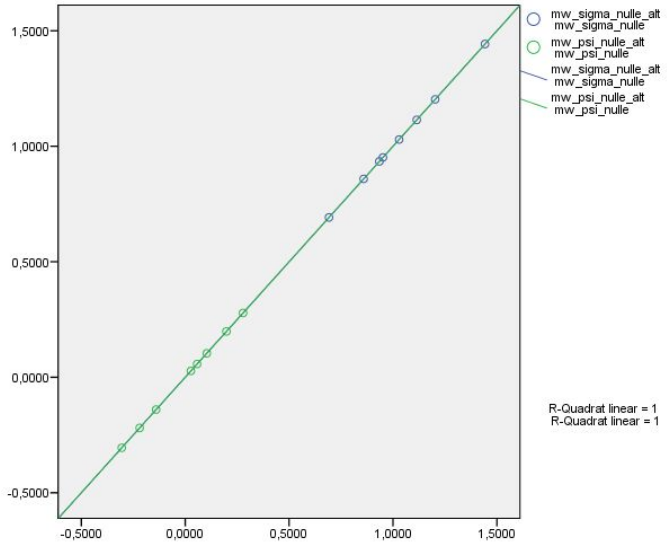
Auch der Durchschnitt beider Normierungen nähert sich mehr aneinander an. Wie aus dem Streudiagramm in Abbildung 6.20 ersichtlich, hängen die Mitte-Normierung und die Null-Eins-Normierung nun mit Bestimmtheitsmaßen von jeweils 0.997 bei Item- und Transferparametern zusammen.

Abbildung 6.20: Parametertransformation bei 5000 Personen und 8 Items



Die Ergebnisse der alten Gamma-Funktionen hängen mit denen der neuen wiederum perfekt zusammen. Abbildung 6.21 zeigt ein überlagertes Streudiagramm mit den beiden Bestimmtheitsmaßen von 1.

Abbildung 6.21: Vergleich alte vs. neue Gamma-Funktionen bei 5000 Personen und 8 Items



Die Standardabweichungen und Varianzen der Null-Eins-normierten Itemparameterschätzungen sind in Tabelle 6.13 zu sehen.

Tabelle 6.13: Statistiken der Null-Eins-normierten Itemparameter bei 5000 Personen und 8 Items

Nr.	Standardabweichung	Varianz
1	0.1851	0.0343
2	0.1123	0.0126
3	0.0809	0.0065
4	0.0494	0.0024
5	0.0299	0.0009
6	0.0802	0.0064
7	0.1200	0.0144
8	0.1836	0.0337

Wie bei den vorangegangenen Simulationen, weisen der erste und der letzte Itemparameter auch hier die größte, der fünfte die kleinste Standardabweichung auf. Dies kann man in Abbildung 6.22 gut erkennen.

Abbildung 6.22: Standardabweichungen der Null-Eins-normierten Itemparameter bei 5000 Personen und 8 Items

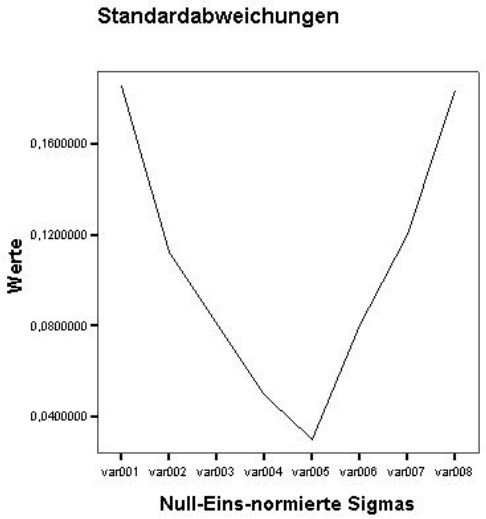


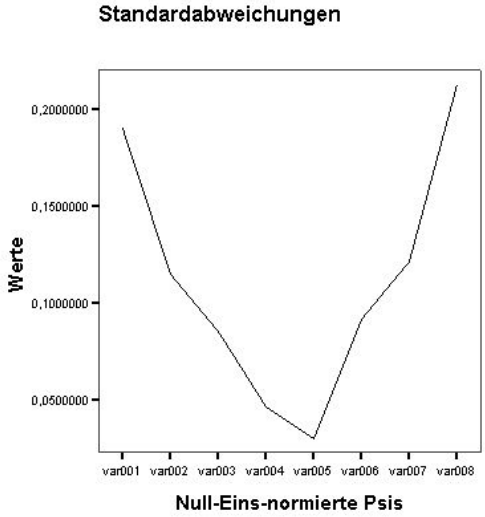
Tabelle 6.14 zeigt die Standardabweichungen und Varianzen für die Null-Eins-normierten geschätzten Transferparameter.

Tabelle 6.14: Statistiken der Null-Eins-normierten Transferparameter bei 5000 Personen und 8 Items

Nr.	Standardabweichung	Varianz
1	0.1900	0.0361
2	0.1149	0.0132
3	0.0854	0.0073
4	0.0466	0.0022
5	0.0302	0.0009
6	0.0917	0.0084
7	0.1215	0.0148
8	0.2127	0.0452

Auch hier ist wieder der U-förmige Verlauf erkennbar. In Abbildung 6.23 kann man erkennen, dass die Streuungen beim ersten und letzten Parameter wieder am höchsten sind. Das Minimum der Streuung liegt beim fünften Parameter.

Abbildung 6.23: Standardabweichungen der Null-Eins-normierten Transferparameter bei 5000 Personen und 8 Items



6.1.3.5 Simulationen mit 100000 Personen und 8 Items

Für acht Items wurden schließlich zu guter Letzt noch 100 Datensätze mit 100000 Personen simuliert. Die Simulationsparameter decken sich wiederum mit den obigen. Die Schätzdauer mit den neuen Gamma-Funktionen betrug durchschnittlich 10 Minuten, mit den alten Gamma-Funktionen wieder etwa eine Minute. In Tabelle 6.15 werden die Durchschnittspa-

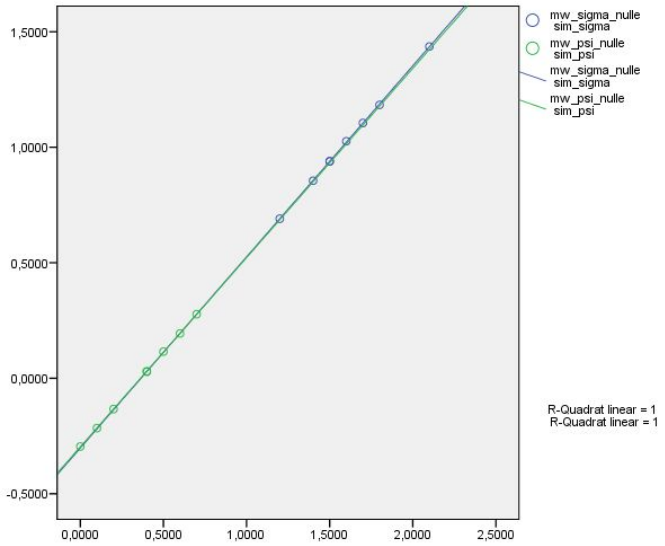
rameter aus den geschätzten Kempf-Modell-Parametern dargestellt.

Tabelle 6.15: Geschätzte Parameter des Kempf-Modells bei 100000 Personen und 8 Items

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.5542	0.2109	1.0259	-0.2955
2	1.9700	0.6271	1.4361	0.1152
3	1.4648	0.3739	0.9396	-0.1335
4	1.6324	0.2897	1.1051	-0.2158
5	1.2106	0.5394	0.6906	0.0305
6	1.4618	0.7894	0.9385	0.2772
7	1.3765	0.7041	0.8551	0.1938
8	1.7093	0.5348	1.1834	0.0281

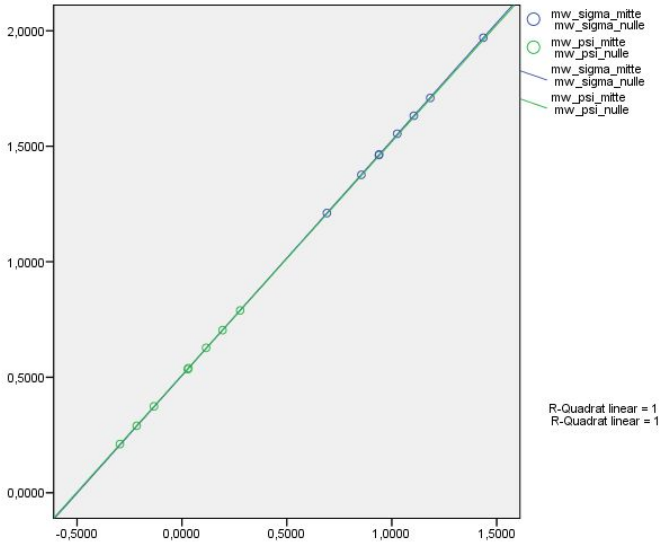
Bei 100000 Personen decken sich die geschätzten mit den ursprünglichen Parametern perfekt. Abbildung 6.24 zeigt das überlagerte Streudiagramm für die Null-Eins-normierten Item- und Transferparameter mit eingezeichneten Regressionsgeraden. Beide Bestimmtheitsmaße sind 1.

Abbildung 6.24: Parameterschätzung bei 100000 Personen und 8 Items



Die Mitte-normierten-Parameter entsprechen den Null-Eins-normierten Parametern ebenfalls perfekt. In Abbildung 6.25 zeigt das Streudiagramm mit Regressionsgeraden Bestimmtheitsmaße von 1 für Item- und Transferparameter.

Abbildung 6.25: Parametertransformation bei 100000 Personen und 8 Items



Die Parameter, die mit den alten bzw. neuen Gamma-Funktionen geschätzt wurden, entsprechen einander auch bei 100000 Personen wieder. Abbildung 6.26 zeigt das überlagerte Streudiagramm mit Bestimmtheitsmaßen von jeweils 1 für Item- und Transferparameter.

Abbildung 6.26: Vergleich alte vs. neue Gamma-Funktionen bei 100000 Personen und 8 Items

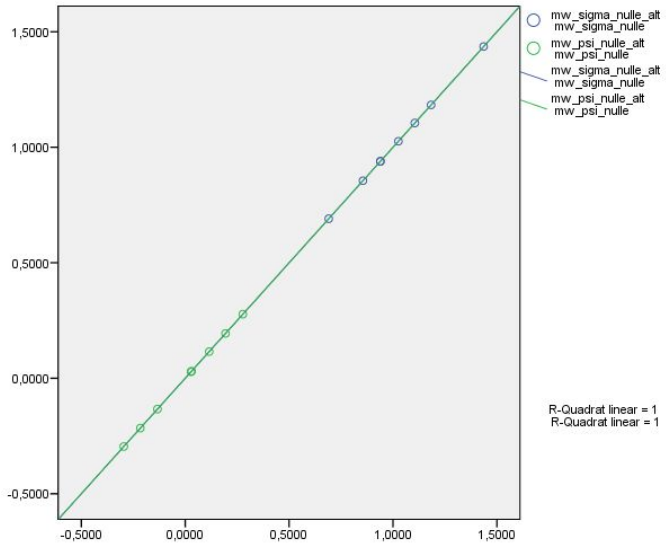


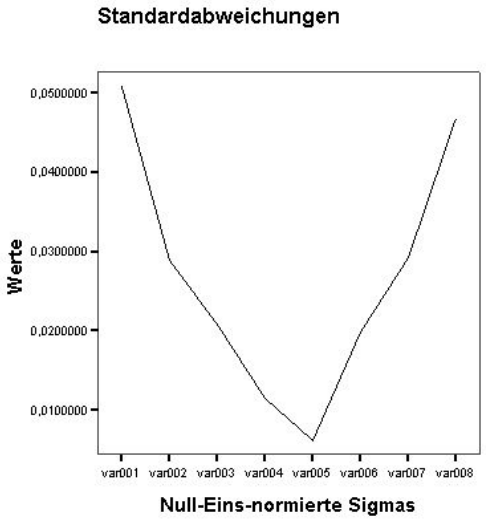
Tabelle 6.16 enthält die Streuungsmaße für die geschätzten Null-Eins-normierten Itemparameter.

Tabelle 6.16: Statistiken der Null-Eins-normierten Itemparameter bei 100000 Personen und 8 Items

Nr.	Standardabweichung	Varianz
1	0.0507	0.0026
2	0.0288	0.0008
3	0.0207	0.0004
4	0.0114	0.0001
5	0.0061	0.0000
6	0.0198	0.0004
7	0.0291	0.0008
8	0.0468	0.0022

Anhand von Abbildung 6.27 kann man erkennen, dass die Standardabweichung beim ersten Parameter am größten ist, beim fünften am kleinsten und sie dann wieder bis zum letzten Parameter ansteigt.

Abbildung 6.27: Standardabweichungen der Null-Eins-normierten Itemparameter bei 100000 Personen und 8 Items



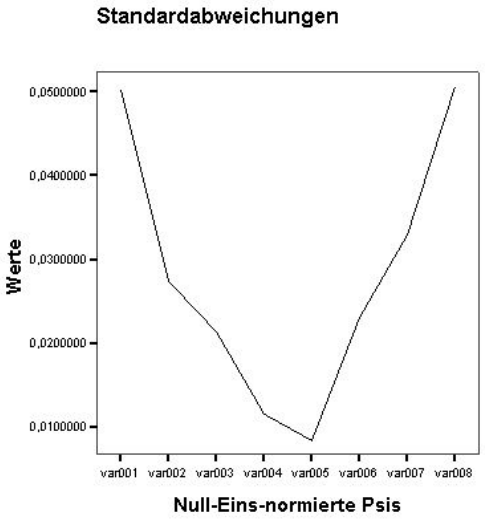
Die Standardabweichungen und Varianzen wurden auch für die Null-Eins-normierten geschätzten Transferparameter berechnet und in Tabelle 6.17 dargestellt.

Tabelle 6.17: Statistiken der Null-Eins-normierten Transferparameter bei 100000 Personen und 8 Items

Nr.	Standardabweichung	Varianz
1	0.0501	0.0025
2	0.0274	0.0007
3	0.0213	0.0005
4	0.0115	0.0001
5	0.0084	0.0000
6	0.0230	0.0005
7	0.0329	0.0011
8	0.0506	0.0026

Abbildung 6.28 zeigt die Standardabweichungen für die Transferparameter. Die Streuungen sind wieder beim ersten und letzten Parameter am größten und beim vierten und fünften am kleinsten.

Abbildung 6.28: Standardabweichungen der Null-Eins-normierten Transferparameter bei 100000 Personen und 8 Items



6.1.4 Ergebnisse einer Simulationsreihe mit 20 Items

Es wurde eine weitere Simulationsreihe mit je 100 Datensätzen mit 20 Items und 500, 1000 und 5000 Personen durchgeführt. Datensätze mit nur 100 Personen wurden zwar simuliert, jedoch wurde die Schätzung bei zwei Drittel der 100 Datensätze wegen zu großer Rechenungenauigkeit abgebrochen. Für 20 Items sind 100 Personen mitunter zu wenig. Da die Rechen-

zeiten bei 20 Items deutlich über der Schätzdauer bei 8 Items liegen, wurde in diesem Fall aus praktischen Gründen auf die Schätzung der Datensätze mit 100000 Personen verzichtet. Noch dazu ergaben sich bereits für 5000 Personen gut mit den Simulationsparametern übereinstimmende Schätzungen.

Es muss jedoch erwähnt werden, dass hier eine besondere Schwierigkeit auftrat. DynTest berechnet die Matrix der N_{ri}, also der Anzahl der Personen, die eine falsche Antwort auf ein Item i nach r richtigen Antworten auf vorherige Items gegeben haben. Der letzte dieser N_{ri} war bei 20 Items oft gleich 0, weil einfach nicht so viele verschiedene Antwortmuster vorhanden waren. Wenn dies der Fall war, wurde der letzte Transferparameter immer extrem hoch und damit falsch geschätzt und sorgte so für einen Ausreißer im Durchschnitt und somit auch im Streudiagramm. Daher wurden diese Simulationen durch andere ersetzt, bei denen der letzte N_{ri} vorteilhafter war, damit jeweils 100 Datensätze zum Vergleichen erhalten blieben. Bei steigender Personenzahl trat diese Schwierigkeit zunehmend weniger auf. Durch diese Selektion jedoch waren die Übereinstimmungen der simulierten mit den geschätzten Parametern etwas genauer als bei acht Items, da dort eine solche Selektion nicht vorgenommen wurde. [8]

Die Parameter für die Simulation wurden abermals von der Autorin willkürlich ohne spezielle Normierungen festgelegt. Tabelle 6.18 enthält die Item- und Transferparameter, mit denen die Datensätze generiert wurden.

[8] Bei acht Items trat dieses Problem in wesentlich geringerem Maßstab auf und der letzte simulierte Transferparameter war auch nicht so niedrig angesetzt wie der letzte bei 20 Items (0.4 und nicht 0.1), so dass Unterschiede in den Streudiagrammen, die die Genauigkeit der Schätzungen anzeigen, dort nicht auffielen. Es war kein solcher „Ausreißer-Parameter“ ersichtlich, sondern nur eine allgemeine Ungenauigkeit. Eine große Streuung beim letzten Parameter war bei acht Items aber auch zu beobachten.

Tabelle 6.18: Ausgangsparameter der Simulation bei 20 Items

Item	Itemschwierigkeit	Transfer
1	1.6	0
2	2.1	0.5
3	1.5	0.2
4	1.7	0.1
5	1.2	0.4
6	1.5	0.7
7	1.4	0.6
8	1.8	0.4
9	2.0	0.1
10	0.9	0.2
11	1.5	0.7
12	1.2	0.2
13	2.1	0.1
14	1.1	0.3
15	1.3	0.8
16	1.4	0.6
17	1.8	0.5
18	0.8	0.4
19	1.5	0.4
20	1.3	0.1

Die Werte der Personenparameter sowie die Verteilung wurde beibehalten (siehe Tabellen 6.1 und 6.2).

6.1.4.1 Simulationen mit 500 Personen und 20 Items

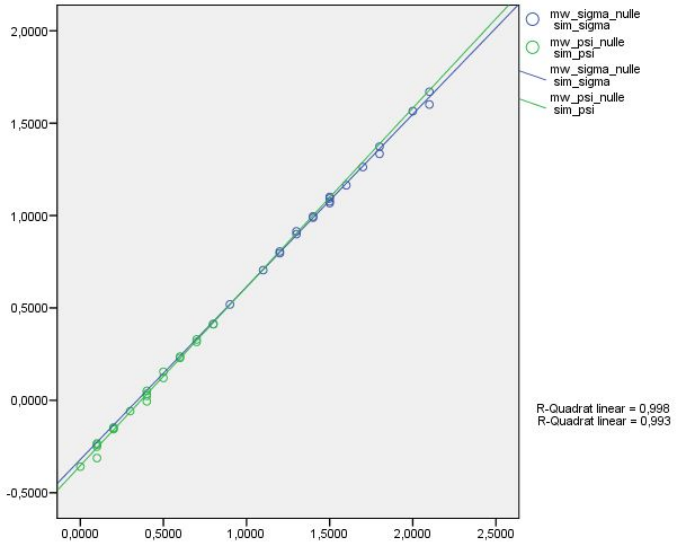
Mit den angegebenen Parametern wurden zunächst 100 Datensätze mit 500 Personen generiert, diese geschätzt und schließlich der Durchschnitt aus den Schätzungen gebildet. Mit den neuen Grundfunktion dauerte die Schätzung pro Datensatz ungefähr acht Minuten, mit den alten Gamma-Funktionen ca. drei Minuten. Die Durchschnittsschätzungen mit den neuen Gamma-Funktionen werden in Tabelle 6.19 dargestellt.

Tabelle 6.19: Geschätzte Parameter des Kempf-Modells bei 500 Personen und 20 Items

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.9871	0.0760	1.1637	-0.3589
2	2.5147	0.6640	1.6008	0.1214
3	1.8459	0.3182	1.0674	-0.1558
4	2.0895	0.2107	1.2623	-0.2399
5	1.5023	0.5487	0.7963	0.0359
6	1.8650	0.8930	1.0922	0.3165
7	1.7315	0.7921	0.9887	0.2368
8	2.2092	0.5625	1.3734	0.0511
9	2.4481	0.2069	1.5650	-0.2328
10	1.1446	0.3078	0.5190	-0.1515
11	1.8665	0.8998	1.1009	0.3297
12	1.4971	0.3073	0.8056	-0.1471
13	2.5682	0.1781	1.6698	-0.2503
14	1.3667	0.4148	0.7045	-0.0585
15	1.6282	1.0000	0.9147	0.4126
16	1.7277	0.7689	0.9951	0.2295
17	2.1442	0.6752	1.3340	0.1548
18	1.0000	0.4854	0.4127	-0.0055
19	1.8242	0.5143	1.0766	0.0245
20	1.6058	0.1190	0.8996	-0.3123

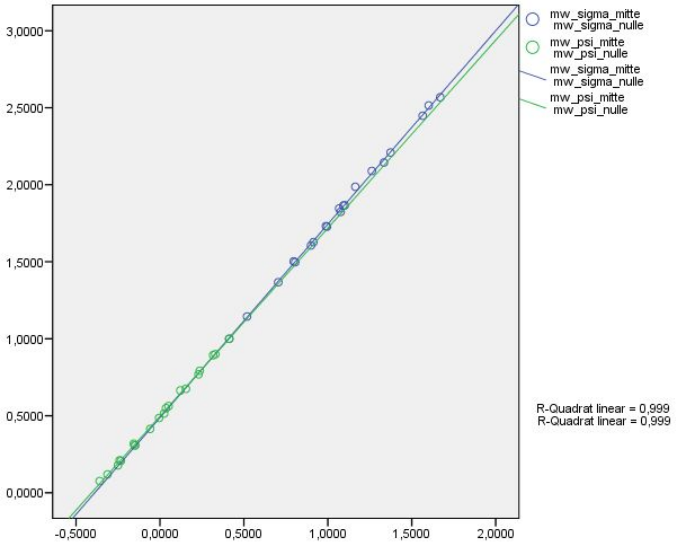
Abbildung 6.29 zeigt, dass die geschätzten Parameter relativ gut mit den geschätzten Parametern übereinstimmen. Es wurde wieder ein überlagertes Streudiagramm mit Regressionsgeraden erstellt. Das Bestimmtheitsmaß für die Itemparameter liegt bei 0.998, das für die Transferparameter bei 0.993.

Abbildung 6.29: Parameterschätzung bei 500 Personen und 20 Items



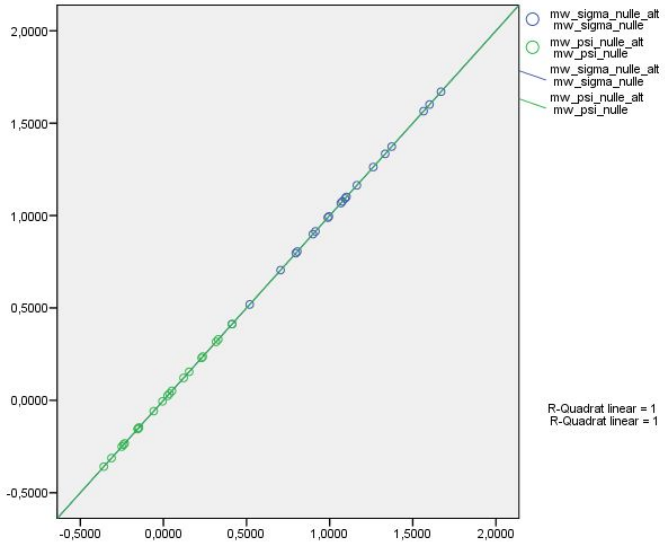
Den Zusammenhang der beiden Normierungsarten zeigt Abbildung 6.30. Die Mitte-normierten und Null-Eins-normierten Item- und Transferparameter verbindet ein Bestimmtheitsmaß von je 0.999.

Abbildung 6.30: Parametertransformation bei 500 Personen und 20 Items



Die Schätzungen mit alten und neuen Gamma-Funktionen stimmen ebenfalls miteinander überein. Abbildung 6.31 zeigt das überlagerte Streudiagramm und Bestimmtheitsmaße von jeweils 1 für die Null-Eins-normierten Item- und Transferparameter.

Abbildung 6.31: Vergleich alte vs. neue Gamma-Funktionen bei 500 Personen und 20 Items



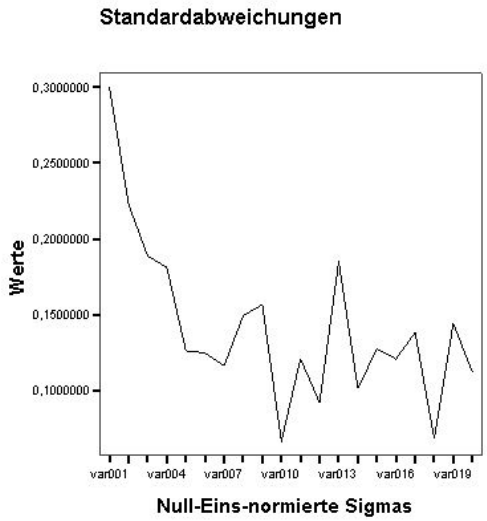
Für die Simulationsreihen mit 20 Items wurden ebenfalls Statistiken für die Streuung der Parameterschätzungen erstellt. Tabelle 6.20 enthält die Standardabweichungen und Varianzen für die Null-Eins-normierten Itemparameter.

Tabelle 6.20: Statistiken der Null-Eins-normierten Itemparameter bei 500 Personen und 20 Items

Nr.	Standardabweichung	Varianz
1	0.2996	0.0898
2	0.2225	0.0495
3	0.1885	0.0356
4	0.1811	0.0328
5	0.1261	0.0159
6	0.1248	0.0156
7	0.1162	0.0135
8	0.1493	0.0223
9	0.1566	0.0245
10	0.0663	0.0044
11	0.1209	0.0146
12	0.0919	0.0084
13	0.1850	0.0342
14	0.1015	0.0103
15	0.1272	0.0162
16	0.1207	0.0146
17	0.1382	0.0191
18	0.0684	0.0047
19	0.1442	0.0208
20	0.1119	0.0125

Auch hier wurde zur graphischen Veranschaulichung mit SPSS ein Liniendiagramm für die Standardabweichungen erstellt (Abbildung 6.32). Die X-Achse markiert wieder die einzelnen Itemparameternummern, wobei aus Schlichtheitsgründen hier nur jede dritte aufgetragen wurde. Auf der Y-Achse sind wieder die Werte für die Standardabweichungen aufgetragen. Man erkennt die höchste Streuung beim ersten Parameter und dann wechselnde Täler und Spitzen. Besonders niedrige Streuung trat beim 10. und 18. Parameter auf.

Abbildung 6.32: Standardabweichungen der Null-Eins-normierten Itemparameter bei 500 Personen und 20 Items



Auch für die Transferparameter wurden Streuungsstatistiken erstellt. Tabelle 6.21 stellt die Standardabweichungen und Varianzen für die Null-Eins-normierten Transferparameter dar.

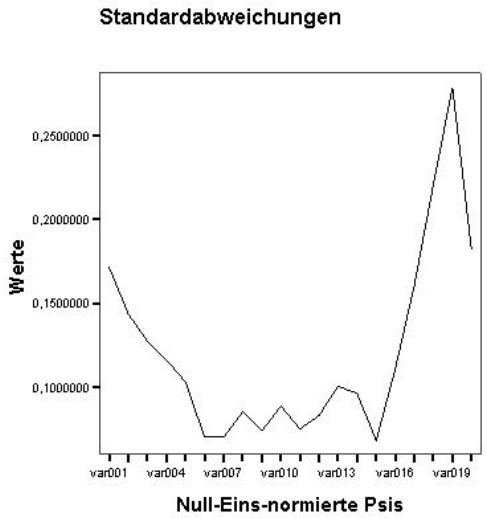
Tabelle 6.21: Statistiken der Null-Eins-normierten Transferparameter bei 500 Personen und 20 Items

Item	Standardabweichung	Varianz
1	0.1714	0.0294
2	0.1437	0.0207
3	0.1275	0.0163
4	0.1164	0.0136
5	0.1031	0.0106
6	0.0703	0.0049
7	0.0704	0.0050
8	0.0855	0.0073
9	0.0739	0.0055
10	0.0888	0.0079
11	0.0753	0.0057
12	0.0832	0.0069
13	0.1006	0.0101
14	0.0963	0.0093
15	0.0684	0.0047
16	0.1110	0.0123
17	0.1607	0.0258
18	0.2211	0.0489
19	0.2787	0.0777
20	0.1819	0.0331

Abbildung 6.33 zeigt eine besonders hohe Streuung beim vorletzten, also 19. Parameter, die dann beim letzten wieder abfällt [9]. Besonders wenig Streuung trat beim sechsten, siebten und 15. Transferparameter auf. Die Form erinnert im weitesten Sinne an ein „U“, weist aber dennoch kleine Spitzen und Täler auf.

[9] Diese hohe Streuung weist auch auf zu kleine vorletzte N_{ri} hin, die jedoch nicht gleich Null waren.

Abbildung 6.33: Standardabweichungen der Null-Eins-normierten Transferparameter bei 500 Personen und 20 Items



6.1.4.2 Simulationen mit 1000 Personen und 20 Items

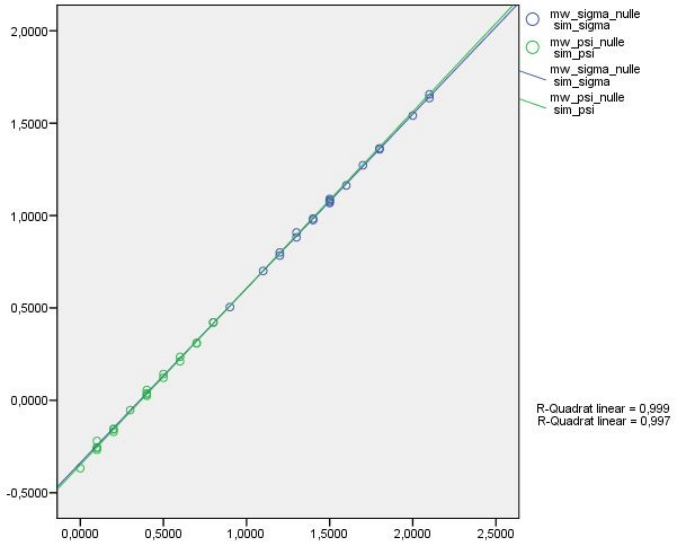
Für 1000 Personen wurden ebenfalls 100 Datensätze simuliert und geschätzt. Die Schätzdauer mit den neuen Gamma-Funktionen betrug hier um die 12 Minuten, mit den alten nur ca. fünf. Tabelle 6.22 enthält die durchschnittlich geschätzten Parameter des Kempf-Modells.

Tabelle 6.22: Geschätzte Parameter des Kempf-Modells bei 1000 Personen und 20 Items

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.9819	0.0304	1.1629	-0.3674
2	2.5717	0.6452	1.6359	0.1218
3	1.8493	0.2704	1.0679	-0.1706
4	2.1036	0.1442	1.2722	-0.2674
5	1.4791	0.5126	0.7831	0.0250
6	1.8460	0.8705	1.0746	0.3106
7	1.7152	0.7461	0.9755	0.2109
8	2.2032	0.5147	1.3579	0.0310
9	2.4315	0.1529	1.5412	-0.2555
10	1.1183	0.2776	0.5050	-0.1547
11	1.8505	0.8621	1.0821	0.3082
12	1.4882	0.2710	0.8000	-0.1571
13	2.5706	0.1478	1.6559	-0.2553
14	1.3586	0.3998	0.7003	-0.0528
15	1.5849	1.0000	0.8818	0.4214
16	1.7143	0.7612	0.9836	0.2353
17	2.1925	0.6464	1.3640	0.1421
18	1.0000	0.5159	0.4215	0.0386
19	1.8468	0.5376	1.0898	0.0563
20	1.6155	0.1939	0.9082	-0.2205

Die Übereinstimmung zwischen simulierten und geschätzten Parametern war auch hier sehr gut. Abbildung 6.34 zeigt im überlagerten Streudiagramm Bestimmtheitsmaße von 0.999 für die Item- und 0.997 für die Transferparameter. Beide Parameter sind wiederum Null-Eins-normiert.

Abbildung 6.34: Parameterschätzung bei 1000 Personen und 20 Items



Die Mitte- und die Null-Eins-Normierung hängen in diesen Fall mit Bestimmtheitsmaßen von jeweils 0.999 miteinander zusammen. In Abbildung 6.35 verdeutlicht dies wieder ein Streudiagramm.

Abbildung 6.35: Parametertransformation bei 1000 Personen und 20 Items

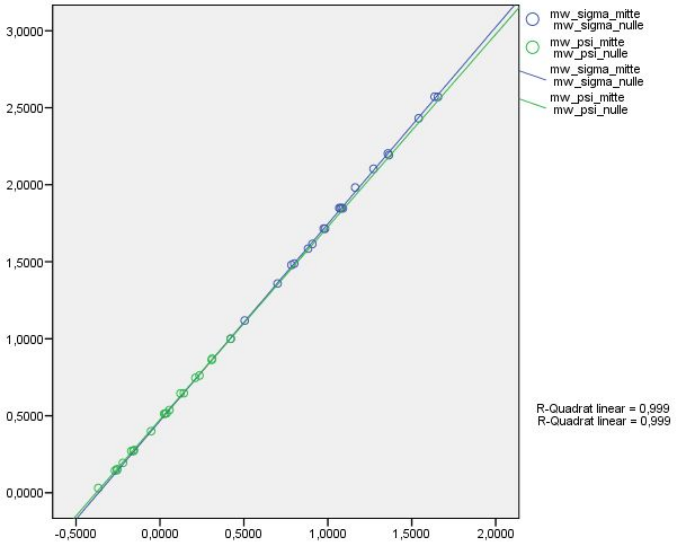
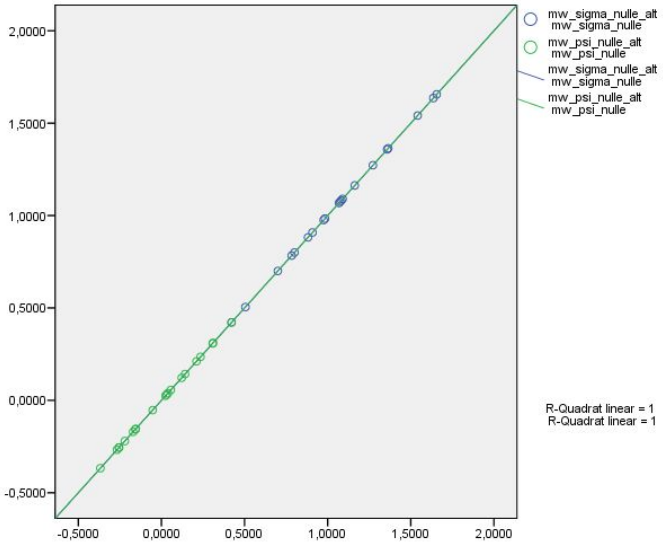


Abbildung 6.36 zeigt, dass die Schätzungen der alten und neuen Gamma-Funktionen wieder identisch sind. Bei beiden Item- und Transferparametern liegen die Bestimmtheitsmaße bei 1.

Abbildung 6.36: Vergleich alte vs. neue Gamma-Funktionen bei 1000 Personen und 20 Items



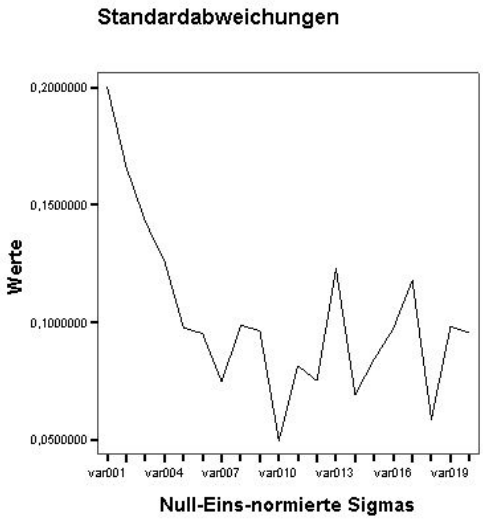
Es wurden auch hier wieder die Standardabweichungen und Varianzen für die Schätzungen der Null-Eins-normierten Itemparameter berechnet und in Tabelle 6.23 dargestellt.

Tabelle 6.23: Statistiken der Null-Eins-normierten Itemparameter bei 1000 Personen und 20 Items

Item	Standardabweichung	Varianz
1	0.2000	0.0400
2	0.1660	0.0276
3	0.1428	0.0204
4	0.1260	0.0159
5	0.0977	0.0096
6	0.0952	0.0091
7	0.0746	0.0056
8	0.0988	0.0098
9	0.0964	0.0093
10	0.0497	0.0025
11	0.0815	0.0067
12	0.0750	0.0056
13	0.1229	0.0151
14	0.0692	0.0048
15	0.0843	0.0071
16	0.0973	0.0095
17	0.1178	0.0139
18	0.0588	0.0035
19	0.0981	0.0096
20	0.0956	0.0091

Die graphische Veranschaulichung der Standardabweichungen ist in Abbildung 6.37 zu sehen. Die größte Streuung kann man beim ersten Parameter erkennen, die kleinste beim zehnten. Ansonsten sind relativ viele Spitzen und Täler sichtbar.

Abbildung 6.37: Standardabweichungen der Null-Eins-normierten Itemparameter bei 1000 Personen und 20 Items



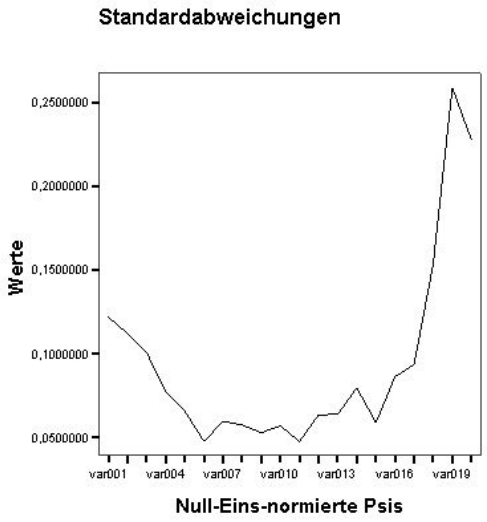
Auch die geschätzten Null-Eins-normierten Transferparameter wurden auf ihre Streuung hin untersucht. Tabelle 6.24 enthält deren Standardabweichungen und Varianzen.

Tabelle 6.24: Statistiken der Null-Eins-normierten Transferparameter bei 1000 Personen und 20 Items

Item	Standardabweichung	Varianz
1	0.1217	0.0148
2	0.1119	0.0125
3	0.1006	0.0101
4	0.0772	0.0060
5	0.0655	0.0043
6	0.0477	0.0023
7	0.0597	0.0036
8	0.0574	0.0033
9	0.0529	0.0028
10	0.0570	0.0032
11	0.0476	0.0023
12	0.0636	0.0040
13	0.0638	0.0041
14	0.0796	0.0063
15	0.0588	0.0035
16	0.0863	0.0074
17	0.0935	0.0087
18	0.1535	0.0236
19	0.2586	0.0669
20	0.2276	0.0518

Abbildung 6.38 zeigt das Liniendiagramm der Standardabweichungen der Null-Eins-normierten Transferparameter. Es ist wieder die größte Streuung beim vorletzten Parameter erkennbar, die kleinste beim sechsten und elften. Hier ist die Form deutlicher „U-“förmig.

Abbildung 6.38: Standardabweichungen der Null-Eins-normierten Transferparameter bei 1000 Personen und 20 Items



6.1.4.3 Simulationen mit 5000 Personen und 20 Items

5000 Personen war bei 20 Items die maximale Personenanzahl. Tabelle 6.25 enthält die Ergebnisse der durchschnittlichen Schätzungen der Item- und Transferparameter des Kempf-Modells.

Tabelle 6.25: Geschätzte Parameter des Kempf-Modells bei 5000 Personen und 20 Items

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	2.0293	0.0007	1.1694	-0.3482
2	2.6544	0.6253	1.6406	0.1218
3	1.8848	0.2461	1.0654	-0.1613
4	2.1428	0.1216	1.2598	-0.2542
5	1.5087	0.4983	0.7844	0.0291
6	1.8856	0.8786	1.0684	0.3155
7	1.7697	0.7570	0.9825	0.2249
8	2.2663	0.4994	1.3544	0.0322
9	2.5315	0.1207	1.5550	-0.2518
10	1.1295	0.2499	0.5042	-0.1538
11	1.8982	0.8761	1.0818	0.3169
12	1.5136	0.2467	0.7934	-0.1553
13	2.6455	0.1248	1.6416	-0.2463
14	1.3823	0.3745	0.6959	-0.0578
15	1.6370	1.0000	0.8880	0.4113
16	1.7653	0.7467	0.9845	0.2221
17	2.2671	0.6287	1.3617	0.1342
18	1.0000	0.5127	0.4113	0.0483
19	1.8820	0.4988	1.0741	0.0362
20	1.6297	0.1074	0.8851	-0.2640

Die Genauigkeit der Übereinstimmung zwischen simulierten und geschätzten Parametern ist hier nun fast perfekt. Abbildung 6.39 zeigt wieder im Rahmen eines überlagerten Streudiagramms ein Bestimmtheitsmaß von 1 für die Null-Eins-normierten Item- und von 0.999 für die Transferparameter. [10]

[10] Bei den Mitte-normierten Transferparameter betrug das Bestimmtheitsmaß auch 1. Dies kam durch Rundungen bei der Bildung des Durchschnitts zustande, obwohl die beiden Normierungsarten mit $r^2 = 1$ perfekt zusammenhängen.

Abbildung 6.39: Parameterschätzung bei 5000 Personen und 20 Items

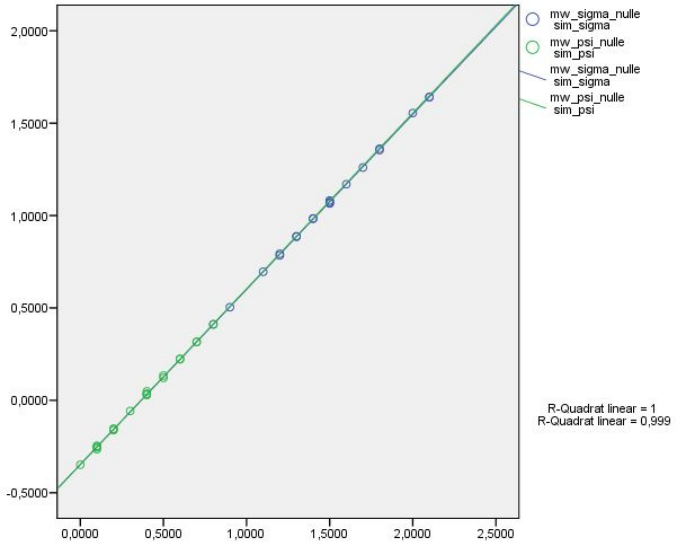


Abbildung 6.40 zeigt einen perfekten Zusammenhang zwischen Mitte- und Null-Eins-Normierung der Item- und Transferparameter mit Bestimmtheitsmaßen von jeweils 1.

Abbildung 6.40: Parametertransformation bei 5000 Personen und 20 Items

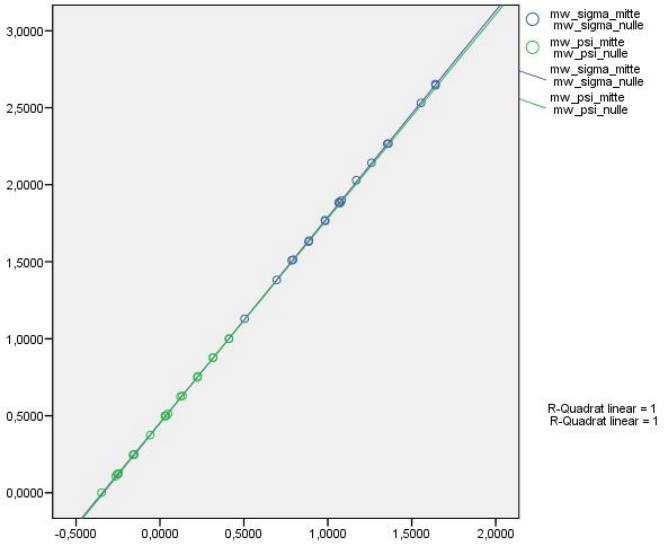
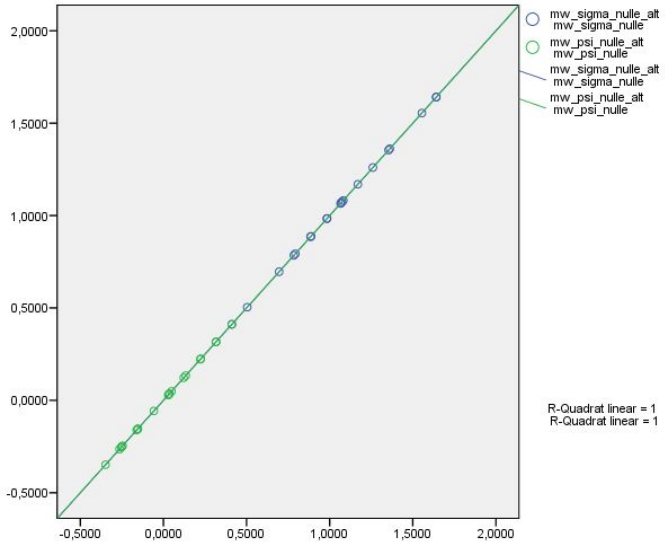


Abbildung 6.41 zeigt, dass die Schätzungen der alten und der neuen Gamma-Funktionen auch hier perfekt übereinstimmen. Beide Bestimmtheitsmaße entsprechen 1.

Abbildung 6.41: Vergleich alte vs. neue Gamma-Funktionen bei 5000 Personen und 20 Items



In Tabelle 6.26 werden die Standardabweichungen und Varianzen für die Null-Eins-normierten Itemparameter angeführt.

Tabelle 6.26: Statistiken der Null-Eins-normierten Itemparameter bei 5000 Personen und 20 Items

Item	Standardabweichung	Varianz
1	0.1014	0.0103
2	0.0823	0.0068
3	0.0647	0.0042
4	0.0645	0.0042
5	0.0542	0.0029
6	0.0442	0.0020
7	0.0384	0.0015
8	0.0485	0.0024
9	0.0451	0.0020
10	0.0244	0.0006
11	0.0410	0.0017
12	0.0275	0.0008
13	0.0428	0.0018
14	0.0329	0.0011
15	0.0426	0.0018
16	0.0428	0.0018
17	0.0602	0.0036
18	0.0336	0.0011
19	0.0605	0.0037
20	0.0596	0.0035

Abbildung 6.42 zeigt die Standardabweichungen der Itemparameter. Die größte Streuung liegt wieder beim ersten Parameter vor, die kleinste beim zehnten. Spitzen und Täler sind auch hier zu sehen.

Abbildung 6.42: Standardabweichungen der Null-Eins-normierten Itemparameter bei 5000 Personen und 20 Items

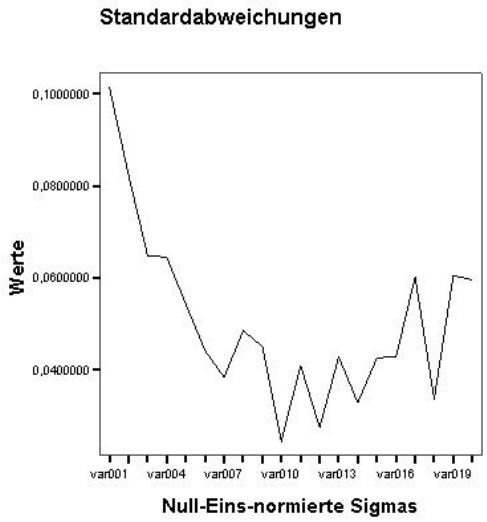


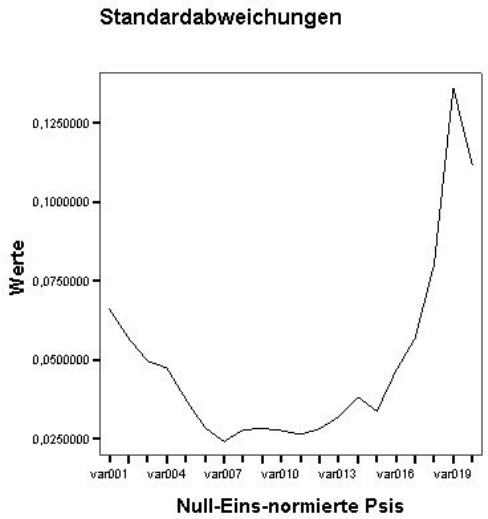
Tabelle 6.27 enthält die Standardabweichungen und Varianzen für die Schätzungen der Null-Eins-normierten Transferparameter.

Tabelle 6.27: Statistiken der Null-Eins-normierten Transferparameter bei 5000 Personen und 20 Items

Item	Standardabweichung	Varianz
1	0.0659	0.0043
2	0.0568	0.0032
3	0.0495	0.0025
4	0.0474	0.0022
5	0.0374	0.0014
6	0.0284	0.0008
7	0.0241	0.0006
8	0.0278	0.0008
9	0.0283	0.0008
10	0.0275	0.0008
11	0.0264	0.0007
12	0.0281	0.0008
13	0.0320	0.0010
14	0.0382	0.0015
15	0.0336	0.0011
16	0.0465	0.0022
17	0.0568	0.0032
18	0.0802	0.0064
19	0.1361	0.0185
20	0.1114	0.0124

In Abbildung 6.43 sind die Standardabweichungen wieder graphisch veranschaulicht. Die größte Streuung tritt beim vorletzten Parameter auf, die kleinste beim siebenten. Die Form erinnert an ein „U“.

Abbildung 6.43: Standardabweichungen der Null-Eins-normierten Transferparameter bei 5000 Personen und 20 Items



6.2 Mathematiksubtest der PISA-Studie

Der vorliegende Datensatz enthält einen Teil der Ergebnisse einer österreichischen Teilstichprobe aus der Studie PISA 2003 und wurde von der Statistik Austria zur Verfügung gestellt. Er beinhaltet insgesamt 20 dichotome Items aus dem Mathematiktest. Die Items wurden von insgesamt 6702 SchülerInnen vollständig bearbeitet. Folgende Items waren im Testheft enthalten: Cube Painting Q1, Cube Painting Q2, Cube Painting Q4, Growing Up Q1, Growing Up Q3, Pipelines Q1, Car Drive Q1, Car Drive Q2, Car Drive Q3, Running Tracks Q1, Running Tracks Q2, Running Tracks Q3, Diving Q1, Exchange Rate Q1, Exchange Rate Q2, Exchange Rate Q3, Height Q1, Making a Booklet Q1, Carbon Dioxide Q1, Carbon Dioxide Q2.

Es musste keine Person ausgeschieden werden, weil sie alle oder kein Item richtig gelöst hatte. Die Parameter wurden erst nach 976918 Iterationen in der Gesamtstichprobe hinreichend genau geschätzt. In der ersten Stichprobe wurde das Genauigkeitskriterium nach 65269 Iterationen, in der zweiten nach 201797 Iterationen erreicht. Die gesamte Schätzung aller Parameter dauerte dementsprechend lange, nämlich 3 Stunden und 50 Minuten.

Ergebnisse

Tabelle 6.28 enthält die geschätzten Parameter des Rasch-Modells für diesen Datensatz.

Die leichtesten Items sind demnach bei weitem Nummer 7 [11], weiters noch Nummer 14 und 15. Die schwersten Items sind das zwölfte, zweite und elfte.

Tabelle 6.28: Geschätzte Parameter des Rasch-Modells der PISA-Studie

Item	Produktnormierte Itemleichtigkeit	Itemschwierigkeit	Personenfähigkeit
1	1.4360	0.6964	0.0297
2	0.1849	5.4087	0.0657
3	0.4731	2.1135	0.1096
4	1.7365	0.5759	0.1633
5	0.6682	1.4966	0.2292
6	0.7509	1.3318	0.3107
7	45.9343	0.0218	0.4123
8	2.9704	0.3367	0.5400
9	0.3020	3.3115	0.7025
10	0.3000	3.3337	0.9123
11	0.1887	5.2995	1.1884
12	0.1740	5.7473	1.5604
13	0.7668	1.3040	2.0771
14	6.3836	0.1567	2.8244
15	4.0976	0.2440	3.9679
16	0.5699	1.7546	5.8661
17	2.4175	0.4136	9.4453
18	1.7840	0.5605	17.7427
19	0.5159	1.9383	48.6298
20	0.8852	1.1296	

In Tabelle 6.29 sind die geschätzten Parameter des Kempf-Modells aufgeführt.[12] Die Itemparameter geben an, dass auch im Kempf-Modell das siebte, 14. und 15. das leichteste sind und zwölf, zwei und elf das schwerste. Diese Ergebnisse decken sich also mit den Itemparametern des Rasch-Modells. Für die Transferparameter ergibt sich folgendes Bild: vom ersten bis zum 14. Transferparameter, also bis man 13 Items richtig gelöst ist der Transfer fast gleich bleibend hoch. Beim 15. bis zum 17. Parameter fällt der Transfer ganz leicht ab

[11] Dieses siebte Item wurde von fast allen Personen gelöst und sollte in der Folge bei der Vorgabe des Tests eliminiert werden. Jetzt lässt sich das Item allerdings nicht einfach so herausstreichen, da dann die ohnehin schon fragwürdige Reihenfolge noch mehr durcheinander geraten und die serielle Abhängigkeit gestört werden würde.

[12] Der Übersichtlichkeit halber wurde in dieser Tabelle die Itemnummer in die erste Spalte geschrieben. Es ist jedoch zu beachten, dass dies für die Transferparameter nicht korrekt ist. Ein Transfer bei Itemnummer 1 heißt in diesem Fall, dass der Transfer so aussieht, nachdem man kein Item vorher richtig gelöst hat, einer bei Item 5, dass man vorher vier Items gelöst hat usw..

und bei den letzten drei Parametern ist der Transfer dann extrem niedrig. Die N_{ri} und $N_{a_{vo}}$ sind dabei unauffällig. Wenn das Kempf-Modell gelten würde, dann würde man den größten Teil der Bearbeitung der Mathematikaufgaben etwas dazu lernen und am Ende der Bearbeitung würde eine vollkommene Lernhemmung auftreten. Dies kann etwa durch auftretende Konzentrationsschwäche erklärt werden.

Tabelle 6.29: Geschätzte Parameter des Kempf-Modells der PISA-Studie

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.6799	0.9798	0.6531	0.1310
2	6.5565	0.9856	4.2898	0.1353
3	3.1306	0.9829	1.7349	0.1333
4	1.5612	0.9742	0.5646	0.1268
5	2.4869	0.9586	1.2549	0.1152
6	2.3114	0.9611	1.1240	0.1170
7	1.0000	0.9668	0.1461	0.1213
8	1.3046	0.9831	0.3732	0.1335
9	4.3363	0.9974	2.6341	0.1441
10	4.3655	1.0000	2.6559	0.1460
11	6.4187	1.0000	4.1871	0.1460
12	6.8967	0.9881	4.5436	0.1372
13	2.2934	0.9831	1.1106	0.1334
14	1.1342	0.9262	0.2461	0.0910
15	1.2205	0.8498	0.3105	0.0341
16	2.7366	0.8103	1.4411	0.0046
17	1.3893	0.7354	0.4364	-0.0513
18	1.5102	0.0000	0.5266	-0.5997
19	2.8022	0.0001	1.4901	-0.5997
20	1.9441	0.0006	0.8501	-0.5993

Tabelle 6.30 zeigt jedoch, das das Kempf-Modell für diesen Datensatz nicht gelten kann. Der errechnete χ^2-Wert von 889.7015 übertrifft bei weitem den kritischen Wert bei $df = 33$ Freiheitsgraden. [13]

[13] Der kritische χ^2-Wert wird nicht vom Programm ausgegeben, er wurde in den folgenden Tabellen jeweils mit $\alpha = 0.05$ von der Verfasserin hinzugefügt.

Tabelle 6.30: Modelltest Kempf-Modell der PISA-Studie

H0	-49332.4969
H1	-30166.0134
+	-18721.6327
Likelihood-Ratio	889.7015
df	33
χ^2-Wert kritisch	43.77

Aufgrund der Größe dieses χ^2-Wertes werden für diesen Datensatz auch noch die einzelnen Parameter der beiden Untergruppen angegeben. [14] Tabelle 6.31 enthält die geschätzten Mitte-normierten Item- und Transferparameter der ersten und zweiten Untergruppe zum direkten Vergleich. Personen mit niedrigen Rohscores bilden die erste, Personen mit hohen Rohscores die zweite Untergruppe. Für die erste Teilstichprobe konnten nur 15 Transferparameter geschätzt werden, da in dieser die entsprechenden letzten fünf Rohscorehäufigkeiten der Personen ($N_{a_{v0}}$) fehlen.

Wenn man beide Untergruppen miteinander vergleicht, ergeben sich bei den Items 10, 11 und 12 - den "Running Tracks"-Aufgaben des Mathematiktests - die deutlichsten Unterschiede in der Itemschwierigkeit. Diese drei Items sind für die Personen aus der ersten Teilstichprobe deutlich schwieriger als für die der zweiten. Die anderen Itemparameter unterscheiden sich nur geringfügig voneinander. Das 12. Item ist in der Gesamtstichprobe leichter als in beiden Untergruppen. Die Transferparameter unterscheiden sich unter anderem bei Nummer 1 und 9 voneinander. Personen der ersten Untergruppe weisen hier deutlich höhere Werte auf als Personen der zweiten. Bei Nummer 14 und 15 verhält es sich umgekehrt und Personen der zweiten Gruppe lernen mehr dazu als Personen der ersten. Die Transferparameter der Untergruppen unterscheiden sich bis auf Nummer 18 und 20 nicht von den Parametern der Gesamtstichprobe. In der Gesamtstichprobe sind die Werte der dieser beiden Parameter deutlich niedriger.

[14] Bei den anderen Datensätzen wird darauf verzichtet, da sich die Likelihoods der Gesamt- und Teilstichproben dort nicht so stark unterscheiden.

Tabelle 6.31: Mitte-normierte Parameter der beiden Teilstichproben der PISA-Studie

Item	Erste Untergruppe		Zweite Untergruppe	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	2.0582	0.9684	1.0001	0.0000
2	8.8526	0.9771	9.5241	0.9981
3	4.0954	0.9737	4.4677	1.0000
4	1.8421	0.9644	1.9657	0.9831
5	3.1279	0.9495	3.6097	0.9297
6	2.8167	0.9542	3.4956	0.9655
7	1.0018	0.9796	1.0000	0.9536
8	1.4226	0.9982	1.7473	0.8619
9	6.4796	0.9867	5.4721	0.5660
10	8.8557	0.9803	3.9682	0.9470
11	17.1722	0.9624	5.9491	0.9722
12	12.2338	0.9688	8.2549	0.8173
13	2.8485	0.8296	3.2092	0.8302
14	1.2113	0.5797	1.1330	0.8252
15	1.3679	0.0018	1.1140	0.8859
16	3.5906		3.7260	0.9987
17	1.7018		1.1287	0.9998
18	1.7231		2.0398	0.4025
19	3.8461		3.6880	0.0010
20	2.2272		3.1575	0.3342

Der Effekt der Transferparameter kann ebenfalls nicht vernachlässigt werden. Bei $df = 19$ wird der χ^2-Wert von 136.2164 signifikant und das Rasch-Modell darf ebenfalls nicht angenommen werden (siehe Tabelle 6.32).

Tabelle 6.32: Modelltest Rasch-Modell der PISA-Studie

Rasch-LH	-49400.6051
Kempf-LH	-49332.4969
Likelihood-Ratio	136.2164
df	19
χ^2-Wert kritisch	30.14

6.3 Zufallsauswahl aus Items von Bahrick & Hall

Diese Testdaten stammen ursprünglich aus einer Erhebung von Bahrick & Hall (1991). Diese legten im Rahmen eines Forschungsprojekts zur „Lifetime Maintenance“ 1074 Personen

102 Items vor. Die Items wurden von Held und Korossy (1998) zu einer Reanalyse herangezogen. Sie wählten von den 102 Items jedoch nur 7 Items zur elementaren Algebra aus. Diese Items und die dazugehörigen Daten wurden der Verfasserin von Weber (siehe auch Weber, 2005) zur Verfügung gestellt und der Gebrauch der Daten von Prof. Bahrick per E-Mail-Kommunikation autorisiert.

- The result of dividing $8y^2+8y+2$ by $2y+1$ is:
  $[A]8y^2+4y+3$ $[B]4y+2$
  $[C]12y+3$ $[D]4y^2+2$ $[E]\frac{8y^2+3}{y}$
- If $9x^3+3x^2$ is divided by $3x^2$, the quotient is:
  $[A]3x+1$ $[B]6x+1$
  $[C]3x$ $[D]6x$ $[E]9x^3$
- Simplify $\frac{x^6}{x}$ if x does not equal 0.
- Factor: $49-x^2$
- Factor: $x^2-5x-24$
- What is/are the factors of $9x+9$?
  $[A]x=9$ $[B](x+1)$ and 9
  $[C]9$ and x $[D]x$ [E]none of the above.
- $\frac{x^2-3x+2}{3x-3}=$ $[A]\frac{x^2+2}{-3}$ $[B]\frac{x-2}{3}$
  $[C]x+1$ $[D]\frac{x+2}{3}$ $[E]x^2-1$

Diese sieben Items decken zwar einen einheitlichen Wissensbereich ab, sind jedoch trotzdem problematisch zu interpretieren. Die Items wurden hier aus einem großen Itempool selektiv ausgewählt, und es ist nicht klar, ob diese Items hintereinander bearbeitet wurden, oder ob dazwischen andere Aufgaben vorgegeben wurden. Der Lerngewinn oder -verlust kann daher nicht eindeutig interpretiert werden (siehe Abschnitt 6.2).

Es wurden 397 Personen in der Gesamtstichprobe ausgeschieden, weil sie alle oder kein Item richtig gelöst hatten. Nach 45553 Iterationen wurde das Genauigkeitskriterium für die Gesamtstichprobe erreicht. In der ersten Stichprobe wurde dieses nach 9468 und in der zweiten Stichprobe nach 27417 Iterationen erreicht. Insgesamt dauerte die Schätzung aller Parameter 9 Minuten mit den neuen Gamma-Funktionen. [15]

[15] Alle folgenden Datensätze wurden ebenfalls mit den neuen Gamma-Funktionen geschätzt, da diese bei realen Datensätzen flexibler waren und mehr Items als die alten schätzen konnten.

Ergebnisse

Die Itemparameter des Rasch-Modells in Tabelle 6.33 zeigen, dass das vierte und siebente Item offenbar am schwierigsten waren, das dritte und das erste am leichtesten. Die Schwierigkeit der Items steigt und fällt ohne Kontinuität.

Tabelle 6.33: Geschätzte Parameter des Rasch-Modells bei Bahrick & Hall

Item	Produktnormierte Itemleichtigkeit	Itemschwierigkeit	Personenfähigkeit
1	1.8757	0.5331	0.1482
2	1.2716	0.7864	0.3728
3	2.1835	0.4580	0.7332
4	0.4718	2.1194	1.3670
5	1.2255	0.8160	2.6848
6	0.6786	1.4737	6.7306
7	0.4894	2.0434	

Die Itemschwierigkeitsparameter des Kempf-Modells (siehe Tabelle 6.34) decken sich mit denen des Rasch-Modells, hier sind ebenfalls das vierte und siebente Item am schwersten, das erste und dritte am leichtesten. Unter der Annahme, dass das Modell gilt, zeigen die Transferparameter für diese Daten eine Fluktuation an Lerngewinn und -hemmung an, da die Parameter nicht monoton fallen oder steigen. Bei Nummer 5 und 3, also nach vier und zwei gelösten Items scheint der Transfer am größten zu sein, am niedrigsten ist der Transfer bei Nummer 7, also nach sechs gelösten Items. Ob diese Fluktuation von der nicht überprüfbaren Bearbeitungsfolge der Items abhängt, oder die Items tatsächlich zu gleichzeitigen Aktivierungs- und Inhibitionsprozessen führen, kann leider nicht nach geprüft werden.

Tabelle 6.34: Geschätzte Parameter des Kempf-Modells bei Bahrick & Hall

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.0005	0.6086	0.3399	-0.0370
2	1.4003	0.7847	0.7244	0.1323
3	1.2597	0.9590	0.5892	0.3000
4	3.0030	0.8779	2.2658	0.2220
5	1.6745	0.9995	0.9881	0.3389
6	2.2807	0.2995	1.5711	-0.3343
7	2.6846	0.0005	1.9596	-0.6219

Die Interpretation der Parameter ist jedoch, abgesehen von den inhaltlichen Aspekten,

nicht legitim, da das Kempf-Modell bei diese Daten eindeutig nicht gilt. In Tabelle 6.35 ist ein χ^2-Wert von 77.99 ist bei $df = 9$ Freiheitsgraden ersichtlich, welches ein signifikantes Ergebnis bedeutet.

Tabelle 6.35: Modelltest Kempf-Modell bei Bahrick & Hall

H0	-1681.0004
H1	-777.3113
+	-864.6929
Likelihood-Ratio	77.9925
df	9
χ^2-Wert kritisch	16.92

Es kann auch keine Reduktion zum Rasch-Modell erfolgen (siehe Tabelle 6.36). Mit einem χ^2-Wert von 44.06 und $df = 6$ Freiheitsgraden darf der Effekt der Transferparameter nicht vernachlässigt werden. Die Daten entsprechen also weder dem Kempf- noch dem Rasch-Modell.

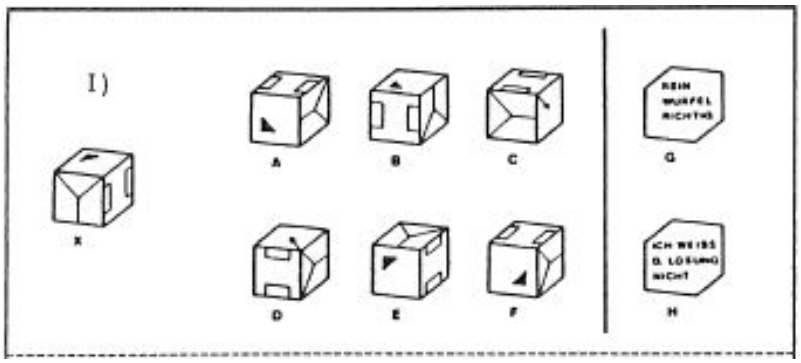
Tabelle 6.36: Modelltest Rasch-Modell bei Bahrick & Hall

Rasch-LH	-1703.0322
Kempf-LH	-1681.0004
Likelihood-Ratio	44.0635
df	6
χ^2-Wert kritisch	12.59

6.4 3DW-Daten von Gittler

Der 3DW (Dreidimensionaler Würfeltest) von Gittler (1990) dient zur Messung des räumlichen Vorstellungsvermögens. Er enthält in seiner Papier-Bleistift-Version ein „Warming-Up-Item“ und 17 eigentlichen Testaufgaben. Die Aufgaben bestehen aus Würfeln, von denen drei Seiten sichtbar sind. Auf jeder Seite befindet sich ein Muster, das nur einmal vorkommen darf. Die Probanden sollen sich nun die Vorgabewürfel in veränderter Lage, d.h. ein- oder mehrfach gedreht und / oder gekippt, vorstellen und diese aus sechs Antwortalternativen aussuchen. Die Antwortmöglichkeiten „kein Würfel richtig“ bzw. „ich weiß die Lösung nicht“ können ebenfalls angekreuzt werden. Bei der Bearbeitung des Tests soll kein Zeitdruck aufkommen. Ein Beispiel eines 3DW-Items wäre Abbildung 6.44.

Abbildung 6.44: Beispielitem des 3DW



Der Datensatz, der zur Analyse mittels DynTest herangezogen wurde, stammt aus der Normierungsstichprobe des 3DW von Gittler. Er besteht aus 866 männlichen und weiblichen Schülern aller Schultypen im Alter zwischen 13 und 19 Jahren (siehe Testmanual, Gittler, 1990). Der Datensatz wurde ebenfalls von Fischer (2003) verwendet.

Es mussten 116 Personen in der Gesamtstichprobe ausgeschieden werden, weil sie alle oder kein Item richtig beantwortet hatten. Es wurden 30188 Iterationen in der Gesamtstichprobe benötigt, um das Genauigkeitskriterium für die Parameterschätzung zu erreichen. Die Parameter wurden nach 21486 Iterationen in der ersten und nach 66056 in der zweiten Stichprobe geschätzt. Insgesamt betrug die Rechendauer 18 Minuten.

Ergebnisse

Das schwierigste Item, den Itemparametern des Rasch-Modells aus Tabelle 6.37 zufolge, ist mit Abstand das fünfte, gefolgt vom zehnten Item. Das leichteste Item ist das erste.

Tabelle 6.37: Geschätzte Parameter des Rasch-Modells des 3DW

Item	Produktnormierte Itemleichtigkeit	Itemschwierigkeit	Personenfähigkeit
1	2.9506	0.3389	0.0516
2	1.4713	0.6797	0.1131
3	1.0334	0.9676	0.1866
4	0.6474	1.5448	0.2751
5	0.2707	3.6940	0.3822
6	2.4559	0.4072	0.5132
7	0.6574	1.5210	0.6754
8	0.8201	1.2194	0.8794
9	0.6992	1.4302	1.1412
10	0.3763	2.6575	1.4857
11	0.6374	1.5689	1.9546
12	1.8351	0.5449	2.6228
13	1.6070	0.6223	3.6399
14	1.7298	0.5781	5.3555
15	0.8644	1.1569	8.8172
16	2.3471	0.4261	19.2487
17	0.5535	1.8066	

Auch hier gelten für die Itemschwierigkeitsparameter des Kempf-Modells die gleichen Ergebnisse wie für die des Rasch-Modells (siehe Tabelle 6.38). Das fünfte und das zehnte Item sind am schwersten, das erste am leichtesten. Die Transferparameter zeigen abermals kein monotones Steigen oder Fallen, bis Nummer 7 bleiben sie in etwa gleich, von Nummer 8 bis 12 fallen sie ab, von 13 bis 16 steigen sie und beim letzten fallen sie abrupt ab. Wenn das Kempf-Modell gelten würde, würde dies abermals inhaltlich durch Auftreten von Lernhemmungen oder Konzentrationsverlust erklärbar sein.

Tabelle 6.38: Geschätzte Parameter des Kempf-Modells des 3DW

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.0003	0.2968	0.2833	-0.0458
2	1.7580	0.3408	0.6377	-0.0252
3	2.4566	0.4059	0.9645	0.0052
4	3.7720	0.3899	1.5798	-0.0023
5	8.5908	0.4119	3.8340	0.0080
6	1.2549	0.3839	0.4024	-0.0051
7	3.7825	0.4724	1.5847	0.0363
8	3.1070	0.2625	1.2688	-0.0619
9	3.5770	0.2553	1.4886	-0.0652
10	6.2814	0.0955	2.7537	-0.1400
11	3.8639	0.0724	1.6228	-0.1508
12	1.5568	0.0003	0.5436	-0.1845
13	1.7172	0.5174	0.6186	0.0574
14	1.6207	0.7929	0.5735	0.1862
15	2.8739	0.9997	1.1597	0.2830
16	1.2900	0.9965	0.4188	0.2815
17	4.4142	0.0166	1.8802	-0.1769

Mit einem χ^2-Wert von 58.09 bei $df = 23$ Freiheitsgraden gilt das Kempf-Modell jedoch auch für diese Daten nicht und die Parameter dürften in dem Fall gar nicht interpretiert werden (siehe Tabelle 6.39).

Tabelle 6.39: Modelltest Kempf-Modell des 3DW

H0	-5185.8579
H1	-2402.1671
+	-2754.6475
Likelihood-Ratio	58.0866
df	23
χ^2-Wert kritisch	35.17

Da dieser Datensatz die Normierungsstichprobe eines Rasch-skalierten Tests ist, ist es nicht verwunderlich, dass die Transfereffekte hier nicht signifikant und somit vernachlässigbar sind. Tabelle 6.40 zeigt, dass dieser Datensatz mit einem χ^2-Wert von 21.73 bei $df = 16$ dem Rasch-Modell zugeordnet werden kann.

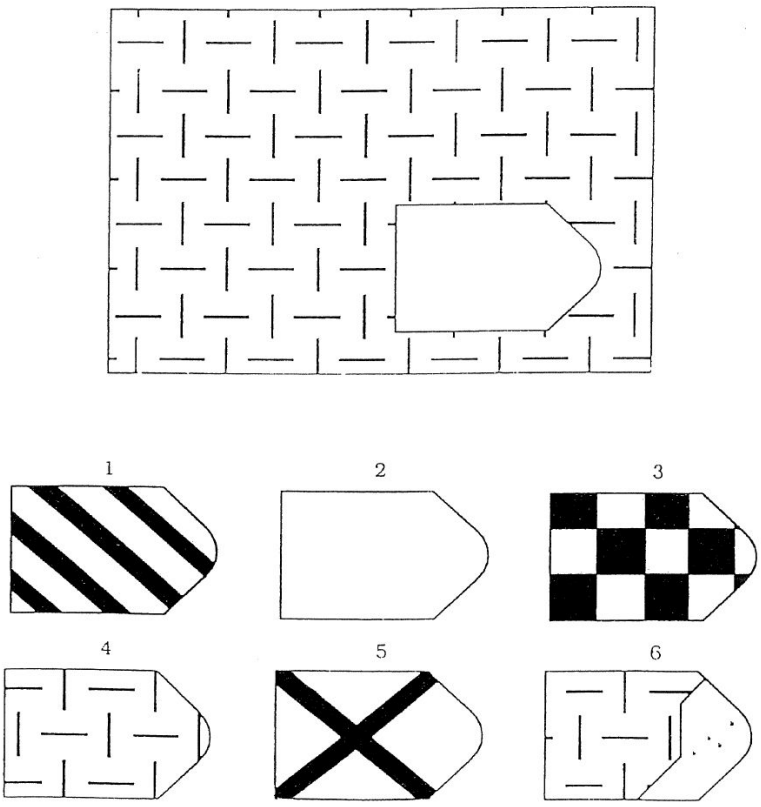
Tabelle 6.40: Modelltest Rasch-Modell des 3DW

Rasch-LH	-5196.7220
Kempf-LH	-5185.8579
Likelihood-Ratio	21.7282
df	16
χ^2-Wert kritisch	26.30

6.5 SPM-Daten von Schmöger

Der SPM (Standard Progressive Matrices) von Raven (deutsche Version von Kratzmeier & Horn, 1987) ist ein sprachfreier Test zur Erfassung der allgemeinen Intelligenz. Der SPM wurde entwickelt, um unterschiedliche Grade kognitiver Fähigkeiten von Kindern hin bis zu Erwachsenen zu messen. Der Test setzt sich aus 5 Aufgabensets (Teile A - E) mit jeweils 12 Items zusammen. Zu bearbeiten sind unvollständige geometrische Figuren oder Muster. Der fehlende Teil des Musters soll mit einem zusätzlichen Teil ergänzt werden, der nach dem Multiple-Choice-Prinzip aus 8 - 10 Alternativen ausgewählt wird. Diese Aufgaben sind am Anfang sehr leicht und werden gegen Ende hin immer schwieriger. Das erste Beispiel (Item A1) wird als Übungsbeispiel verwendet (siehe Abbildung 6.45). Die Testbearbeitung erfolgt ohne Zeitdruck.

Abbildung 6.45: Übungsbeispiel des SPM



Im Rahmen mehrerer Forschungspraktika I (2005 - 2007) im Psychologiestudium der Universität Wien wurden unter Schmöger Testdaten von Erwachsenen im Alter von 20 bis 50 Jahren und von Kindern zwischen 7 und 12 Jahren erhoben. Für die Analyse mit dem dynamischen Testmodell von Kempf wurden die beiden großen Datensätze in ihre Untertests á 12 Items aufgeteilt. Es zeigte sich, dass die Parameter für die Erwachsenen lediglich bei einem von fünf, für die Kinder bei zwei von fünf Untertests schätzbar waren. Bei den anderen Subtests wurde die Schätzung auch mit den neuen Gamma-Funktionen wegen zu großer Ungenauigkeit abgebrochen. Im Folgenden finden sich die Ergebnisse für den Erwachsenen-Subtest C und die Kinder-Subtests C und E.

6.5.1 SPM Subtest C, Erwachsene

Dieser Datensatz umfasst 343 Personen und 12 Items. 148 Personen mussten in der Gesamtstichprobe ausgeschieden werden, weil sie alle oder kein Item richtig gelöst hatten. Es wurden 4228 Iterationen benötigt, um das Genauigkeitskriterium für die Parameterschät-

zung der Gesamtstichprobe zu erreichen. In der ersten Stichprobe wurden für dies 4349, in der zweiten 7222 Iterationen gebraucht. Insgesamt rechnete DynTest für diesen Datensatz 3 Minuten.

Ergebnisse

Die Itemparameter des Rasch-Modells sind zwar nicht streng monoton nach Schwierigkeit geordnet, es zeigt sich jedoch (siehe Tabelle 6.41), dass die Items tatsächlich zum Ende hin schwieriger werden. Das erste Item ist das leichteste, das letzte das schwierigste.

Tabelle 6.41: Geschätzte Parameter des Rasch-Modells des SPM, Erwachsene, Subtest C

Item	Produktnormierte Itemleichtigkeit	Itemschwierigkeit	Personenfähigkeit
1	9.5761	0.1044	0.0540
2	3.2669	0.3061	0.1310
3	5.0338	0.1987	0.2404
4	0.7333	1.3638	0.3969
5	1.8165	0.5505	0.6255
6	0.5796	1.7254	0.9702
7	1.6814	0.5947	1.5139
8	0.5397	1.8529	2.4278
9	0.7033	1.4219	4.1189
10	0.4045	2.4722	7.8202
11	0.2469	4.0509	19.7009
12	0.1291	7.7471	

Diese Ordnung wird auch bei den Itemschwierigkeitsparametern des Kempf-Modells beibehalten. Tabelle 6.42 zeigt, dass das dritte und erste Item am leichtesten, das letzte am schwierigsten ist. Abbildung 6.46 veranschaulicht zusätzlich graphisch den Verlauf der Transferparameter. Diese bleiben bis zu sieben gelösten Items, also bis Nummer 8 mit kleineren Schwankungen in etwa gleich hoch, sinken bei Nummer 9 und 10 stark und steigen bei den letzten beiden wieder stark an.

Tabelle 6.42: Geschätzte Parameter des Kempf-Modells des SPM, Erwachsene, Subtest C

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.0010	0.9523	0.4158	0.3266
2	1.0397	0.9035	0.4868	0.2373
3	1.0003	0.9171	0.4146	0.2623
4	1.3969	0.8255	1.1408	0.0945
5	1.1337	0.9417	0.6588	0.3072
6	1.4721	0.7381	1.2786	-0.0657
7	1.1482	0.9997	0.6855	0.4134
8	1.6524	0.8393	1.6088	0.1197
9	1.2745	0.5349	0.9167	-0.4377
10	1.3624	0.0003	1.0776	-1.4168
11	1.9435	0.6396	2.1419	-0.2460
12	3.6756	0.9952	5.3142	0.4053

Dieses Ergebnis ist wiederum inhaltlich schwierig zu interpretieren. Es dürften hier abermals inhibitorische und steigernde Lerneffekte bzw. Konzentrationsschwächen auftreten.

Abbildung 6.46: Transferparameter des SPM, Erwachsene, Subtest C

In Tabelle 6.43, kann man erkennen, dass für diese Daten das Kempf-Modell gilt. Der χ^2-Wert von 22.92 ist bei $df = 20$ nicht signifikant. Die geringe Stichprobengröße lässt jedoch eine eher ungenaue Schätzung vermuten.

Tabelle 6.43: Modelltest Kempf-Modell des SPM, Erwachsene, Subtest C

H0	-570.9492
H1	-225.1317
+	-334.3530
Likelihood-Ratio	22.9291
df	20
χ^2-Wert kritisch	31.41

Eine Reduktion zum Rasch-Modell ist nicht möglich (siehe Tabelle 6.44). Die Transferparameter können mit einem χ^2-Wert von 21.66 bei $df = 11$ nicht vernachlässigt werden.

Tabelle 6.44: Modelltest Rasch-Modell des SPM, Erwachsene, Subtest C

Rasch-LH	-581.7806
Kempf-LH	-570.9492
Likelihood-Ratio	21.6628
df	11
χ^2-Wert kritisch	19.68

6.5.2 SPM Subtest C, Kinder

Dieser Datensatz umfasst 625 Kinder und die gleichen 12 Items wie bei den Erwachsenen. 15 Personen wurden in der Gesamtstichprobe ausgeschieden, weil sie alle oder kein Item richtig gelöst hatten. Das Genauigkeitskriterium für die Parameterschätzung der Gesamtstichprobe wurde nach 6055 Iterationen erreicht. In der ersten Stichprobe benötigte dies 17284 Iterationen, in der zweiten Stichprobe 25644. Die gesamte Rechendauer betrug 6 Minuten.

Ergebnisse

Die Itemparameter des Rasch-Modells beim gleichen Subtest, aber bei den Kindern, zeigen ähnliche Ergebnisse wie bei den Erwachsenen (siehe Tabelle 6.45). Die Schwierigkeiten steigen wieder nicht streng monoton an, das erste Item ist wieder das leichteste, das letzte das schwerste. Hier ist jedoch die Schwierigkeit des letzten Items viel größer als bei den Erwachsenen.

Tabelle 6.45: Geschätzte Parameter des Rasch-Modells des SPM, Kinder, Subtest C

Item	Produktnormierte Itemleichtigkeit	Itemschwierigkeit	Personenfähigkeit
1	10.9466	0.0914	0.0376
2	5.3898	0.1855	0.1107
3	2.0134	0.4967	0.2307
4	1.2233	0.8174	0.4130
5	2.6919	0.3715	0.6837
6	1.1489	0.8704	1.0884
7	1.4707	0.6800	1.7141
8	0.6437	1.5535	2.7428
9	1.2344	0.8101	4.6146
10	0.2623	3.8129	8.6788
11	0.1902	5.2583	21.7327
12	0.0382	26.1967	

Für die Itemparameter des Kempf-Modells ergibt sich - ersichtlich in Tabelle 6.46 - Ähnliches. Die Schwierigkeit nimmt auch hier zum Ende hin sehr stark zu. Die Transferparameter bleiben auch für die ersten sieben Summenscores, also bis Nummer 8 in etwa gleich, dann fallen sie ab und steigen im Unterschied zu den Erwachsenen nicht wieder an. Unter Annahme des Kempf-Modells fände also ab acht vorangegangen gelösten Items eine Lernhemmung statt.

Tabelle 6.46: Geschätzte Parameter des Kempf-Modells des SPM, Kinder, Subtest C

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.0000	0.7956	0.1748	0.0928
2	1.1869	0.8013	0.2498	0.0951
3	1.7490	0.7289	0.4754	0.0660
4	2.3840	0.7575	0.7302	0.0775
5	1.5715	1.0000	0.4041	0.1748
6	2.6484	0.8337	0.8363	0.1081
7	2.2650	0.9999	0.6824	0.1748
8	4.1570	0.8496	1.4416	0.1144
9	2.4678	0.0000	0.7638	-0.2265
10	8.1625	0.0009	3.0490	-0.2261
11	10.8062	0.0008	4.1090	-0.2262
12	52.2283	0.0046	20.7315	-0.2246

Im Unterschied zu den Erwachsenen wird in dieser Stichprobe der Modelltest jedoch mit

einem χ^2-Wert von 84.83 bei $df = 18$ signifikant (siehe Tabelle 6.47). Somit hat das Kempf-Modell für diese Testdaten keine Gültigkeit und die Parameter dürfen eigentlich nicht interpretiert werden.

Tabelle 6.47: Modelltest Kempf-Modell des SPM, Kinder, Subtest C

H0	-2374.0687
H1	-1507.4017
+	-824.2506
Likelihood-Ratio	84.8329
df	18
χ^2-Wert kritisch	28.87

Ein weiterer Unterschied zu den Erwachsenen ist die Signifikanz der Transferparameter (siehe Tabelle 6.48). Diese sind mit einem χ^2-Wert von 18.53 knapp nicht signifikant und können daher vernachlässigt werden. Eine Reduktion zum Rasch-Modell wäre hier also zulässig.

Tabelle 6.48: Modelltest Rasch-Modell des SPM, Kinder, Subtest C

Rasch-LH	-2383.3349
Kempf-LH	-2374.0687
Likelihood-Ratio	18.5324
df	11
χ^2-Wert kritisch	19.68

6.5.3 SPM Subtest E, Kinder

Dieser Datensatz umfasst die selben 625 Kinder und 12 Items eines anderen Subtests des SPM. Es wurden in der Gesamtstichprobe 164 Personen ausgeschieden, weil sie alle oder kein Item richtig gelöst hatten. Die Parameter der Gesamtstichprobe konnten nach 47051 Iterationen hinreichend genau geschätzt werden. Das Genauigkeitskriterium für die erste Stichprobe wurde nach 20414, für die zweite Stichprobe nach 38612 Iterationen erreicht. Die Schätzung aller Parameter insgesamt dauerte 15 Minuten.

Ergebnisse

Tabelle 6.49 zeigt, dass die Itemparameter des Rasch-Modells hier bis auf zwei Ausnahmen nach Schwierigkeit geordnet sind, wobei nicht das letzte, sondern das vorletzte Item das schwierigste ist.

Tabelle 6.49: Geschätzte Parameter des Rasch-Modells des SPM, Kinder, Subtest E

Item	Produktnormierte Itemleichtigkeit	Itemschwierigkeit	Personenfähigkeit
1	7.5921	0.1317	0.0527
2	2.7193	0.3678	0.1303
3	3.6938	0.2707	0.2443
4	1.8981	0.5268	0.4110
5	1.7924	0.5579	0.6562
6	1.2430	0.8045	1.0215
7	1.1953	0.8366	1.5822
8	0.8547	1.1699	2.4911
9	0.3627	2.7571	4.1090
10	0.2516	3.9743	7.5297
11	0.1213	8.2459	18.2361
12	0.2743	3.6461	

Dasselbe Bild ergibt sich für die Itemschwierigkeitsparameter des Kempf-Modells (siehe Tabelle 6.50). Interessanterweise wurden bei diese Daten statt zwölf Transferparametern nur elf gefunden. Das liegt daran, dass sowohl der letzte N_{ri} als auch der letzte und vorletzte $N_{a_{vo}}$ (d.h. die Rohscorehäufigkeit der Personen) gleich 0 ist. Somit kann kein letzter Transferparameter berechnet werden. Die ersten sechs Transferparameter sind in etwa konstant hoch, die letzten fünf sind konstant niedrig, wobei dies durch geringe Häufigkeiten (aber nicht gleich 0) in den letzten fünf N_{ri} bedingt ist. Unter Gültigkeit des Kempf-Modells würde also eine starke Lernhemmung auftreten.

Tabelle 6.50: Geschätzte Parameter des Kempf-Modells des SPM, Kinder, SubtestE

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.0000	0.9229	0.3861	0.3249
2	1.2136	0.9670	0.5556	0.3599
3	1.1454	0.9524	0.5015	0.3483
4	1.3693	0.9938	0.6792	0.3812
5	1.4124	1.0000	0.7133	0.3861
6	1.6293	0.8099	0.8855	0.2353
7	1.5303	0.0000	0.8069	-0.4074
8	1.6586	0.0002	0.9087	-0.4073
9	2.5697	0.0002	1.6317	-0.4073
10	3.3492	0.0001	2.2503	-0.4073
11	5.8354	0.0013	4.2230	-0.4064
12	2.9155		1.9060	

Das Kempf-Modell gilt hier (siehe Tabelle 6.51) jedoch mit einem χ^2-Wert von 38.36 bei $df = 13$ wiederum nicht und die Parameter können so nicht interpretiert werden.

Tabelle 6.51: Modelltest Kempf-Modell des SPM, Kinder, Subtest E

H0	-1720.1641
H1	-541.1408
+	-1159.8449
Likelihood-Ratio	38.3567
df	13
χ^2-Wert kritisch	22.36

Das Rasch-Modell kann allerdings ebenfalls nicht angenommen werden (siehe Tabelle 6.52). Die Transferparameter sind mit einem χ^2-Wert von 58.02 und $df = 11$ Freiheitsgraden signifikant und dürfen somit nicht vernachlässigt werden.

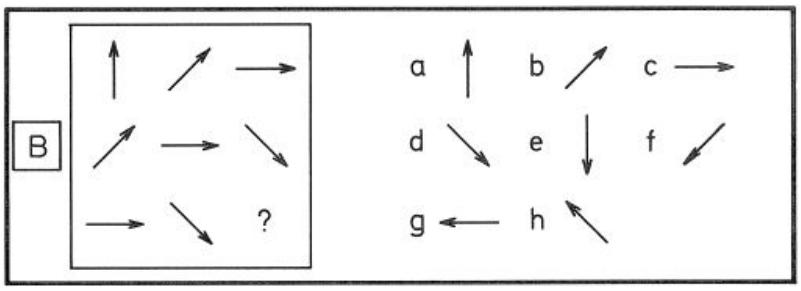
Tabelle 6.52: Modelltest Rasch-Modell des SPM, Kinder, Subtest E

Rasch-LH	-1749.1756
Kempf-LH	-1720.1641
Likelihood-Ratio	58.0230
df	11
χ^2-Wert kritisch	19.68

6.6 WMT-Daten von Weber

Der WMT (Wiener Matrizen-Test) von Formann & Piswanger (1979) ist dem SPM vom Konzept her ähnlich und ebenfalls ein Test zur Erfassung von sprachfreier Intelligenz. Er umfasst 24 Rasch-homogene Items. Die Aufgaben bestehen wie beim SPM aus Figuren, die nach einem bestimmten System angeordnet sind. Die letzte, dazu passende Figur soll nach dem Multiple-Choice Prinzip aus acht Figuren ausgewählt werden (siehe Abbildung 6.47). Die Personen haben bei der Bearbeitung der Items keinen Zeitdruck.

Abbildung 6.47: Übungsbeispiel des WMT



Der erste von zwei WMT-Datensätzen (siehe Abschnitt 6.7) wurde im Zuge seiner Diplomarbeit von Weber (1999) erhoben. Verwendet wurde hierzu die Computertestversion des WMT. Die Stichprobe umfasste 521 Lehrlinge und AHS Schüler/innen im Alter von 15 bis 18 Jahren.

Es musste keine Person in der Gesamtstichprobe ausgeschieden werden. Das Genauigkeitskriterium für die Parameterschätzung bei der Gesamtstichprobe wurde nach 12245 Iterationen erreicht. Die Schätzung der Parameter der ersten Stichprobe benötigte 4674 Iterationen, die der zweiten Stichprobe 9483 Iterationen. Die gesamte Rechenzeit betrug 8 Minuten.

Ergebnisse

Die Itemparameter des Rasch-Modells in Tabelle 6.53 sind etwas auf- und ab schwankend, aber man kann die Tendenz erkennen, dass die Schwierigkeit bei den letzten Items ansteigt. Das letzte Item ist das schwerste, das vierte Item das leichteste.

Tabelle 6.53: Geschätzte Parameter des Rasch-Modells des WMT von Weber

Item	Produktnormierte Itemleichtigkeit	Itemschwierigkeit	Personenfähigkeit
1	2.0292	0.4928	0.0272
2	6.9266	0.1444	0.0590
3	4.8787	0.2050	0.0963
4	7.5535	0.1324	0.1402
5	2.1322	0.4690	0.1920
6	0.9964	1.0036	0.2533
7	3.3796	0.2959	0.3263
8	1.3046	0.7665	0.4134
9	1.3165	0.7596	0.5180
10	1.2135	0.8241	0.6443
11	2.1972	0.4551	0.7978
12	2.7921	0.3582	0.9859
13	0.7352	1.3601	1.2190
14	0.3571	2.8001	1.5114
15	0.3014	3.3180	1.8842
16	0.7092	1.4100	2.3690
17	1.0143	0.9859	3.0160
18	0.3147	3.1773	3.9090
19	1.0053	0.9947	5.1995
20	0.2151	4.6502	7.1908
21	0.3250	3.0770	10.5909
22	0.2546	3.9284	17.5246
23	0.3426	2.9190	38.5902
24	0.1916	5.2196	

Die Itemschwierigkeitsparameter des Kempf-Modells in Tabelle 6.54 sehen etwas anders aus, als die des Rasch-Modells. Im Großen und Ganzen schwanken die Parameter mehr, das letzte Item ist aber wiederum das schwerste. Das zweite Item ist am leichtesten. Wie bei dem Datensatz zuvor wird auch hier ein Transferparameter zu wenig gefunden. Das liegt wiederum an den fehlenden letzten beiden $N_{a_{vo}}$. Zusätzlich dazu sind die letzten 5 N_{ri} generell gleich 0, was die hohen letzten Transferparameter ab Nummer 20 erklärt. Immer wenn N_{ri} fehlen, werden die Transferparameter so hoch geschätzt. Die Transferparameter sind grundsätzlich aber konstant hoch. Lediglich bei Nummer 12 und zwischen Nummer 16 bis 19 treten Inhibitionen auf, wenn vorher 17 richtig gelöst wurden (also bei Nummer 18) liegt der Transferparameter am niedrigsten.

Tabelle 6.54: Geschätzte Parameter des Kempf-Modells des WMT von Weber

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.3763	0.8423	0.5168	0.0648
2	1.0004	0.8479	0.1986	0.0697
3	1.0757	0.8617	0.2624	0.0812
4	1.0146	0.8905	0.2107	0.1057
5	1.3971	0.8607	0.5344	0.0804
6	2.0080	0.8700	1.0514	0.0883
7	1.2014	0.8819	0.3687	0.0983
8	1.7291	0.8176	0.8153	0.0440
9	1.7016	0.7522	0.7920	-0.0114
10	1.7423	0.8300	0.8265	0.0545
11	1.3207	0.7597	0.4697	-0.0051
12	1.2022	0.5806	0.3694	-0.1566
13	2.2579	0.8046	1.2628	0.0329
14	3.7680	0.7698	2.5407	0.0035
15	4.3200	0.9705	3.0078	0.1733
16	2.2483	0.5123	1.2547	-0.2144
17	1.8038	0.3469	0.8786	-0.3544
18	4.0515	0.0004	2.7806	-0.6476
19	1.7765	0.4179	0.8554	-0.2943
20	5.4808	0.9996	3.9902	0.1980
21	3.7815	0.9977	1.2628	0.1963
22	4.6930	0.9990	2.5407	0.1975
23	3.6603	0.9962	3.0078	0.1951
24	5.9393		1.2547	

Für diese Testdaten wird jedoch die Likelihood-Ratio mit einem χ^2-Wert von 79.98 bei $df = 35$ signifikant, d.h. das Kempf-Modell gilt in diesem Fall wieder nicht (siehe Tabelle 6.55).

Tabelle 6.55: Modelltest Kempf-Modell des WMT von Weber

H0	-5619.5994
H1	-2959.9312
+	-2619.6798
Likelihood-Ratio	79.9768
df	35
χ^2-Wert kritisch	49.77

Der zweite Likelihood-Ratio-Test (siehe Tabelle 6.56) zeigt jedoch, dass die Lerneffekte

hier mit einem χ^2-Wert von 29.27 und $df = 23$ Freiheitsgraden nicht signifikant sind und somit eine Reduktion zum Rasch-Modell legitim wäre, was sich auch mit dem Anspruch des WMT auf Rasch-Homogenität deckt.

Tabelle 6.56: Modelltest Rasch-Modell des WMT von Weber

Rasch-LH	-5634.2348
Kempf-LH	-5619.5994
Likelihood-Ratio	29.2709
df	23
χ^2-Wert kritisch	35.17

6.7 WMT-Daten von Formann, Waldherr & Piswanger

Der zweite WMT-Datensatz stammt von Formann, Waldherr & Piswanger (im Druck). Im Zuge einer Revidierung des WMT wurden Testdaten bestehend aus 21 Items des ursprünglichen WMT und 277 Personen erhoben. Dies ist jedoch noch nicht die Endversion des WMT2, dieser soll aus nur 18 Items bestehen. Von den 21 Items konnten jedoch lediglich die ersten 16 geschätzt werden, bei Hinzunahme von weiteren Items wurde die Rechenungenauigkeit auch mit den neuen Gamma-Funktionen jeweils zu groß. [16]

Es wurden 77 Personen in der Gesamtstichprobe von DynTest ausgeschieden, die alle oder kein Item richtig gelöst hatten. Das Genauigkeitskriterium für die Parameterschätzung der Gesamtstichprobe wurde nach 6288 Iterationen erreicht, das der ersten Stichprobe nach 3066 und das der zweiten Stichprobe nach 4732. Insgesamt betrug die Rechendauer 2 Minuten.

Ergebnisse

Tabelle 6.57 zeigt, dass die Itemparameter des Rasch-Modells etwas auf und ab schwanken, die schwierigeren Items finden sich jedoch im Großen und Ganzen wieder gegen Ende. Das schwierigste Item ist das 14., das leichteste das zweite.

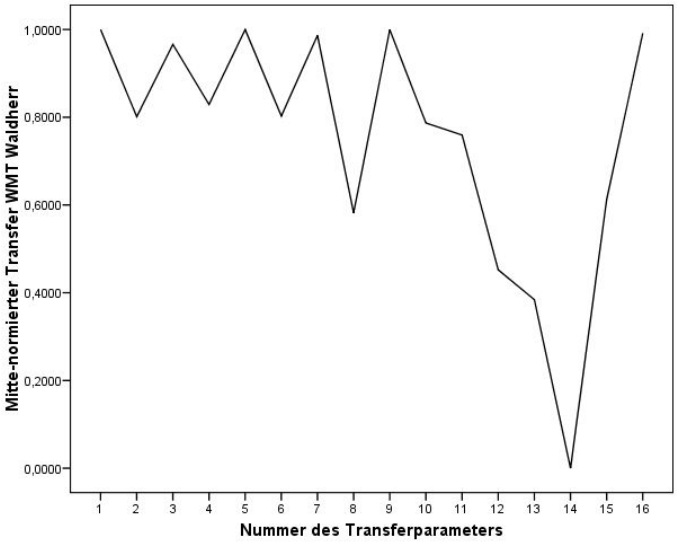
[16] Für die Aufrechterhaltung der seriellen Abhängigkeit ist es nötig, die Items von hinten nach vorne zu eliminieren.

Tabelle 6.57: Geschätzte Parameter des Rasch-Modells des WMT von Formann, Waldherr & Piswanger

Item	Produktnormierte Itemleichtigkeit	Itemschwierigkeit	Personenfähigkeit
1	1.6706	0.5986	0.0488
2	5.4280	0.1842	0.1093
3	2.8767	0.3476	0.1844
4	2.6180	0.3820	0.2784
5	0.6819	1.4664	0.3967
6	1.4816	0.6749	0.5474
7	2.3991	0.4168	0.7419
8	1.4816	0.6749	0.9972
9	0.9174	1.0901	1.3406
10	0.7574	1.3203	1.8175
11	0.8812	1.1349	2.5101
12	0.7308	1.3683	3.5830
13	0.3727	2.6829	5.4215
14	0.2512	3.9805	9.1833
15	0.2744	3.6449	20.6461
16	0.3547	2.8191	

Genau das gleiche Resultat ergibt sich für die Itemschwierigkeitsparameter des Kempf-Modells in Tabelle 6.58. Abbildung 6.48 veranschaulicht die Transferparameter. Sie sind in etwa gleich hoch bis Nummer 11, also bis man zehn vorangegangene Items richtig gelöst hat, mit einem kleinen Einbruch bei Nummer 8, sinken dann bis zum Tiefpunkt bei Nummer 14 ab und steigen dann wiederum stark an.

Abbildung 6.48: Transferparameter des WMT von Formann, Waldherr & Piswanger



Es dürften nach anfänglichem konstantem Transfer somit wieder inhibitorische Prozesse ablaufen. Abermals ist die inhaltliche Interpretation problematisch.

Tabelle 6.58: Geschätzte Parameter des Kempf-Modells des WMT von Formann, Waldherr & Piswanger

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	1.5127	0.9999	0.8238	0.2720
2	1.0001	0.8012	0.2722	0.0583
3	1.2151	0.9660	0.5036	0.2356
4	1.1585	0.8290	0.4426	0.0881
5	2.2048	0.9998	1.5685	0.2719
6	1.3977	0.8027	0.7001	0.0599
7	1.2722	0.9864	0.5650	0.2575
8	1.2832	0.5820	0.5769	-0.1776
9	1.7209	0.9994	1.0478	0.2715
10	1.9276	0.7869	1.2702	0.0428
11	1.7339	0.7595	1.0618	0.0134
12	1.8334	0.4521	1.1689	-0.3174
13	2.7085	0.3843	2.1105	-0.3904
14	3.5158	0.0001	2.9791	-0.8037
15	3.2015	0.6122	2.6409	-0.1451
16	2.6412	0.9916	2.0380	0.2631

Die Likelihood-Ratio ist mit einem χ^2-Wert von 23.75 bei $df = 28$ nicht signifikant (siehe

Tabelle 6.59), somit kann angenommen werden, dass das Kempf-Modell für diese Daten gilt.

Tabelle 6.59: Modelltest Kempf-Modell des WMT von Formann, Waldherr & Piswanger

H0	-924.6001
H1	-578.9519
+	-333.7752
Likelihood-Ratio	23.7458
df	28
χ^2-Wert kritisch	41.34

Jedoch ist auch der Effekt der Transferparameter mit einem χ^2-Wert von 16.19 bei $df = 15$ nicht signifikant (siehe Tabelle 6.60). Eine Reduktion zum Rasch-Modell ohne Lerneffekte wäre für diesen verkürzten Datensatz also auch möglich, dies deckt sich wieder mit der Forderung nach Rasch-Homogenität des WMT.

Tabelle 6.60: Modelltest Rasch-Modell des WMT von Formann, Waldherr & Piswanger

Rasch-LH	-932.6952
Kempf-LH	-924.6001
Likelihood-Ratio	16.1903
df	15
χ^2-Wert kritisch	25.00

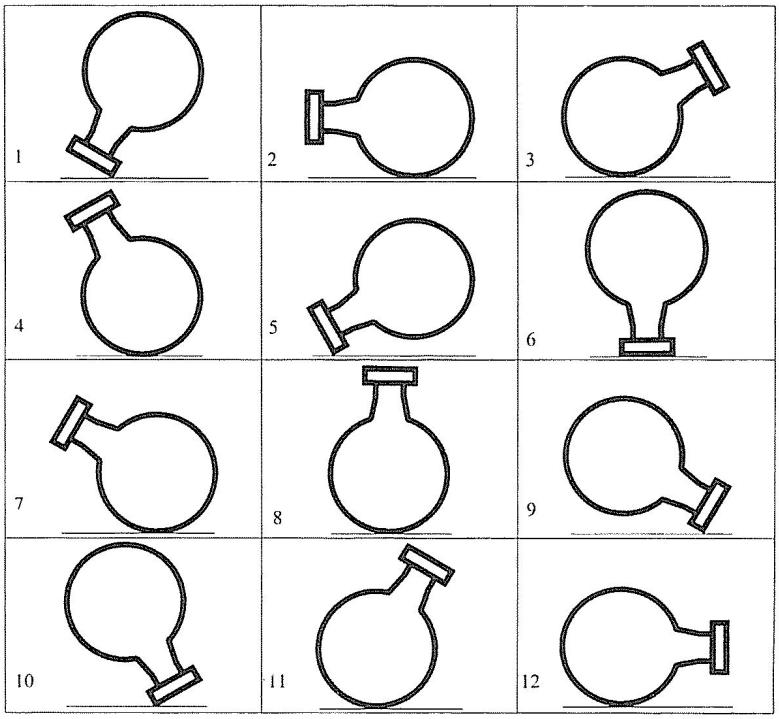
6.8 Water-Level Tasks von Formann

Die Water-Level Tasks wurden 1948 von Piaget und Inhelder erfunden, um die mentale Entwicklung der Raumvorstellung bei Kindern zu erfassen. Speziell für die Fähigkeit, sich das Verhalten von Flüssigkeiten im Raum vorstellen zu können, ist die Wahrnehmung, mentale Repräsentation und mentale Entwicklung eines horizontalen und vertikalen Raum- und Achsensystems notwendig. Um dies erfassen zu können, entwickelte Piaget einen Papier-Bleistift-Test. Es werden auf dem Papier verschiedene runde und zugestoppelte Flaschen in verschiedenen Neigungswinkeln präsentiert (siehe Abbildung 6.49). Die Testperson soll nun pro Flasche den Wasserstand als Linie einzeichnen, wie er aussehen würde, wenn die Flasche halb gefüllt wäre. Die Wasserfläche soll mit gestrichelten Linien eingezeichnet werden.

Der vorliegende Datensatz wurde im Sommersemester 2006 im Rahmen eines Forschungspraktikums II von Prof. Formann im zweiten Abschnitt des Psychologiestudiums in Wien erhoben. Die Stichprobe umfasste insgesamt 367 Personen beiderlei Geschlechts im Alter

zwischen 16 und 72 Jahren. Es wurden Hauptschüler, Lehrlinge, Maturanten und Hochschulabsolventen getestet. Den Personen wurde ein Testbuch mit 12 Seiten vorgelegt, auf jeder Seite befand sich eine leere Flasche. Die Neigung der Flaschen wurde, bezogen auf die Vertikale, mit Schritten von 30° verändert (also 0°, 30°, 60°, 90° usw. bis 330°). Abbildung 6.49 zeigt, wie diese Flaschen durch Zufallsauswahl den Seiten eins bis 12 zugeordnet wurden (siehe Formann, 2003).

Abbildung 6.49: Items der Water-Level Tasks nach Piaget



Die 12 Flaschen sind jedoch nicht gleich schwer. Die leichtesten 4 Items waren die Flaschen mit der Neigung von 0, 90, 180 bzw. 270°, in Abbildung 6.49 wären das die Items 2, 6, 8 und 12. Sie wurden nicht in die Bewertung miteinbezogen. Die übrigen 8 Items wurden mit dreierlei Toleranzabweichungen zur Horizontalen als *richtig* und *falsch* gewertet - mit 4, 7 und 10°.

Für die Analyse mittels DynTest erweisen sich die Daten in zwei Punkten als besonders problematisch:

Zum einen bleiben von den ursprünglichen 367 Personen nur wenige übrig, die *nicht* entweder alle oder kein Item richtig gelöst hatten. Damit möglichst wenige Personen ausgeschieden werden müssen, wurde von den drei Toleranzabweichungen diejenige ausgewählt,

bei der die meisten Personen zumindest ein Item falsch gelöst hatten. Das war das Toleranzniveau von 7° Abweichung von der Horizontalen. Trotzdem wurden 317 Personen ausgeschieden und es blieben lediglich 50 übrig. Dies ist jedoch sehr sehr wenig für eine genaue Parameterschätzung.

Zum anderen ist es für das dynamische Testmodell von Kempf nicht sinnvoll, Items, die vorgegeben wurden, anschließend aus dem Datensatz zu streichen, da die Personen ja auch durch diese Items etwas ge- oder verlernt haben können und (siehe Abschnitt 4.1) der eigentlich vollständige vorangegangene partielle Antwortvektor in die Transferparameter mit einfließen sollte.

Diese Kritikpunkte lassen nur eine - wenn überhaupt - sehr vorsichtige Interpretation der Ergebnisse der Parameterschätzung zu.

Das Genauigkeitskriterium für die Gesamtstichprobe wurde nach 7192 Iterationen erreicht. Die Schätzung in der ersten Stichprobe benötigte 282, in der zweiten Stichprobe 14009 Iterationen. Insgesamt dauerte die Schätzung aller Parameter 2 Minuten.

Ergebnisse

Die schwierigsten Items sind den Itemparametern des Rasch-Modells in Tabelle 6.61 zufolge die ersten beiden. Die Schwierigkeit ist tendenziell fallend, das letzte Item stellt das leichteste dar.

Tabelle 6.61: Geschätzte Parameter des Rasch-Modells der WLT

Item	Produktnormierte Itemleichtigkeit	Itemschwierigkeit	Personenfähigkeit
1	0.2680	3.5305	0.1134
2	0.3107	3.2181	0.2898
3	1.5094	0.6625	0.5680
4	0.8443	1.1844	1.0217
5	2.2114	0.4522	1.8215
6	1.0569	0.9462	3.4793
7	1.5094	0.6625	8.5512
8	2.5271	0.3957	

Dieses Bild wird durch die Itemschwierigkeitsparameter des Kempf-Modells bestätigt (siehe Tabelle 6.62), auch hier sind die ersten Items die schwersten und das letzte das leichteste. Die Transferparameter unterliegen einer starken Fluktuation. Der Lerngewinn bzw. -verlust steigt und fällt. Der niedrigste Transfer tritt nach zwei vorangegangen gelösten Items

auf, also bei Nummer 3, der höchste bei Nummer 8. Letzteres liegt wieder an einem letzten N_{ri} gleich 0, also wurde der letzte Transferparameter unsinnig hoch geschätzt.

Tabelle 6.62: Geschätzte Parameter des Kempf-Modells der WLT

Item	Mitte-normiert		Null-Eins-normiert	
	Itemschwierigkeit	Transfer	Itemschwierigkeit	Transfer
1	6.8659	0.2576	3.8670	-0.0414
2	6.1834	0.1483	3.4634	-0.1060
3	1.4557	0.0009	0.6672	-0.1932
4	2.3740	0.2035	1.2104	-0.0733
5	1.0022	0.3973	0.3990	0.0413
6	1.9069	0.0658	0.9341	-0.154
7	1.3805	0.5477	0.6228	0.1302
8	1.0009	0.9991	0.3983	0.3972

Tabelle 6.63 zeigt, dass das Kempf-Modell mit einem χ^2-Wert von 19.08 bei $df = 10$ knapp nicht gelten kann. Die Parameter dürfen also sowieso nicht interpretiert werden.

Tabelle 6.63: Modelltest Kempf-Modell der WLT

H0	-129.3574
H1	-43.0744
+	-76.7439
Likelihood-Ratio	19.0782
df	10
χ^2-Wert kritisch	18.31

Die Transfereffekte sind jedoch nach Tabelle 6.64 mit einem χ^2-Wert von 2.09 und $df = 7$ Freiheitsgraden nicht signifikant, d.h. eine Reduktion zum Rasch-Modell wäre möglich. Für die Water-Level Tasks treten also in dieser Stichprobe keine Lerneffekte auf und sie sind Rasch-homogen.

Tabelle 6.64: Modelltest Rasch-Modell der WLT

Rasch-LH	-130.3999
Kempf-LH	-129.3574
Likelihood-Ratio	2.0850
df	7
χ^2-Wert kritisch	14.07

7 Diskussion und Kritik

Dynamische Modelle können Lernen während einer Testbearbeitung abbilden. Personen- bzw. item- und operationsspezifische Lernmodelle berücksichtigen nicht die vorangegangenen Reaktionen der Testpersonen, dies können nur reaktionskontingente Modelle. An sich sind diese drei Herangehensweisen stimmig und nachvollziehbar für verschiedene Fragestellungen. Einmal interessieren die Personenfähigkeiten, ein anderes Mal die Schwierigkeit der Items bzw. der dazu benötigten Operationen und schließlich das (Lösungs-)Verhalten der Person. Besonders eingehend wurde letztere Gruppe in der vorliegenden Arbeit vorgestellt. Innerhalb der reaktionskontingenten Lernmodelle sind wiederum verschiedene Modellansätze unterscheidbar. Das eine basiert auf dem Prinzip der Markov-Ketten und der LCA, ein zweites auf dem LLTM und das dritte - für diese Arbeit wichtigste - bildet eine Verallgemeinerung des Rasch-Modells. Die Herangehensweisen an reaktionskontingentes Lernen sind somit wieder breit gefächert. Der Leser mag selbst das für ihn passendste bzw. angenehmste Modell wählen.

Im Zuge der Anwendung und der Adaptierung des Computerprogramms ergaben sich für das dynamische Testmodell von Kempf einige Kritikpunkte. Einige Eigenschaften des Modells bzw. auch des Programms sind problematisch.

Die wohl grundsätzlichste Kritik am dynamischen Testmodell von Kempf muss die Interpretierbarkeit der Transferparameter betreffen. Kempf & Hampapa (1975) geben als Stärke des Modells an, dass für die Transferparameter nicht wichtig ist, welche, sondern nur wie viele Items bearbeitet wurden. Genau dieser Punkt stellt nach Meinung der Autorin aber die größte Schwäche des Modells dar. Gerade bei psychologischen Leistungstests können die Transferparameter in vielen Fällen nicht sinnvoll interpretiert werden. Was nützt es zu wissen, dass beispielsweise nach der Bearbeitung von 4 Items ein Lernabfall stattfindet, wenn man nicht weiß, nach *welchen* vier Items. Angenommen die Person hat die ersten und die letzten zwei Items eines 20 Item-langen Tests gelöst, dann würde das denselben Lerngewinn oder -verlust bedeuten wie bei einer Person, die die Items 10-13 richtig gelöst hat. Wie kann

dieser Abfall der Transferparameter begründet werden, da ja möglicherweise völlig andere Items beteiligt waren? Nach Meinung der Autorin wäre allenfalls eine Interpretation von Persönlichkeits- oder Einstellungsfragebogen möglich, in dem man mit den Transferparametern die Tendenz zu einer bestimmten Meinung oder Persönlichkeitseigenschaft misst. Dann könnte ein Abfallen der Transferparameter wirklich mit einer Art „Karthasis“ oder Meinungsänderung in Verbindung gebracht werden, ein Steigen der Transferparameter würde dann die Tendenz zur “Verstärkung“ einer Meinung bedeuten.

Kempf gibt als Idealfall nur steigende oder nur sinkende Transferparameter an. Er begründet auf- und absteigende Transferparameter mit wechselnden Prozessen der Lernhemmung bzw. Konzentrationsschwäche und Lerneffekten. In der Praxis zeigte sich jedoch, dass die Lernparameter bei keinem Test nur stiegen oder nur fielen, sondern sich die Werte auf und ab bewegten. Bei kontinuierlichem Steigen und Sinken wäre die inhaltliche Interpretation der Transferparameter leichter. Ein Test müsste auf jeden Fall Items haben, die von *allen* Personen kontinuierlich und in *der selben* Reihenfolge bearbeitet werden, um optimale Bedingungen für die serielle Abhängigkeit und somit die Interpretierbarkeit der Parameter zu schaffen. Wenn Items ausgelassen werden, oder ein vorangegangenes Item etwa durch ein Zurückblättern erneut bearbeitet werden kann, beeinflusst dies das kontinuierliche Lernen. Der Test muss auch eindimensional dieselbe Fähigkeit messen, da sonst durch seine Bearbeitung verschiedene Fähigkeiten angesprochen werden und Lernen dadurch nicht kontinuierlich im selben Bereich stattfinden kann. Das Testmodell von Kempf kann aufgrund seiner dynamischen Komponente auch für Kurzzeitlerntests (siehe Abschnitt 2) angewendet werden. Kontinuierliches Dazulernen kann dabei auch durch zusätzliches Feedback oder Hilfestellungen seitens des Testleiters über richtig oder falsch gelöste Items gefördert werden. Wenn kein Feedback gegeben wird, kann eine Person lediglich durch „Warm-Werden“ oder Einarbeiten in die geforderte Fähigkeit bzw. das Gebiet, das der Test abfragt, dazulernen, und stetiges Dazulernen ist weniger leicht. Die Anwendung von DynTest auf einen Kurzzeitlerntest mit Feedback wäre daher ein sinnvolles Ziel für zukünftige Untersuchungen. Es darf nur kein Test verwendet werden, der adaptiv vorgegeben wird, da dann nicht alle Personen die gleichen Items in der gleichen Reihenfolge bearbeiten.

Kempf selbst (1975) gab mehrere Kritikpunkte des Computerprogramms zur Parameterschätzung zu bedenken. Zum einen kann es bei den Schätzgleichungen des Fortran-Programms zu Problemen kommen. Die Delta-Funktionen sind wesentlich kleiner als die Gamma-Funktionen. Wenn nun die Anzahl der Items sehr groß ist und/oder die Itemschwierigkeit

große Variation zeigt, können große numerische Ungenauigkeiten bei der Berechnung der G-Funktionen und deren erster partieller Ableitungen auftreten. Im schlimmsten Fall kommt es zu einem berechneten Wert von $G(k;s) < 0$ und somit zu unsinnigen Parameterschätzern.

Eine weitere Schwierigkeit tritt dann auf, wenn - wie bereits erwähnt - der/die letzte(n) N_{ri}, also die Häufigkeiten mit der Personen ein Item i falsch beantworten, nachdem sie r richtig beantwortet haben, gleich Null oder sehr klein sind. Dann sind zu wenig Personen vorhanden, um den/die letzten Transferparameter zu schätzen und der Parameter nimmt einen sehr großen Wert an, der aber so nicht interpretiert werden kann. Es wäre eine zukünftige Aufgabe, eine Abbruchbedingung einzubauen, nach deren Erfüllung der Parameter gar nicht geschätzt wird. Dies könnte sich jedoch noch schwierig gestalten, da nicht ganz klar ist, wie groß die Häufigkeit sein muss, um genaue Schätzungen vornehmen zu können.

Im Zuge der Testung des Programms stellte sich heraus, dass bei Weitem nicht jeder Datensatz für DynTest geeignet war. In einigen Datensätzen konnten die Parameter überhaupt nicht geschätzt werden, da entweder die Rechenungenauigkeit zu groß war oder das Programm wegen ungeeigneter Daten die Schätzung zu früh abbrach. Wie bereits oben erwähnt, mussten bei einem Datensatz sukzessive Items eliminiert werden, erst dann konnte die Schätzung durchgeführt werden. Die Schätzung der Parameter dauerte mitunter jedoch bei einigen einzelnen Simulationsdatensätzen bis zu mehreren Stunden, was sich ebenfalls mühsam gestaltete.

Nicht alle in der vorliegenden Arbeit verwendeten Datensätze waren laut LPCM-Win 1.0 wohl konditioniert, trotzdem wurden von DynTest alle Parameter geschätzt. Das Rasch-Modell in LPCM-Win 1.0 teilt die Gesamtstichprobe allerdings nach anderen Gesichtspunkten in Untergruppen auf und es erfüllte immer nur eine dieser Untergruppen das Kriterium der Wohlkonditioniertheit nicht. Daher war für das Kempf-Modell diese fehlende Wohlkonditioniertheit vernachlässigbar.

In den Simulationen waren ca. 500 Personen für eine einigermaßen genaue Schätzung notwendig, darunter traten größere Abweichungen in den Parameterschätzern gegenüber den simulierten Parametern auf. Kempf (1975) wies selbst auf einen Grund für ungenaue Schätzungen hin. Es ist möglich, dass existierende Ungenauigkeiten in den Transferparameterschätzungen durch weitere Ungenauigkeiten der Itemparameterschätzungen ausgeglichen werden und die logarithmierte Likelihood ein Maximum erreicht, obwohl beide Parameter starke Abweichungen von den „korrekten“ Werten zeigen. Die Interpretation der Parameter soll auch aus diesem Grund immer sehr vorsichtig erfolgen. Dieses Phänomen in Form ei-

ner gleichzeitigen starken Abweichung der Item- und Transferparameter von den simulierten Parametern konnte allerdings bei den Simulationsdatensätzen nicht beobachtet werden.

Alles in Allem kann das Kempf-Modell aber auf jeden Fall dazu genutzt werden, um festzustellen ob *überhaupt* Lernprozesse während eines Tests auftreten. Wenn das Kempf-Modell gilt, heißt das, dass diese Prozesse in signifikantem Maße auftreten. *Welche* Lernprozesse das aber sind, muss gut überdacht und mit Vorsicht interpretiert werden.

8 Zusammenfassung

Während der Bearbeitung eines Tests können Lern-, aber auch Verlerneffekte auftreten. Um dynamisches Lernen, also Lernen während einer Testbearbeitung, zu messen, wurden verschiedene Modelle konstruiert. Man kann eine Unterscheidung treffen zwischen Modellen, die personenspezifisches, item- bzw. operationsspezifisches und reaktionskontingentes Lernen erfassen.

Als Beispiel für personenspezifische Lernmodelle dient das Modell von Klauer & Sydow (1992). Diese gehen davon aus, dass Lernen von der Anzahl der von den Personen bearbeiteten Items abhängt. Ein Item wird von der Person entweder selber gelöst, oder es erfolgt eine Hilfestellung durch den Testleiter. Durch beides findet Lernen statt. Die Modellstruktur basiert auf der logistischen Funktion des Rasch-Modells, zu dem noch zusätzlich ein Lernzuwachsparameter eingeführt wird.

Als Vertreter der operations- und itemspezifischen Lernmodelle ist das operationsspezifische linear logistische Denkmodell von Spada (1976) zu nennen, das eine Erweiterung des linear logistischen Denkmodells von Scandura (1973) darstellt und auf das LLTM von Fischer & Formann (1972) zurückgeht. Bei Spada findet Lernen durch Üben von Operationen statt. Er führt einen Parameter ein, der den Effekt des Übens einer Operation auf die Operationsschwierigkeit eines Items beschreibt. Der Effekt des Übens hängt von der Übungshäufigkeit einer Operation ab.

Die reaktionskontingenten Lernmodelle beschreiben Lernen in Abhängigkeit von vorangegangenen Reaktionen der Person. Sie beinhalten die gemischten und latenten Markov-Modelle (z.B. Langeheine & Van de Pol, 1990), in denen Personen mit einer bestimmten Wahrscheinlichkeit von einem Zustand bei einem Zeitpunkt zu einem anderen Zustand zu einem anderen Zeitpunkt wechseln können, das Modell von Verhelst & Glas (1993), das auf dem LLTM mit inkomplettem Design basiert und in dem zusätzliche Parameter für die Lernrate und die gegebenen Reinforcements eingeführt werden, und das dynamische Testmodell von Kempf (1974).

Das dynamische Testmodell von Kempf (1974) basiert in seiner Modellstruktur auf der BTL-Darstellung des Rasch-Modells. Zusätzlich zu den Itemschwierigkeits- und Personenfähigkeitsparametern werden noch Lern- oder Transferparameter eingeführt. Diese beschreiben den Lerneffekt pro (partiellem) Rohscore in Abhängigkeit von der Anzahl der bisher gelösten Items. Dabei ist es nicht wichtig, welche Items eine Person gelöst hat, sonder nur wie viele. Das Kempf-Modell stellt eine Verallgemeinerung des Rasch-Modells dar - wenn alle Transferparameter gleich Null sind, gilt das Rasch-Modell.

Kempf & Hampapa bzw. Kempf & Mach (1975) entwickelten ein Fortran-Programm zur Schätzung der Item- und Transferparameter. Dieses Programm („DynTest“) wurde neu adaptiert und erweitert. Es besteht neben der Hauptroutine aus insgesamt 19 Subroutinen. Um die Benutzerfreundlichkeit zu erhöhen, wurde in Java eine Graphische Benutzeroberfläche (GUI) geschaffen, in die die nötigen Parameter, wie etwa die Anzahl der Personen, die Anzahl der Items, der Name des Datensatzes an sich, die gewünschte Ausgabedatei, oder Genauigkeitsanforderungen an die Schätzung eingegeben werden können. Nach der Schätzung öffnet sich das Ausgabefenster automatisch.

Am Originalprogramm wurde Folgendes verändert: Die Erweiterung der Personenanzahl von 450 auf 1000000 und der Itemanzahl von 20 auf 100, die Schätzung der Item- und Personenparameter des Rasch-Modells (siehe Fischer, 1974) und ein Modelltest zur Prüfung, ob das Kempf-Modells zum Rasch-Modell reduziert werden kann.

Im Programm werden zunächst die Item- und Personenparameter des Rasch-Modells geschätzt. Mittels Gradientenmethode werden dann die Item- und Transferparameter des Kempf-Modells iterativ geschätzt. Und schließlich werden zwei Modellgeltungstest durchgeführt. Für den Test auf Geltung des Kempf-Modells wird die Stichprobe zu diesem Zweck in zwei Untergruppen mit hohem und niedrigem Score aufgeteilt und die Likelihoods dieser beiden Gruppen durch einen Likelihoodquotiententest miteinander verglichen. Ob das Kempf-Modell auf das Rasch-Modell reduziert werden kann, wird durch einen Likelihoodquotiententest mit den Gesamtlikelihoods der beiden Modelle überprüft. Die Item- und Personenparameter des Rasch-Modells, die Item- und Transferparameter des Kempf-Modells in drei verschiedenen Normierungen für die Gesamtstichprobe und beide Untergruppen und die beiden Modellgeltungstests für das Kempf- und das Rasch-Modell werden in der Ausgabe aufgeführt.

Zur Anwendung und genaueren Untersuchung des Fortran-Programms wurden einerseits

zwei Simulationsreihen mit 8 und 20 Items durchgeführt. Mit 8 Items wurden für 100, 500, 1000, 5000 und 100000 Personen und mit 20 Items für 500, 1000 und 5000 Personen jeweils 100 Datensätze simuliert und der Durchschnitt aus den Kempf-Modell-Parameterschätzern ermittelt. Geschätzt wurden alle Datensätze einmal mit den alten Gamma-Funktionen und einmal mit den neuen. Dadurch konnte festgestellt werden, dass sich zwischen beiden Schätzmethoden keine Unterschiede in der Genauigkeit ergeben. Weiters wurden auch die Standardabweichungen und Varianzen für die Item- und Transferparameterschätzer berechnet. Die größte Streuung bei den Parametern ergab sich meistens für das erste und die beiden letzten Parameter. Es trat auch die Schwierigkeit auf, dass manche Datensätze den letzten Score nicht aufwiesen, mit dem die Häufigkeit angegeben wird, dass Personen ein Item falsch beantworten, nachdem sie vorher eine Anzahl von Items richtig beantwortet hatten. Wenn dieser Score gleich 0 oder einfach sehr klein ist, werden Transferparameter von über 0.9 geschätzt, die aber so nicht stimmen (können). Allgemein wurde aber mit steigender Personenanzahl die Schätzung der Parameter genauer, ab 500 Personen wies der Zusammenhang zwischen simulierten und geschätzten Parametern bereits ein r^2 von über 0.95 für Item- und Transferparameter auf, die Schätzung war also schon ab 500 Personen relativ gut.

Andererseits wurden folgende echte Datensätze herangezogen und mit DynTest analysiert. Für einen Datensatz des Mathematiksubtest der PISA-Studie (20 Items, 6702 Personen) musste sowohl die Annahme der Geltung des Kempf-Modells als auch die Reduktion zum Rasch-Modell verworfen werden. Beide Modellgeltungstests wurden hoch signifikant.

Für Daten von Bahrick & Hall (7 Items, 1074 Personen) galt das Kempf-Modell auch nicht, es konnte ebenfalls nicht auf das Rasch-Modell reduziert werden.

3-DW Testdaten von Gittler (17 Items, 866 Personen) konnte auf das Rasch-Modell reduziert werden, das Kempf-Modell fand keine Geltung.

Es wurden SPM-Daten von Schmöger von Erwachsenen und Kindern analysiert, lediglich die Parameter in 3 Subtests konnten geschätzt werden. Für die Erwachsenen im Subtest C (12 Items, 343 Personen) galt das Kempf-Modell, es konnte nicht auf das Rasch-Modell reduziert werden. Für die Kinder (12 Items, 626 Personen) galt das Kempf-Modell im Subtest C nicht, das Rasch-Modell jedoch schon, im Subtest E (12 Items, 626 Personen) galt weder das Kempf- noch das Rasch-Modell.

Die Analyse eines WMT-Datensatzes von Weber (24 Items, 521 Personen) ergab keine Geltung des Kempf-Modells, jedoch eine zulässige Reduktion zum Rasch-Modell.

Bei anderen, revidierten WMT-Daten von Formann, Waldherr & Piswanger (21 Items, 277

Personen) konnten lediglich 16 Items analysiert werden. Für diesen Datensatz galten sowohl das Kempf- als auch das Rasch-Modell.

Bei einem Water-Level Tasks-Datensatz von Formann (8 Items, 367 Personen) blieben lediglich 50 Personen über, die nicht alle oder kein Item richtig gelöst hatten. Das Kempf-Modell galt bei diesem Datensatz nicht, eine Reduktion zum Rasch-Modell war jedoch möglich.

Der Hauptkritikpunkt des Modells bzw. Programms ist zum einen die Frage nach der inhaltlichen Interpretation der Transferparameter. Eine inhaltlich sinnvolle Interpretation ist unter anderem nur bei Tests möglich, die immer in der gleichen Reihenfolge lückenlos von allen Personen bearbeitet werden. Ein Feedback nach jeder Bearbeitung wäre ebenfalls sinnvoll, um Lerneffekte zu verstärken. Zum anderen kann nicht jeder Datensatz für die Analyse herangezogen werden, da in vielen Fällen die Rechenungenauigkeit zu groß ist, oder die Analyse vorzeitig abgebrochen wird, weil die Daten einen partiellen Score oder Vektor nicht aufweisen, der aber für die Schätzung der Transferparameter benötigt wird. Zusätzliche Programmmodifikationen, Verbesserungen und Erweiterungen sind möglich und bleiben eine Herausforderung für die Zukunft.

Literaturverzeichnis

[1] Andersen, E.B. (1971). Asymptotic Properties of Conditional Likelihood Ratio Tests. *Journal of the American Statistical Association*, 66, 630-633.

[2] Bahrick, H. P. & Hall, L.K. (1991). Lifetime maintenance of high school mathematics content. *Journal of Experminental Psychology: General*, 120, 20-33.

[3] Blumen, I.M., Kogan, M. & McCarthy, P.J. (1955). *The industrial mobility of labor as a probability process.* Ithaca: Cornell University Press.

[4] Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM alogrithm. *Psychometrika*. 46, 443-459.

[5] Bradley, R.A. & Terry, M.E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparison. *Biometrika*, 39, 324-345.

[6] Buckingham, B.R. (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology*, 12, 271-275.

[7] Budoff, M., Meskin, J. & Harrison, R.H. (1971). Educational test of the learning potential hypothesis. *American Journal of Mental Deficiency*, 76, 159-169.

[8] Converse, P.E. (1964). The nature of belief systems in mass publics. In: D.E. Apter (Ed.). *Ideology and discontent* (pp.206-261). New York: The Free Press.

[9] Converse, P.E. (1970). Attitudes and non-attitudes: Continuation of a dialogue. In: E.R. Tufte (Ed.). *The quantitative analysis of social problems.* (pp. 168-189). Reading: Addison-Wesley.

[10] Feuerstein, R., Rand, Y. & Hoffmann, M.B. (1979). *The dynamic assessment of retarded performers: the learning potential assessment device, theory, instruments and techniques.* Baltimore: University Park Press.

[11] Fischer, G.H. (1972). Conditional maximum-likelihood estimation of item parameters for a linear logistic model. *Research Bulletin. No. 9*, Vienna: University of Vienna, Institute of Psychology.

[12] Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

[13] Fischer, G.H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, 48, 3-26.

[14] Fischer, G.H. (2003). The Precision of Gain Scores Under an Item Response Theory Perspective: A Comparison of Asymptotic and Exact Conditional Inference About Change. *Applied Psychological Measurement*, 27(1), 3-26.

[15] Fischer, G.H., & Formann, A.K. (1972). An algorithm and a FORTRAN program for estimating the item parameters of the linear logistic test model. *Research Bulletin No. 24*, Vienna: University of Vienna, Institute of Psychology.

[16] Fischer, G.H. & Molenaar, I.W. (Eds.). (1995). *Rasch models: their foundations, recent developments and applications*. New York: Springer.

[17] Formann, A.K. (2003). Modeling Data from Water-Level Tasks: A Test Theoretical Analysis. *Perceptual and Motor Skills*, 96, 1153-1172.

[18] Formann, A.K., Piswanger, K. (Hrsg.) (1979). *Wiener Matrizen-Test. Ein Rasch-skalierter sprachfreier Intelligenztest*. Weinheim: Beltz.

[19] Formann, A.K., Waldherr, K., Piswanger, K. (Hrgs.) (im Druck). *Revision des Wiener Matrizen-Tests* (Arbeitstitel).

[20] Glas, C.A.W. (1988). The Rasch model and multi-stage testing. *Journal of Educational Statistics*, 13, 45-52.

[21] Goodman, L.A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, 56, 841-868.

[22] Guthke, J. (1990). Learning tests as an alternative or completion of intelligence tests: a critical review. *European Journal of Psychology of Education*, 5, 117-133.

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

[24] Gittler, G. (1990). *3DW. Dreidimensionaler Würfeltest. Ein raschskalierter Test zur Messung des räumlichen Vorstellungsvermögens*. Weinheim: Beltz.

[25] Flammer, A. & Schmid, H. (1982). Lerntests: Konzept, Realisierungen, Bewährung. Eine Übersicht. *Schweizerische Zeitschrift für Psychologie*, 41, 114-138.

[26] Held, T. & Korossy, K. (1998). Data analysis as a heuristic for establishing theoretically founded item structures. *Zeitschrift für Psychologie*, 206, 169-188.

[27] Jensen, A.R. (1961). Learning abilities in Mexican-American and Anglo-American children. *California Journal of Educational Research*, 12(4), 147-159.

[28] Kempf, W.F. (1974). Dynamische Modelle zur Messung sozialer Verhaltensdispositionen. In: W.F. Kempf (Hrsg.). *Probabilistische Modelle in der Sozialpsychologie* (pp. 13-55). Bern: Huber.

[29] Kempf, W.F. (Hrsg.) (1974). *Probabilistische Modelle in der Sozialpsychologie*. Bern: Huber.

[30] Kempf, W.F. & Hampapa, P. (1975). The numerical solution of a set of conditional estimation equations arising in a dynamic test model. In: Kempf, W.F., Hampapa, P. & Mach, G. (Eds.). *Conditional maximum likelihood estimation for a dynamic test model* (pp. 5-32). Arbeitsbericht 13, Institute for Science Education at the University of Kiel.

[31] Kempf, W.F. & Mach, G. (1975). A Fortran program for CML estimation in a dynamic test model. In: Kempf, W.F., Hampapa, P. & Mach, G. (eds.). *Conditional maximum likelihood estimation for a dynamic test model* (pp. 33-61). Arbeitsbericht 13, Institute for Science Education at the University of Kiel.

[32] Kempf, W.F. (1977). A dynamic test model and its use in the microevaluation of instructional material. In: Spada, H. & Kempf, W. F. (Eds.). *Structural models of thinking and learning* (pp. 295-318). Proceedings of the IPN-Symposium 7, Kiel 1975. Bern: Huber.

[33] Kempf, W.F., Hampapa, P. & Mach, G. (Eds.). (1975). *Conditional maximum likelihood estimation for a dynamic test model*. Arbeitsbericht 13, Institute for Science Education at the University of Kiel.

[34] Kern, B. (1930). *Wirkungsform der Übung.* Münster: Helios.

[35] Klauer, K.C. & Sydow, H. (1992). Interindividuelle Unterschiede in der Lernfähigkeit. Zur Analyse von Lernprozessen bei Kurzzeitlerntests. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 3, 175-190.

[36] Kratzmeier, H. & Horn, R. (1987). *Standard Progressive Matrices* (2. Auflage). Weinheim: Beltz.

[37] Langeheine, R. & Van de Pol, F. (1990). Veränderungsmessung bei kategorialen Daten. *Zeitschrift für Sozialpsychologie*, 21, 88-100.

[38] Luce, R.D. (1959). *Individual Choice Behavior*. New York: Wiley.

[39] Luce, R.D., Bush, R.R. & Galanter, E.(Eds.) (1963). *Handbook of mathematical psychology*. New York: Wiley.

[40] Macready, G.B. & Dayton, C.M. (1980). The nature and use of state mastery learning models. *Applied Psychological Measurement*, 4, 493-516.

[41] Morgan, T.M., Aneshensel, C.S. & Clark, V.A. (1983). Parameter estimation for mover-stayer models: Analyzing depression over time. *Sociological Methods & Research*, 11, 345-366.

[42] Piaget, J. & Inhelder, B. (1948). *La représentation de l'espace chez l'enfant [Spatial representation in children]*. Paris: Presses Univer. de France.

[43] Rogosa, D.R. & Wilett, J.B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, 50, 203-228.

[44] Rohwer, W.D., Jr. (1971). Learning, race, and school success. *Review of Educational Research*, 41(3), 191-210.

[45] Rost, J. (2002). Mixed and latent Markov models as item response models. *Methods of Psychological Research (MPR-online)*, 7, 53-72.

[46] Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2. Auflage). Bern: Huber.

[47] Scandura, J.M. (1973). *Structural learning I. Theory and research.* New York: Gordon & Breach.

[48] Severson, R.A. (1976). Environmental and emotionally-based influences upon the learning process. *American Psychological Association Convention*, Washington D.C.

[49] Spada, H. (1976). *Modelle des Denkens und des Lernens*. Bern: Huber.

[50] Spada, H. & Kempf, W. F. (Eds.) (1977). *Structural models of thinking and learning*. Proceedings of the IPN-Symposium 7, Kiel 1975. Bern: Huber.

[51] Sternberg, S.H. (1959). A path dependent linear model. In: Bush, R.R. & Estes, W.K. (Eds.). *Studies in mathematical learning theory* (pp. 308-339). Stanford: Stanford University Press.

[52] Sternberg, S.H. (1963). Stochastik learning theory. In: Luce, R.D., Bush, R.R. & Galanter, E. (Eds.). *Handbook of mathematical psychology, Vol. II* (pp 1-120). New York: Wiley.

[53] Van de Pol, F., Langeheine, R. & de Jong, W. (1996). *PANMARK 3. User's manual. PANel analysis using MARKov chains. A latent class program.* Voorburg: The Netherlands.

[54] Verhelst, N.D. & Glas, C.A.W. (1995). Dynamic generalizations of the Rasch model. In: Fischer, G.H. & Molenaar, I.W. (Eds.). *Rasch models: their foundations, recent developments and applications* (pp.181-202). New York: Springer.

[55] Weber, M. (1999). *Motivationale Aspekte einer umfassenden computer-unterstützten Leistungsdiagnostik von Lehrlingskandidaten*. Unveröffentlichte Diplomarbeit. Universität Wien.

[56] Weber, M. (2005). *Die Anwendbarkeit probabilistischer Modelle im Rahmen der Wissensraumtheorie*. Unveröffentlichte Dissertation. Universität Wien.

[57] Wiseman, S. (1954). Symposium on the effects of coaching and practice in intelligence tests. IV. The Manchester experiment. *British Journal of Educational Psychology*, 24, 5-8.

[58] Wygotski, L.S. (1964 Russ. 1934). *Denken und Sprechen*. Berlin: Akademie-Verlag.

[59] Zimmermann, D.W. & Williams, R.H. (1982a). The relative error magnitude in three measures of change. *Psychometrika*, 47, 141-147.

[60] Zimmermann, D.W. & Williams, R.H. (1982b). On the high predictive potential of change and growth measures. *Educational and Psychological Measurement*, 42, 961-968.

[61] Zubin, J. (1950). Symposium on statistics for the clinican. *Journal of Clinical Psychology*, 6, 1-6.

[62] URL: http://www.eclipse.org/ Stand: 16.5.2008

[63] URL: http://www.silverfrost.com/16/plato3.asp Stand: 16.5.2008

Lebenslauf

Zur Person

Mag. rer. nat. Marlis Posch

geboren am **5. Juli 1981** in Wien, Österreich

ledig, keine Kinder

Ausbildung

2005/06	Lehrgang zur Klinischen- und Gesundheitspsychologin, KlinGes, Wien
seit 2004	Dissertationsstudium der Psychologie, Universität Wien
1999-2004	Diplomstudium der Psychologie, Universität Wien
1999	Matura am BG/BRG Schwechat

Berufliche Tätigkeiten

2004-2008	Univ. Ass. i.A. am Institut für psychologische Grundlagenforschung, Fakultät für Psychologie, Universität Wien
seit WS 2005/06	Lehrveranstaltungsleiterin für die Übungen zur Psychologischen Methodenlehre und Statistik I und II, Universität Wien
Sommer 2004	Psychologische Leitung im Sommercamp "Fit statt dick", Pressbaum
2002/03	Praktikum im Verein SOPS - Sozialpädagogische Betreuungs- und Beratungsstelle, Schwechat
Sommer 2000/01	Kundenaquisition für die BA-CA, Wien

Wien, 10. September 2008