Fachbereich Erziehungswissenschaft und Psychologie
der Freien Universität Berlin

Ein probabilistisches Testmodell zur Erfassung intraindividueller Variabilität

Dissertation
zur Erlangung des akademischen Grades
Doktor der Philosophie (Dr. phil.)

vorgelegt von

Dipl. - Psych.
Hosoya, Georg

Berlin, 2012

Datum der Disputation: 19.02.2013

Erstgutachter: Prof. Dr. Michael Eid

Zweitgutachter: Ass.Prof. Mag.Dr. Rainer W. Alexandrowicz

Danksagung

Bedanken möchte ich mich bei Prof. Dr. Jürgen Bortz (†), der es mir ermöglicht hat, mich auf eigene Verantwortung mit einem ziemlich unorthodoxen Thema zu beschäftigen, Prof. Dr. Michael Eid, der mir mit seinen kritischen und konstruktiven Fragen geholfen hat, einige Ungenauigkeiten zu beseitigen und Dipl.-Psych. Claudia Crayen sowie Prof. Dr. Tanja Lieschetzke, die mir freundlicher Weise den in dieser Arbeit verwendeten Datensatz zur Erprobung des hier entwickelten Modells zur Verfügung gestellt haben.

Inhaltsverzeichnis

1. Einleitung und Überblick

Gegenstand der vorliegenden Arbeit ist die Entwicklung eines probabilistischen Testmodells zur Erfassung von intraindividueller Variabilität im Rahmen von Ambulatory Assessments (Fahrenberg & Myrtek, 1996) oder anderen massiv-längsschnittlichen Studien, bei denen multivariate, diskrete Zeitreihen von Personen auf manifesten Indikatoren für einen latenten Trait anfallen. Potentielle Anwendungsmöglichkeiten des Modells liegen zum Beispiel in der Psychologischen Diagnostik oder auch der Therapie-Evaluation (vgl. z.B. Ebner-Priemer, Eid, Kleindienst, Stabenow & Trull, 2009). Intraindividuelle Variabilität ist schon seit Stern (Stern, 1900, zitiert nach Eid & Diener, 1999) Gegenstand psychologischer Betrachtung und es existiert eine Reihe von Ansätzen zum methodischen Umgang mit dem Phänomen. Diese Ansätze werden in Kapitel 2 kurz beleuchtet. Bei der Sichtung der verwendeten Ansätze fällt auf, dass zur Thematik „intraindividuelle Variabilität“ probabilistische Ansätze der Psychodiagnostik rar sind. Kurz gesagt, es fehlt ein moderner Ansatz der Item-Response-Theorie zur Überprüfung der Hypothese, dass intraindividuelle Variabilität eine Eigenschaft ist, die gemessen werden kann und die Personen *interindividuell* voneinander unterscheidet. Diese konkrete Fragestellung wurde vor allem von Eid & Diener (1999) mit Hilfe von Latent-State-Trait-Modellen und der intraindividuellen Standardabweichung angegangen. Dabei hat sich gezeigt, dass intraindividuelle Variabilität, gemessen an der intraindividuellen Standardabweichung, durchaus reliabel und valide zwischen Personen hinsichtlich der Variabilität differenziert. Was bisher nicht vorliegt, ist ein Testmodell, dass es ermöglicht, probabilistische Aussagen über die Variabilität im Antwortverhalten von Personen zu treffen, wobei dieses Antwortverhalten von einem latenten Parameter im Sinne eines Traits abhängt, der mit der

Variabilität von manifesten, multivariaten Zeitreihen in Verbindung steht. Oder einfach ausgedrückt: Es fehlt ein Rasch-Modell zur psychometrischen Skalierung intraindividueller Variabilität als Trait.

Zur Entwicklung eines probabilistischen Testmodells stehen mindestens drei formale Ansätze zur Verfügung: Der klassische Ansatz von Rasch (1961), multinomiale Logit-Modelle (vgl. Skondral & Rabe-Hesketh, 2004, p. 74) und die Maximum-Entropie-Methode von Jaynes (Jaynes, 1957a, 1957b, 2003). In der vorliegenden Arbeit wird der letztere Modellierungsansatz verwendet, da dieser im Rahmen der psychologischen Methodik noch nicht näher exploriert wurde und dieser Ansatz einige formale Eigenschaften aufweist, die die Anwendung relativ einfach machen. Um die Verbindung zwischen dem klassischen Ansatz von Rasch und der Maximum-Entropie-Methode von Jaynes zu verdeutlichen, werden bekannte Rasch-Modelle unter der Anwendung der Methode in Kapitel 3 hergeleitet. In der aktuellen Forschung findet der Ansatz von Jaynes vor allem im Rahmen modernerer Ansätze der künstlichen Intelligenz unter dem Stichwort *probabilistic graphical models* (Koller & Friedman, 2009) eine Anwendung. Die näheren Beziehungen zu Probabilistischen Grafischen Modellen werden in dieser Arbeit nicht näher behandelt, da der Autor erst gegen Ende der Arbeit auf den breiten Literaturkorpus im Bereich des *machine learings* gestoßen ist, der sich dieser Thematik widmet.

Die Fragestellungen der Arbeit beziehen sich vor allem auf die Anwendbarkeit der Maximum-Entropie-Methode zur Generierung eines neuen probabilistischen Modells zur Erfassung intraindividueller Variabilität und die Untersuchung der Eigenschaften des resultierenden Modells. Ein zweiter Block von Fragestellungen bezieht sich auf die konkrete Anwendbarkeit des Modells auf einen realen, im Rahmen eines Ambulatory Assessments angefallenen Datensatzes, mit dem Ziel, intraindividuelle Variabilität probabilistisch zu skalieren, die Passung des Modells zu evaluieren und die Anwendbarkeit des Modells zu erproben. Die zwei Fragestellungs-Blöcke werden in den jeweiligen Kapiteln 4 und 5 konkretisiert.

In Kapitel 4 wird der klassische Ansatz von Jaynes verwendet, um ein probabilistisches Testmodell zur Erfassung der intraindividuellen Variabilität auf Basis der absoluten suk-

zessiven Differenzen von intraindividuellen, multivariaten Zeitreihen zu generieren. Dabei resultiert ein probabilistisches Modell, dass einen Markov-Prozess erster Ordnung definiert und dem Partial-Credit-Modell (PCM) von Masters (Masters, 1982) ähnelt. Die Eigenschaften des Modells, wie die Likelihood-Funktion, die Erwartungwerte und die Varianz der manifesten Variable, die Kategorien-Charakteristik-Kurven und die Übergangsmatrix des Markov-Prozesses werden dargestellt und der Modellparameter zur Erfassung der Personenfähigkeit wird simulativ auf Bias und empirische Varianz untersucht. Die Parameterschätzung erfolgt mit der Markov-Chain-Monte-Carlo-Methode (MCMC) (Metropolis, Rosenbluth, Teller & Teller, 1953; Gill, 2008; Gelman & Hill, 2007). Ferner wird gezeigt, wie die Bewertung der Reliabilität der Erfassung der Variabilität auf Basis von Andrichs Reliabilitäts-Index (Andrich, 1988) geschehen kann und es wird ein Verfahren zur Bewertung der Modell-Passung auf Basis von standardisierten Residuen (Wright & Stone, 1969) auf das generierte Modell übertragen.

In Kapitel 5 wird das Modell auf einen Ambulatory-Assessment-Datensatz von Crayen, Eid, Lischetzke, Courvoisier und Vermunt (in Druck) angewendet, um an einer realen Stichprobe zu überprüfen, ob sich die Variabilität auf drei Kurzskalen des Mehrdimensionalen Befindlichkeitsfragebogens (MDBF) (Steyer, Schwenkmezger & Eid, 1997) reliabel erfassen lässt und inwiefern das Modell die Daten hinreichend genau abbildet. Zudem werden die anhand der Stichprobe geschätzten Parameter des Modells explorativ mit drei Subskalen des NEO-FFI (Borkenau & Ostendorf, 1993) (Neurotizismus, Extraversion und Gewissenhaftigkeit) korreliert.

In Kapitel 6 werden die Ergebnisse der Arbeit vor den theoretischen Hintergründen diskutiert.

2. Psychologischer Hintergrund

2.1. Konzeptuelle und definitorische Rahmen zur intraindividuellen Variabilität

In der Differentiellen Psychologie ist es von Interesse, inwiefern sich Individuen hinsichtlich eines empirisch abgrenzbaren Merkmals unterscheiden und inwiefern diese Unterschiede mit anderen Konstrukten zusammenhängen. Die Erfassung dieser Merkmale mit Hilfe von mathematischen Modellen ist traditionell Gegenstand der Psychometrie. Im Haupt-Fokus des Interesses stehen meist Merkmale, die als Traits konzipiert sind. Ein Trait ist eine relativ stabile Verhaltensdisposition eines Individuums, die unabhängig von der Situation ist und deren quantitative Ausprägung sich mittels manifester Indikatoren für diesen Trait mit Hilfe von mathematischen Modellen erfassen lässt. In einigen Modellen, wie z.B. in denen der klassischen und probabilistischen Testtheorie, gilt der Rohwert der Testleistung als manifester Indikator der latenten Merkmalsausprägung, sofern die Indikatoren ein gemeinsames, modelltheoretisch abgesichertes, empirisch abgrenzbares Merkmal erfassen.

Ein weiterer Gegenstand, der die Psychologie spätestens seit Stern (Stern, 1900, zitiert nach Eid & Diener, 1999) beschäftigt, ist die Variabilität und Dynamik menschlichen Verhaltens und Erlebens. Zur quantitativen Untersuchung dieses Phänomens werden substanzwissenschaftliche Theorien und mathematische Modelle benötigt, mit denen sich Hypothesen, die aus den Theorien abgeleitet wurden, prüfen lassen. Ein Blick in die Literatur zeigt, dass eine Vielzahl unterschiedlicher Theorien und modelltheoretischer Ansätze zur Untersuchung des Phänomens intraindividuelle Variabilität bestehen. Ein

klassischer Artikel zur Definition des Konstrukts intraindividuelle Variabilität ist derjenige von Fiske und Rice (1955), in dem einige klassische Forschungsfragen, die mit dem Gegenstand verbunden sind, formuliert werden.

2.1.1. Intraindividuelle Variabilität nach Fiske und Rice (1955)

1955 veröffentlichten Fiske und Rice (1955) einen wegweisenden Artikel über intraindividuelle Reaktionsvariabilität, der den Stand der Forschung zu dem damaligen Zeitpunkt zusammenfasst. In dem Artikel definieren Fiske und Rice (1955) das Konstrukt „intraindividuelle Variabilität“ und unterscheiden drei Typen der intraindividuellen Variabilität.

Typ-I-Variabilität, auch als reine intraindividuelle Variabilität bezeichnet, ist die Differenz zweier Reaktionen eines Individuums zu zwei Messzeitpunkten unter sonst gleichen Bedingungen. Fiske und Rice (1955) räumen ein, dass diese Definition relativ abstrakter Natur und eher von theoretischem Interesse ist, da es relativ schwierig ist, ein Individuum unter mehreren, sonst identischen Bedingungen zu beobachten.

Ferner schränken Fiske und Rice (1955) die Definition dahingehend ein, dass innerhalb der Zeitreihe der Reaktionen des Individuums keine Trends, Autokorrelationen oder Oszillationen vorliegen sollten, um von Typ-I-Variabilität sprechen zu können. Zudem ist die Ordnung, bzw. die zeitliche Abfolge der Reaktionen irrelevant. Unter einer linearen Modellannahme wäre diese Art der Variabilität etwa so zu verstehen wie eine stationäre, Gaußsche Zeitreihe mit einem Mittelwert von μ und einer Streuung von σ, wobei diese Parameter Charakteristiken der Person sind, und sonst keine weiteren endogenen oder exogenen Variablen, wie z.B. die Situation oder internale Stimuli einen Einfluss auf die Reaktionen besitzen.

Werden innerhalb eine Zeitreihe Trends, Autokorrelationen und Oszillationen entdeckt oder zeigen sich signifikante Mittelwertsunterschiede in unterschiedlichen Segmenten der Zeitreihe, so wäre demnach nicht mehr von Typ-I-Variabilität zu sprechen, und es wäre zu fragen, welche internen oder externen Variablen mit den beobachteten Fluktuationen und Mittelwertsunterschieden korrespondieren.

Typ-II-Variabilität erlaubt Trends, Autokorrelationen und Oszillationen der Zeitreihe,

jedoch ist die Situations-Invarianz ebenfalls eine Bedingung, d.h. externale Einflüsse, wie z.B. der Einfluss der Situation auf die Reaktionen eines Individuums spielen hier - wie bei der Variabilität des Typs I - keine Rolle. Fiske und Rice (1955) konzipieren das Konstrukt als Reaktionen des Individuums auf internale Stimuli, bzw. als Reaktionen des Organismus auf vorhergehende Reaktionen, eine Annahme, die sich mit autoregressiven Modellen prüfen lässt.

Typ-III-Variabilität besteht aus den Komponenten des Typs-I und des Typs-II, allerdings ist nun zusätzlich eine Variation der situationalen Bedingung erlaubt. Das heißt, die Variabilität der Reaktionen eines Organismus ist einerseits auf intraindividuelle Variabilität des Typs-I, Reaktionen auf vorhergehende und internale Stimuli und auf externale Situationen zurückzuführen. Anders ausgedrückt: Die mögliche Kontextabhängigkeit menschlichen Verhaltens wird in dieser Definition berücksichtigt.

Ausgehend von den Definitionen zur intraindividuellen Variabilität formulierten Fiske und Rice (1955) einige zu klärende, programmatische Fragestellungen:

1. Ist es möglich, aus der klassischen Fehlervarianz der Psychometrie eine Varianzkomponente zu extrahieren, die mit dem Individuum zusammenhängt?

2. Ist es möglich, eine faktorielle Struktur intraindividueller Variabilität nachzuweisen?

3. Trägt das Wissen um die intraindividuelle Variabilität eines Individuums dazu bei, dessen Verhalten bzw. Erleben vorherzusagen?

4. Was ist die Bedeutung des Konzeptes innerhalb des Studiums der Persönlichkeit und der Persönlichkeitstheorien?

5. Hängt Variabilität mit einer Integration der Persönlichkeit zusammen?

6. Existiert eine physiologische Basis für das Phänomen?

Eid und Diener (1999) untersuchten diese Fragestellungen systematisch mit Latent-State-Trait-Modellen anhand von intraindividuellen Standardabweichungen als manifeste

Indikatoren, worauf in einem späteren Abschnitt über Modelle zur Erfassung intraindividueller Variabilität näher eingegangen wird.

2.1.2. Der theoretisch-konzeptionelle Rahmen von Ram und Gerstorf (2009) zur intraindividuellen Variabilität

Ram und Gerstorf (2009) geben einen guten Überblick über den derzeitigen Stand der Forschung aus überwiegend methodischer Perspektive und schlagen aufbauend auf Fiske und Rice (1955) einen konzeptuellen, heuristischen Rahmen vor, um bestehende Ansätze und Modelle zur Untersuchung des Phänomens zu ordnen.

Bezugnehmend auf Nesselroade (1991) grenzen Ram und Gerstorf (2009) intraindividuelle Variabilität (*intraindividual variability*) von intraindividueller Veränderung (*intraindividual change*) ab. Intraindividuelle Veränderung beseht nach Nesselroade (1991) aus bleibenden, längerfristigen Veränderungen menschlichen Verhaltens und Erlebens, denen das Konzept der Entwicklung zugrunde liegt. So handelt es sich z.B. bei der Veränderung kognitiver Leistungsprozesse im Verlauf der Lebensspanne, Altern, Reifung und Wachstum nicht um intraindividuelle Variabilität. Intraindividuelle Veränderungen sind längerfristige Prozesse, wogegen es sich bei intraindividueller Variabilität um Veränderungen, Oszillationen und Fluktuationen auf einer kurzfristigeren Mikro-Zeit-Skala (Minuten, Stunden, Tage, Wochen) handelt.

Konzeptionell differenzieren Ram und Gerstorf (2009) ferner zwischen *dynamischen Charakteristiken* einer Person und *dynamischen Prozessen*. Dynamische Charakteristiken sind Eingenschaften einer Person, wie z.B. deren inhärente Kapazität zur Veränderung, der ein Trait-Charakter zugesprochen wird. Hierzu gehören z.B. die Konstrukte Plastizität, Labilität, Rigidität und Robustheit. Dynamische Prozesse hingegen sind systematische Veränderungen im Verhalten und Erleben über die Zeit, die von Ram und Gerstorf in drei Klassen eingeteilt werden: 1. Prozesse zur Wahrung der Stabilität, 2. Inkrementelle Veränderungsprozesse und 3. Transformationale Veränderungsprozesse. Prozesse zur Wahrung der Stabilität dienen dazu, die funktionale Einheit und Stabilität eines Systems aufrechtzuerhalten. Diese Prozesse setzen dann ein, wenn ein System

durch innere oder äußere Einflüsse aus dem Gleichgewicht gebracht wurde. Inkrementelle Veränderungsprozesse sind durch kleinschrittige, langfristige, direktionale Veränderungen gekennzeichnet, wie dies z.B. beim Lernen nach dem Verstärkungs-Paradigma der Fall ist. Ziel dieser Veränderungen ist die schrittweise Verfeinerung und Elaboration schon vorhandener Charakteristiken eines Systems. Transformationale Veränderungsprozesse hingegen sind nach Ram und Gerstorf (2009) durch eine rapide, plötzliche Veränderung eines Systems in einen qualitativ anderen Zustand gekennzeichnet.

Dynamische Charakteristiken und dynamische Prozesse werden von Ram und Gerstorf (2009) in Verbindung mit den Konzepten *net-intraindividual variability* und *time structured intraindividual variability* gebracht. Diese Konzepte lehnen sich an die klassische Definition von Fiske und Rice (1955) zur intraindividuellen Variabilität an. *Net-intraindividual variability* entspricht in etwa der intraindividuellen Variabilität des Typs-I nach Fiske und Rice. Hierbei handelt es sich um eine Variabilität, bei der die Reihenfolge oder die zeitliche Abfolge einer Zeitreihe keine Rolle spielt. Das heißt unter der Annahme der net-intraindividuellen Variabilität sind die Werte einer intraindividuellen Zeitreihe identisch und unabhängig verteilt. Diese Art der Variabilität lässt sich durch verschiedene Indices, wie z.B. die intraindividuelle Standardabweichung einer manifesten Zeitreihe kennzeichnen. Net-intraindividuelle Variabilität wird von Ram und Gerstorf (2009) in einen engen Zusammenhang mit den dynamischen Charakteristiken einer Person gebracht, wie z.B. der Spanne des Ausdrucksverhaltens einer Person über verschiedene soziale Situationen hinweg. Ein weiteres Beispiel für die dynamischen Charakteristiken einer Person wäre z.B. die Schwankungen oder die Stabilität des Affekts, die sich eindeutig auf die Person und nicht etwa auf andere Faktoren, wie z.B. die Situation zurückführen lassen. Liegen in einer intraindividuellen Zeitreihe Trends, Autokorrelationen und systematische Muster vor, so handelt es sich um zeit-strukturierte intraindividuelle Variabilität, die nach Ram und Gerstorf (2009) einen dynamischen Prozess darstellt. Auch bei der Definition von zeit-strukturierter intraindividueller Variabilität beziehen sich Ram und Gerstorf (2009) auf Fiske und Rice (1955) und bringen das Konzept mit der intraindividuellen Variabilität des Typs-II nach Fiske und Rice (1955) in Verbindung.

Zusammenfassend kann also gesagt werden, dass Ram und Gerstorf (2009) intraindividuelle Variabilität von intraindividueller Veränderung abgrenzen und intraindividuelle Variabilität in zwei Teilkonzepte, net-intraindividuelle Variabilität und zeit-strukturierte intraindividuelle Variabilität aufspalten. Net-intraindividuelle Variabilität steht in Zusammenhang mit dynamischen Charakteristiken eines Individuums und entspricht in etwa der Typ-I-Variabilität nach Fiske und Rice (1955). Zeit-strukturierte intraindividuelle Variabilität steht in Zusammenhang mit der Typ-II-Variabilität nach Fiske und Rice (1955) und bildet dynamische Prozesse im Gegensatz zu dynamischen Charakteristiken einer Person ab. Es fällt auf, dass die Konzeption von Ram und Gerstorf (2009) eine starke Analogie zu zeitreihenanalytischen und autoregressiven Modellen aufweist. Die Variabilität, die sich mit Hilfe eines autoregressiven Modells aufklären lässt, entspricht in etwa der Konzeption der zeit-strukturierten intraindividuellen Variabilität. Die Residuen der Analyse, d.h. die Fehlerkomponente, die sich nicht durch systematische Trends, Oszillationen und Autokorrelationen aufklären lässt, entspricht in etwa der Konzeption von net-intraindividueller Variabilität.

Zur Modellierung intraindividueller Variabilität existiert eine Vielzahl von Modellen und Ansätzen, die Ram und Gerstorf (2009) in ihrem Übersichtsartikel berichten. Diese Modelle werden den Domänen net-intraindividuelle Variabilität und zeit-strukturierte intraindividuelle Variabilität zugeordnet.

Zur mathematischen Beschreibung von zeit-strukturierter intraindividueller Variabilität und von dynamischen Prozessen werden nach Ram und Gerstorf (2009) unter anderem autoregressive Modelle (Box & Jenkins, 1976), spektralanalytische Methoden (Jenkins & Watts, 1968) und Modelle zur Beschreibung nicht-linearer, dynamischer Systeme (Gottman, Murray, Swanson, Tyson & Swanson, 2002) herangezogen. Multivariate, zeit-strukturierte, intraindividuelle Variabilität wird unter anderem mit Hilfe der Dynamischen Faktoranalyse (P. Molenaar, 1985; Nesselroade & Ram, 2004), der Multivariaten Spektralanalyse (Jenkins & Watts, 1968), Hidden-Markov-Modellen (Elliott, Aggoun & Moore, 1995) und State-Space-Modellen angegangen. Zur Modellierung von zyklischen und regulativen Prozessen kommen Oszillatormodelle (Boker, 2001) und Modelle zur

Beschreibung gedämpfter Oszillatoren (Chow, Ram, Boker, Fujita & Clore, 2005) zum Einsatz.

Die Erfassung von net-intraindividueller Variabilität erfolgt nach Ram und Gerstorf (2009) mit Indices, wie z.B. der intraindividuellen Standardabweichung, die über eine intraindividuelle Zeitreihe berechnet werden, um die unstrukturierten Fluktuationen über eine Zeitreihe zu charakterisieren. Da die Verwendung der intraindividuellen Standardabweichung zur Beschreibung der dynamischen Charakteristiken einer Person per Definition voraussetzt, dass die Messwerte identisch und unabhängig (iid) verteilt sind, empfehlen Ram und Gerstorf (2009) die intraindividuellen Zeitreihen vor der Verwendung solcher Indices von eventuellen Mustern und Autokorrelationen zu bereinigen. Net-intraindividuelle Variabilität ist demnach als „netto“ -intraindividuelle Variabilität zu verstehen, die von allen systematischen Effekte mit Hilfe von Modellen zur Beschreibung von zeit-strukturierter intraindividueller Variabilität bereinigt wurde.

2.2. Modelle zur Erfassung intraindividueller Variabilität

Ram und Gerstorf (2009) geben einen guten Überblick über gängige Methoden zur Modellierung intraindividueller Variabilität. In diesem Abschnitt werden einige der Modelle genauer betrachtet. Eine erschöpfende Behandlung der Vielzahl der Modelle würde den Rahmen dieser Arbeit sprengen. Der interessierte Leser, bzw. die interessierte Leserin sei an den Artikel von Ram und Gerstorf (2009) verwiesen, der einen guten Einstieg in die Literatur bietet. In diesem Abschnitt werden vorrangig Modelle und Methoden betrachtet, die dazu verwendet werden, um intraindividuelle Variabilität zu modellieren.

Im Rahmen der traditionellen Item-Response-Theorie existieren zwar Modelle zur Veränderungsmessung (Rost & Spada, 1983; Fischer & Ponocny, 1995), aber Modelle zur expliziten Erfassung von intraindividueller Variabilität im Rahmen multidimensionaler Zeitreihen sind rar. Ansätze zur Fusionierung von item-response-theoretischen Modellen mit dynamischen Faktormodellen zeigen sich jedoch im Forschungsumfeld von Nesselroade. So legte Zhang (2007) ein dynamisches Faktormodell vor, dass die kategoriale Natur der Daten berücksichtigt, anstatt diese als linear und kontinuierlich zu betrachten und

Ram (2005) kombinierte Andrichs Rating-Skalen-Modell mit einem Oszillator-Modell um Schwankungen im Affekt zu modellieren.

2.2.1. Ein IRT-Oszillator-Modell zur Erfassung von zyklischen Affektverläufen (Ram et al., 2005)

Ram et al. (2005) formulieren ein Modell zur Erfassung interindividueller Variabilität in Zyklen von positivem und negativem Affekt unter Berücksichtigung der kategorialen Natur der manifesten Indikatoren für diese Konstrukte. Als Basismodell kam das Rating-Skalen-Modell von Andrich (1978b) zum Einsatz, allerdings wurde dieses dahingehend erweitert, dass die Reaktion eines Individuums v zum Zeitpunkt t auf Item i modelliert wird:

$$P(X_{vit} = x) = \frac{\exp[x(\theta_{vt} - \beta_i) + \tau_x]}{\sum_{l=0}^{m} \exp[l(\theta_{vt} - \beta_i) + \tau_l]}. \tag{2.1}$$

Die hier verwendete Notation für das Rating-Skalen-Modell weicht von derjenigen im Original-Artikel von Ram et al. (2005) ab und orientiert sich an diejenigen in Fischer (1995a). Ram et al. (2005) gehen also davon aus, dass die Reaktion einer Person v auf ein Item i zu Zeitpunkt t von dem latenten Zustand θ_{vt} zum Zeitpunkt t, den Item-Schwierigkeiten β_i und den Schwellen-Parametern τ_x abhängt. Die Parameter θ_{vt} des Rating-Skalen-Modells werden nach Maßgabe eines sinusoidalen Modells zur Messung zyklischer Veränderung zerlegt:

$$\theta_{vt} = \mu_v + \alpha_v t_{vt} + R_v[cos(\omega_v \cdot t_{vt} + \phi_v] + \epsilon_{vt}. \tag{2.2}$$

Dies bedeutet, dass für jede Person v ein zyklisches, sinusoidales Oszillator-Modell formuliert wird. μ_v ist der Mittelwert der Oszillation, α_v repräsentiert einen generellen, personenspezifischen, linearen Trend in Abhängigkeit der Zeit t_{vt}, R_v ist die Amplitude der personenspezifischen Oszillation, ω_v ist die Frequenz der personenspezifischen Oszillation und, ϕ_v ist die Phase der Oszillation und t_{vt} ist der personenspezifische Zeit-Index. Die Varianz und Kovarianz der personenspezifischen Parameter wird im Sinne der Multilevel-Analyse modelliert. Zur Parameterschätzung kommt WinBUGS (Lunn, Thomas, Best & Spiegelhalter, 2000) zum Einsatz. Hinter dem Modell steckt einfach gesagt die Annahme,

dass sich der quantitative, zyklische Verlauf menschlichen emotionalen Erlebens durch eine Sinuskurve darstellen lässt, wobei die Parameter zur Beschreibung dieses Graphen von der Person abhängen und das kategoriale Antwortformat berücksichtigt wird. Somit ist es mit diesem Modell potentiell möglich, interindividuelle Unterschiede hinsichtlich eines sinusoidalen, zyklischen Verlaufs zu erfassen. Durch die Verwendung von Andrichs Rating-Skalen-Modell (Andrich, 1978b) ist es darüber hinaus möglich, Aussagen über die Schwierigkeiten der verwendeten Testaufgaben und die Schwellenparameter des kategorialen Antwortformates zu treffen. Empirisch untersuchten Ram et al. (2005) mit diesem Modell an einer Stichprobe von 179 College-Studenten die Fragestellung, ob sich positiver und negativer Affekt in wöchentlichen Zyklen bewegt, fanden aber keine Hinweise auf einen solchen Effekt. Bemerkenswert ist jedoch die kreative Koppelung zweier Modelle aus unterschiedlichen Forschungstraditionen.

2.2.2. Latent-State-Trait-Modelle (Steyer et al., 1999)

Die Frage nach der intraindividuellen Variabilität und der interindividuellen Differenzierungskraft intraindividueller Variabilität ist stellenweise in der Literatur mit der State-Trait-Debatte verknüpft. Aus trait-theoretischer Perspektive ist intraindividuelle Variabilität in einer beobachteten Variable lediglich Fehlervarianz. Das, was interessiert, ist das relativ zeitstabile Verhalten und Erleben einer Person, das sich in stabilen intraindividuellen Mittelwerten auf Test-Skalen ausdrückt, welche zwischen Personen hinsichtlich der mittleren Merkmalsausprägung differenzieren. Aus einer situationistischen Perspektive ist das Verhalten einer Person hauptsächlich von der Situation bestimmt.

Ein Ansatz, der sowohl die stabilen Aspekten des Verhaltens und Erlebens einer Person, als auch situative Faktoren berücksichtigt, ist die Latent-State-Trait-Theorie (Steyer, Schmitt & Eid, 1999).

Steyer et al. (1999) zeigten, dass LST-Modelle Werkzeuge darstellen, die es erlauben Trait-Komponenten und situationsspezifische State-Komponenten menschlichen Verhaltens zu separieren und der methodischen Analyse zugänglich zu machen.

Latent-State-Trait-Modelle können als eine Erweiterung der Klassischen Testtheorie

(vgl. Fischer, 1974; Nunally, 1978) verstanden werden. In der Klassischen Testtheorie wird die Ausprägung auf einer manifesten Variable Y_i in eine True-Score Komponente τ_i und einen Fehleranteil ϵ_i zerlegt:

$$Y_i = \tau_i + \epsilon_i. \tag{2.3}$$

In Folge der Unabhängigkeit der wahren Werte und der Messfehler ergibt sich der Satz der Varianzzerlegung der manifesten Variable in die Varianz der True-Score-Variable und der Fehlervarianz:

$$\mathrm{Var}(Y_i) = \mathrm{Var}(\tau_i) + \mathrm{Var}(\epsilon_i). \tag{2.4}$$

Die Reliabilität ist als Anteil der True-Score Varianz an der Varianz der manifesten Variable definiert:

$$\mathrm{Rel} = \frac{\mathrm{Var}(\tau_i)}{\mathrm{Var}(Y_i)}. \tag{2.5}$$

Mit diesem klassischen Modell lassen sich keine Hypothesen über die Situationsabhängigkeit von Messwerten prüfen, da diese in dem Modell nicht vorgesehen sind. Steyer et al. (1999) erweiterten dieses Modell um situationsabhängige Komponenten:

$$Y_{ik} = \tau_{ik} + \epsilon_{ik}. \tag{2.6}$$

Die manifeste Varianz einer beobachteten Variable i zu Zeitpunkt k wird in einem situationsabhängigen True-Score τ_{ik}, sowie einen Fehleranteil ϵ_{ik} zerlegt. Ferner erfolgt eine lineare Dekomposition der latenten State-Variable (τ_{ik}) in einen Trait-Anteil ξ_{ik} und ein latentes State-Residuum ζ_{ik}:

$$\tau_{ik} = \xi_{ik} + \zeta_{ik}. \tag{2.7}$$

Das Modell impliziert folgende Unabhängigkeiten:

$$\mathrm{Cov}(\epsilon_{ik}, \zeta_{ik}) = \mathrm{Cov}(\epsilon_{ik}, \tau_{ik}) = \mathrm{Cov}(\epsilon_{ik}, \xi_{ik}) = \mathrm{Cov}(\zeta_{ik}, \xi_{ik}) = 0. \tag{2.8}$$

Aus diesen Eigenschaften folgen die Varianzzerlegungen

$$\mathrm{Var}(Y_{ik}) = \mathrm{Var}(\tau_{ik}) + \mathrm{Var}(\epsilon_{ik}) \tag{2.9}$$

und

$$\mathrm{Var}(\tau_{ik}) = \mathrm{Var}(\xi_{ik}) + \mathrm{Var}(\zeta_{ik}). \tag{2.10}$$

Ein Vorteil dieses Latent-State-Trait-Modells ist es, dass nicht nur Aussagen über die Reliabilität, sondern auch über die Konsistenz und Spezifizität getroffen werden können. Die Konsistenz ist der Anteil der Varianz der latenten Trait-Variable an der beobachteten Varianz:

$$\mathrm{Con}(Y_{ik}) = \frac{\mathrm{Var}(\xi_{ik})}{\mathrm{Var}(Y_{ik})}. \tag{2.11}$$

Sie drückt aus, wie viel Varianz der manifesten Variablen auf stabile Trait-Effekte zurückgeht. Die Spezifizität ist der Anteil der Varianz der beobachteten Variablen, der auf die State-Residuen zurückzuführen ist:

$$\mathrm{Spe}(Y_{ik}) = \frac{\mathrm{Var}(\zeta_{ik})}{\mathrm{Var}(Y_{ik})}. \tag{2.12}$$

Die Spezifizität gibt den Anteil der Variabilität der manifesten Variablen an, der auf situationsspezifische oder messgelegenheitsspezifische Effekte zurückzuführen ist. Die Summe von Konsistenz und Spezifizität ergeben die Reliabilität der Gesamtmessung:

$$Rel(Y_{ik}) = Con(Y_{ik}) + Spe(Y_{ik}). \tag{2.13}$$

Auf der Basis dieser Grundmodelle definieren Steyer et al. (1999) fünf Grundmodellklassen der Latent-State-Trait-Theorie: latente Trait-Modelle, latente State-Modelle mit und ohne Methodenfaktoren sowie latente State-Trait-Modelle mit und ohne Methodenfaktoren. Abbildung 2.1 zeigt ein Multitrait-Multistate-Modell der Latent-State-Trait-Theorie. Die manifeste Variable Y_{ik} wird in traitspezifische Anteile ξ_i und messgelegenheitsspezifische Anteile ζ_k zerlegt. Ferner ist eine Kovariation zwischen den latenten Trait-Variablen ξ_i vorgesehen, was es ermöglicht, die internale Struktur der manifesten Variablen unter Berücksichtigung situationsspezifischer Effekte zu überprüfen. Eine Anwendungsmöglichkeit besteht darin, die Stabilität eines Instrumentes über die Zeit zu evaluieren, indem ξ_1 und ξ_2 als latente Trait-Variablen definiert werden, die das Testverhalten in zwei Testhälften „beeinflussen“. Eine hohe Korrelation ψ_{12} spräche also dafür, dass beide Testhälften etwas Gemeinsames erfassen. Mit diesem Modell ist es nicht nur möglich, die Spezifizität und Konsistenz zu errechnen, um zu beurteilen, ob eher trait-spezifische, oder state-spezifische Anteile für die Variabilität der manifesten Variable verantwortlich sind, es

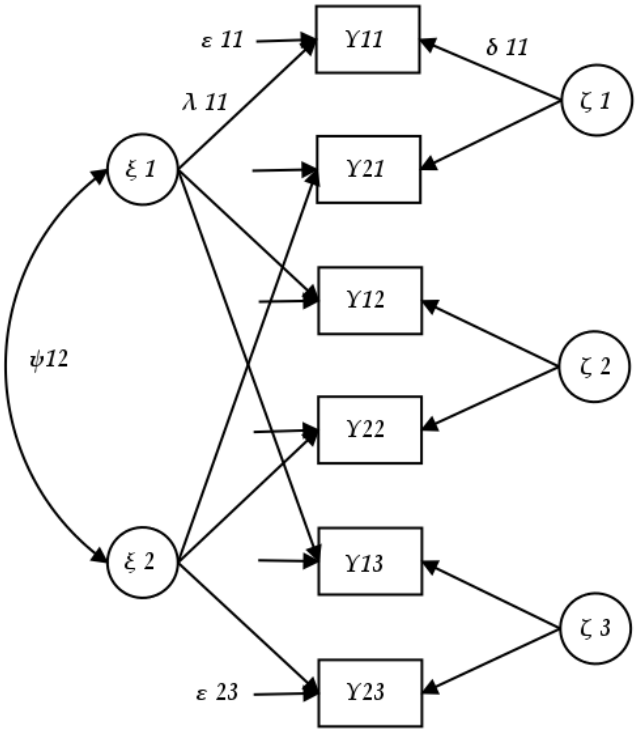

Abbildung 2.1.: Multitrait-Multistate-Modell

kann auch über die Ladungsparameter, bzw. Pfadkoeffizienten λ_{ik} und δ_{ik} die die Enge des Zusammenhangs zwischen der latenten und der jeweiligen manifesten Variable beschreiben werden.

Nach Eid und Diener (1999) ist das in Abbildung 2.1 dargestellte Modell wie folgt definiert:

$$Y_{ik} = \lambda_{ik}\xi_i + \delta_{ik}\zeta_k + \epsilon_{ik}. \tag{2.14}$$

Konsistenz, Spezifizität und Reliabilität werden wie folgt berechnet:

$$\mathrm{Con}(Y_{ik}) \quad = \quad \frac{\lambda_{ik}^2 \mathrm{Var}(\xi_i)}{\mathrm{Var}(Y_{ik})} \tag{2.15}$$

und

$$\mathrm{Spe}(Y_{ik}) \quad = \quad \frac{\delta_{ik}^2 \mathrm{Var}(\zeta_k)}{\mathrm{Var}(Y_{ik})}. \tag{2.16}$$

Die Reliabilität ergibt sich aus der Summe der Konsistenz und Spezifizität.

Bezüglich der intraindividuellen Variabilität lässt sich feststellen, dass dieses Modell die Variation in den Daten von Messzeitpunkt zu Messzeitpunkt einerseits auf stabile Trait-Komponenten, andererseits auf variable, situative Umstände zurückführt. Allerdings beinhaltet dieser Modelltypus keinen direkten psychometrischen Index, der die Variabilität eines Individuums innerhalb einer Zeitperiode charakterisiert, da die Variabilität lediglich auf variable messgelegenheitsspezifische Effekte und stabile Personen-Effekte zurückgeführt wird. Vor dem Hintergrund der Definitionen nach Fiske und Rice (1955) wird hier tendenziell Typ-III-Variabilität erfasst.

Zur Klärung der von Fiske und Rice (1955) aufgeworfenen Fragen wendeten Eid und Diener (1999) diesen Modell-Typus auf *Variabilitäts-Indices* in Form der intraindividuelle Standardabweichungen von längsschnittlichen Affekt-Ratings an, um zu überprüfen, ob der so erfassten intraindividuellen Affekt-Variabilität eines Individuums eher State- oder Trait-Charakter zukommt. Ferner bewerteten Eid und Diener (1999) die Reliabilität, Validität und Dimensionalität der intraindividuellen Standardabweichung mittels konfirmatorischer Faktoranalysen und kommen zu dem Schluss, dass intraindividuelle Affekt-Variabilität ein mehrdimensionales Konstrukt ist, dass reliabel und valide zwischen Personen differenziert, wobei die Variabilitäts-Faktoren positiv miteinander korreliert sind.

2.2.3. Traits als Verteilungen von States

Ein weiter Ansatz, der sowohl die Situation, als auch personenbezogene Traits als mögliche Quelle von manifestem Verhalten berücksichtigt, wird von William Fleeson (2001) vertreten. Fleeson (2001) sieht Traits als Verteilungen von States und charakterisiert sowohl State-, als auch Trait-Informationen durch *eine* Verteilung. Manifeste Verteilungen von States können gewonnen werden, indem das Verhalten von Personen längsschnittlich über eine längeren Zeitraum mit einem manifesten Indikator erhoben wird. So ist

es z.B. denkbar, im Rahmen eines ambulanten Assessments eine Gruppe von Personen mit mobilen Computern auszustatten, die programmiert wurden, um nach einem festgesetzten Erhebungs-Rhythmus die individuell eingeschätzte Lage auf den Big-Five Skalen zu erheben. Der Trait-Charakter der so gewonnenen Dichten würde sich in einer relativen Stabilität der individuellen Mittelwerte und der Streuungen zeigen, sofern eine Normalverteilung verwendet wird, um die empirischen Dichten zu modellieren. Prinzipiell handelt es sich hier um die Erfassung von Typ-I-Variabilität nach Fiske und Rice (1955), da die Reihenfolge der Messwerte in dieser Konzeption keine Rolle spielt und zudem die Variabilität lediglich auf die Person zurückgeführt wird.

Aus methodischer Perspektive birgt Fleesons (2001) Ansatz allerdings eine Schwierigkeit. Wie Brendan et al. (2006) zeigen, sind die intraindividuellen Streuungen und die Mittelwerte unter Umständen miteinander konfundiert, was besonders bei schiefen manifesten Verteilungen der Fall ist. Wird nun die intraindividuelle Standardabweichung mit einem externen Kriterium, z.B. zum Zweck der Validierung korreliert, so können künstlich hohe Korrelationen auftreten, die eigentlich auf extreme Mittelwerte zurückzuführen sind. Es ist daher nötig, den möglichen Zusammenhang zwischen Standardabweichung und Mittelwert bei einer Korrelation mit externen Kriterien zu kontrollieren. Weiterhin ist es in dem Ansatz von Fleeson nicht möglich, situationsabhängige Komponenten zu erfassen.

2.2.4. Dynamische Faktormodelle

Sobald mehrere Messwerte einer oder mehrerer Personen auf einer oder mehreren Variablen vorliegen, so handelt es sich um Daten, die sich potentiell dazu eignen, die Variabilität von Personen auf diesen Variablen zu charakterisieren. Formal betrachtet handelt es sich um multivariate Zeitreihen, die sowohl für eine, als auch für mehrere Personen vorliegen können. In der Literatur wird die Frage nach der Dimensionalität, bzw. internalen Struktur vor allem mit faktorenanalytischen Modellen und Strukturgleichungsmodellen angegangen, welche zumeist Erweiterungen der Cattell'schen P-Technik (Cattell, Cattell & Rhymer, 1947) darstellen. Die folgenden Ausführungen zu dynamischen Faktormodel-

len lehnen sich eng an Nesselroade & Ram (2002) an.

Cattell's P-Technik (Cattell et al., 1947)

Liegen Daten von nur einer Person auf mehreren Variablen zu mehreren Messzeitpunkten vor, so ist es möglich faktorenanalytisch zu explorieren, welche bestimmten Variablengruppen einer multivariaten Zeitreihe gemeinsam über die Zeit kovariieren. Abbildung

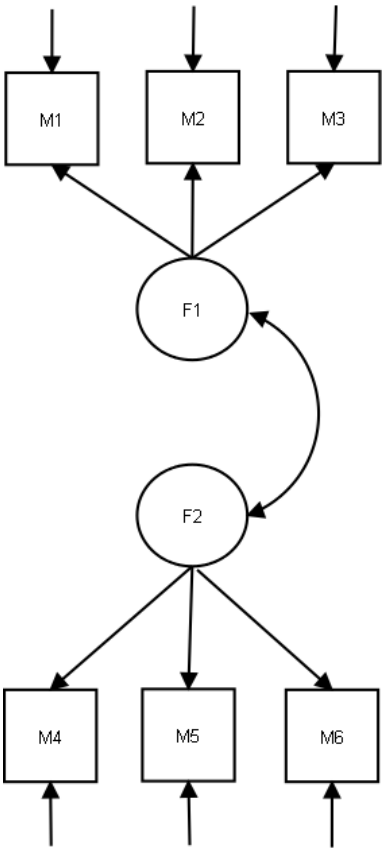

Abbildung 2.2.: Cattell's P-Technik Faktormodell

2.2 zeigt eine schematische Darstellung des Modells das hinter Cattell's P-Technik steht. Die Kovarianz der Variablen $M1 - M3$ und $M4 - M6$ wird jeweils auf das „Wirken" der latenten Faktoren $F1$ und $F2$ zurückgeführt. Zudem ist eine Kovariation der Faktoren $F1$ und $F2$ erlaubt. Liegen theoretische Hypothesen darüber vor, welche Variablen-Gruppen innerhalb einer Person über die Zeit kovariieren, lassen sich diese mit Hilfe

des Cattell'schen P-Faktor Modells prüfen. Zudem ist es möglich, Hypothesen über die individuelle, internale Struktur der latenten Variablen untereinander zu überprüfen.

Bei Cattells P-Technik handelt es sich um ein ideographisches Verfahren, das schon relativ früh in der klinischen Forschung zur Psychotherapieevaluation und zur Einzelfallanalyse zum Einsatz gekommen ist. Luborsky und Minz (1970) sowie Jones und Nesselroade (1990) geben einen Überblick über die Anwendungen der P-Technik. Ein historischer Überblick des Einsatzes des Verfahrens - speziell im klinischen Kontext - gibt Luborsky (1995).

Nesselroade und Ram (Nesselroade & Ram, 2004) führen zwei Hauptkritikpunkte an der P-Technik an, die mit der mangelnden Modellflexibilität zusammenhängen. Zum einen ist in dem Modell keine autoregressive Komponente auf Faktorebene vorgesehen, zum anderen ist das Modell so gestaltet, dass die Ladungen über alle Messgelegenheiten hinweg konstant sind. Es ist jedoch denkbar, dass die Ladungen von Messzeitpunkt zu Messzeitpunkt schwanken. Ferner lässt sich mit dem Modell der P-Technik keine zeitliche Seriabilität abbilden. Es ist jedoch theoretisch denkbar, dass die Ausprägung auf dem latenten Faktor $F1$ zu einem Zeitpunkt t_1 eine Auswirkung auf die Ausprägung auf diesem Faktor zu einem Messzeitpunkt t_2 hat.

Dynamische Modelle, die faktor- und zeitreihenanalytische Methoden miteinander kombinieren, adressieren die angesprochenen Probleme.

Das Direkt-Autoregressive-Faktor-Wert Modell (ARFS-Modell)

Eine dynamische Erweiterung der Cattell'schen P-Technik ist das *direct autoregressive factor score model* (Nesselroade et al., 2002). Wie aus Abbildung 2.3 ersichtlich, existiert pro Messzeitpunkt ein eigenes Messmodell, wodurch es möglich ist, unterschiedliche Ladungs-Strukturen pro Messzeitpunkt abzubilden. Zudem sind zeitverzögerte Beziehungen vorhergehender Faktorwerte mit nachfolgenden Faktorwerten erlaubt. Es ist also überprüfbar, ob und inwiefern die Ausprägung auf einem latenten Faktor zu einem Messzeitpunkt t mit der Ausprägung zu einem späteren Messzeitpunkt $t+1$ zusammenhängt. Dieses Modell erlaubt also die Analyse autoregressiver Beziehungen der Faktor-Werte

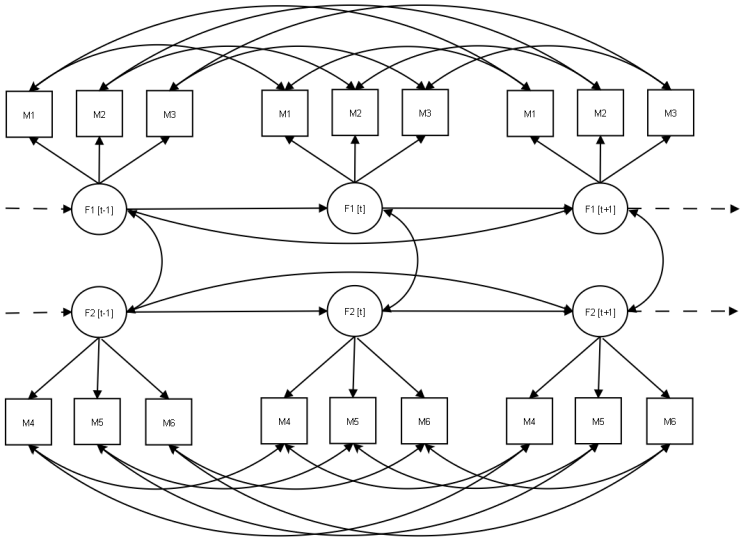

Abbildung 2.3.: Das Direkt-Autoregressive-Faktor-Wert-Modell (ARFS) (Nesselroade, 2002)

untereinander und adressiert damit vor allem Fragen, die mit der Variabilität des Typs-II nach Fiske und Rice (1955) zusammenhängen. Durch die Korrelation der Messfehler gleichartiger manifester Variablen von Messzeitpunkt zu Messzeitpunkt, wird der Messwiederholungscharakter des Modells berücksichtigt. Dieses Modell erlaubt zudem, die Unterschiede der Zusammenhänge der Faktoren auf Strukturebene pro Messzeitpunkt zu analysieren. Bei diesem Modell handelt es sich ebenfalls um ein ideographisches Verfahren das sich vor allem zur Einzelfallanalyse eignet.

Nach Nesselroade und Ram (2004) wurde dieser Typus von Modell zur Analyse zeitverzögerter Beziehungen in der Psychophysiologie (Kettunen & Ravaja, 2000; P. Molenaar, 1987), der Stimmungsforschung (Nesselroade et al., 2002; Shifrin, Hooker, Wood & Nesselroade, 1997) und der Therapieforschung (P. Molenaar, 1987) eingesetzt.

Das White-Noise-Factor-Score-Modell (WNFS-Modell)

Ein weiteres Modell aus der Klasse der dynamischen Faktormodelle ist das *white-noise-factor-score-model* von Molenaar (1985). Wie aus Abbildung 2.4 hervorgeht, ist das

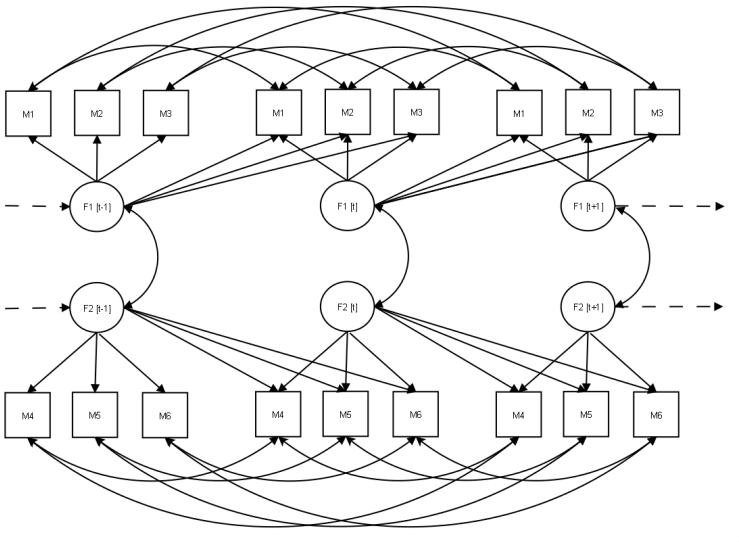

Abbildung 2.4.: Das White-Noise-Factor-Score-Modell (WNFS-Modell) (Molenaar, 1985)

WNFS-Modell dem ARFS-Modell sehr ähnlich, allerdings werden hier keine autoregressiven Komponenten auf Faktorebene modelliert. Vielmehr besteht die Möglichkeit, dass Faktorwerte zu einem bestimmten Zeitpunkt t sich direkt auf die manifesten Variablen zum Zeitpunkt t und auf spätere Messzeitpunkte $t+1$ auswirken. Die Stärke dieser Zusammenhänge wird durch die entsprechenden Pfadkoeffizienten erfasst. Wie beim ARFS-Modell kann die Beziehung der latenten Variablen auf Strukturebene separat pro Messzeitpunkt modelliert werden, allerdings bestehen keine Beziehungen auf latenter Ebene von Messzeitpunkt zu Messzeitpunkt und somit keine autoregressive Komponente auf Faktor-Ebene.

Erweiterungen der dynamischen Faktormodelle für kategoriale Daten

Das WNFS- und DARFS-Modell sind vor allem dann indiziert, wenn längsschnittlichen, intervallskalierte Daten vorliegen. Werden Informationen mit Hilfe von Rating-Skalen erhoben, wird den Skalen zumeist Intervallskalenniveau unterstellt. Neuere Trends im Bereich der dynamischen Faktormodelle gehen dahin, die kategoriale Natur von Rating-Skalen direkt zu berücksichtigen. So stellte Zhang (2007) kategoriale Erweiterungen des DARFS-Modells und des WNFS-Modells vor, wobei zur Schätzung der Parameter die Monte-Carlo-Markov-Chain-Methode vorgeschlagen wird.

2.2.5. Indices zur Erfassung intraindividueller Variabilität und affektiver Instabilität

Im Rahmen der Forschung zur intraindividuellen Affektvariabilität (vgl. z.B. Ebner-Priemer et al., 2009) bedient man sich unter anderem manifester Indices zur Erfassung der Variabilität affektiven Erlebens, die auf Moskowitz und Zuroff (2004) zurückzuführen sind: Pulse, Flux und Spin.

Moskowitz und Zuroff (2004) untersuchten die intraindividuellen Variabilität interpersonellen Verhaltens im Rahmen des interpersonellen Zirkumplex-Modells (vgl. z.B. Kiesler, 1983). Nach diesem Modell lässt sich interpersonelles Verhalten durch zwei orthogonale Dimensionen - Dominanz und Verträglichkeit - beschreiben, die ein kartesisches Koordinatensystem bilden. Das Verhalten eines Individuums in einer zwischenmenschlichen Begegnung lässt sich durch eine Koordinate bzw. einen Vektor in dem zweidimensionalen Koordinatensystem charakterisieren. Unter der Hypothese der intraindividuellen Variabilität ist zu vermuten, dass interpersonelles Verhalten durchaus variabel sein kann und eine Person sich nicht lediglich immer nur dominant und verträglich verhält. Zur Charakterisierung der zeitlichen Dynamik interpersonellen Verhaltens definieren Moskowitz und Zuroff (2004) drei neue Konstrukte, die sich potentiell zur Beschreibung von Persönlichkeit eignen: Flux, Puls und Spin. Flux bezeichnet die Standardabweichungen

auf einer der Zirkumplex-Dimensionen:

$$\text{Flux(X)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^2) - \left(\sum_{i=1}^{n}\frac{1}{n}x_i\right)^2}. \tag{2.17}$$

Pulse bezeichnet die Variabilität (Standardabweichung) der Längen des durch die Koordinaten (x, y) definierten Vektors:

$$\text{Pulse(X,Y)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{x_i^2+y_i^2}\right)^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{x_i^2+y_i^2}\right)\right]^2}. \tag{2.18}$$

Spin ist die Variabilität des Winkels zwischen den Vektoren von Messzeitpunkt zu Messzeitpunkt:

$$\text{Spin(X,Y)} = \sqrt{-\log\left[\left(\frac{1}{n}\sum_{i=1}^{n}cos\ \alpha_i\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}sin\ \alpha_i\right)^2\right]}. \tag{2.19}$$

Moskowitz und Zuroff (2004) fanden mittels einer Latent-State-Trait-Analyse eine hohe Stabilität des Flux der Variablen „submissives Verhalten“, „verträgliches Verhalten“ und „streitsüchtiges Verhalten“. Ferner zeigte sich eine mittlere bis hohe Stabilität im Puls und Spin, was darauf hindeutet, dass diese Variablen relativ stabile Aspekte der Persönlichkeit erfassen, bzw. dass die intraindividuelle Variabilität dieser Variablen interindividuell stabil ist. Eine Korrelation der Flux-Variablen mit den Persönlichkeitsvariablen des NEO-FFI (Costa & McCrae, 1992) zeigte einen Zusammenhang zwischen Extraversion und dem Flux der Variable „verträgliches Verhalten“. Die NEO-FFI Skala „Verträglichkeit“ ging mit einem geringen Spin und höherem Flux auf der Variable „streitsüchtiges Verhalten“ einher. Die Persönlichkeitsvariable Neurotizismus zeigte einen hohen Zusammenhang mit dem Flux der Variable „unterwürfiges Verhalten“, „streitsüchtiges Verhalten“ und insgesamt kommen Moskowitz und Zuroff (2004) zu dem Schluss, dass Flux, Puls und Spin im interpersonellen Verhalten reliabel erfasst werden können und eine Bereicherung darstellen, um interindividuelle Unterschiede zu beschreiben.

Moskowitz und Zuroffs (2004) Ansatz wurde von Russel (2007) zur Untersuchung der Variabilität des interpersonalen Verhaltens von Personen mit der Diagnose Borderline Persönlichkeitsstörung (BPS) eingesetzt. Zusätzlich zu den von Moskowitz und Zuroff

(2004) verwendeten Skalen zur Erfassung des interpersonalen Verhaltens (2004) wurden 9 Items zur Erfassung der Affekt-Valenz verwendet, welche die Pole *pleasant* und *unpleasant affect* des Zirkumplex-Modells der Emotionen (Larsen & Diener, 1992; Russell, 2003) abdecken. Hinsichtlich der Affekt-Variabilität zeigte sich in einem Vergleich mit einer Kontrollgruppe in der Gruppe der Personen mit der Diagnose BPS eine höhere Variabilität der Affekt-Valenz und der Variable *pleasant affect*.

Ebner-Priemer et al. (2009) geben einige Hinweise zur Verwendung von Indices zur Erfassung intraindividueller Variabilität und kritisieren die Verwendung der intraindividuellen Standardabweichung (Flux) als Index im Rahmen von Zirkumplex-Modellen, da dieser Index es nicht erlaubt, Trends innerhalb einer intraindividuellen Zeitreihe zu berücksichtigt. Ein sprunghafter, stationärer Prozess mit einer bestimmten Standardabweichung ließe sich nicht von einer kontinuierlichen Zuwachs, bzw. einer stetigen Verringerung der Merkmalsausprägung unterscheiden. Jedoch ist diese Unterscheidung für die psychodiagnostische Interpretation wichtig. Ebner-Priemer (2009) schlagen daher weitere Indices vor: die Mittlere Quadratische Differenz (MSSD für *mean squared successive difference*), die Mittlere Quadratische Differenz in einer euklidischen Metrik (MSSD-Euklid) und in der City-Block-Metrik (MSSD-City-Block).

Die MSSD ist wie folgt definiert:

$$\text{MSSD(X)} = \frac{1}{n-1}\sum_{i=2}^{n}(x_i - x_{i-1})^2. \tag{2.20}$$

Durch die Berücksichtigung des Wertes x_{i-1} ist es mittels dieses Index möglich, Veränderungen in den Zeitreihen von Zeitpunkt $i-1$ zu Zeitpunkt i aufzudecken, die zeitliche Abfolge wird also berücksichtigt.

Der MSSD-Euklid-Index basiert auf der Berechnung der euklidischen Distanz der Vektoren von Zeitpunkt $i-1$ zu Zeitpunkt i in zweidimensionalen, kartesischen Raum eines Zirkumplex-Modells. Der Mittelwert dieser Distanzen ist der MSSD-Euklid-Index:

$$\text{MSSD-Euklid} = \frac{1}{n-1}\sum_{i=2}^{n}\left[(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2\right]. \tag{2.21}$$

Der MSSD-City-Index basiert auf der City-Block-Metrik der Distanzen der Vektoren

von Zeitpunkt $i-$ zu Zeitpunkt i:

$$\text{MSSD-city} = \frac{1}{n-1} \sum_{i=2}^{n} \left(|x_i - x_{i-1}| - |y_i - y_{i-1}| \right)^2 . \tag{2.22}$$

Ebner-Priemer et al. (2009) empfehlen die Verwendung der MSSD-Indices im Rahmen von Untersuchungen zur intraindividuellen Affekt-Variabilität, wenn die zeitliche Abfolge der Ereignisse eine Rolle spielt. Diese Indices erfassen also eher zeitstrukturierte intraindividuelle Variabilität.

2.3. Zusammenfassende Betrachtung

Bei der Zusammenschau des psychologischen Hintergrundes ist zu verzeichnen, dass es an griffigen definitorischen Rahmen, zu nennen wäre hier von allem Ram und Gerstorf (2009), nicht mangelt. Auch existieren zahlreiche Möglichkeiten zur Modellierung intraindividueller Variabilität. Zu nennen wären vor allem die dynamischen, faktorenanalytischen Ansätze, die aus der P-Technik von Cattell erwachsen sind und die Anwendung von Latent-State-Trait-Modellen auf intraindividuelle Standardabweichungen. Hier wurde nur ein kleiner Einblick in die Modellierungsmöglichkeiten gegeben und z.B. die an Bedeutung gewinnende Modellgruppe der Hidden-Markov-Modelle wurde lediglich erwähnt. Was allerdings auffällt ist, dass Ansätze der Item Response Theorie (IRT) zur Erfassung intraindividueller Variabilität relativ rar sind. Ein Ansatz ist derjenige des Oszillator-Modells von Ram et al. (Ram et al., 2005), allerdings sind die damit prüfbaren Hypothesen relativ spezifisch. Was bisher fehlt ist ein IRT-Modell, das es erlaubt die Variabilität auf multivariaten, intraindividuellen Zeitreihen parametrisch als IRT-Modell zu skalieren und zu überprüfen, ob es sich bei der intraindividuelle Variabilität selbst um einen Trait handelt, der interindividuell zwischen Personen differenziert. Probabilistische IRT-Modelle sind dann angezeigt, wenn die abhängigen Variablen kategorialer Natur sind, wie dies bei der Erhebung psychologischer Daten mit Items häufig der Fall ist. Zudem bieten IRT-Modelle den Vorteil, dass sie es erlauben zu überprüfen, ob es aus psychometrischer Perspektive gerechtfertigt ist, einen Summenwert über manifeste Variablen als Indikator der Merkmalsausprägung einer Person zu bilden, um zu vermei-

den, dass in die Bildung des Summenwerts Variablen eingehen, die keine gemeinsame latente Variable erfassen. Oder anders ausgedrückt: psychometrische Testmodelle erlauben es, Homogenitätshypothesen bezüglich des Item-Materials zu prüfen. Testmodelle bieten weitere wünschenswerte Möglichkeiten der Bewertung einer psychologischen Messung, wie z.B. die Berechenbarkeit der globalen und individuellen Messgenauigkeit und die Skalierung von Items und Personen auf einer latenten Dimension.

Aus dem psychologischen Hintergrund ergibt sich zunächst die übergeordnete Fragestellung der Herleitbarkeit eines probabilistischen Testmodells zur Erfassung intraindividueller Variabilität. Um diese globale Fragestellung näher zu konkretisieren, muss allerdings zunächst auch auf einen modelltheoretischen Hintergrund zurückgegriffen werden. Um probabilistische IRT-Modelle zu definieren bedarf es eines Modellierungs-Kontextes, aus dem sich die Modelle formal und strukturiert herleiten lassen. Ein klassischer Ansatz zur Generierung probabilistischer Testmodelle ist derjenige von Georg Rasch (Rasch, 1961).

Allerdings schreibt Andersen (Andersen, 1995b):

> Georg Rasch was a lazy journal and book reader, as can be seen, e.g., from the sparseness of references in his papers. He once admitted quite frankly to me that he never really read papers even in major statistical journals.

Von daher ist es Rasch wahrscheinlich nicht aufgefallen, dass ein Ansatz der probabilistischen Modellierung bereits 1957 in einem völlig anderen Kontext publiziert wurde, der große strukturelle Ähnlichkeit mit der von Rasch vorgestellten Modellklasse aufweist. Der amerikanische Physiker E.T. Jaynes publizierte zwei Artikel in der *Physical Review* (Jaynes, 1957a, 1957b), in dem gezeigt wird, dass sich wichtige Verteilungen der Thermodynamik auf Basis rein informationstheoretischer Überlegungen ohne jegliches physikalische Argument herleiten lassen. Die verwendete Methode der Modellherleitung wird u.a. als Maximum-Entropie-Methode bezeichnet und findet z. B. in so diversen Bereichen wie der komputationellen Linguistik (Berger, Della Pietra & Della Pietra, 1996), dem maschinellen Lernen (Koller & Friedman, 2009) und den Neurowissenschaften (Schneidman, Berry, Ronen & Bialek, 2006) eine Anwendung. Die Maximum-Entropie-Methode kann

als Formalismus verstanden werden, der es generell erlaubt, Wahrscheinlichkeitsverteilungen, bzw. probabilistische Modelle herzuleiten. Zur Beleuchtung des modelltheoretischen Hintergrundes werden im folgenden Kapitel kurz die probabilistischen Modelle im Sinne von Rasch vorgestellt, die Maximum-Entropie-Methode wird beschrieben und es wird gezeigt, dass sich die Modelle sensu Rasch durch die Anwendung der Methode auf eine psychologische Fragestellung herleiten lassen. Ein Grund für den Rückgriff auf die Maximum-Entropie-Methode besteht darin, dass sie einen formalen Kontext bietet, um neue, bisher unbekannte probabilistische Modelle herzuleiten, was für die übergeordnete psychometrische Fragestellung der Entwicklung eines probabilistischen Testmodells zur Erfassung intraindividueller Variabilität in massiv längsschnittlichen Designs von Bedeutung ist.

3. Modelltheoretischer Hintergrund

Um die übergeordnete Fragestellung der Formulierung eines probabilistischen Testmodells zur Erfassung intraindividueller Variabilität zu bearbeiten ist es notwendig, den psychometrischen Hintergrund von probabilistischen Testmodellen in der Psychologie näher zu beleuchten, die Maximum-Entropie-Methode kurz vorzustellen und zu zeigen, dass sich bekannte probabilistische Testmodelle der Psychologie aus der Anwendung der Maximum-Entropie-Methode herleiten lassen. Sollte dies gelingen, ist zu vermuten, dass sich die Maximum-Entropie-Methode dazu eignet, auch neue probabilistische Testmodelle zu definieren. Von daher ist der zielführende Gedanke dieses Kapitels derjenige zu zeigen, dass probabilistische Testmodelle sensu Rasch und Maximum-Entropie-Modelle unter bestimmten Umständen miteinander kompatibel sind, bzw. sich Rasch-Modelle aus der Anwendung der Maximum-Entropie-Methode ergeben. In diesem Kapitel wird quasi das Handwerkszeug der Modelldefinition vorgestellt, welches im folgenden Kapitel zur Anwendung kommen soll.

3.1. Probabilistische Testmodelle

Der dänische Mathematiker Georg Rasch (Rasch, 1960) führte in den 60er Jahren eine Reihe von mathematischen Modellen in die Psychologie ein, die es erlauben, das Antwortverhalten von Personen auf Testaufgaben probabilistisch zu modellieren. Im deutschsprachigen Raum wurde dieser Ansatz vor allem von G.H. Fischer (Fischer, 1974; Fischer & Molenaar, 1995) vertreten.

Die Modelle von Rasch eignen sich unter anderem dazu, die Fähigkeiten von Personen und die Schwierigkeiten von Aufgaben auf Basis von in Testsitzungen erhobenen Daten

auf einer Dimension zu skalieren. In einer Testsitzung werden Personen mit Testaufgaben (*items*) konfrontiert, auf welche zu reagieren ist. In der Regel werden die Reaktionen der Personen dichotom - z.B. bei Leistungstests - oder polytom - z.B. bei der Einstellungsmessung - kodiert. Es fällt also eine Datenmatrix X der Dimensionalität $N \cdot k$ an, wobei N für die Anzahl der getesteten Personen und k für die Anzahl der Testaufgaben steht. x_{vi}, ein Eintrag in der Datenmatrix, ist die numerisch kodierte Reaktion einer Person v auf eine Testaufgabe i. In den Modellgleichungen der Literatur findet sich auch gelegentlich die Schreibweise x für eine numerisch kodierte, kategoriale Reaktion. Bei dichotomem Antwortformat gilt in der Regel $x_{vi} \in \{0, 1\}$ und bei polytomem Antwortformat gilt in der Regel $x_{vi} \in \{0, \ldots, m\}$, wobei $m - 1$ der Anzahl der Kategorien eines polytomen Antwortformats entspricht.

Eine sehr allgemeine, klassische Formulierung für sog. polytome Rasch-Modelle gibt Rasch (1961) in dem sogenannten *Berkeley paper* (Rasch, 1961):

$$P\{x|\theta_v, \sigma_i\} = \frac{1}{\gamma(\theta_v, \sigma_i)} \exp[\phi(x)\theta_v + \psi(x)\sigma_i + \chi(x)\theta_v\sigma_i + \rho(x)]. \tag{3.1}$$

$P\{x|\theta_v, \sigma_i\}$ ist die Wahrscheinlichkeit der diskreten, kategorialen Reaktion x bei gegebenen Parametern θ_v und σ_i, wobei die θ_v Eingenschaften der Personen - deren Fähigkeit - und die σ_i Eigenschaften der Items - deren Schwierigkeit - charakterisieren. Es wird angenommen, dass es sich bei dem quantitativ erfassbaren Merkmal θ_v um einen eindimensionalen Trait handelt. Die Funktionen $\phi(x)$, $\psi(x)$, $\chi(x)$ und $\rho(x)$ sind Funktionen der beobachteten Daten x (Reaktionen) und werden Scoring-Funktionen genannt. $\gamma(\theta_v, \sigma_i)$ ist eine normalisierenden Konstante, die gewährleistet, dass die Summe der durch das Modell vorhergesagten Einzelwahrscheinlichkeiten für die diskreten Reaktionen Eins ergibt. Rasch studierte insbesondere Modelle ohne den Produktterm $\chi(x)\theta_v\sigma_i$, da diese Modelle nach Rasch besondere Eigenschaften aufweisen, auf die in einem späteren Abschnitt am Beispiel des dichotomen Rasch-Modells konkreter eingegangen wird.

Andersen (1995a) gibt eine explizitere Formulierung des allgemeinen Basismodells von Rasch. Andersens Formulierung ermöglicht es, eine Menge von probabilistischen Testmodellen für polytome Antwortformate und auch das dichotome Rasch-Modell als Spezial-

fälle abzuleiten (Mair & Hatzinger, 2007):

$$P(X_{vi} = x) = \frac{\exp[\phi(x)\theta_v + \psi(x)\beta_i + \chi(x)\theta_v\beta_i + \tau_x]}{\sum_{l=0}^{m} \exp[\phi(l)\theta_v + \psi(l)\beta_i + \chi(l)\theta_v\beta_i + \tau_l]}. \tag{3.2}$$

In Gleichung 3.2 ist X_{vi} die Antwort eines Individuums v auf ein Item i, x ist der Index der möglichen Antworten, welcher von 0 bis m läuft. $\phi(x)$, $\psi(x)$ und $\chi(x)$ sind Scoring-Funktionen, θ_v ist ein Parameter, der der Person zugeordnet ist, β_i ist ein item-bezogener Parameter und τ_x ist ein kategorienspezifischer Parameter.

Durch das Einführen einiger Restriktionen, wie dem Weglassen des Produktterms $\chi(x)\theta_v\beta_i$ und der Einschränkung der Scoring-Funktionen auf $\phi(x)$ folgt:

$$P(X_{vi} = x) = \frac{\exp[\phi(x)(\theta_v + \beta_i) + \tau_x)]}{\sum_{l=0}^{m} \exp[\phi(l)(\theta_v + \beta_i) + \tau_l]}. \tag{3.3}$$

Die Wahrscheinlichkeit $P(X_{vi} = x)$, dass eine Person v auf Item i die Kategorie x_{vi} wählt, wird in Abhängigkeit der Personen-Fähigkeit θ_v, der item-spezifischen Komponenten β_i (Item-Leichtigkeiten), kategorien-spezifischen Komponenten τ_x (Schwellen-Parametern) und der Scoring-Funktion $\phi(x)$ modelliert.

Wird die Scoring-Funktion $\phi(x)$ als

$$\phi(x) = x_{vi}, \text{ mit } x_{vi} = 0, \ldots, m, \tag{3.4}$$

definiert, so resultiert das Rating-Skalen-Modell von Andrich (1978b):

$$P(X_{vi} = x_{vi}) = \frac{\exp[x_{vi}(\theta_v + \beta_i) + \tau_x]}{\sum_{l=0}^{m} \exp[l(\theta_v + \beta_i) + \tau_l]}. \tag{3.5}$$

Eine wichtige Annahme dieses Modells ist, dass die Kategorien-Parameter τ_x für alle Items identisch sind. Die Item-Parameter β_i entsprechend den Leichtigkeiten der Items.

Im Partial-Credit-Modell von Masters (1982) wird die Annahme der über die Items identischen Kategorien-Parameter gelockert:

$$P(X_{vi} = x_{vi}) = \frac{\exp[x_{vi}\theta_v + \beta_{ix}]}{\sum_{l=0}^{m} \exp[l\theta_v + \beta_{il}]}. \tag{3.6}$$

In diesem Modell existieren kategorienspezifische Parameter β_{ix}, die über die Items hinweg variieren können oder anders ausgedrückt, die β_{ix} sind item- und kategorienspezi-

fische Konstanten. Nach Mair und Hatzinger (2007) beschreiben die β_{ix} item- und kategorienspezifische Leichtigkeiten. Zur Identifikation des Modells muss eine Restriktion über die kategorienspezifischen Parameter gesetzt werden. Dies kann einerseits über eine Summen-Normierung geschehen:

$$\sum_{x=1}^{m} \beta_{ix} = 0. \tag{3.7}$$

Dies bedeutet, dass die Summe der kategorienspezifischen Parameter für jedes Item Null gesetzt wird. Andererseits ist es ebenfalls möglich, den Parameter der ersten Kategorie für jedes Item Null zu setzten: $\beta_{i1} = 0$. Wird die Summen-Normierung über die Kategorien-Parameter gewählt, so lassen sich die Schwellen-Parameter τ_{ix} der polytomen Modelle durch eine einfache lineare Transformation berechnen (vgl. Skondral & Rabe-Hesketh, 2004):

$$\tau_{ix} = \beta_{i[x+1]} - \beta_{ix}.$$

Die Summe der Schwellen-Parameter τ_{ix} über ein Item ergibt die Leichtigkeit des Items.

Die Anwendung des Partial-Credit-Modells ist z.B. dann sinnvoll, wenn die Annahme des Rating-Skalen-Modells der Identität der Kategorien-Parameter über die Items in einem konkreten Anwendungsfall nicht zutrifft. Ferner geben Mair und Hatzinger (2007) eine Modellhierarchie an, an dessen Spitze das Partial-Credit-Modell von Masters steht. Das Rating-Skalen Modell von Andrich (1978b) und das dichotome Rasch-Modell lassen sich nach Mair und Hatzinger (2007) als Spezialfälle des Partial-Credit-Modells darstellen.

3.1.1. Eigenschaften probabilistischer Testmodelle am Beispiel des dichotomen Rasch-Modells

Die angesprochenen Modelle haben einige interessante Eigenschaften, die sie von den klassischen Modellen der Messfehlertheorie unterscheiden. Diese Eigenschaften seien im Folgenden am Beispiel des dichotomen Rasch-Modells dargestellt.

Modelldefinition

Das dichotome Rasch-Modell folgt im Prinzip mit $m = 1$ als Spezialfall der allgemeinen Formulierung des Rating-Skalen-Modells in Gleichung 3.5. Es entfallen die Parameter τ_x und x_{vi} darf nur die Werte 0 und 1 annehmen. Zudem wird das Vorzeichen des Schwierigkeitsparameters β_i umgekehrt, damit dieser eine Item-Schwierigkeit und keine Item-Leichtigkeit darstellt:

$$p(X_{vi} = x_{vi}) = \frac{\exp(x_{vi}(\theta_v - \beta_i))}{1 + \exp(\theta_v - \beta_i)}. \tag{3.8}$$

Die Wahrscheinlichkeit $p(X_{vi} = x_{vi})$, dass eine Person v eine Testaufgabe i löst ($X_{vi} = 1$), bzw. nicht löst ($X_{vi} = 0$), ist eine logistische Funktion der Differenz der Personen-Fähigkeit θ_v und der Itemschwierigkeit β_i. Anhand des dichotomen Rasch-Modells lassen sich einige generelle Eigenschaften von Rasch-Modellen erläutern. Eine Modellannahme ist, dass es sich bei der latenten Fähigkeitsdimension θ um einen eindimensionalen Trait handelt. Zudem werden Itemschwierigkeiten und Personen-Fähigkeiten gemeinsam auf einer Dimension skaliert. In diesem Sinne handelt es sich bei dem dichotomen Rasch-Modell um ein Messmodell, dass die Schwierigkeiten der Testaufgaben und die Fähigkeiten der Personen auf einer Skala abbildet.

Spezifische Objektivität und Separierbarkeit der Parameter

Eine wichtige Eigenschaft von Rasch-Modellen besteht in der sog. spezifischen Objektivität der Messung (Rasch, 1961). Fischer (1995a) beschreibt spezifische Objektivität als die Genrealisierbarkeit des Vergleichs zweier Objekte auf der Basis eines Satzes von Indikatoren für ein spezielles Merkmal. Liefern unterschiedliche Indikatoren identische Ergebnisse hinsichtlich der relativen Merkmalsausprägungen der interessierenden Objekte, so gilt die Messung als spezifisch objektiv. Nach Fischer (1995a) erläuterte Rasch (1960) spezifische Objektivität gerne am Beispiel der klassischen Mechanik. Dieses Beispiel, welches auch von Andrich (Andrich, 1988, p. 19) berichtet wird, sei hier zur Veranschaulichung wiedergegeben. Nehmen wir an, es existiert ein Satz von Objekten O_v mit den Massen M_v. Werden in einer experimentellen Situation Kräfte F_i auf diese Objekte appliziert, ist die

Beschleunigung A_{vi} beobachtbar. Nach dem zweiten Newton'schen Axiom gilt:

$$A_{vi} = M_v^{-1} F_i. \tag{3.9}$$

Der Vergleich zweier Objekte O_v und O_w bezüglich ihrer Massen kann durch folgenden Quotienten durchgeführt werden:

$$\frac{A_{vi}}{Awi} = \frac{M_v^{-1} F_i}{M_w^{-1} F_i} = \frac{M_w}{M_v}. \tag{3.10}$$

Dies bedeutet nach Rasch, dass das Verhältnis der Massen der Objekte äquivalent zum Verhältnis der beobachteten Beschleunigungen und unabhängig von den applizierten Kräften F_i ist. Der Vergleich der Massen der Objekte anhand der applizierten Kräfte ist in dem Sinne spezifisch objektiv, als dass lediglich die Beschleunigungen beobachtet werden müssen, um die Objekte hinsichtlich der Massen zu vergleichen.

Ersetzen wir nun die Massen gedanklich durch Personen-Fähigkeiten, die Kräfte durch Items und die Beschleunigung durch Lösungswahrscheinlichkeiten, so zeigt sich eine Analogie zum dichotomen Rasch-Modell. Formal zeigt sich die Modelleigenschaft der spezifischen Objektivität beim dichotomen Rasch-Modell, wenn die Logits der Lösungswahrscheinlichkeiten berechnet werden:

$$log\left(\frac{p(X_{vi}=1)}{p(X_{vi}=0)}\right) = \theta_v - \beta_i. \tag{3.11}$$

Die Logits der Lösungswahrscheinlichkeit stehen in einem linearen Verhältnis zu der Differenz der Personen-Fähigkeiten und der Itemschwierigkeiten.

Werden nun die Differenzen der Logits der Lösungswahrscheinlichkeiten zweier Personen v und w mit den Fähigkeits-Parametern θ_v und θ_w gebildet, so verschwinden die Item-Parameter aus der Gleichung und es zeigt sich, dass die Personen-Fähigkeiten auf einer Differenzenskala abgebildet werden:

$$log\left(\frac{p(X_{vi}=1)}{p(X_{vi}=0)}\right) - log\left(\frac{p(X_{wi}=1)}{p(X_{wi}=0)}\right) \quad = \quad \theta_v - \theta_w. \tag{3.12}$$

Dies bedeutet für das Rasch-Modell, dass bei Modellgeltung der Vergleich der Personen-Fähigkeiten auf einer Logit-Skala nicht von den Item-Parametern abhängt, sofern die Items aus einem Itemsatz stammen, für die das Rasch-Modell in der Zielpopulation gilt.

Eine weitere formale Grundlage dieser sogenannten Separierbarkeit der Personen- und Item-Parameter im Rasch-Modell ist das bedingte Rasch-Modell. Das bedingte Rasch-Modell gibt die Wahrscheinlichkeiten von personenbezogenen Antwortvektoren $p(\mathbf{x_v}|r_v)$ bei gegebener Anzahl gelöster Aufgaben r_v an. Die folgende Darstellung des Modells lehnt sich an diejenige von Molenaar (1995) an.

Mit $\xi_v = \exp(\theta_v)$ und $\epsilon_i = \exp(-\beta_i)$ folgt das dichotome Rasch-Modell in delogarithmierter Schreibweise:

$$p(X_{vi} = 1) = \frac{\xi_v \epsilon_i}{1 + \xi_v \epsilon_i} \tag{3.13}$$

und

$$p(X_{vi} = 0) = \frac{1}{1 + \xi_v \epsilon_i}. \tag{3.14}$$

In dem Fall, dass ein Test nur aus zwei Items besteht, ergibt sich die Wahrscheinlichkeit, dass eine Person bei gegebenem Rohwert von $x_{v.} = 1$ Aufgabe 1 löst und Aufgabe 2 nicht löst unter Anwendung des Multiplikations- und Additionstheorems zu:

$$p(X_{v1} = 1, X_{v2} = 0 | r_v = 1) = \frac{\frac{\xi_v \epsilon_i}{(1+\xi_v\epsilon_i)(1+\xi_v\epsilon_i)}}{\frac{\xi_v\epsilon_i + \xi_v\epsilon_i}{(1+\xi_v\epsilon_i)(1+\xi_v\epsilon_i)}} = \frac{\epsilon_1}{\epsilon_1 + \epsilon_2}. \tag{3.15}$$

Das Bemerkenswerte an dieser Stelle ist, dass die Personen-Parameter aus der Gleichung gekürzt werden können und die bedingte Wahrscheinlichkeit eines Antwortvektors bei gegebenem Rohwert nicht von der Personen-Fähigkeit abhängt, sondern lediglich von der Schwierigkeit der Aufgaben.

Die Verallgemeinerung dieses Gedankens auf eine Testlänge von k Items führt zum bedingten Rasch-Modell (vgl. E. Molenaar, 1995):

$$\begin{aligned} p(\mathbf{x_v}|r_v) &= \frac{\prod_{i=1}^{k} \epsilon_i^{x_{vi}}}{\sum_{\mathbf{y}|r_v} \prod_{i=1}^{k} \epsilon_i^{y_i}} \\ &= \frac{\exp\left(-\sum_{i=1}^{k} x_{vi}\beta_i\right)}{\sum_{\mathbf{y}|r_v} \exp\left(-\sum_{i=1}^{k} y_i\beta_i\right)}. \end{aligned}$$

Die Summe im Nenner läuft über alle möglichen Antwortmuster $\mathbf{y}$ der Länge k, die einen Rohwert von r_v ergeben. Das y_i im Nenner bezeichnet das jeweilige Element k des Antwortvektors $\boldsymbol{y}$, der zu einem Rohwert von r_v führt. Dieser Ausdruck wird auch als

elementarsymmetrische Grundfunktion r_v-ter Ordnung, oder kurz als γ_{r_v} bezeichnet. In delogarithmierter Schreibweise sieht der Ausdruck folgendermaßen aus:

$$\begin{aligned}
\gamma_0 &= 1 \\
\gamma_1 &= \epsilon_1 + \epsilon_2 + \ldots + \epsilon_k \\
\gamma_2 &= \epsilon_1\epsilon_2 + \epsilon_1\epsilon_3 + \ldots + \epsilon_{k-1}\epsilon_k \\
&\vdots \\
\gamma_k &= \epsilon_1\epsilon_2 \ldots \epsilon_k.
\end{aligned}$$

Die Bedeutung dieser Funktion lässt sich anhand des Multiplikations- und Additionstheorems der Wahrscheinlichkeitsrechnung veranschaulichen. Es existiert nur ein Antwortmuster, dass zu einem Rohwert von 0 führt. Ein Rohwert von 1 ist bei Lösung des zweiten Items (ϵ_1) *oder* durch Lösung des zweiten Items (ϵ_2) *oder* durch Lösung des dritten Items (ϵ_3) ... oder durch Lösung des k-ten Items erzielbar und so fort.

Ein Rohwert von 2 ist bei Lösung des ersten Items (ϵ_1) *und* des zweiten Items (ϵ_2) *oder* durch Lösung des ersten Items (ϵ_2) *und* des dritten Items (ϵ_2) *oder* durch Lösung des ersten Items (ϵ_1) *und* des dritten Items (ϵ_3) usw. erzielbar.

Die numerische Berechnung der elementarsymmetrischen Grundfunktion sieht auf den ersten Blick trivial aus, ist allerdings ein relativ komplexes Problem, wenn man sich vor Augen führt, dass bei 20 Items schon $2^{20} = 1048576$ unterschiedliche Antwortmuster existieren. Als klassische Lösungen für die Bestimmung von γ_{rv} finden sich in der Literatur die Summations- und Differenzen-Methode (vgl. Fischer, 1995b).

Die Item-Parameter lassen sich durch die Maximierung der bedingten Likelihood unter Annahme der stochastischen Unabhängigkeit der Antwortvektoren $\boldsymbol{x_v}$ ohne Bezugnahme auf die Personen-Parameter schätzen.

$$cL = \prod_{v=1}^{N} \frac{\exp\left(-\sum_{i=1}^{k} x_{vi}\beta_i\right)}{\sum_{\mathbf{y}|r_v} \exp\left(-\sum_{i=1}^{k} y_i\beta_i\right)} \rightarrow \max. \tag{3.16}$$

Bei diesem Vorgehen sind die Spaltensummen einer Datenmatrix, d.h. die Häufigkeiten der Lösungen eines Items suffiziente Statistiken zur Schätzung der Item-Parameter (E. Molenaar, 1995).

Ebenso ist es möglich, die Personen-Parameter unabhängig von den verwendeten Indikatoren zu schätzen, wenn das bedingte Rasch-Modell auf Basis der Summe der Lösungen auf den Items, anstatt auf der Summe der Lösungen der Personen über alle Aufgaben formuliert wird.

Die Eigenschaft der Separierbarkeit der Parameter ist eine Folge dessen, dass das Rasch-Modell zur Exponentialfamilie gehört und somit suffiziente Statistiken zur Schätzung der Parameter besitzt. Pitman (1936) und Koopman (1936) haben allgemein gezeigt, dass Funktionen, die zur Exponentialfamilie gehören suffiziente Statistiken zur Schätzung der Parameter aufweisen. Nach Molenaar (1995) ist ein Standard-Resultat der Exponentialfamilie, dass die Parameter, für die suffiziente Statistiken vorliegen, bei der Parameterschätzung nach der Conditional-Maximum-Likelihoood-Methode nicht benötigt werden, sofern ein bedingtes Modell auf Basis der entsprechenden suffizienten Statistiken verwendet wird.

Suffiziente Statistiken zur Parameterschätzung

Nach Mair und Hatzinger (2007) ist die Conditional-Maximum-Likelihoood-Methode für die Bedeutung der spezifischen Objektivität bei Rasch-Modellen von theoretischem Interesse. Ein weiterer Ansatz der Parameterschätzung ist die Joint-Maximum-Likelihood-Methode.

Wie u.a. auch Andrich (1988) zeigt, gilt unter der Annahme der stochastischen Unabhängigkeit der Antworten x_{vi}:

$$L = \prod_{v=1}^{N}\prod_{i=1}^{k} p(X_{vi} = x_{vi}) \tag{3.17}$$

$$= \prod_{v=1}^{N}\prod_{i=1}^{k} \frac{\exp(x_{vi}(\theta_v - \beta_i))}{1 + \exp(\theta_v - \beta_i)}. \tag{3.18}$$

L ist Likelihood-Funktion über eine Datenmatrix der Dimensionalität $N \cdot k$. Wird L logarithmiert und partiell nach den Parametern differenziert, so folgen die Maximum-

Likelihood-Schätzgleichung zur Schätzung der Parameter (vgl. E. Molenaar, 1995):

$$logL = \sum_{v=1}^{N}\sum_{i=1}^{k} x_{vi}(\theta_v - \beta_i) - \sum_{v=1}^{N}\sum_{i=1}^{k} log(1 + \exp(\theta_v - \beta_i)) \quad (3.19)$$

$$= \sum_{v=1}^{N} x_{v.}\theta_v - \sum_{i=1}^{k} x_{.i}\beta_i - \sum_{v=1}^{N}\sum_{i=1}^{k} log(1 + \exp(\theta_v - \beta_i)). \quad (3.20)$$

$x_{v.}$ ist der Personen-Rohwert ($\sum_{i=1}^{k} x_{vi}$=$x_{v.}$) , $x_{.i}$ ist der Item-Rohwert ($\sum_{v=1}^{N} x_{vi}$=$x_{.i}$). In Gleichung 3.20 wird der Sachverhalt der Suffizienz der Randsummen dadurch deutlich, dass der jeweilige Parameter innerhalb eines Summanden bei der jeweiligen suffizienten Statistik zur Schätzung des Parameters steht. Der Personen-Rohwert $x_{v.}$ ist somit eine suffiziente Statistik zur Schätzung des Parameters θ_v und der Item-Rohwert $x_{.i}$ ist eine suffiziente Statistik zur Schätzung des Parameters β_i. Noch deutlicher wird die Tatsache der Suffizienz, wenn partiell nach den Parametern differenziert wird:

$$\frac{\partial logL}{\partial \theta_v} = \sum_{i=1}^{k} x_{vi} - \sum_{i=1}^{k} \frac{\exp(\theta_v - \beta_i)}{(1 + \exp(\theta_v - \beta_i))} \quad (3.21)$$

$$\frac{\partial logL}{\partial \beta_i} = \sum_{v=1}^{N} x_{vi} - \sum_{v=1}^{N} \frac{\exp(\theta_v - \beta_i)}{(1 + \exp(\theta_v - \beta_i))}. \quad (3.22)$$

Durch Nullsetzen und umstellen erhalten wir die sogenannten Joint-Maximum-Likelihood-Schätzgleichungen:

$$\sum_{i=1}^{k} x_{vi} = \sum_{i=1}^{k} \frac{\exp(\theta_v - \beta_i)}{(1 + \exp(\theta_v - \beta_i))} \quad (3.23)$$

$$\sum_{v=1}^{N} x_{vi} = \sum_{v=1}^{N} \frac{\exp(\theta_v - \beta_i)}{(1 + \exp(\theta_v - \beta_i))}. \quad (3.24)$$

Für die Schätzung der Parameter werden also nur die Randsummen einer Datenmatrix benötigt. Der Gradient der Schätzung strebt gegen die Bedingung der Gleichheit der unter dem Modell erwarteten Randsummen mit den beobachteten Randsummen einer Datenmatrix. Das Vorliegen von suffizienten Statistiken zur Parameterschätzung ist eine sehr vorteilhafte Eigenschaft. Im Falle von Rasch-Modellen bedeutet dies praktisch, dass die suffiziente Statistik, wie z.B. der Summenwert, die gesamte Information enthält, die zur Schätzung des entsprechenden latenten Parameters benötigt wird.

Die Informationsfunktion

Die sogenannte Informationsfunktion eines Tests basiert auf der zweiten partiellen Ableitung der log-Likelihood-Funktion nach dem Parameter θ_v (vgl. hierzu z.B. Hoijtink und Boomsma, 1995 und Rost, 2004, p. 358):

$$\frac{\partial^2 logL}{\partial \theta_v^2} = -\sum_{i=1}^{k} \frac{\exp(\theta_v - \beta_i)}{(1+\exp(\theta_v - \beta_i))^2} \tag{3.25}$$

$$= -\sum_{i=1}^{k} \frac{\exp(\theta_v - \beta_i)}{(1+\exp(\theta_v - \beta_i))} \cdot \frac{1}{(1+\exp(\theta_v - \beta_i))} \tag{3.26}$$

$$= -\sum_{i=1}^{k} p(X_{vi}=1)p(X_{vi}=0). \tag{3.27}$$

Der negative Erwartungswert der zweiten partiellen Ableitung wird als die Informationsfunktion eines Tests bezeichnet (Rost, 2004, p. 357):

$$I = -E\left\{\frac{\partial^2 logL}{\partial \theta_v^2}\right\} \tag{3.28}$$

$$= \sum_{i=1}^{k} p(X_{vi}=1)p(X_{vi}=0). \tag{3.29}$$

Die Informationsfunktion ist als Folge der stochastischen Unabhängigkeit über alle k Items additiv. Die untere Schranke der Varianz eines Parameterschätzers $\hat{\theta}_v$ für einen Gesamttest ergibt sich aus dem Kehrwert der Informationsfunktion:

$$Var(\hat{\theta}_v) = \frac{1}{I} \tag{3.30}$$

$$= \frac{1}{\sum_{i=1}^{k} p(X_{vi}=1)p(X_{vi}=0)}. \tag{3.31}$$

Nach der Cramér-Rao-Ungleichung (Rao, 1945) ist der Kehrwert der Informationsfunktion die untere Schranke der Varianz eines Schätzers.

Die Informationsfunktion ist von praktischem Interesse. Nach der Informationsfunktion misst ein einzelnes Item besonders gut in dem Skalenbereich der der Itemschwierigkeit β_i entspricht, da die Informationsfunktion eines einzelnen Items am Ort (θ_v=β_i) ein Maximum besitzt. Zudem ist am Ort des Maximums die Varianz des Schätzers der Personen-Fähigkeit minimal. Durch die Additivität der Informationsfunktion über alle Items eines

Tests ist es möglich, a priori einen Test aus einem Rasch-homogenen Item-Satz zusammenzustellen, der in einem definierten Merkmalsbereich besonders gut differenziert. Ein Anwendungsfall ist das computerisierte adaptive Testen (CAT). Wird eine Testsitzung an einem Computer durchgeführt, lässt sich die Personen-Fähigkeit im Laufe der Testsitzung fortwährend schätzen. Die Item-Vorgabe erfolgt dann in Abhängigkeit des geschätzten Fähigkeitsniveaus der Person, was sich positiv auf die Testökonomie auswirkt. Da die Item-Vorgabe exakt an die Fähigkeiten einer Person angepasst werden kann, ergibt sich eine erhöhte Präzision der Schätzung der Personen-Fähigkeit θ_v. Voraussetzung für diese Praxis ist allerdings das Vorliegen eines Itemsatzes, für den nachgewiesen wurde, dass das Rasch-Modell in der spezifischen Zielpopulation gilt.

Prüfbarkeit

Die bisher beschriebenen Eigenschaften sind Eigenschaften des Modells. Um zu überprüfen, ob ein probabilistisches Testmodell auf einen gegebenen Datensatz passt, stehen eine Reihe von Ansätzen zur Verfügung. Nach Rost (2004) existiert keine verbindliche Taxonomie der Vielzahl von Modelltests, die für Rasch-Modelle vorliegen. Glas und Verhelst (1995) geben einen Überblick über gängige Verfahren und Rost (2004) schlägt eine Einteilung der Modellgeltungs-Tests in drei Kategorien vor: Globale Modellgeltungstests, Tests der Itemhomogenität und Tests der Personenhomogenität. Im Folgenden wird für jede Klasse von Tests ein Beispiel gegeben, um die Funktionsweise dieser Tests zu veranschaulichen.

Ein Beispiel für einen globalen Modellgeltungstest ist der Test gegen die saturierte Likelihood (vgl. Rost, 2004). Bei der Erhebung von Daten im Rahmen einer Testung fallen Datenmatrizen an, die aus Antwortmustern von Personen bestehen. Bei einem Test mit dichotomem Antwortformat der Länge $k = 10$ existieren beispielsweise $2^k = 2^{10}$ mögliche unterschiedliche Antwortmuster $\boldsymbol{x}$, von denen jedes potentiell mehrmals beobachtbar ist. Die Berechnung der saturierten Likelihood setzt an den beobachteten Häufigkeiten $n(\mathbf{x})$ der jeweiligen Antwortmuster an. Zur Bestimmung der saturierten Likelihood werden die in einer Testung beobachteten, relativen Häufigkeiten der Antwortmuster $n(\mathbf{x})/N$

berechnet.

N ist hierbei die Anzahl der Personen, die getestet wurden. Die relative Häufigkeit der Antwortmuster wird auch als Pattern-Wahrscheinlichkeit bezeichnet. Zur Berechnung der saturierten Likelihood werden die Pattern-Wahrscheinlichkeiten für alle beobachteten Pattern gebildet und über die beobachteten Pattern multipliziert:

$$Lsat = \prod_{\mathbf{x}} \left(\frac{n(\mathbf{x})}{N} \right)^{n(\mathbf{x})}. \tag{3.32}$$

Im Prinzip beruht die saturierte Likelihood auf einem Modell, für das keine Restriktionen vorliegen, denn die Berechnung der saturierten Likelihood setzt direkt an den beobachteten Daten an und es werden keine Parameter geschätzt. Die Freiheitsgrade der saturierten Likelihood sind $df_{sat} = 2^k - 1$. Insgesamt existieren 2^k unterschiedliche Pattern-Wahrscheinlichkeiten, aber eine dieser Wahrscheinlichkeiten kann berechnet werden, wenn die anderen Pattern-Wahrscheinlichkeiten bekannt sind.

Werden nun Restriktionen eingeführt um die Daten zu erklären, so können die Modelle, welche die Restriktionen abbilden, keine höhere Likelihood besitzen, als die saturierte Likelihood. Von daher bietet es sich an, zur globalen Testung eines konkreten Modells die saturierte Likelihood heranzuziehen.

Konkret kommen Likelihood-Ratio-Tests zum Einsatz, die auf dem Likelihood-Quotienten

$$LR = \frac{L_m}{L_{sat}} \tag{3.33}$$

aufbauen. L_m ist die Likelihood des zu testenden Modells und L_{sat} ist die Likelihood des saturierten Modells.

Sind die asymptotischen Voraussetzungen erfüllt, so folgt die Teststatistik

$$-2 \cdot \log(LR) \tag{3.34}$$

bei Geltung der Nullhypothese, dass das Modell perfekt passt, einer χ^2-Verteilung mit $df = df_{sat} - df_m$ Freiheitsgraden. Üblicher Weise wird die Nicht-Signifikanz eines solchen Tests als Hinweis auf Modellgeltung gewertet. Dieses Vorgehen ist aus zwei Gründen problematisch. Erstens bedeutet die Nicht-Signifikanz eines klassischen Signifikanz-Tests

nicht zwingend, dass die Nullhypothese in der Population gilt. Zweitens müssen die asymptotischen Voraussetzungen des Tests erfüllt sein, ansonsten folgt die Teststatistik nicht der zentralen χ^2-Verteilung. Die Erfüllung der asymptotischen Voraussetzungen bedeutet, dass die Stichprobe hinreichend groß sein muss, so dass möglichst alle möglichen Antwortmuster auch beobachtet werden können. Führt man sich vor Augen, dass bei einem Test mit nur 10 Items bereits 1024 Antwortmuster möglich sind, so wird klar, dass in den wenigsten Fällen genügend Personen getestet werden, um einen gültigen Modellgeltungstest durchzuführen. Größere Ausfälle an potentiell beobachtbaren Antwortmustern führen zu einem Testergebnis, dass auf asymptotischen Annahmen aufbaut, dessen Voraussetzungen nicht erfüllt sind.

Zur Lösung dieser Problematik, welche eine Vielzahl von Modellgeltungs-Tests betrifft, schlug von Davier (1997) ein Bootstrap-Verfahren (Efron & Tibshirani, 1993) vor. Von der Idee her werden in einem ersten Schritt die Parameter eines zu testenden Modells geschätzt. Die geschätzten Parameter werden verwendet, um Daten aus dem Modell stochastisch zu simulieren. Für die simulierten Daten wird ein Modellgeltungstest durchgeführt und die entsprechende Teststatistik wird aufgezeichnet. Dieser Vorgang der Simulation und Berechnung einer Teststatistik wird sehr häufig wiederholt, so dass eine Verteilung der Teststatistik bei Modellgeltung folgt. Die empirische, auf dem beobachteten Datensatz basierende Teststatistik wird nun anhand der simulierten Verteilung evaluiert. Liegt die Teststatistik des empirischen Datensatzes im Zentrum der simulierten Verteilung, so gilt dies als Hinweis auf Modellpassung. Diese Verfahren sind in dem Programm WINMIRA (Davier von, 2000) zur Schätzung von Rasch- und Mixed-Rasch-Modellen implementiert.

Ein weiterer bekannter Modellgeltungstest ist der Andersen-Test (Andersen, 1973), ein Likelihood-Ratio-Test, welcher auf dem bedingten Rasch-Modell aufbaut. Dieser Test prüft die Nullhypothese, dass die Itemschwierigkeiten in den durch Score-Gruppen definierten Subsets des Gesamtdatensatzes identisch sind. Im dichotomen Rasch-Modell sind beispielsweise keine Interaktionen zwischen Items und Personen vorgesehen, daher beschreibt dieses Modell eine Situation, bei dem die Itemschwierigkeiten in jedem belie-

bigen Subset der Personen identisch sind. Nach Rost (2004) handelt es sich bei dem Test von Andersen um einen Test auf Personenhomogenität.

Zur Durchführung des Tests werden die Itemschwierigkeiten in jeder Rohwert-Gruppe separat mit dem bedingten Rasch-Modell geschätzt. In der Praxis wird die Stichprobe auch lediglich in nur zwei Gruppen geteilt, um die Identität der Item-Parameter in den Subgruppen zu überprüfen.

Andersen (1973) hat gezeigt, dass der auf dem Likelihoodquotienten λ basierende Ausdruck

$$-2\log(\lambda) \quad = \quad -2\log\left(\frac{cL_{\text{ges}}}{\prod_{r=1}^{k-1} cL_r}\right)$$

asymptotisch mit $df = (k-1)\cdot(k-2)$ Freiheitsgraden χ^2 verteilt ist, wenn die Nullhypothese der identischen Itemschwierigkeiten gilt. Im Zähler des Quotienten steht die bedingte Likelihood cL_{ges}, die sich ergibt, wenn die Item-Parameter über die Gesamtdaten geschätzt werden. Im Nenner steht das Produkt der Likelihoods für separate Parameterschätzungen in den jeweiligen Rohwert-Gruppen. Inhaltlich bedeutet dies: sind die Item-Parameter in den Rohwert-Gruppen identisch, so ist $\lambda = 1$ und eine separate Schätzung der Item-Parameter liefert keine zusätzliche Information hinsichtlich der Daten gegenüber der Schätzung über den Gesamtdatensatz. Damit wird $-2\ \log(\lambda)$ gleich Null. Der Test prüft also die Nullhypothese der identischen Itemschwierigkeiten in den Rohwert-Gruppen. Eine Nicht-Signifikanz des Tests wird als Indikator für die Modellgeltung gewertet. Auch bei diesem Test stellt sich die Fragen, die auch für den Likelihood-Ratio-Test gegen das saturierte Modell auftraten: Wann sind die asymptotischen Voraussetzungen erfüllt und wann macht es Sinn, eine Nicht-Signifikanz auf Hinweis der Modellgeltung zu werten? Zudem zeigte Stelzl (1979), dass der Andersen-Test insensitiv auf Itemheterogenität reagiert.

Ein Test zur Prüfung der Itemhomogenität wurde von Martin-Loef (Martin-Loef, 1973) vorgeschlagen. Die Annahme der Itemhomogenität besagt nach Rost (2004), dass alle Items in einem Test die selbe latente Eigenschaft θ_v erfassen, was sich daran zeigt, dass die Personen-Parameter unabhängig von Subset der Items sein sollten, das zur Schätzung

der Personen-Fähigkeit herangezogen wird. Eine Abwandlung dieser Überlegung ist die Grundlage für die Konstruktion des Martin-Loef-Tests. Martin-Loef verwendete nicht die Item-Parameter, sondern die jeweiligen suffizienten Statistiken als Ausgangsbasis für den Test. Sollten die Items homogen sein, so müssten Personen mit hohen Scores in einer Testhälfte auch hohe Scores in der anderen Testhälfte erzielen. Auch bei der Anwendung des Martin-Loef-Tests fällt eine Teststatistik an, deren Verteilung bei Geltung der Nullhypothese asymptotisch χ^2 verteilt ist.

Während die vorstehenden Ansätze zur Testung der Modellgeltung vornehmlich Abwandlungen bekannter χ^2-Verfahren und Likelihood-Rato-Tests darstellen, besteht eine weitere Möglichkeit der Bewertung der Modellpassung in der Berechnung von Fit-Statistiken, die auf standardisierten Residuen aufbauen (siehe z.B. Wright, 1969 oder von Davier, 1996). Nach Rost (2004) ist ein einfaches Residualmaß

$$z_{vi} = \frac{x_{vi} - \langle x_{vi} \rangle}{\sqrt{var(x_{vi})}}. \tag{3.35}$$

z_{vi} ist das standardisierte Residuum des Modells hinsichtlich der Itemantwort x_{vi}. $\langle x_{vi} \rangle$ ist der Erwartungswert der Antwort unter dem Modell und $var(x_{vi})$ ist die Varianz der Itemantwort unter dem Modell. Diese z-Werte können quadriert und über alle Personen v und Items i addiert werden, wobei χ^2-verteilte Teststatistiken anfallen. Somit lassen sich standardisierte Residuen nutzen, um detaillierte Analysen hinsichtlich der Modellgeltung durchzuführen. Beispielsweise ließen sich die standardisierten Residuen über jedes einzelne Item aggregieren, um die Modellpassung hinsichtlich der Items zu bewerten. Es wäre ebenfalls möglich, die Residuen über die Personen zu aggregieren, um Aussagen über die Fehlpassung von personenbezogenen Antwortmustern zu treffen. Schließlich wäre eine globale Modellkontrolle denkbar, indem über alle Personen- und Items aggregiert wird. In Gleichung 3.35 besteht die einzige Schwierigkeit darin, die Varianz der Antwort unter dem Modell zu berechnen. Diese ergibt sich jedoch aus der zweiten partiellen Ableitung der Modellgleichung nach den jeweiligen Parametern, bzw. aus der Informationsfunktion. Die Berechnung des Erwartungswerts einer Antwort unter dem Modell ist durch die erste partielle Ableitung des logarithmierten Nenners des Modells Modells nach dem entsprechenden Parameter möglich. Diese Sachverhalte werden in einem späteren Abschnitt im

Bezug auf das in dieser Arbeit zu generierende Testmodell genauer beleuchtet.

Reliabilität

Neben der Bewertung der Modellpassung ist bei der Testkonstruktion die Bewertung der Reliabilität eines Tests von Interesse. Gemeinhin werden die bei der Parameterschätzung anfallenden Standardfehler der Personen-Parameter als Indikatoren der Messgenauigkeit verwendet. Der Unterschied zum Ansatz der Klassischen Testtheorie liegt hierbei darin, dass die Messgenauigkeit der individuellen Merkmalsausprägung nicht über alle Bereiche der Trait-Skala homogen ist, sondern die Tatsache berücksichtigt wird, dass extreme Merkmalsausprägungen weniger genau erfasst werden. Andererseits existiert ein Maß der Messgenauigkeit, dass dem Reliabiltäts-Konzept in der klassischen Testtheorie ähnelt und das als Separabilität bezeichnet wird (Hoijtink & Boomsma, 1995; Andrich, 1988). Der Begriff Separabilität wurde vermutlich daher gewählt, da eine hohe Reliabilität eines Gesamt-Tests im klassischen Sinne anzeigt, dass ein Test zwischen Personen trennt.

Einen Reliabilitäts-Index im Rahmen von Rasch-Modellen, der sich an das Reliabiltäts-Konzept der Klassischen Testtheorie anlehnt, berichten Hoijtink & Boomsma (1995) :

$$Rel_1 = \frac{\sigma^2(\theta)}{\sigma^2(\hat{\theta})} = \frac{\sigma^2(\hat{\theta}) - \sigma^2(e)}{\sigma^2(\hat{\theta})}. \tag{3.36}$$

Dieser Koeffizient wird auch als Andrich-Reliabilität bezeichnet (Andrich, 1988). $\sigma^2(\theta)$ ist hierbei die Varianz der latenten Trait-Verteilung und $\sigma^2(\hat{\theta})$ ist die Varianz der geschätzten Personen-Parameter. $\sigma^2(e)$ ist die Fehlervarianz der Schätzung. Die Reliabilität entspricht dem Anteil der Varianz der wahren Werte an den beobachteten Werten. Ein praktischer Problempunkt bei der Berechnung der Reliabilität ist es, die Varianz der latenten Trait-Verteilung zu schätzen. Dies kann dadurch geschehen, dass die latente Trait-Verteilung selbst - z.B. im Rahmen eines Multilevel-Ansatzes - modelliert wird.

Wird die latente Trait-Verteilung nicht modelliert, kann die Reliabilität auf Basis der Schätzfehler-Varianz der geschätzten Personen-Parameter und der mittleren Varianz der Parameterschätzer bewertet werden (Andrich, 1988). In diesem Fall entspricht $\sigma^2(\hat{\theta})$ der Varianz der geschätzten Personen-Parameter und $\sigma^2(e)$ wird über die mittlere

Schätzfehler-Varianz der Personen-Parameter ermittelt.

3.1.2. Zusammenfassende Betrachtung

Zusammenfassend kann konstatiert werden, dass probabilistische Testmodelle im Gegensatz zur klassischen Testtheorie die kategoriale Natur von Test-Daten in der Modellierung berücksichtigen. Die Merkmalsausprägung wird latent auf einer Differenzen-Skala abgebildet, die sinnvolle Vergleiche zwischen den Merkmalsausprägungen verschiedener Personen erlaubt. Die Homogenitäts-Annahmen der Modelle sind prüfbar und die individuelle sowie die globale Genauigkeit der Merkmalserfassung sind bewertbar. Ferner bieten Rasch-Modelle die Möglichkeit des dynamischen, adaptiven Testens, was bei klassischen Testmodellen nicht unbedingt möglich ist.

3.2. Maximum-Entropie-Modelle

Während der klassische Ansatz von Rasch, Rasch-Modelle herzuleiten, ist in der vorliegenden Arbeit von Interesse, inwiefern sich die Maximum-Entropie-Methode eignet, probabilistische Testmodelle zu definieren. Die Hintergründe dieser Methode werden im folgenden kurz dargestellt. Der US-amerikanische Physiker Edwin T. Jaynes veröffentlichte im Jahr 1957 zwei Artikel in der Zeitschrift *Physical Review*, in denen er zeigte, dass sich eine wichtige Wahrscheinlichkeitsverteilung der statistischen Mechanik, namentlich der kanonische Zustand eines kanonischen Ensembles, sich unter Anwendung einer Methode herleiten lässt, die als Maximum-Entropie-Methode bezeichnet wird.

Jaynes (1957a, 1957b) beschreibt die Methode als eine Möglichkeit, Wahrscheinlichkeitsverteilungen aus informationstheoretischen Überlegungen heraus herzuleiten. Die Methode wird manchmal im Zusammenhang mit dem Prinzip der Indifferenz (Keynes, 1921) genannt. Das Prinzip der Indifferenz ist eine Regel der Zuweisung epistemischer Wahrscheinlichkeiten hinsichtlich des Eintretens eines Ereignisses x_i, $i \in 1, \ldots, n$, wenn keine Informationen hinsichtlich des Ereignisses vorhanden sind. Nach dem Prinzip der Indifferenz sind in diesem Fall alle möglichen Ereignisse a priori gleich wahrscheinlich

($p_i = 1/n$) (Keynes, 1921, p. 42). Jaynes erweiterte dieses Prinzip dahingehend, dass er einen Formalismus entwickelte, der es ermöglicht, die Wahrscheinlichkeitsverteilungen hinsichtlich des Eintretens von Ereignissen oder des Zustands von Systemen auf Basis der verfügbaren Information (Daten) zu ermitteln. Als Kriterium für die Herleitung von Wahrscheinlichkeitsverteilungen auf der Basis partieller Information bedient sich Jaynes einer Größe aus der Informationstheorie, der Shannon-Entropie (Shannon, 1948). Kurz gesagt ist das Rational hinter der Anwendung der Maximum-Entropie-Methode dasjenige der formalen Konstruktion von Wahrscheinlichkeitsverteilungen die einerseits maximale Informationsentropie besitzen und andererseits mit der partiellen Information, die meist die Form von Erwartungswerten über Funktionen von Daten annimmt, kongruent sind. Nach Jaynes (1957a) sind die auf diese Weise gefundenen Verteilungen *least biased* im Sinne von „am wenigsten voreingenommen“, da sie eben diejenigen Verteilungen sind, die nur diejenigen Informationen berücksichtigen, die durch die beobachteten Daten gegeben sind. Abgesehen von den Informationen in den Daten und der Wahl der Nebenbedingungen, besitzen diese Verteilungen maximale Informationsentropie und werden somit manchmal als „objektiv“ beschrieben. Der Begriff *bias* ist hier nicht im Sinne der Schätztheorie gemeint. Wahrscheinlichkeiten werden von Jaynes als Grad des Wissens über den Zustand eines Systems aufgefasst, wobei dieses Wissen allerdings auf empirischen Beobachtungen fußt und die vorhandenen Unsicherheiten in Form von maximaler Informationsentropie der entsprechenden Verteilungen berücksichtigt werden. Als Konsequenz der Möglichkeit der Herleitung physikalischer Verteilungen allein auf Basis informationsthoretischer und wahrscheinlichkeitstheoretischer Argumentation schließt Jaynes (1957a), dass die Informationsentropie - als Maß der Unsicherheit in einer Wahrscheinlichkeitsverteilung - ein weitaus fundamentaleres Konzept darstellt, als beispielsweise Energie.

3.2.1. Definition und Eigenschaften der Maximum-Entropie-Verteilung

Ausgangspunkt von Jaynes (2003, p. 355) Darstellung der Herleitung einer sehr allgemeine Maximum-Entropie-Verteilung für diskrete Daten ist eine Variable X, die n verschiedene diskrete Werte x_i annehmen kann. Die zielführende Grundfrage ist nun, wie

wahrscheinlich die konkrete Realisierung dieser Variable ist, wenn beobachtete Daten vorliegen. Die Wahrscheinlichkeit der Realisierung einer konkreten Ausprägung der Variable wird als p_i bezeichnet. Die Funktionen der Variable werden als $f_k(x)$ bezeichnet, wobei k den Index der Funktionen darstellt, welcher von 1 bis m läuft.

Gesucht sind also die Wahrscheinlichkeit p_i der Realisationen der Variable

$$X = x_i,\ i = (1, \ldots, n), \tag{3.37}$$

auf der Basis der Erwartungswerte der Funktionen

$$f_k(x),\ k = (1, \ldots, m). \tag{3.38}$$

Die Erwartungswerte der Funktionen unter der gesuchten Verteilung definiert Jaynes (2003) wie folgt:

$$\langle f_k(x) \rangle = \sum_{i=1}^{n} p_i f_k(x_i). \tag{3.39}$$

Zudem wird eine Größe F_k eingeführt die auf beobachteten Daten basiert. Konkret könnten dies z. B. beobachtete Mittelwerte sein. Nach Jaynes sollen die beobachteten Werte F_k den Erwartungswerten $\langle f_k \rangle$ unter der gesuchten Verteilung entsprechen, so dass gilt:

$$F_k = \langle f_k(x) \rangle \tag{3.40}$$

$$= \sum_{i=1}^{n} p_i f_k(x_i). \tag{3.41}$$

Die von Jaynes verwendete Notation in Gleichung 3.41 ist etwas ungewöhnlich, von daher sei sie hier näher erläutert. F_k stellt eine beobachtete Statistik, wie z.B. einen Mittelwert über beobachtete Daten dar und $\langle f_k(x) \rangle$ ist der Erwartungswert dieser Statistik unter der gesuchten Verteilung. $\sum_{i=1}^{n} p_i f_k(x_i)$ ist eine auf der Wahrscheinlichkeitstheorie basierende Formulierung dieses Erwartungswertes in Form von einerseits den Wahrscheinlichkeiten p_i der i möglichen Realisationen der Zufallsvariablen und die $f_k(x_i)$ sind Funktionen der manifesten, beobachteten Variablen x_i. Die beobachteten Statistiken F_k stellen Nebenbedingungen (*constraints*) dar, die in der gesuchten Wahrscheinlichkeitsverteilung berücksichtigt werden müssen. Nun existiert theoretisch eine Menge von Verteilungen, die mit den Nebenbedingungen kompatibel sind. Um diese Menge einzuschränken, wird

das Kriterium der maximalen Informationsentropie eingeführt. Es wird also diejenige Verteilung gesucht, die einerseits mit den beobachteten Werten kompatibel ist, andererseits jedoch maximale Informationsentropie besitzt und somit nur diejenige Information im Shannon'schen Sinne berücksichtigt, die mit den Beobachtungen kompatibel ist. Die zusätzliche Nebenbedingung der maximalen Informationsentropie wird benötigt, um die Menge der möglichen Verteilungen einzuschränken, sodass letztlich lediglich nur *eine* Verteilung resultiert, die einerseits maximale Informationsentropie besitzt, andererseits mit den zusätzlichen in Gleichung 3.41 formulierten Nebenbedingungen kompatibel ist. Das Kriterium der maximalen Entropie wird nach Jaynes (2003) folgendermaßen dargestellt:

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \; log(p_i) \rightarrow max. \tag{3.42}$$

Das Erstrebenswerte an der Wahl dieser Nebenbedingung liegt einerseits darin, den Raum der möglichen Wahrscheinlichkeitsverteilungen einzuschränken und andererseits von einem informationstheoretischen Hintergrund her diejenige Verteilung zu finden, die lediglich die empirisch beobachteten Informationen berücksichtigt und gegenüber allen anderen Annahmen indifferent ist. Nach Jaynes ist die resultierende Verteilung objektiv in dem Sinne, dass in diese keine subjektiven Annahmen einfließen, was sich in der Wahl der maximalen Informationsentropie als Kriterium ausdrückt. Andererseits jedoch wird das, was beobachtet wird (die Statistiken F_k) als Optimierungskriterium berücksichtigt. In diesem Sinne stellt Gleichung 3.41 ein Optimierungsproblem dar: die beobachteten Statistiken F_k sollen möglichst gut mit den unter der zu findenden Verteilung erwarteten Werten $\langle f_k \rangle$ übereinstimmen, wobei die Verteilung selbst maximale Informationsentropie besitzen sollte. Zur Lösung dieses Optimierungsproblems der Maximierung der Entropie unter Nebenbedingungen bedient sich Jaynes (2003) der Methode der Lagrange Multiplikatoren. Es werden so viele Lagrange Multiplikatoren eingeführt, wie Nebenbedingungen

vorhanden sind:

$$0 = \delta \left[H - (\lambda_0 - 1) \sum_i p_i - \sum_{j=1}^{m} \lambda_j \sum_i p_i f_j(x_i) \right] \tag{3.43}$$

$$= \sum_i \left[\frac{\partial H}{\partial p_i} - (\lambda_0 - 1) - \sum_{j=1}^{m} \lambda_j f_j(x_i) \right] \delta p_i. \tag{3.44}$$

Als Lösung für p_i erhält Jaynes:

$$p_i = \exp \left\{ -\lambda_0 - \sum_{j=1}^{m} \lambda_j f_j(x_i) \right\}. \tag{3.45}$$

Eine weitere Bedingung, die eingehalten werden muss, um von einer Wahrscheinlichkeitsverteilung zu sprechen, ist diejenige, dass sie Summe der Einzelwahrscheinlichkeiten p_i Eins ergeben muss:

$$1 = \sum_i p_i = \exp\{-\lambda_0\} \sum_i \exp \left\{ -\sum_{j=1}^{m} \lambda_j f_j(x_i) \right\}. \tag{3.46}$$

Von daher wird eine Zustandssumme Z definiert, die gewährleistet, dass die Summe der Einzelereignisse 1 ergibt:

$$Z(\lambda_1, \ldots, \lambda_m) = \sum_{i=1}^{n} \exp \left\{ -\sum_{j=1}^{m} \lambda_j f_j(x_i) \right\}. \tag{3.47}$$

Somit reduziert sich Gleichung 3.46 zu

$$\lambda_0 = \log Z(\lambda_1, \ldots, \lambda_m). \tag{3.48}$$

Setzen wir λ_0 in Gleichung 3.45 ein, so erhalten wir:

$$p_i = \frac{1}{\sum_{i=1}^{n} \exp \left\{ -\sum_{j=1}^{m} \lambda_j f_j(x_i) \right\}} \cdot \exp \left\{ -\sum_{j=1}^{m} \lambda_j f_j(x_i) \right\} \tag{3.49}$$

$$= \frac{\exp \left\{ -\sum_{j=1}^{m} \lambda_j f_j(x_i) \right\}}{\sum_{i=1}^{n} \exp \left\{ -\sum_{j=1}^{m} \lambda_j f_j(x_i) \right\}}.$$

Diese Wahrscheinlichkeitsverteilung nennt Jaynes kanonische Maximum-Entropie-Verteilung. Sie beschreibt eine Klasse von Wahrscheinlichkeitsverteilungen, die von den Parametern λ_j und den Funktionen der Daten $f_j(x_i)$ abhängen. Um diejenige konkrete Verteilung zu finden, die einerseits mit den beobachteten Informationen in F_k kongruent ist,

müssen die Parameter λ_j so gewählt werden, dass die Erwartungswerte der Funktionen $\langle f_j(x_i)\rangle$ den Statistiken F_k der beobachteten Daten (z.B. beobachteten Mittelwerten) entsprechen. Aus Gleichung 3.46 folgt:

$$F_k = \exp\{-\lambda_0\} \sum_i f_k(x_i) \exp\left\{ -\sum_{j=1}^{m} \lambda_j f_j(x_i) \right\}, \tag{3.50}$$

was nach Jaynes (2003, p. 356)

$$F_k = -\frac{\partial \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_k} \tag{3.51}$$

entspricht. Die Erwartungswerte unter dem Modell sind demzufolge:

$$\langle f_k(x_i)\rangle = -\frac{\partial \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_k}. \tag{3.52}$$

Um nun die konkrete Verteilung zu ermitteln, welche einerseits mit den beobachteten Statistiken über F_k kongruent ist, und andererseits maximale Informationsentropie besitzt, muss das durch Gleichung 3.51 definierte System in Abhängigkeit der Lagrange Multiplikatoren gelöst werden, so dass die Erwartungswerte unter der Verteilung den beobachteten Nebenbedingungen entsprechen. Die Lagrange Multiplikatoren entsprechen im Prinzip Modellparametern. Es resultiert also *eine* konkrete Verteilung, die einerseits mit den beobachteten Statistiken kongruent ist, andererseits maximale Informationsentropie besitzt.

Jaynes (2003, p. 358-361) entwickelt weitere, interessante formale Eigenschaften der kanonischen Maximum-Entropie-Verteilung. So entspricht nach Janynes die maximal erreichbare Entropie

$$H_{max} = S(F_1, \ldots, F_m) = \log \mathrm{Z}(\lambda_1, \ldots, \lambda_m) + \sum_{k=1}^{m} \lambda_k F_k. \tag{3.53}$$

Ferner lassen sich die Parameter λ_k explizit berechnen, sofern folgender Ausdruck evaluierbar ist:

$$\lambda_k = \frac{\partial S(F_1, \ldots, F_m)}{\partial F_k}. \tag{3.54}$$

Die Kovarianz zweier Funktionen f_k und f_j unter der Verteilung ergibt sich aus

$$\langle f_j f_k\rangle - \langle f_j\rangle \langle f_k\rangle = -\frac{\partial^2 \log \mathrm{Z}}{\partial \lambda_j \partial \lambda_k}. \tag{3.55}$$

Dementsprechend ist die Varianz einer Funktion f_k

$$\left\langle {f_k}^2 \right\rangle - \langle f_k \rangle^2 = -\frac{\partial^2 \log \mathrm{Z}}{\partial \lambda_k^2}. \tag{3.56}$$

Die hier nach Jaynes (2003, Kap.10) dargestellten Sachverhalte sind relativ komplex, daher werden diese an zwei einfachen Beispielen erläutert. Nehmen wir an, wir werfen eine Münze zehn Mal ($N = 10$) und würden gerne auf Basis der Beobachtung induktiv ermitteln, wie hoch die Wahrscheinlichkeit ist, dass die Münze Kopf zeigt. Die möglichen Realisationen der Zufallsvariable X sind $x_1 = 0$ (Zahl) und $x_2 = 1$ (Kopf). Ferner beobachten wir in dem Experiment einen Mittelwert von $F = 0.54$. Intuitiv würden wir 0.54 als die Wahrscheinlichkeit auf Basis unserer Beobachtung ansehen, dass die Münze Kopf zeigt. Das Problem der Ermittlung der Wahrscheinlichkeit kann allerdings auch formal mit der Maximum-Entropie-Methode angegangen werden, indem das Problem in einer Wahrscheinlichkeitsverteilung p_i enkodiert und deren Parameter λ ermittelt wird. Diese Vorgehen bietet zusätzlich den Vorteil, dass die Unsicherheit bezüglich der Wahrscheinlichkeiten der Realisation der Zufallsvariable über die Standardfehler, bzw. die Posterior-Verteilungen der Parameter bewertet werden kann. Das Problem kann wie folgt in Gleichung 3.49 übersetzt werden. Die Anzahl der möglichen Ereignisse ist $n = 2$, für die Anzahl der Funktionen gilt $m = 1$, da nur eine manifeste Statistik $F = 0.54$ bekannt ist. Für die Variable X gilt $X \in \{0, 1\}$. Zudem nehmen wir an, dass die Funktion der Daten $f(x_i)$ mit x_i identisch ist. Mit diesen Spezifikationen folgt aus Gleichung 3.49 direkt:

$$p_i = \frac{\exp\{-\lambda \cdot x_i\}}{1 + \exp\{-\lambda\}}. \tag{3.57}$$

Es ist beachtenswert, dass diese Gleichung mit $x_i = 1$ die Link-Inverse eines generalisierten, gemischten Modells mit binomialer Fehlerstruktur darstellt. Die Schreibweise p_i ist vielleicht etwas ungewohnt und wird von daher zur Erleichterung des Verständnisses angepasst:

$$p(X = x_i) = \frac{\exp\{-\lambda \cdot x_i\}}{1 + \exp\{-\lambda\}}. \tag{3.58}$$

Um nun den konkreten Parameter λ zu ermitteln, kann Gleichung 3.52 angewendet werden, um den Erwartungswert $\langle x_i \rangle$ unter dem Modell zu ermitteln. Die Zustandssumme

Z ist hierbei der Nenner des resultierenden Binomialmodells.

$$\langle x_i \rangle = -\frac{\partial \log(1 + \exp\{-\lambda\})}{\partial \lambda} \tag{3.59}$$

$$= \frac{\exp\{-\lambda\}}{1 + \exp\{-\lambda\}}. \tag{3.60}$$

Gleichung 3.60 bringt zum Ausdruck, dass der Erwartungswert der Variable X in einem Zusammenhang mit dem Parameter λ des Modells steht. Je größer λ, desto geringer die Wahrscheinlichkeit des Auftretens der Realisation $x_2 = 1$. Ist $\lambda = 0$, so sind beide Realisationen gleich wahrscheinlich. Um nun den Parameter λ zu ermitteln, kommt Gleichung 3.51 zum Einsatz:

$$0.54 = \frac{\exp\{-\lambda\}}{1 + \exp\{-\lambda\}}. \tag{3.61}$$

Die linke Seite der Gleichung beinhaltet die beobachtete Statistik $F = 0.54$, die rechte Seite der Gleichung ist der Erwartungswert dieser Statistik unter dem Modell. Um λ zu ermitteln, muss die Gleichung nach λ umgestellt werden:

$$\lambda = -\log\left\{\frac{0.54}{1 - 0.54}\right\} \tag{3.62}$$

$$= -0.16. \tag{3.63}$$

Wird dar Parameter in die Modellgleichung eingesetzt, so erhalten wir:

$$p(X = x_i) = \frac{\exp(0.16 \cdot x_i)}{1 + \exp(0.16)}. \tag{3.64}$$

Demzufolge ist die Wahrscheinlichkeit von $x_i = 1$:

$$p(X = 1) = \frac{\exp(0.16)}{1 + \exp(0.16)} = 0.54. \tag{3.65}$$

Die Gegenwahrscheinlichkeit ist:

$$p(X = 0) = \frac{1}{1 + \exp(0.16)} = 0.46. \tag{3.66}$$

Diese konkrete Verteilung besitzt einerseits maximale Informationsentropie und andererseits ist sie mit der verfügbaren Information in Form der beobachteten Statistik F_k kongruent. Die Wahl einer anderen Verteilung als derjenigen mit maximaler Informationsentropie unter Nebenbedingungen würde bedeuten, dass in die Wahl der Verteilung

(subjektive) Informationen eingegangen sind, die durch die Datenlage nicht gerechtfertigt sind. Gleichung 3.61 stellt im Prinzip den Gradient des Modells zur Schätzung der Parameter mit der Maximum-Likelihood-Methode dar. Das hier dargestellt Beispiel ist sehr einfach. In realen Anwendungen, die sehr viel mehr Funktionen der Daten beinhalten können, wird das Problem der Bestimmung der Parameter mit Gradientenverfahren, dem EM-Algorithmus oder der MCMC-Methode angegangen.

Über die Ermittelung der Standardfehler oder der Posterior-Verteilungen des Modells ließe sich Inferenzstatistik betreiben. Beispielsweise kann Gleichung 3.56 verwendet werden, um die Varianz der Variable X unter dem Modell zu ermitteln:

$$\begin{aligned} \left\langle x_i^2 \right\rangle - \left\langle x_i \right\rangle^2 &= \frac{1}{1+\exp(0.16)} \cdot \frac{\exp(0.16)}{1+\exp(0.16)} && (3.67) \\ &= p(X=0)p(X=1) && (3.68) \\ &= 0.248. && (3.69) \end{aligned}$$

Gleichung 3.56 definiert im Prinzip die Diagonale der Hesse-Matrix, von daher lässt sich der Standardfehler von λ_k wie folgt berechnen:

$$\begin{aligned} \mathrm{se}(\hat{\lambda}) &= \frac{1}{\sqrt{p(X=0) \cdot p(X=1) \cdot N}} && (3.70) \\ &= 0.6634, && (3.71) \end{aligned}$$

wobei N der Anzahl der Beobachtungen entspricht. Es ist beachtenswert, dass der Standardfehler des Schätzers mit steigender Anzahl von Beobachtungen N sinkt.

Die Maximum-Entropie-Methode funktioniert nicht nur bei binären, bzw. dichotomen Ereignissen, auch mehrkategorielle Ereignisse lassen sich durch die Anwendung der Methode probabilistisch modellieren. Ein naheliegendes Beispiel ist der Würfelwurf. Im Folgenden sei nur kurz die Definition eines Maximum-Entropie-Modells dargestellt, wenn der beobachtete Mittelwert von N Würfen mit einem Würfel $F = 3.54$ beträgt. Die Modelldefinition erfolgt über Gleichung 3.49. Es existieren $n = 6$ mögliche Ereignisse und zudem ist lediglich eine Statistik F_k bekannt, von daher ist $m = 1$. Ferner gilt $f(x_i) = x_i$ und $x_i \in \{1, ..., 6\}$. Es folgt:

$$p(X = x_i) = \frac{\exp\{-\lambda \cdot x_i\}}{\sum_{i=1}^{6} \exp\{-\lambda \cdot x_i\}}. \qquad (3.72)$$

Gleichung 3.72 beschreibt ein Maximum-Entropie-Modell, das für m=6 mögliche Ereignisse definiert ist. Es wird davon ausgegangen, dass die einzige verfügbare Information hinsichtlich des zugrundeliegenden Prozesses der beobachtete Mittelwert $F = 3.54$ ist, von daher besitzt das Modell nur lediglich einen Parameter. Ein wesentlicher Unterschied zu dem Modell für dichotome Ereignisse liegt in der Zustandssumme. Diese läuft nun über alle $m = 6$ möglichen Ereignisse, um zu gewährleisten, dass die Summe der Wahrscheinlichkeiten der Einzelereignisse 1 ergibt. Zur Ermittelung der Wahrscheinlichkeiten der Einzelereignisse muss der Parameter wiederum so gewählt werden, dass der erwartete Wert $\langle x_i \rangle$ unter dem Modell der beobachteten Statistik $F = 3.54$ entspricht. Aus Gleichung 3.51 resultiert:

$$3.54 = \frac{\sum_{i=1}^{6} x_i \cdot \exp\{-\lambda \cdot x_i\}}{\sum_{i=1}^{6} \exp\{-\lambda \cdot x_i\}} \tag{3.73}$$

$$= \sum_{i=1}^{6} x_i \cdot p(X = x_i). \tag{3.74}$$

Als Lösung der Gleichung - z.B. mit einem Gradientenverfahren - in Abhängigkeit von λ resultiert:

$$\lambda = -0.014. \tag{3.75}$$

Bei einem Parameter von $\lambda = -0.014$ entspricht der erwartete Wert unter dem Modell der beobachteten Statistik F. Die Wahrscheinlichkeiten $\boldsymbol{p}$ werden berechnet, indem der Parameter in Gleichung 3.72 eingesetzt wird:

$$\boldsymbol{p} = [0.161, 0.163, 0.166, 0.168, 0.170.0.173]\,. \tag{3.76}$$

Dies bedeutet, dass nach dem Modell auf Basis der beobachteten, mittleren Augenzahl von 3.54 das Auftreten von höheren Zahlen leicht favorisiert wird. So ist z.B. nach dem Modell die Wahrscheinlichkeit des Auftretens des Ereignisses $P(X = 1) = 0.161$, wohingegen die Wahrscheinlichkeit des Auftretens des Ereignisses $P(X = 6) = 0.173$ beträgt. Auch bei diesem Modell kann über die Berechnung des Standardfehlers oder der Posterior-Verteilung von λ eine Aussage über die Genauigkeit der Parameterschätzung getroffen und Inferenzstatistik betrieben werden.

Abschließend sei nur kurz dargestellt, dass der Maximum-Entropie-Formalismus sich nicht nur auf eine Variable X anwenden lässt, sondern es auch möglich ist, die Zustände mehrerer Variablen simultan zu modellieren. Es wird also die gemeinsame Dichte (*joint density*) mehrerer Variablen unter Berücksichtigung der verfügbaren Informationen in der Form von Beobachtungen modelliert. In diesem Fall kann die kanonische Maximum-Entropie-Verteilung wie folgt dargestellt werden:

$$p(X = x_1, \ldots, x_n) = \frac{\exp\left\{-\sum_{j=1}^{m} \lambda_j f_j(x_j)\right\}}{\sum_{x \in X} \exp\left\{-\sum_{j=1}^{m} \lambda_j f_j(x_j)\right\}}. \tag{3.77}$$

In dieser Darstellungsform ist das Modell auch als *undirected graphical model* (Koller & Friedman, 2009) bekannt und ein Reihe von Modellen, wie z.B. das Ising-Modell der statistischen Mechanik oder die Boltzmann-Verteilung ergeben sich als Spezialfall. Es wird nicht lediglich nur die Verteilung einer Variable abgebildet, sondern es werden die Zustände eines Systems modelliert, dass aus n Variablen besteht. Die Variablen selbst werden in diesem Kontext als Knoten eines Netzwerks aufgefasst. Die Abhängigkeiten zwischen den Variablen und Informationen über die Zustände der einzelnen Variablen können in den Funktionen $f_j(x_j)$ enkodiert werden. In graphentheoretischer Terminologie enkodieren diese Funktionen Zustände von Cliquen. Die Zustandssumme läuft in dieser Form über alle möglichen Zustände des Systems, welches aus mehren Variablen bestehen kann. Nähere Ausführungen diesbezüglich würden den Rahmen der Arbeit sprengen. In (Koller & Friedman, 2009) finden sich detaillierte Ausführungen. Derzeit wird der von Jaynes dargestellte Formalismus unter anderem in der Künstlichen Intelligenz eingesetzt. Die von Jaynes beschriebene Modellklasse ist in diesen Bereichen als *undirected graphical model*, *markov network* und *log-linear model* bekannt (siehe Koller, 2009, Kap. 4). Die hier nach Jaynes dargestellten mathematischen Sachverhalte finden sich auch in (Koller & Friedman, 2009). Der mathematische Beweis, dass es sich bei Gleichung 3.49 um eine Verteilung mit maximaler Informationsentropie bei Berücksichtigung von Nebenbedingungen in der Form von Erwartungswerten handelt, findet sich in Sektion 20.3.4. Hier werden auch die Beziehungen zwischen der Maximum-Entropie-Methode und der Maximum-Likelihood-Methode (sog. konvexe Dualität) näher beleuchtet. Die von Jaynes

gelieferten Ergebnisse hinsichtlich der Erwartungswerte und der Varianz der Funktionen in Gleichung 3.49 werden ebenfalls in (Koller & Friedman, 2009), Sektion 20.2.3. (*Properties of the Likelihood-Function*) mathematisch bewiesen. Kapitel 8 in (Koller & Friedman, 2009) geht ausführlich auf die Exponentialfamilie und deren Eigenschaften und Bedeutung ein.

Eine interessante Beobachtung ist, dass der von Jaynes vorgeschlagene Formalismus mit Ausnahme der Notation eine hohe Ähnlichkeit mit demjenigen besitzt, welcher auch von Rasch (1961) dargelegt wurde. Ein wichtiger Unterschied besteht jedoch darin, dass Jaynes eine Herleitung für die Gleichung 3.49 auf Basis von informations- und wahrscheinlichkeitstheoretischen Überlegungen gibt.

3.2.2. Rasch-Modelle als Maximum-Entropie-Modelle

Vergleichen wir die kanonische Maximum-Entropie-Verteilung (Gleichung 3.49) mit der generellen Definition von Rasch-Modellen (Rasch, 1961) (Gleichung 3.1), so zeigen sich frappierende formale Ähnlichkeiten. Zur Verdeutlichung werden Raschs ursprüngliche Definition von Rasch-Modellen und die kanonische Maximum-Entropie-Verteilung nochmals gegenübergestellt:

$$P\left\{x|\theta_v,\sigma_i\right\} = \frac{1}{\gamma(\theta_v,\sigma_i)}\ \exp[\phi(x)\theta_v + \psi(x)\sigma_i + \chi(x)\theta_v\sigma_i + \rho(x)]. \tag{3.78}$$

und

$$p_i = \frac{\exp\left\{-\sum_{j=1}^{m}\lambda_j f_j(x_i)\right\}}{\sum_{i=1}^{n}\exp\left\{-\sum_{j=1}^{m}\lambda_j f_j(x_i)\right\}}. \tag{3.79}$$

$\gamma(\theta_v,\sigma_i)$ in Gleichung 3.78 entspricht der Zustandssumme $Z(\lambda_1,\ldots,\lambda_m)$ in Gleichung 3.79. Die Funktionen $\phi(x)$, $\psi(x)$ und $\chi(x)$ finden ihre Entsprechung in den Funktionen $f_j(x_i)$ in Gleichung 3.79. Die Parameter σ_i und θ_v in Gleichung 3.78 werden in Gleichung 3.79 durch die Lagrange Multiplikatoren λ_j repräsentiert. Es scheint also die Vermutung nahe zu liegen, dass Rasch-Modelle ebenfalls als Maximum-Entropie-Modelle betrachtet werden können, d.h. Rasch-Modelle beschreiben ebenfalls Wahrscheinlichkeitsverteilungen, die einerseits mit beobachteten Daten kongruent sind und andererseits maximale Informationsentropie besitzen.

Zur Anwendung in der Psychometrie ist es zweckmäßig, die Notation zur Darstellung kanonischen Maximum-Entropie-Verteilung (Gleichung 3.79) in ein in der psychologischen Literatur gängigeres Format zu überführen.

$$p(X = x) = \frac{\exp\left\{\sum_{j=1}^{k} \lambda_j f_j(x)\right\}}{\sum_{l=1}^{m} \exp\left\{\sum_{j=1}^{k} \lambda_j f_j(x)\right\}}, \tag{3.80}$$

mit $x \in \{1, \ldots, m\}$. Das Ergebnis x ist nun explizit als Realisation einer Zufallsvariable X deklariert. Das negative Vorzeichen der Summen im Exponenten von Gleichung 3.79 ist nicht unbedingt nötig und kann weggelassen werden. Die Konsequenz ist lediglich, dass die Vorzeichen der Parameter sich ändern, was bei der Interpretation der Parameter berücksichtigt werden muss. Von daher wird in Gleichung 3.80 auf das negative Vorzeichen verzichtet, um die Ähnlichkeiten mit dem Ansatz von Rasch noch deutlicher zu machen. Zudem startet die Summe im Nenner nicht bei 0, sondern bei 1, wobei m der Anzahl der Kategorien eines Items entspricht. Ferner wurde der Index der möglichen Ereignisse als l und der Index der Funktionen der Daten als j deklariert. Mittels der Gleichung 3.80 ist es möglich, die Wahrscheinlichkeit der Ereignisse $X = 1, X = 2, \ldots, X = m$ auf Basis der Scoring-Funktionen $f_j(x)$ und den Parametern λ_j zu modellieren. Im Kontext von psychometrischen Modellen enkodieren die möglichen Ereignisse $x \in \{1, \ldots, m\}$ die Wahl der Kategorie $1, 2, \ldots, m$. Der Ausdruck $p(X = x)$ in Gleichung 3.80 entspricht in Gleichung 3.79 den Wahrscheinlichkeiten p_i.

An dieser Stelle ist es angebracht, eine Beziehung zu der für die Parameterschätzung wichtigen Gleichung 3.41 herzustellen, die hier nochmals aufgeführt wird:

$$F_k = \langle f_k(x) \rangle = \sum_{i=1}^{n} p_i f_k(x_i). \tag{3.81}$$

Die Wahrscheinlichkeiten p_i in Gleichung 3.81 entsprechen den durch das Modell in Gleichung 3.79 beschriebenen Wahrscheinlichkeiten. Die Funktionen $f_k(x_i)$ in Gleichung 3.81 stehen im psychometrischen Kontext für die Ausprägungen der Scoring-Funktionen in Abhängigkeit der Variable x_i, $\langle f_k(x) \rangle$ ist der Erwartungswert der Funktion k unter dem Modell und F_k ist eine beobachtete (suffiziente) Statistik zur Schätzung des korrespondie-

renden Modellparameters. Insgesamt bringt Gleichung 3.81 das Desideratum zum Ausdruck, dass die unter Modell erwarteten Ausprägungen der Funktionen $\langle f_k(x) \rangle$ möglichst gut mit den entsprechenden beobachteten Statistiken F_k korrespondieren und im günstigsten Fall mit diesen übereinstimmen sollten.

Das dichotome Rasch-Modell als Maximum-Entropie-Modell

Im folgenden wird gezeigt, wie das dichotome Rasch-Modell aus der kanonischen Maximum-Entropie-Verteilung folgt. Zur Herleitung des Rasch-Modells muss die Anzahl der möglichen Ereignisse m (Item gelöst oder nicht gelöst) und die Anzahl der Funktionen der Daten k definiert werden. Da das dichotome Rasch-Modell bei dichotomem Antwortformat verwendet wird, gilt $m = 2$. Da die Spaltensummen und Zeilensummen einer Datenmatrix potentiell als suffiziente Statistiken fungieren, ist die Anzahl der Funktionen $k = 2$. Ein weiterer Grund für diese Wahl ist, dass in dem Modell erstens die Personen-Fähigkeiten und zweitens die Itemschwierigkeiten erfasst werden sollen. Als Kodierung des Antwortformats wird $x \in \{0, 1\}$ gewählt. Somit folgt aus der kanonischen Maximum-Entropie-Verteilung:

$$p(X = x) = \frac{\exp(\lambda_1 x + \lambda_2 x)}{\exp(\lambda_1 \cdot 0 + \lambda_2 \cdot 0) + \exp(\lambda_1 \cdot 1 + \lambda_2 \cdot 1)} \tag{3.82}$$

$$= \frac{\exp((\lambda_1 + \lambda_2)x)}{1 + \exp(\lambda_1 + \lambda_2)}. \tag{3.83}$$

Diese Gleichung sieht dem dichotomen Rasch-Modell schon recht ähnlich. Allerdings ergeben sich zwei wesentliche Unterschiede. Erstens ist das Modell nicht für mehrere Items und Personen definiert, da lediglich die zwei Parameter λ_1 und λ_2 existieren. Zudem ist das Vorzeichen für λ_2, dem potentiellen Item-Parameter, positiv. Rasch wählte ein negatives Vorzeichen, damit hohe Item-Parameter eine hohe Schwierigkeit des Items signalisieren und nicht eine hohe Leichtigkeit, daher wird auch hier das Vorzeichen für den Item-Parameter umgekehrt. Zudem wird das Modell über eine gesamte Datenmatrix ausgeweitet. Hierbei werden design-spezifische Überlegungen berücksichtigt: die manifeste Antwort x_{vi} einer Person v auf ein Item i wird auf die Person v und das Item i zurückgeführt. Zudem wird die Bezeichnung der Parameter an ein übliches Format an-

gepasst. θ_v ist ein Personen-Parameter, der deren Fähigkeit repräsentiert und β_i ist die Schwierigkeit des Items:

$$p(X_{vi} = x_{vi}) = \frac{\exp((\theta_v - \beta_i)x_{vi})}{1 + \exp(\theta_v - \beta_i)}. \tag{3.84}$$

Der wesentliche Punkt bei dieser Operation liegt darin, dass anstatt lediglich nur ein Ereignis x auf der Basis einer Beobachtungsreihe zu modellieren, der Modellierungsraum auf eine Datenmatrix über die Personen v und Items i eines Tests ausgeweitet wird. Zudem wurde die Notation in eine für die Darstellung von Rasch-Modellen gängigere Form überführt. Es ist zu beachten, dass kein Maximum-Entropie-Modell für eine Gesamtdatenmatrix formuliert wird, sondern lediglich nur für eine einzelne Antwort x_{vi}, wobei die Parameter zur Beschreibung der Verteilung vom zugrundeliegenden Design abhängen. Oder anders ausgedrückt: der Parameter λ_{vi} - ein linearer Prädiktor - wird in Komponenten zu Lasten eines Items i und einer Person v linear zerlegt:

$$p(X_{vi} = x_{vi}) = \frac{\exp(\lambda_{vi} \cdot x_{vi})}{1 + \exp(\lambda_{vi})}, \tag{3.85}$$

mit

$$\lambda_{vi} = \theta_v - \beta_i. \tag{3.86}$$

Die lineare Zerlegung von λ_{vi} entspricht dem generellen Vorgehen im Falle generalisierter, lineare Modelle, wobei die logistische Funktion von λ_{vi} einer Logit-Link-Inversen entspricht und eine Maximum-Entropie-Verteilung für binäre, bzw. dichotome Ereignisse darstellt und der Exponentialfamilie angehört.

Um die Ähnlichkeiten zwischen dem Maximum-Entropie-Formalismus und Raschs Vorgehensweise näher zu untersuchen, werden im folgenden die Schätzgleichungen und die Informationsfunktion des Rasch-Modells unter Anwendung des Maximum-Entropie-Formalismus hergeleitet. Nach Gleichung 3.52 gilt:

$$\langle x_{vi} \rangle = \frac{\partial \log(1 + \exp(\theta_v - \beta_i))}{\partial \theta_v} \tag{3.87}$$

$$= \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)} \tag{3.88}$$

und

$$\langle x_{vi} \rangle = -\frac{\partial \log(1 + \exp(\theta_v - \beta_i))}{\partial \beta_i} \tag{3.89}$$

$$= \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}. \tag{3.90}$$

Die Gleichungen entsprechen den Erwartungswerten der Variable x_{vi} unter dem Modell. Das Kriterium der Parameterschätzung nach der Maximum-Entropie-Methode besteht darin, dass die Parameter so geschätzt werden müssen, dass die erwarteten Werte $\langle x_{vi} \rangle$ unter der Verteilung den manifesten Werten entsprechen. Werden die Erwartungswerte unter dem Modell und die manifesten Werte x_{vi} zeilen- bzw. spaltenweise nach Maßgabe der jeweiligen Parameter aufsummiert, so ergeben sich die bekannten, z.B. von Molenaar (1995) beschriebenen Joint-Maximum-Likelihood-Schätzgleichungen:

$$\sum_{v=1}^{N} x_{vi} = \sum_{v}^{N} \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)} \tag{3.91}$$

und

$$\sum_{i=1}^{k} x_{vi} = \sum_{i}^{k} \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}. \tag{3.92}$$

Um die Summierung zu rechtfertigen, muss allerdings auf die Annahme der stochastischen Unabhängigkeit der Antworten x_{vi} zurückgegriffen werden. Die Spalten- und Zeilensummen sind suffizient zur Schätzung der Parameter, d.h. es werden keine Informationen über die Reihung der manifesten Daten x_{vi} benötigt.

Die Varianz der manifesten Antworten x_{vi} und damit die Informationsfunktion, die zur Bestimmung der Genauigkeit der Erfassung eines Parameters benötigt wird, folgt im Rahmen des Maximum-Entropie-Formalismus aus Gleichung 3.56. Im Falle des dichotomen Rasch-Models ist

$$\log \mathrm{Z} = \log \left\{ 1 + \exp(\theta_v - \beta_i) \right\}. \tag{3.93}$$

Dementsprechend ist

$$\frac{\partial^2 \log \mathrm{Z}}{\partial \theta_v^2} = \frac{\exp(\theta_v - \beta_i)}{(1 + \exp(\theta_v - \beta_i))} \cdot \frac{1}{(1 + \exp(\theta_v - \beta_i))} \tag{3.94}$$

$$= p(X_{vi} = 1) \cdot p(X_{vi} = 0) \tag{3.95}$$

die Varianz von x_{vi} unter dem Modell. Wird diese Größe unter der Annahme der stochastischen Unabhängigkeit der Antworten über die Items eines Tests aufsummiert, so ergibt sich die bekannte Informationsfunktion eines Tests.

Es ist zu Verzeichnen, dass die Anwendung der Maximum-Entropie-Methode auf Probleme der psychologischen Testung unter Annahme eines dichotomen Itemformates und der Berücksichtigung von Itemschwierigkeiten und Personen-Fähigkeiten relativ einfach zu Ergebnissen führt, die mit denjenigen formal äquivalent sind, die auch Rasch lieferte. Zudem ist der Ansatz im Falle des dichotomen Rasch-Modells mit demjenigen der generalisierten gemischten Modelle kompatibel. Vom Prinzip her entspricht die Verteilung der binären Antwort x_{vi} einer Maximum-Entropie-Verteilung, deren Parameter (der lineare Prädiktor in der Terminologie der gemischten, linearen Modelle) in design-spezifische Komponenten (Personen und Items) zerlegt wird.

Das Partial-Credit-Modell als Maximum-Entropie-Modell

Was hier für das dichotome Rasch-Modell gezeigt wurde gilt, ebenso für das Partial-Credit-Modell von Masters (1982), d.h. es ist möglich das Partial-Credit-Modell durch die Anwendung des Maximum-Entropie-Formalismus herzuleiten. Zielführend ist die Definition der Funktionen der Daten $f_j(x)$ in Gleichung 3.80. Zunächst wird eine Funktion $f_1(x)$ benötigt, um die Merkmalsausprägung einer Person zu erfassen. Diese ist im Einklang mit dem Partial-Credit-Modell eine lineare Scoring-Funktion $f_1(x_1) = x$. Zudem wird für jede Item-Kategorien-Kombination eine *weitere* Funktion $f_j(x_j)$ benötigt. Diese Funktionen werden $f_j(x_j) = 1$ gesetzt, sofern eine item-spezifische Kategorie gewählt wurde, ansonsten gilt $f_j(x_j) = 0$. Bei den Kategorien-Parametern handelt es sich also um item- und kategorienspezifische Konstanten, ganz im Einklang mit der Ur-Formulierung von Rasch.

Aus Gleichung 3.80 folgt mit diesen Spezifikationen nach Anpassung der Notation direkt:

$$p(X_{vi} = x_{vi}) = \frac{\exp\{\theta_v x_{vi} + \beta_{ix}\}}{\sum_{l=1}^{m} \exp\{\theta_v l + \beta_{il}\}}, \tag{3.96}$$

mit $x_{vi} \in \{1, \ldots, l, \ldots, m\}$. Die Parameter θ_v und β_{ix} entsprechen den λ_j in der Maximum-

Entropie-Verteilung. x_{vi} entspricht der Funktion $f_1(x) = x$. Die den Kategorien-Parametern β_{ix} zugeordneten Funktionen $f_j(x)$ werden alle 1 gesetzt, sofern die entsprechende Kategorie gewählt wurde, da es sich um item- und kategorienspezifische Konstanten handelt, ansonsten ist deren Wert 0. Der Nenner des Modells entspricht in der kanonischen Maximum-Entropie-Verteilung der Zustandssumme Z, wobei l, der Index der Kategorien von 1 bis m läuft.

Es ist an dieser Stelle vielleicht angebracht, die Beziehungen zur Gleichung 3.80 stärker herauszuarbeiten. Hilfreich ist hierbei die Vergegenwärtigung, dass die Zustandssumme Z alle möglichen Zustände eines Systems enkodiert. Im Falle des PCM sind die Zustände die möglichen Ausprägungen der Scoring-Funktion. Wird die Zustandssumme für beispielsweise ein dreikategorielles Item ausgeschrieben, so erhalten wir:

$$\begin{aligned} Z &= \exp\{\theta_v \cdot 1 + 1 \cdot \beta_{i1} + 0 \cdot \beta_{i2} + 0 \cdot \beta_{i3}\} + && (3.97)\\ &\quad \exp\{\theta_v \cdot 2 + 0 \cdot \beta_{i1} + 1 \cdot \beta_{i2} + 0 \cdot \beta_{i3}\} + && (3.98)\\ &\quad \exp\{\theta_v \cdot 3 + 0 \cdot \beta_{i1} + 0 \cdot \beta_{i2} + 1 \cdot \beta_{i3}\}. && (3.99) \end{aligned}$$

In dieser Darstellung werden die Funktionen $f_j(x)$ des PCM deutlich. f_1 entspricht der Scoring-Funktion für den Personen-Parameter θ_v, welche die Werte 1, 2 oder 3 annehmen kann. Die Funktion f_2, welche dem Parameter β_{i1} zugeordnet ist, wird 1 gesetzt, sofern Kategorie 1 gewählt wurde und ansonsten ist deren Ausprägung 0. Analoges gilt für die Funktionen f_3 und f_4, welche den Parametern β_{i2} und β_{i3} zugeordnet sind. Dieses Vorgehen ähnelt der Dummy-Codierung im ALM. In der oben verwendeten Parametrisierung handelt es sich bei den Kategorien-Parametern um Kategorien-Leichtigkeiten, zudem ist für die Identifikation des Modell eine Summen-Normierung über die Kategorien-Parameter nötig. Die suffizienten Statistiken zur Parameterschätzung sind die manifesten Statistiken, die mit den Erwartungswerten der Funktionen korrespondieren. Im Falle des PCM wären dies die Summe der Werte der Scoring-Funktion für eine Person v für den Parameter θ_v und die Summe der Wahl einer entsprechenden Kategorie x auf einem Item i für einen Kategorien-Parameter β_{ix}. Hohe Parameter β_{ix} gehen mit einer häufigen Wahl der entsprechenden Wahl der Kategorie in beobachteten Daten F (der verfügbaren Infor-

mation) einher. Die in dieser Arbeit gewählte Parametrisierung mit Summen-Normierung ist etwas ungewöhnlich, von daher seien die Zusammenhänge zwischen dem PCM und der kanonischen Verteilung dargestellt, wenn eine übliche Parametrisierung gewählt wird, bei der die Scoring-Funktion für die Ratings nicht bei 1, sondern bei 0 startet. Die entsprechende Zustandssumme ist:

$$\begin{aligned} Z &= \exp\{\theta_v \cdot 0 + 0 \cdot \beta_{i1} - 0 \cdot \beta_{i2} - 0 \cdot \beta_{i3}\} + && (3.100)\\ &\quad \exp\{\theta_v \cdot 1 + 0 \cdot \beta_{i1} - 1 \cdot \beta_{i2} + 0 \cdot \beta_{i3}\} + && (3.101)\\ &\quad \exp\{\theta_v \cdot 2 + 0 \cdot \beta_{i1} - 1 \cdot \beta_{i2} - 1 \cdot \beta_{i3}\}. && (3.102) \end{aligned}$$

Die wesentlichen Unterschied zu der zuvor dargestellten Zustandssumme besteht darin, dass nun die Scoring-Funktion f_1 bei 0 startet, der Parameter β_{i1} ist dadurch, dass alle Werte der Funktion f_2 gleich Null sind, auf Null gesetzt. Zudem sind die Werte der Scoring-Funktionen anders definiert. Die β_{i2} zugeordnete Funktion wird -1 gesetzt, sofern Kategorie 2 ($x_{vi} = 1$) gewählt wurde, bei der Wahl von Kategorie 3 werden β_{i2} und β_{i3} -1 gesetzt, ansonsten sind die Ausprägungen dieser Funktionen gleich 0. In dieser Form stellen die Parameter β_{ix} die Schnittpunkte der Kategorien-Response-Funktionen des Modells dar. Im weiteren Verlauf der Arbeit wird die erstere Variante unter Anwendung der Summen-Normierung verwendet, es wäre aber ebenso möglich, die zweite Variante zu verwenden. Im übrigen lassen sich die Parameter der beiden Varianten ineinander überführen, worauf in einem späteren Abschnitt eingegangen wird.

Nach dem Maximum-Entropie-Formalismus ergibt sich der Erwartungswert der Funktionen der Daten unter dem Modell durch das einmalige differenzieren der logarithmierten Zustandssumme Z nach dem entsprechenden Parameter.

Die Zustandssumme Z des Partial-Credit-Modells ist:

$$Z(\theta_v, \beta_{il}) = \sum_{l=1}^{m} \exp\{\theta_v \cdot l + \beta_{il}\}. \tag{3.103}$$

Differenzieren nach θ_v ergibt:

$$\begin{aligned} \frac{\partial log Z(\theta_v, \beta_{il})}{\partial \theta_v} &= \frac{\sum_{l'=1}^{m} l' \cdot \exp\{\theta_v \cdot l' + \beta_{il'}\}}{\sum_{l=1}^{m} \exp\{\theta_v \cdot l + \beta_{il}\}} && (3.104)\\ &= \langle x_{vi} \rangle. && (3.105) \end{aligned}$$

$\langle x_{vi} \rangle$ ist der Erwartungswert der Antwort x_{vi} unter dem Modell. Dieses Ergebnis bedeutet implizit, dass die Summen der Scorings einer Person eine suffiziente Statistik zur Schätzung des Personen-Parameters θ_v darstellt. Die Notation l' im Zähler wurde gewählt, um den Laufindex im Nenner vom Laufindex im Zähler notationell zu trennen. Da das Kriterium der Parameterschätzung darin besteht, dass die Erwartungswerte unter dem Modell möglichst gut mit den korrespondierenden beobachteten Statistiken (F_k) übereinstimmen sollten, folgt, dass die Summe der erwarteten Scorings einer Person unter dem Modell möglichst gut mit den beobachteten Summen übereinstimmen sollten. Ferner legt dieses Ergebnis nahe, dass sich θ_v unabhängig von den Kategorien-Parametern schätzen lassen sollte, da das resultierende Modell zur Exponentialfamilie gehört.

Das zweimalige Differenzieren der Zustandssumme Z nach dem Parameter θ_v ergibt die Varianz der manifesten Antwort x_{vi} und damit indirekt die Informationsfunktion:

$$\frac{\partial^2 log Z(\theta_v, \beta_{il})}{\partial \theta_v^2} = \frac{\sum_{l'=1}^{m} l'^2 \cdot \exp\{l' \cdot \theta_v + \beta_{il'}\}}{\sum_{l=1}^{m} \exp\{l \cdot \theta_v - \beta_{il}\}} - \tag{3.106}$$

$$\left[\frac{\sum_{l'=1}^{m} l' \cdot \exp\{l' \cdot \theta_v + \beta_{il'}\}}{\sum_{l=1}^{m} \exp\{l \cdot \theta_v + \beta_{il}\}}\right]^2 \tag{3.107}$$

$$= \langle x_{vi}^2 \rangle - \langle x_{vi} \rangle^2 . \tag{3.108}$$

Dieses grundlegende Ergebnis wird auch von Rost berichtet (Rost, 2004, p. 359) berichtet. Um den Erwartungswerte und die Varianzen der manifesten Antworten x_{vi} zu erhalten, ist es also lediglich notwendig, die Zustandssumme Z ein, bzw. zwei Mal nach dem entsprechenden Parameter zu differenzieren. Da das Partial-Credit-Modell von Masters (1982) das Obermodell des Ratings-Skalen-Modells von Andrich (1978b) ist, gilt der hier dargestellt Formalismus ebenfalls für dieses. Der Unterschied zwischen den Modellen besteht lediglich darin, dass die Nebenbedingungen bezüglich der Kategorien-Parameter von der Gestalt sind, dass die Abstände der Kategorien-Leichtigkeiten für jedes Item identisch sind.

In Bezug auf das Partial-Credit-Modell ist es interessant zu bemerken, dass nach Rost (Rost, 2004, p. 210) lange Zeit darüber Unklarheit herrschte, wie aus einer multidimensionalen Personenvariable θ_{vx} durch eine lineare Restriktion der Form

$$\theta_{vx} = \phi(x) \cdot \theta_v + \psi(x) \tag{3.109}$$

eine eindimensionale Personenvariable θ_v erzeugt werden kann. In Gleichung 3.109 ist $\phi(x)$ nach Rost ein Bestandteil der Item-Parameter. Nach Rost konnte Andrich (1978a, 1978c) aufbauend auf Arbeiten von Andersen (1977) zeigen, dass es sich bei $\phi(x)$ nicht um einen zu schätzenden Parameter, sondern um die Anzahl der Schwellen handelt, die von der 0-ten bis zur x-ten Kategorie eines Items überschritten wird:

$$\phi(x) = x. \tag{3.110}$$

Dieses Resultat ist mit dem Maximum-Entropie-Ansatz kompatibel. $\phi(x)$ ist in diesem Kontext die Funktion $f_1(x)$, welcher der Personen-Parameter θ_v zugeordnet ist, die $\psi(x)$ sind im Prinzip die Komponenten im Exponent des Modells, die auf die Items zurückzuführen sind, allerdings fehlen in der obigen Darstellung die den Item-Komponenten implizit zugeordneten Funktionen.

Das bedingte Rasch-Modell als Maximum-Entropie-Modell

Auch das bedingte Rasch-Modell ist mit dem Maximum-Entropie-Formalismus kompatibel. Das bedingte Rasch-Modell beschreibt die Wahrscheinlichkeit eines Antwortvektors $\mathbf{x_v}$ bei gegebenem Rohwert $x_{v.} = r_v$ unter Berücksichtigung der Schwierigkeiten der Items β_i. Um das bedingte Rasch-Modell unter Anwendung der Maximum-Entropie-Methode zu definieren, muss die Anzahl der Funktionen $f_j(x_j)$ der Anzahl der Items in einem Test entsprechen. Somit liegt für jedes Item nach dem Formalismus auch ein Parameter λ_j vor. Die Zustandssumme im Nenner der kanonischen Maximum-Entropie-Methode muss so gestaltet werden, dass sie alle möglichen Zustände des Systems, d.h. alle möglichen Antwortvektoren bei gegebener Vektorsumme abbildet, damit die Summe der Wahrscheinlichkeiten der möglichen Antwortvektoren bei gegebener Vektorsumme Eins ergibt. Unter Berücksichtigung dieser Nebenbedingungen ergibt sich unter Anpassung der Notation unmittelbar:

$$p(\mathbf{x_v}|r_v) \quad = \quad \frac{\exp\left(-\sum_{i=1}^{k} x_{vi}\beta_i\right)}{\sum_{\mathbf{y}|r_v} \exp\left(-\sum_{i=1}^{k} y_i\beta_i\right)}.$$

Die Item-Parameter β_i entsprechen den Lagrange-Multiplikatoren λ_j und die Antwort x_{vi} einer Person v auf Item i entspricht den Funktionen $f_j(x_i)$. Die Zustandssumme im Nenner läuft über alle möglichen Antwortmuster $\boldsymbol{y}$, die einen Rohwert von r_v ergeben. y_i ist hier wiederum der Identifikator für einen konkreten Eintrag y_i in einem Antwortvektor $\boldsymbol{y}$, der zu einem Rohwert von r_v führt. In Bezug auf Gleichung 3.81 entsprechen die durch das bedingte Rasch-Modell vorhergesagten Patternwahrscheinlichkeiten den p_i. Die Funktionen $f_j(x_j)$ sind *itembezogene* Funktionen, die die Werte 0 oder 1 annehmen können. Der Wert 0 wird gewählt, wenn das Item nicht gelöst wurde und der Wert 1 wird gewählt, wenn das Item gelöst wurde. $\langle f_j \rangle$ ist der Erwartungswert der entsprechenden Item-Lösungen unter dem Modell und F_k, die verfügbare Information, ist der jeweilige Spaltenmittelwert einer beobachteten Datenmatrix.

Nach dem Maximum-Entropie-Formalismus wäre es nun möglich, die Schätzgleichungen für das bedingte Rasch-Modell herzuleiten, indem der Nenner des Modells logarithmiert und nach den Parametern differenziert wird. Hierbei müsste sich zeigen, dass die Spaltensummen einer Datenmatrix, d.h. die Häufigkeiten der Lösung auf einem Item, suffiziente Statistiken zur Schätzung der Item-Parameter darstellen. Auf eine Durchführung wird an dieser Stelle verzichtet.

3.3. Zusammenfassende Betrachtung des modelltheoretischen Hintergrundes

In diesem Kapitel wurden probabilistische Testmodelle sensu Rasch vorgestellt und einige deren Eigenschaften wurden beschrieben. Demgegenüber wurde der Maximum-Entropie-Formalismus von Jaynes gestellt und auf starke formale Ähnlichkeiten zwischen der kanonischen Maximum-Entropie-Verteilung und dem von Rasch (1961) dargestellten allgemeinen Obermodell für Rasch-Modelle (Gleichung 3.1) wurde hingewiesen. Anhand des dichotomen Rasch-Modells, des Partial-Credit-Modells und des bedingten Rasch-Modells wurde dargelegt, dass die Anwendung der Maximum-Entropie-Methode auf Fragestellungen der Item-Response-Theorie zu Ergebnissen führt, die mit schon bekannten Ergebnis-

sen der Item-Response-Theorie kompatibel sind. Zu nennen wäre die Möglichkeit der Herleitung der Erwartungswerte einer Reaktion x_{vi} auf Basis der Differenzierung der Zustandssumme der Modelle nach einem Parameter (Gleichung 3.52) und die Möglichkeit der Herleitung der Varianz der Erwartungswerte einer Antwort durch zweimaliges differenzieren der logarithmierten Zustandssumme nach einem Parameter (Gleichung 3.56). Die Varianz des manifesten Datums unter dem Modell lässt sich dazu nutzen, um die Informationsfunktion eines Tests zu generieren. Die formalen Ähnlichkeiten sind insofern nicht überraschend, da die kanonische Maximum-Entropie-Verteilung, als auch die genannten Rasch-Modelle Verteilungen beschreiben, die zur Exponentialfamilie gehören. Von daher sind die bekannten Ergebnisse zur Exponentialfamilie, wie das Vorliegen von suffizienten Statistiken zur Parameterschätzung, auf beide Modellierungsansätze übertragbar und somit repliziert auch der Maximum-Entropie-Formalismus die schon bekannten Ergebnisse hinsichtlich der Varianz, Kovarianz und der Erwartungswerte von manifesten Variablen unter dem Modell. Ein wesentlicher Unterschied zwischen Rasch-Modellen und dem Maximum-Entropie-Formalismus besteht jedoch darin, dass Jaynes ein Rational für die Herleitung der kanonischen Maximum-Entropie-Verteilung gibt. Die kanonische Maximum-Entropie-Verteilung wird auf Basis zweier Nebenbedingungen formal hergeleitet: einerseits maximale Informationsentropie der diskreten Verteilung und andererseits Kongruenz mit beobachteten Statistiken in der Form, als dass die Erwartungswerte der Daten unter der Verteilung mit dem manifesten Pendant kongruent sein sollten. Der Grund, warum die Methode in diese Arbeit Eingang gefunden hat ist derjenige, dass es möglich sein sollte, noch nicht bekannte Item-Response-Modelle unter Anwendung der Methode für Fragestellungen der Psychologie herzuleiten. Zur Modellherleitung müssen lediglich die Zustandssumme Z und die Funktionen der Daten der Fragestellung und der Datenlage gemäß definiert werden. Da nun das Handwerkszeug zur Modellgenerierung kurz dargestellt wurde, können die Fragestellungen der Arbeit konkretisiert werden.

4. Modellentwicklung

Ausgangspunkt der Modellentwicklung ist die Frage, inwiefern die Entwicklung eines neuen, probabilistischen Testmodells mittels der Maximum-Entropie-Methode möglich ist, das es erlaubt, die intraindividuelle Variabilität einer Person zu skalieren. Wünschenswert ist, dass die Eigenschaft der Variabilität einer Person v auf mehreren Items in einer latenten Variable η_v abgebildet wird, die sinnvolle Vergleiche der Personen hinsichtlich der Merkmalsausprägung auf einer Differenzenskala erlaubt. Die Berücksichtigung mehrerer Items als Indikatoren für die Variabilität einer Person hat den Vorteil, dass es möglich ist, ein Konstrukt in seiner inhaltlichen Breite abzubilden. Zudem liegt die Vermutung nahe, dass die Verwendung mehrerer Items die Reliabilität der Erfassung der latenten Variable erhöht. Die Verwendung eines Messmodells hat den Vorteil, dass es möglich ist, die Passung des Modells auf empirische Daten zu bewerten und eventuelle Abweichungen der Daten von dem Modell, bzw. Quellen der Fehlpassung zu identifizieren. Günstig wäre es ebenfalls, wenn die latente Variable η_v eine suffiziente Statistik zur Schätzung des Parameters besitzt und spezifisch objektive Vergleiche hinsichtlich der Merkmalsausprägung erlaubt. Im folgenden werden die Fragestellungen zur Modellentwicklung konkretisiert.

4.1. Fragestellungen zur Modellentwicklung

Aus dem psychologischen Hintergrund der Arbeit geht hervor, dass explizit probabilistische Messmodelle zur Skalierung intraindividueller Variabilität und zur Klärung der Frage der interindividuellen Differenzierbarkeit intraindividueller Variabilität rar sind. Hieraus ergibt sich direkt die übergeordnete Fragestellung der Entwicklung eines probabilistischen Testmodells zur Skalierung intraindividueller Variabilität. Da intraindividuelle

Variabilität ein komplexes Konstrukt ist, ist es nötig, den Anwendungsbereich des zu generierenden Modells zu definieren. Gewünscht wird also ein probabilistisches Testmodell, dass es ermöglicht, die intraindividuelle Variabilität auf multivariaten, diskreten Zeitreihen auf eine latente, quantitative Personenvariable η_v zurückzuführen. Das Testmodell sollte suffiziente Statistiken zur Parameterschätzung besitzen, prüfbar sein und die individuelle und globale Messgenauigkeit sollte bewertbar sein. Ferner ist es wünschenswert, dass das Modell zur Klasse der Rasch-Modelle gehört und die latente Personenvariable η_v auf einer sinnvoll interpretierbaren Differenzenskala abgebildet wird.

Bei der Modellentwicklung ist es günstig, auf einen übergeordneten Modellierungs-Kontext zurückzugreifen, um aus diesem deduktiv ein Modell für eine konkrete, inhaltliche Fragestellung zu entwickeln. Eine Möglichkeit der Modell-Herleitung liegt in dem klassischen Ansatz von Rasch, der in der psychologischen Literatur gut ausgearbeitet ist. Eine weitere, zu prüfende Möglichkeit besteht in der Anwendung der Maximum-Entropie-Methode zur Modellgenerierung. Im modelltheoretischen Hintergrund wurde auf die formalen Ähnlichkeiten zwischen dem Ansatz von Rasch und der Maximum-Entropie-Methode hingewiesen und es wurde gezeigt, dass sich klassische Ansätze der Psychometrie, wie z.B. das Partial-Credit-Modell von Masters (1982), das dichotome Rasch-Modell und auch das bedingte Rasch-Modell aus der Anwendung der Methode ergeben und die Anwendung der Formeln von Jaynes zur Berechnung der Varianzen und der Erwartungswerte zu den schon bekannten Ergebnissen aus der psychometrischen Literatur führen, die wichtig für die Parameterschätzung und die Berechnung der Test-Information sind. Vor dem Hintergrund dieser Überlegungen und der theoretischen Hintergründe ergeben sich folgende Fragestellungen zur Modellentwicklung:

I.1. *Ist es möglich ein neues, probabilistisches Testmodell zur Skalierung intraindividueller Variabilität unter Anwendung der Maximum-Entropie-Methode herzuleiten?* Diese Fragestellung ist von zentralem Interesse. Sollte es sich zeigen, dass die Maximum-Entropie-Methode verwendet werden kann, um neue psychometrische Modelle zu generieren, würde dies ein Mittel der kohärenten Modelldefinition aus einem wahrscheinlichkeits- und informationstheoretischen Kontext heraus erschlie-

ßen.

I.2. *Welche Wahrscheinlichkeits-Vorhersagen trifft das Modell hinsichtlich der Wahl einer bestimmten Kategorie durch ein Individuum?* Es ist notwendig zu überprüfen, ob das Modell sinnvolle Ergebnisse hinsichtlich der Wahl einer Kategorie durch ein Individuum liefert und ob diese Ergebnisse kongruent mit dem Desideratum der Skalierung von intraindividueller Variabilität sind.

I.3. *Wie sind die Erwartungswerte und die Varianzen der manifesten Variablen unter dem generierten Modell?* Die Untersuchung der Erwartungswerte und der Varianzen der manifesten Statistiken unter dem Modell sind Vorbedingungen zur Durchführung der Schätzung der Modellparameter nach der Maximum-Likelihood-Methode, der Bestimmung der Genauigkeit der Parameterschätzung und der Bestimmung der Test-Information.

I.4. *Besitzt das generierte Modell suffiziente Statistiken zur Parameterschätzung?* Suffiziente Statistiken sind für psychometrische Modelle von zentraler Bedeutung. Häufig wird argumentiert, dass nur dann, wenn ein Messmodell auch suffiziente Statistiken besitzt, es gerechtfertigt ist, diese Statistik als manifesten Indikator der latenten Merkmalsausprägung zu verwenden. Ein Beispiel hierfür ist der Summenwert, bzw. der Rohwert der Personen bei Rasch-Modellen.

I.5. *Wie hängt die latente Trait-Variable zur Erfassung der intraindividuellen Variabilität mit den manifesten Variablen zusammen?* Es ist zu untersuchen, inwiefern die latente Trait-Variable des zu generierenden Testmodells mit der entsprechenden manifesten Statistik zusammenhängt. A priori ist zu erwarten, dass die manifeste Statistik in einem monotonen Verhältnis zur latenten Trait-Variable steht.

I.6. *Erlaubt das generierte Modell Messungen auf einer Differenzenskala?* Eine zentrale Eigenschaft von Messmodellen ist, dass sie sinnvolle Vergleiche von Personen hinsichtlich der Merkmalsausprägung erlauben sollen. Bei Rasch-Modellen handelt es sich bei der Skala des Vergleichs um eine Differenzenskala. Es ist zu prüfen, ob diese

Eigenschaft auch bei dem hier zu definierenden Modell vorliegt.

I.7. *Konvergiert die Schätzung der Parameter des Modells?* In der vorliegenden Arbeit wird die MCMC-Methode zur Parameterschätzung angewendet. Es ist von Interesse zu überprüfen, inwiefern verschiedene Startwerte der Markov-Ketten die Parameterschätzung beeinflussen oder ob die verschiedenen Markov-Ketten auf einer Zielverteilung operieren. Ist dies der Fall, so ist zu vermuten, dass die Likelihood-Fläche des generierten Modells ein absolutes Maximum besitzt.

I.8. *Wie sind der Bias und die Varianz der Parameterschätzer?* Die Bedeutung von Bias wird in der Kontroverse um frequentistische und Bayesianische Methoden unterschiedlich bewertet. In der vorliegenden Arbeit wird der Bias der MCMC-Schätzung nach klassischen Prinzipien evaluiert, um Aufschluss über bestimmte Schwierigkeiten der Parameterschätzung in Bereichen einer extremen Merkmalsausprägung zu erhalten. Ferner soll untersucht werden, inwiefern die Varianz der Parameterschätzer von der Merkmalsausprägung abhängt. A priori lässt sich auf Basis der Befunde für Rasch-Modelle (siehe z.B. Hoijtink & Boomsma, 1995) absehen, dass die Varianz der Schätzer in Bereichen der extremen Merkmalsausprägung höher ist, als in Bereichen mittlerer Merkmalsausprägung.

I.9. *Wie ist die Passung des Modells bewertbar?* Die theoretischen und praktischen Vorteile probabilistischer Messmodelle sind nur dann voll nutzbar, wenn das Modell einigermaßen passt. Absoluter Modell-Fit ist eine unrealistische Annahme. Dennoch ist es notwendig, den relativen Fit von probabilistischen Testmodellen zu bewerten, um z.B. item- oder personenbezogene Quellen der Fehlpassung zu identifizieren.

I.10. *Wie ist die Reliabilität der Gesamtmessung bestimmbar?* Ein wichtiges Anliegen der Psychometrie ist es, Personen reliabel auf einem empirisch abgrenzbaren Merkmal voneinander zu differenzieren. Es ist zu untersuchen, wie die Bewertung der Reliabilität bei dem zu generierenden Modell geschehen kann.

I.11. *Wie ist die personenbezogene Messgenauigkeit bestimmbar?* Neben der Bewertung

der Gesamt-Reliabilität (Separabilität) eines Instrumentes ist es für diagnostische Zwecke günstig, die Genauigkeit der Erfassung des Merkmals eines bestimmten Individuums zu bewerten. Es ist zu untersuchen, wie dies bei dem zu generierenden Modell geschehen kann.

4.2. Vorgehen zur Prüfung der modelltheoretischen Fragestellungen

Um Fragestellung I.1. zu prüfen, wird die Maximum-Entropie-Methode angewendet, um ein neues probabilistisches Testmodell zu definieren. Die absoluten sukzessiven Differenzen von Ratings mehrerer Personen auf mehreren Items werden als manifeste Indikatoren der Variabilität verwendet. Die Berücksichtigung mehrerer Items erlaubt es, Homogenitätshypothesen bezüglich der Items zu prüfen. Die Berücksichtigung mehrerer Personen ermöglicht die Bewertung der Reliabilität und der Passung des Modells. Die Fragestellungen I.2. und I.3. sind inhaltlich eng miteinander verknüpft. Es ist zu prüfen, ob die Anwendung des Maximum-Entropie-Formalismus es erlaubt, die Erwartungswerte und Varianzen der manifesten Variablen in Übereinstimmung mit gängigen Ergebnissen aus der Statistik und der Psychometrie zu berechnen. Die Erwartungswerte der manifesten Variable sind für die Parameterschätzung mit der Maximum-Likelihood-Methode von Bedeutung, da in der Regel die Parameter eines Modells so geschätzt werden, dass die beobachteten Daten (z.B. Summenwerte oder Mittelwerte) mit den unter dem Modell erwarteten Daten (den Erwartungswerten unter dem Modell) möglichst kongruent sind. Die Varianzen der manifesten Variablen unter dem Modell sind von Belang, wenn es gilt die Informationsfunktion eines Tests zu ermitteln. Methodisch wird die Fragestellung I.2. geprüft, indem der Maximum-Entropie-Formalismus und insbesondere die Gleichungen 3.52 und 3.56 auf das definierte Modell angewendet werden. Diese Gleichungen erlauben es, den Erwartungswert und die Varianz einer Funktion $f_j(x_j)$, die einem Parameter zugeordnet ist, zu bewerten und sind von daher von Belang bei der Parameterschätzung, insbesondere was den Gradienten und die Genauigkeit der Parameterschätzung angeht.

Die Ergebnisse werden mit gängigen Ergebnissen aus der Psychometrie verglichen. Zur Bearbeitung der Fragestellung I.2. werden die durch das Modell beschriebenen, vorhergesagten Wahrscheinlichkeiten der Wahl einer bestimmten Kategorie bei unterschiedlichen Parameter-Konfigurationen gebildet. Es wird geprüft, ob die Wahrscheinlichkeiten mit dem Desideratum der Skalierung von Variabilität kongruent sind. Zur weiteren Klärung der Frage werden die Kategorien-Response-Funktionen des Modells berechnet, um zu überprüfen, ob ein hoher Personen-Parameter mit einer erhöhten Variabilität in den durch das Modell vorhergesagten Zeitreihen einhergeht. Fragestellung I.3. wird formal angegangen, in dem die erste und die zweite Ableitung der Modellgleichung nach dem Personen-Parameter gebildet wird. Nach dem Maximum-Entropie-Formalismus entspricht die erste Ableitung dem Erwartungswert der manifesten Daten unter dem Modell und die zweite Ableitung entspricht der Varianz der manifesten Daten. Diese Statistiken können dazu verwendet werden, um standardisierte Residuen zu berechnen, welche zur Prüfung der Modellpassung nützlich sind. Zur Bewertung der Fragestellung I.4. wird die Likelihood-Funktion des Modells gebildet, an der sich das Vorliegen von suffizienten Statistiken zur Schätzungen von Parametern zeigen sollte und das Ergebnis wird in Beziehung zum Maximum-Entropie-Formalismus gesetzt. Fragestellung I.5. impliziert, dass das zu entwickelnde Modell personenbezogene Parameter besitzen soll, die in Zusammenhang mit der Variabilität in manifest-multivariaten Zeitreihen in der Form von Ratings stehen. Es ist von Interesse, inwiefern die latente Trait-Variable mit den manifesten Ratings unter dem Modell zusammenhängt. Zur Untersuchung der Fragestellungen werden manifeste Zeitreihen aus dem Modell bei unterschiedlichen Ausprägungen der latenten Variable simuliert und es wird grafisch dargestellt, inwiefern die mittleren absoluten Differenzen der Ratings mit der latenten Variable zusammenhängen. Der Terminus „Rasch-Modell“ impliziert die Möglichkeit spezifisch objektiver Vergleiche von Personen hinsichtlich der Merkmalsausprägung. Um die Frage der Messung auf einer Differenzenskala (Fragestellung I.6.) zu klären, werden die Logits der Kategorien-Wahrscheinlichkeiten des Modells untersucht, um zu überprüfen, ob die Modellparameter in einer linearen Beziehung zu den Logits benachbarter Kategorien stehen und ob die Item-Parameter bei dem Ver-

gleich der Personen hinsichtlich der Merkmalsausprägung eine Rolle spielen. Fragestellung I.7. berührt den technischen Aspekt der Parameterschätzung, der auch inhaltliche Relevanz besitzt. Es ist zu prüfen, ob die Parameterschätzungen auch bei unterschiedlichen Startwerten konvergieren, was impliziert, dass das Modell eine absolutes Maximum der Likelihood-Fläche besitzt. Ferner interessiert, wie die Merkmalsausprägung auf der latenten Variable mit dem Schätzfehler des jeweiligen Parameters zusammenhängt. Zum Einsatz kommt die MCMC-Methode. Die Prüfung des empirischen Bias und der empirischen Varianz der Parameterschätzer (Fragestellung I.8.) geschieht ebenfalls auf simulativem Weg. Es werden Daten aus dem Modell simuliert und per MCMC-Methode aus den simulierten Daten extrahiert, um den empirischen Bias und die empirische Varianz der Parameterschätzer zu bestimmen.

Die Anwendung von psychometrischen Messmodellen in der Praxis setzt unter anderem voraus, dass die Passung eines Modells auf beobachtete Daten bewertbar ist (Fragestellung I.9.). In der vorliegenden Arbeit wird die Passung auf Basis standardisierter Residuen bewertet, da sich die Residuen formal relativ einfach unter Verwendung der Erwartungswerte und der Varianzen der manifesten Variablen unter dem Modell ergeben. Zur Berechnung der Residuen lassen sich die Ergebnisse der Bearbeitung der Fragestellung I.3. nutzen.

Zur Bewertung der Reliabilität (Fragestellung I.10.) auf Global-Ebene wird die Andrich-Reliabilität (Andrich, 1988) vorgeschlagen, welche an die Definition der Reliabilität in der klassischen Testtheorie angelehnt ist. Zur Berechnung der Andrich-Reliabilität werden die Varianz der geschätzten Trait-Parameter und deren mittlere Standardfehler verwendet. Praktisch wird diese Fragestellung genauer im Teil der Modellanwendung überprüft.

Die Messgenauigkeit auf Individual-Ebene (Fragestellung I.11.) lässt sich im Rahmen probabilistischer Testmodelle mittels der individuellen Standardfehler der Schätzer der Trait-Variable vornehmen. Da in der vorliegenden Arbeit die MCMC-Methode verwendet wird, wird vorgeschlagen, die Posterior-Verteilungen der Parameter zur Bewertung der individuellen Messgenauigkeit zu verwenden. Auch diese Fragestellung wird praktisch im Teil zur Modellanwendung überprüft.

4.3. Modelldefinition

Die Modelldefinition erfolgt durch die Definition der Funktionen in Gleichung 3.80. Eine wichtige Entscheidung besteht darin, welche Statistik der manifesten Daten als Funktion f_1 in der Maximum-Entropie-Verteilung sein soll. Im Falle des dichotomen Rasch-Modells war dies die Scoring-Funktion $f_1(x) = x$, mit $x \in 0, 1$, im Falle des Partial-Credit-Modells wurde die Scoring-Funktion als $f_1(x) = x$ mit $x \in 1, \ldots, m$ gewählt, üblicher wäre die Wahl $x \in 0, \ldots, m-1$, aber wie in einem vorigen Abschnitt dargelegt wurde, lassen sich die beiden Wahlen durch entsprechende Parametrisierungen und Restriktionen auf den Kategorien-Parametern ineinander überführen.

Zur Erfassung der intraindividuellen Variabilität kämen als Scoring-Funktion beispielsweise die intraindividuelle Standardabweichung, die absolute sukzessive Differenz, die quadrierte absolute Differenz oder ähnliche Indices in Betracht, die inhaltlich eine Variabilität in einer Zeitreihe abbilden. Da die intraindividuelle Standardabweichung nicht in der Lage ist, Abhängigkeiten in einer Zeitreihe abzubilden wird hier die absolute Differenz verwendet. Aus der Logik der Maximum-Entropie-Methode folgt, dass bei dieser Wahl die Summe der absoluten Differenzen einer intraindividuellen Zeitreihe eine suffiziente Statistik zur Schätzung der latenten Variable darstellt. Dies wird in einem späteren Abschnitt untersucht werden. Da das Modell für Items des Likert-Typs anwendbar sein soll, ist zu berücksichtigen, dass es sich um ein multinomiales Antwortformat handelt. Die Berücksichtigung geschieht durch das Einfügen von Kategorien-Parametern in die Modellgleichung. Es kann einerseits - wie beim Partial-Credit-Modell - pro Item und Kategorie ein eigener Parameter verwendet werden (Masters, 1982) oder es kann - wie beim Rating-Skalen-Modell - angenommen werden, dass die Schwellenabstände zwischen den Items konstant sind und das Item selbst eine feste Schwierigkeit besitzt (Andrich, 1978b). Die erstere Variante besitzt mehr Freiheitsgrade, von daher wird diese Variante gewählt. Die Variante von Andrich (1978b) resultiert durch eine einfache Restriktion der Kategorien-Parameter. Bei Andrichs Variante muss in der Definition der Kategorien-Leichtigkeiten in der Zustandssumme lediglich die Anzahl der Kategorien und nicht zusätzlich noch die Anzahl der Items berücksichtigt werden.

Werden in der kanonischen Maximum-Entropie-Verteilung die Funktionen der Daten so gewählt, dass f_1 der absoluten Differenz der Itemantworten einer Person v auf ein Item i entspricht und werden zudem item- und kategorienspezifische Parameter im Sinne des PCM eingefügt, so resultiert aus der kanonischen Maximum-Entropie-Verteilung folgendes Modell:

$$p(X_{vi[t]} = x_{vi[t]}|x_{vi[t-1]}) = \frac{\exp\left\{|x_{vi[t]} - x_{vi[t-1]}| \cdot \eta_v + \beta_{ix}\right\}}{\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v + \beta_{il}\right\}}, \tag{4.1}$$

wobei $x_{vi} \in \{1, \ldots, l, \ldots, m\}$. Die Schreibweise $p(X_{vi[t]} = x_{vi}|x_{vi[t-1]})$ wurde gewählt, da das Modell m verschiedene Wahrscheinlichkeitsverteilungen definiert, die jeweils von der zuletzt gewählten Kategorie $x_{vi[t-1]}$ abhängen. Praktisch bedeutet dies, dass das Modell es erlaubt, die Wahrscheinlichkeiten der Reaktionen einer Person v zum Zeitpunkt t auf Item i vorherzusagen, sofern die Parameter des Modells und der zuletzt gewählte Wert $x_{vi[t-1]}$ bekannt sind. Bei $x_{vi[t-1]}$ handelt es sich um *keine* Variable des Modells, sondern es ist eine beobachtete Reaktion zum Zeitpunkt $[t-1]$. Das Resultat eines bedingten Modells ist in der Wahl der Zustandssumme begründet. Diese läuft über alle möglichen Reaktionen $x_{vi[t]}$ bei *bekanntem* $x_{vi[t-1]}$. In diesem Sinne ist es möglich, Abhängigkeiten zwischen $x_{vi[t]}$ und $x_{vi[t-1]}$ abzubilden. $x_{vi[t]}$ ist der numerische Identifikator der gewählten Kategorie zum Zeitpunkt t und $x_{vi[t-1]}$ ist der Identifikator der gewählten Kategorie zum Zeitpunkt $t-1$, oder einfacher gesagt, der zuletzt gewählten Kategorie. $|x_{vi[t]} - x_{vi[t-1]}|$ ist die absolute Differenz der Antwort einer Person v auf ein Item i von Zeitpunkt $t-1$ zu t oder auch die Scoring-Funktion $f_1(x_{vi[t]})$. m ist die Anzahl der Kategorien. Die Parameter β_{ix} repräsentieren item- und kategorienspezifischen Leichtigkeiten, wobei der Index i das Item und der Index x die Kategorie markiert. Anders ausgedrückt handelt es sich um item- und kategorienspezifische Konstanten. Zur Identifikation des Modells muss eine Restriktion über die item-spezifischen Schwellen gelegt werden, da pro Item auf alle Kategorien-Parameter eine konstanter Wert addiert werden kann, ohne die vorhergesagten Wahrscheinlichkeiten zu verändern. Die Restriktion kann durch die Fixierung der Parameter β_{i1} auf Null und den Start der Scoring-Funktion bei 0 geschehen

oder eine Summennormierung der Parameter über die Items kann durchgeführt werden:

$$\sum_{x=1}^{m} \beta_{ix} = 0. \tag{4.2}$$

Das bedeutet, dass pro Item eine Summennormierung für die Kategorien-Parameter durchgeführt wird. Wie in einem späteren Abschnitt gezeigt wird, lassen sich aus diesen Parametern durch einfache lineare Transformationen die Schnittpunkte der Kategorien-Response-Funktionen des Modells berechnen. Die Summe im Nenner des Modells ist die Zustandssumme Z, die gewährleistet, dass die Summe der Einzelwahrscheinlichkeiten 1 ergibt ($\sum_{l=1}^{m} p(X_{vi[t]} = x_{vi}|x_{vi[t-1]}) = 1$). Es handelt sich hierbei also um die Definition eines Markov-Prozesses erster Ordnung, bei dem die Wahrscheinlichkeiten der Wahl einer Kategorie $x_{vi[t]}$ zu einem Zeitpunkt t von der latenten Variable η_v, der zuletzt gewählten Kategorie $x_{vi[t-1]}$ und den Kategorien-Parametern β_{ix} abhängen, welche Kategorien-Leichtigkeiten darstellen. Aus dem Blickwinkel der klassischen Item-Response-Modelle handelt es sich bei dem durch die Anwendung der Maximum-Entropie-Methode resultierenden Modell um das Partial-Credit-Modell von Masters, wobei die Scoring-Funktion für η_v der absoluten Differenz der Item-Antworten vom Zeitpunkt $[t-1]$ zu $[t]$ entspricht.

Im Folgenden sei der Bezug zum Maximum-Entropie-Formalismus anhand der Zustandssumme erläutert, die alle möglichen Zustände des durch das Modell beschriebenen Systems enkodiert. Bei $m = 3$, also bei drei verwendeten Kategorien, sieht die Zustandssumme wie folgt aus:

$$\begin{aligned} Z \;=\; & \exp\left\{\eta_v \cdot |0 - x_{vi[t]}| + 1 \cdot \beta_{i1} + 0 \cdot \beta_{i2} + 0 \cdot \beta_{i3}\right\} + && (4.3)\\ & \exp\left\{\eta_v \cdot |1 - x_{vi[t]}| + 0 \cdot \beta_{i1} + 1 \cdot \beta_{i2} + 0 \cdot \beta_{i3}\right\} + && (4.4)\\ & \exp\left\{\eta_v \cdot |2 - x_{vi[t]}| + 0 \cdot \beta_{i1} + 0 \cdot \beta_{i2} + 1 \cdot \beta_{i3}\right\}. && (4.5) \end{aligned}$$

Den λ_j der allgemeinen Maximum-Entropie-Gleichung (Gleichung 3.80) entsprechen die Parameter η_v, β_{i1}, β_{i2} und β_{i3}. Diesen Parametern sind vier Funktionen $f_j(x_j)$ zugeordnet. $f_1(x_1)$ ist die Scoring-Funktion für die latente Variable η_v, $f_2(x_2)$ ist die Scoring-Funktion für Kategorie 1, deren Wert 1 ist, sofern die erste Kategorie gewählt wurde, $f_3(x_3)$ ist die Scoring-Funktion für Kategorie 2, deren Wert 1 ist, sofern die zweite Kategorie gewählt wurde und $f_4(x_3)$ ist die Scoring-Funktion für Kategorie 3, deren Wert 1

ist, sofern die dritte Kategorie gewählt wurde. Ansonsten sind die Werte der kategorienbezogenen Scoring-Funktionen gleich 0. Zur Erläuterung - und auch zur einfacheren Darstellung des Modells - ist es hilfreich, eine Zustandsmatrix (S) einzuführen:

$$\boldsymbol{S} = \begin{pmatrix} |1 - x_{vi[t-1]}| & 1 & 0 & 0 \\ |2 - x_{vi[t-1]}| & 0 & 1 & 0 \\ |3 - x_{vi[t-1]}| & 0 & 0 & 1 \end{pmatrix}.$$

Die Zeilen der Matrix entsprechen den 3 möglichen Zuständen des Systems. Entweder wird Kategorie 1 gewählt oder es wird Kategorie 2 gewählt oder es wird Kategorie 3 gewählt. Spalte 1 der Matrix repräsentiert die Scoring-Funktion $f_1(x)$ und die übrigen Spalten entsprechen den kategorien-bezogenen Scoring-Funktionen $f_2(x)$ bis $f_4(x)$. Pro zuletzt gewählter Kategorie $x_{vi[t-1]}$ existiert eine eigene Matrix S möglicher Folgezustände, da eine bedingte Wahrscheinlichkeitsverteilung modelliert wird.

Es ist wichtig, das definierte Modell von einem Modell zu unterscheiden, bei dem die *evaluierte*, d.h. ausgerechnete absolute Differenz als Scoring-Funktion verwendet wird. Das hier generierte Modell berechnet die Wahrscheinlichkeit der Wahl der Kategorie $x_{vi[t]}$ bei *gegebenem* $x_{vi[t-1]}$, wobei *eine* Wahrscheinlichkeitsverteilung für $x_{vi[t]}$ *pro* zuletzt gewähltem Wert $x_{vi[t-1]}$ vorliegt. Wird die berechnete Differenz als Scoring-Funktion verwendet, geht die Information über den zuletzt gewählten Wert verloren und in die Berechnung der Zustandssumme geht lediglich die berechnete, absolute Differenz ein. Dies hat zur Folge, dass letzteres Modell lediglich eine Variante des Partial-Credit-Modells von Masters (Masters, 1982) darstellt, wobei die Scoring-Funktion der evaluierten absoluten Differenz entspricht und somit nur *eine* Verteilung für den berechneten Betrag $|x_{vi[t]} - x_{vi[t-1]}|$ vorliegt, welche alle möglichen *Sprungbeträge* abbildet. Das erstere, hier verwendete Modell jedoch bildet m *bedingte* Verteilungen für $p(X_{vi[t]} = x_{vi[t]}|x_{vi[t-1]})$ ab, wodurch die Information über den zuletzt gewählten Wert nicht verloren geht und die Wahrscheinlichkeiten für die Wahl der nächsten Kategorien modelliert werden. Diese wichtige Unterscheidung sei an einem kleinen Beispiel verdeutlicht. Nehmen wir an, die Anzahl der Kategorien eines Items ist $m = 3$ und der zuletzt gewählte Wert ist $x_{vi[t-1]} = 2$. Unter dem hier definierten bedingten Modell sieht die Zustandssumme wie

folgt aus:

$$Z_1 = \exp(|1-2|) \cdot \eta_v + \beta_{i1} + \exp(|2-2|) \cdot \eta_v + \beta_{i2} + \exp(|3-2|) \cdot \eta_v + \beta_{i3}. \quad (4.6)$$

Bei der Verwendung lediglich der berechneten absoluten Differenz nimmt die Zustandssumme folgende Form an:

$$Z_2 = \exp(0) \cdot \eta_v + \beta_{i1} + \exp(1) \cdot \eta_v + \beta_{i2} + \exp(2) \cdot \eta_v + \beta_{i3}. \quad (4.7)$$

Wie ersichtlich, stellen Z_1 und Z_2 zwei völlig unterschiedliche Sachverhalte dar. Bei der Verwendung von Z_1 addieren sich die möglichen Zustände von $x_{vi[t]}$ bei *gegebenen* $x_{vi[t-1]}$ zu 1, wohingegen sich bei Verwendung von Z_2 die Zustände über alle möglichen Sprungbeträge zu 1 summieren. Diese Unterscheidung ist eine wichtige Tatsache zum Verständnis des Modells. Das hier verwendeten Modell bildet einen Markov-Prozess ab, während dies bei der Verwendung der berechneten absoluten Differenz nicht der Fall wäre.

4.4. Die bedingten, erwarteten Kategorien-Wahrscheinlichkeiten unter dem Modell

Um die Bedeutung des Modells weiter zu verdeutlichen ist es günstig, die die vorhergesagten, bedingten Kategorien-Wahrscheinlichkeiten des Modells zu betrachten.

Tabelle 4.1 zeigt exemplarisch 5 unterschiedliche Konfigurationen der unter dem Modell erwarteten Kategorien-Wahrscheinlichkeiten bei unterschiedlichen Parametern η, $\boldsymbol{\beta}$ und zuletzt gewähltem Wert $x_{vi[t-1]}$ für ein Item mit $m = 4$ Kategorien. Die in den sechs Sektionen dargestellten Wahrscheinlichkeiten sind Übergangsmatrizen für einen Markov-Prozesses erster Ordnung, wobei die Wahrscheinlichkeiten der Wahl einer Kategorie von der zuletzt gewählten Kategorie $x_{vi[t-1]}$ und den Modellparametern abhängt.

Betrachten wir Abschnitt 1 in Tabelle 4.1, so zeigen sich die unter dem Modell erwarteten Kategorien-Wahrscheinlichkeiten in Abhängigkeit der zuletzt gewählten Kategorie $x_{vi[t-1]}$, wenn sowohl der Parameter η_v, als auch die Kategorien-Parameter $\boldsymbol{\beta}$ gleich Null sind. In diesem Fall ist die erwartete Wahl einer Kategorie zum Zeitpunkt t unabhängig vom zuletzt gewählten Wert $x_{vi[t-1]}$. Da die Kategorien-Parameter ebenfalls alle gleich

Tabelle 4.1.: Erwartete Kategorien-Wahrscheinlichkeiten $p(X_{vi[t]} = x_{vi[t]}|x_{vi[t-1]})$ für unterschiedliche Parameter η_v und β bei einem vier-kategoriellen Item

		$x_{vi[t]}$=1	$x_{vi[t]}$=2	$x_{vi[t]}$=3	$x_{vi[t]}$=4
Abschnitt 1	$x_{vi[t-1]}$=1	.25	.25	.25	.25
$\eta_v = 0$, $\boldsymbol{\beta} = [0, 0, 0, 0]$	$x_{vi[t-1]}$=2	.25	.25	.25	.25
	$x_{vi[t-1]}$=3	.25	.25	.25	.25
	$x_{vi[t-1]}$=4	.25	.25	.25	.25
Abschnitt 2	$x_{vi[t-1]}$=1	.13	.37	.37	.13
$\eta_v = 0$, $\boldsymbol{\beta} = [-0.5, 0.5, 0.5, -0.5]$	$x_{vi[t-1]}$=2	.13	.37	.37	.13
	$x_{vi[t-1]}$=3	.13	.37	.37	.13
	$x_{vi[t-1]}$=4	.13	.37	.37	.13
Abschnitt 3	$x_{vi[t-1]}$=1	.57	.35	.08	.00
$\eta_v = -1.5$, $\boldsymbol{\beta} = [0, 0, 0, 0]$	$x_{vi[t-1]}$=2	.06	.76	.17	.01
	$x_{vi[t-1]}$=3	.01	.17	.76	.06
	$x_{vi[t-1]}$=4	.00	.08	.35	.57
Abschnitt 4	$x_{vi[t-1]}$=1	.20	.32	.32	.16
$\eta_v = 0$, $\boldsymbol{\beta} = [-0.2, 0.3, 0.3, -0.4]$	$x_{vi[t-1]}$=2	.20	.32	.32	.16
	$x_{vi[t-1]}$=3	.20	.32	.32	.16
	$x_{vi[t-1]}$=4	.20	.32	.32	.16
Abschnitt 5	$x_{vi[t-1]}$=1	.69	.25	.06	.01
$\eta_v = -1.5$, $\boldsymbol{\beta} = [-0.2, 0.3, 0.3, -0.4]$	$x_{vi[t-1]}$=2	.10	.72	.16	.02
	$x_{vi[t-1]}$=3	.02	.16	.73	.08
	$x_{vi[t-1]}$=4	.01	.06	.29	.64
Abschnitt 6	$x_{vi[t-1]}$=1	.01	.06	.29	.64
$\eta_v = 1.5$, $\boldsymbol{\beta} = [-0.2, 0.3, 0.3, -0.4]$	$x_{vi[t-1]}$=2	.15	.06	.25	.55
	$x_{vi[t-1]}$=3	.61	.23	.05	.01
	$x_{vi[t-1]}$=4	.69	.25	.06	.01

Null sind, zeigen sich keine Unterschiede hinsichtlich der Wahl einer bestimmten Kategorie. Anders ausgedrückt, es liegt keine Information vor, die es erlauben würde, eine Vorhersage über eine zu wählende Kategorie zu treffen, die über eine zufällige Wahl hinaus geht. Alle möglichen Zustände sind gleich wahrscheinlich.

In Abschnitt 2 der Tabelle 4.1 wurden lediglich die Kategorien-Parameter $\boldsymbol{\beta}$ variiert. Es zeigt sich, dass mit sinkendem Parameter β_{ix} ebenfalls die Wahrscheinlichkeit sinkt, dass eine bestimmte Kategorie x gewählt wird. Es ist also sinnvoll anzunehmen, dass ein Parameter β_{ix} eine Kategorien-Leichtigkeit darstellt.

Abschnitt 3 dient der Verdeutlichung der Funktion des Parameters η_v. Ist der Parameter η_v gleich Null, so hängen die vorhergesagten Wahrscheinlichkeiten lediglich von den Kategorien-Parametern ab. Ein negativer Parameter η_v jedoch führt dazu, dass der erwartete Markov-Prozess sich stabilisiert, was sich darin zeigt, dass in der Diagonalen der entsprechenden Übergangsmatrix hohe vorhergesagte Wahrscheinlichkeiten aufzufinden sind. Dies bedeutet, dass der Verbleib in einer bestimmten Kategorie von Zeitpunkt $[t-1]$ zu Zeitpunkt $[t]$ wahrscheinlicher ist, als der Kategorien-Wechsel. Die durch diese Übergangsmatrix beschriebene Zeitreihe ist also relativ stabil und zeigt weniger Variabilität, als beispielsweise die Zeitreihe, die durch die Übergangsmatrix in Abschnitt 1 und 2 probabilistisch beschrieben ist. Auffällig ist, dass die Übergangsmatrix symmetrisch ist, was daran liegt, dass die Kategorien-Parameter β_{ix} alle Null sind.

Abschnitt 4 dient der Verdeutlichung der Tatsache, dass der durch das Modell beschriebene Markov-Prozess nicht unbedingt zu einer symmetrischen Verteilung der vorhergesagten Kategorien-Wahrscheinlichkeiten führen muss. Interessant sind hier die Kategorien 1 und 4. Da der Kategorien-Parameter für Kategorie 4 kleiner ist, als derjenige für Kategorie 1 ist, ist die vorhergesagte Wahrscheinlichkeit der Wahl der Kategorie 4 ebenfalls geringer.

In Abschnitt 5 ist der Parameter η_v als -1.5 gewählt, was dazu führt, dass wiederum ein relativ stabiler Markov-Prozess beschrieben wird. Dies ist daran zu erkennen, dass die unter dem Modell erwarteten Übergangswahrscheinlichkeiten in der Diagonalen der Matrix am höchsten sind. Zudem wurden die Kategorien-Parameter $\boldsymbol{\beta}$ dem Abschnitt 4 in

Tabelle 4.1 entsprechend gewählt, was dazu führt, dass die durch das Modell beschriebene Übergangsmatrix nicht symmetrisch ist.

Abschnitt 6 veranschaulicht die vorhergesagten Kategorien-Wahrscheinlichkeiten, wenn der Parameter η_v größer als Null ist. Hier wurde der Wert $\eta_v = 1.5$ gewählt. Es zeigt sich, dass dieser Parameter eine Zeitreihe beschreibt, in der extreme Sprünge auftreten. Nehmen wir z.B. an, der zuletzt gewählte Parameter $x_{vi[t-1]}$ ist 4, so ist die Wahrscheinlichkeit in die am weitesten entfernte Kategorie $x_{vi[t]}$ zu wechseln am höchsten (.69).

Es ist also zu konstatieren, dass mit steigendem Kategorien-Parameter β_{ix} auch die Wahrscheinlichkeit der Wahl einer entsprechenden Kategorie steigt. Zudem führt ein sinkender Parameter η_v dazu, dass der durch das Modell beschriebene Markov-Prozess sich stabilisiert, also die durch das Modell beschriebene Zeitreihe eine geringere Variabilität aufweist. Ein Parameter $\eta_v > 0$ führt zu extrem sprunghaften Zeitreihen, deren Variabilitäten deutlich von einer Zeitreihe abweichen, die lediglich durch die Kategorien-Parameter beschrieben wird ($\eta_v = 0$).

Inhaltlich ist zu verzeichnen, dass die Kategorien-Wahrscheinlichkeiten einerseits auf Basis der Leichtigkeits-Parameter $\boldsymbol{\beta}$ und andererseits durch den personenbezogenen Parameter η_v modelliert werden. Jeder einzelne Parameter β wird nur auf der ihm zugeordneten Kategorie wirksam, wobei sich die Wirksamkeit des Parameters η_v über alle Kategorien hinweg erstreckt. Die Bedeutung des Parameters η_v und die Möglichkeit der Transformation der Kategorien-Parameter $\boldsymbol{\beta}$ in Schwellen-Parameter $\boldsymbol{\tau}$, welche die Schnittpunkte der Kategorien-Funktionen darstellen, wird in einem späteren Abschnitt näher beleuchtet.

4.5. Die Kategorien-Charakteristik-Kurven

Eine weitere Möglichkeit der Verdeutlichung der Funktion des Modells - neben der Darstellung der Übergangsmatrizen - besteht darin, die Kategorien-Charakteristik-Kurven in Abhängigkeit des Parameters η_v und der Kategorien-Parameter zu berechnen und grafisch darzustellen.

Abbildung 4.1 zeigt die Kategorien-Charakteristik-Kurven des Modells in Abhängigkeit

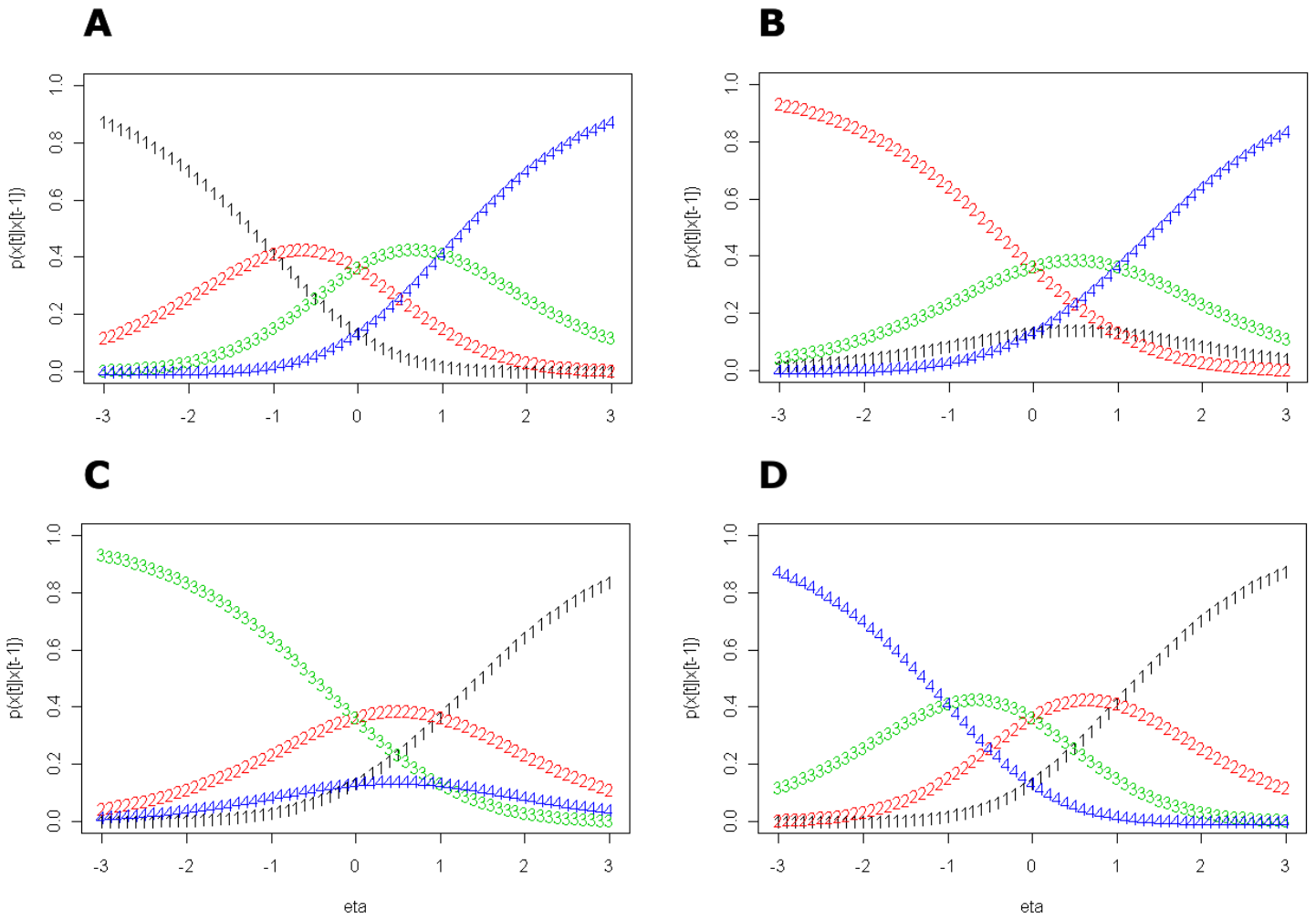

Abbildung 4.1.: Kategorien-Charakteristik-Kurven in Abhängigkeit von η_v für ein Item mit $\boldsymbol{\beta} = [-0.5, 0.5, 0.5, -0.5]$. A: $x_{vi[t-1]} = 1$; B:$x_{vi[t-1]} = 2$; C: $x_{vi[t-1]} = 3$; D: $x_{vi[t-1]} = 4$.

von η_v für ein Item mit den Kategorien-Parametern $\boldsymbol{\beta} = [-0.5, 0.5, 0.5, -0,5]$. Die Abszissen stellen die Ausprägungen der latenten Variable η_v dar, die Ordinaten bezeichnen Wahrscheinlichkeiten für die Wahl einer bestimmten Kategorie. In das Koordinatensystem sind unterschiedliche Kurven eingetragen. Die Kurve, die mit der Zahl 1 markiert ist, stellt die Wahrscheinlichkeit des Auftretens der Kategorie 1 in Abhängigkeit der latenten Variable η_v dar. Dies gilt für die Kurven, die mit den Zahlen 2, 3 und 4 bezeichnet sind, analog.

Pro Item existieren in diesem Fall 4 Diagramme, die die Kategorien-Charakteristik-Kurven in Abhängigkeit der latenten Variable und des zuletzt gewählten Wertes beschreiben. Diagramm A bezeichnet die Wahrscheinlichkeit der Wahl einer Kategorie, wenn $x_{vi[t-1]} = 1$, d.h. die zuletzt gewählte Kategorie war Kategorie 1. Diagramm B stellt die Kategorien-Charakteristik-Kurven für $x_{vi[t-1]} = 2$ dar, Diagramm C stellt die Kategorien-Charakteristik-Kurven für $x_{vi[t-1]} = 3$ dar und Diagramm D stellt die Kategorien-Charakteristik-Kurve für $x_{vi[t-1]} = 4$ dar.

Der Punkt, der durch Abbildung 4.1 verdeutlicht werden soll ist derjenige, dass mit sinkendem Parameter η_v die Wahrscheinlichkeiten sinken, dass eine andere Kategorie als die zuletzt gewählte auftritt. Dies ist z.B. in Diagramm A an der Kategorien-Charakteristik-Kurve für Kategorie 1 zu erkennen. Mit steigendem Parameter η_v sinkt die Wahrscheinlichkeit der wiederholten Wahl der Kategorie 1 während die Wahrscheinlichkeiten der Wahl distaler Kategorien steigt. Ein geringer Parameter η_v geht also mit einer geringen erwarteten Variabilität in der durch das Modell beschriebenen Zeitreihe einher, während ein hoher Parameter η_v eine Zeitreihe beschreibt, in der eine relativ hohe Variabilität vorherrscht. Die Kategorien-Wahrscheinlichkeiten addieren sich bei einem festen Wert von η_v zu Eins, da es sich um disjunkte Kategorien handelt. Der Einfluss der Kategorien-Parameter auf die Kategorien-Wahrscheinlichkeiten spiegelt sich am Ort $\eta_v = 0$ wieder. An diesem Punkt hängen die Wahrscheinlichkeiten lediglich von den Kategorien-Parametern ab und es ist zu beobachten, dass die Kategorien-Wahrscheinlichkeiten die ordinalen Verhältnisse der Kategorien-Parameter an diesem Punkt reflektieren. So gleicht die Auftretenswahrscheinlichkeit der Kategorie 1 der Auftretenswahrscheinlichkeit der

Kategorie 4 am Punkt $\eta_v = 0$. Dies spiegelt die Identität der numerischen Ausprägung der jeweiligen Kategorien-Parameter wieder ($\beta_{i1} = -0.5$ und $\beta_{i4} = -0.5$). Analoges gilt für die Diagramme B, C und D. Im Anhang befindet sich eine R-Funktion, mit der die Kategorien-Charakteristik-Kurven für beliebige Kategorien- und Schwellenparameter grafisch ausgegeben werden können.

An dieser Stelle kann die Frage auftreten kann, wie es sich mit der Ordnung der Schwellenparameter verhält. Bei den $\boldsymbol{\beta}$ handelt es sich nicht um Schwellenparameter, sondern um Leichtigkeits-Parameter, die Schnittpunkte der Kategorien-Funktionen (Schwellenparameter) können allerdings für jedes der Diagramme berechnet werden, was in einem späteren Abschnitt detaillierter betrachtet wird. Betrachten wir Diagramm A, so zeigt sich, dass die Schnittpunkte der benachbarten Kategorien (1 und 2, 2 und 3, 3 und 4) aufsteigend geordnet sind. Ebenso zeigt sich eine Ordnung der Schnittpunkte in Diagramm D. Beide Diagramme stellen eine Situation dar, bei dem die zuletzt gewählte Kategorie eine Extrem-Kategorie der Rating-Skala darstellt. Im Falle von Diagramm A sind nur Sprünge in Richtung oberes Ende und in Diagramm D sind nur Sprünge in Richtung unteres Skalen-Ende möglich. Die Diagramme B und C repräsentieren Situationen, bei denen es sich bei der zuletzt gewählten Kategorie um keine Extrem-Kategorie, sondern um eine Kategorie im mittleren Bereich der Skala handelt. In diesem Fall sind die Schnittpunkte der Kategorien-Funktionen nicht geordnet, denn ein Sprung vom Betrag 1 ist jetzt auf zweifache Weise möglich. Einmal in Richtung oberes, oder in Richtung unteres Ende der Rating-Skala. Von daher tritt hier eine Ordnung der Schwellen im Sinne z.B. des RSM in diesen Fällen nicht auf.

4.6. Die Erwartungwerte und die Varianz der manifesten Variable unter dem Modell

Die Berechnung der Erwartungswerte von $x_{vi[t]}$ unter dem Modell und die Berechnung der erwarteten Varianz von $x_{vi[t]}$ geschieht unter Verwendung der Maximum-Entropie-Methode mit den Gleichungen 3.52 und 3.56. Die Zustandssumme läuft über alle mögli-

chen Zustände des Systems und für die Scoring-Funktion $f_1(x)$ gilt $f_1(x_{vi[t]}) = |x_{vi[t]} - x_{vi[t-1]}|$. Die zu modellierende Wahrscheinlichkeit ist $p(X_{vi[t]} = x_{vi[t]}|x_{vi[t-1]})$. Die bedingte Notation wurde gewählt um deutlich zu machen, dass in die Modellierung als zusätzliche Information der zuletzt gewählte Wert $x_{vi[t-1]}$ eingeht.

Wird die logarithmierte Zustandssumme Z partiell nach der latenten Variable η_v differenziert, so erhalten wir den Erwartungswert der Scoring-Funktion $f_1(x)$, also den Erwartungswert der absoluten Differenz:

$$\begin{aligned}
\frac{\partial logZ}{\partial \eta_v} &= \left\langle |x_{vi[t]} - x_{vi[t-1]}| \right\rangle |x_{vi[t-1]} \\
&= \frac{\sum_{l'=1}^{m} |l' - x_{vi[t-1]}| \cdot \exp\left\{|l' - x_{vi[t-1]}| \cdot \eta_v - \beta_{il'}\right\}}{\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v - \beta_{il}\right\}} \\
&= \sum_{l'=1}^{m} \left[|l' - x_{vi[t-1]}| \cdot \frac{\exp\left\{|l' - x_{vi[t-1]}| \cdot \eta_v - \beta_{il'}\right\}}{\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v - \beta_{il}\right\}} \right] \qquad (4.8) \\
&= \sum_{l'=1}^{m} \left[|l' - x_{vi[t-1]}| \cdot p(X_{vi[t]} = l'|x_{vi[t-1]}) \right]
\end{aligned}$$

.

In obiger Gleichung kann über den Ausdruck

$$p(X_{vi[t]} = l'|x_{vi[t-1]}) = \frac{\exp\left\{|l' - x_{vi[t-1]}| \cdot \eta_v - \beta_{il'}\right\}}{\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v - \beta_{il}\right\}} \qquad (4.9)$$

die Wahrscheinlichkeit der Wahl der Kategorie l' ermittelt werden. Interessiert der Erwartungswert $\langle x_{vi} \rangle |x_{vi[t]}$ der manifesten Reaktion und nicht derjenige des Sprungs, so kann in Gleichung 4.8 $|l' - x_{vi[t-1]}|$ durch l' ersetzt werden, da Gleichung 4.9 die Wahrscheinlichkeit der Wahl einer bestimmten Kategorie l' beschreibt, welche der Wahrscheinlichkeit $p(X_{vi[t]} = x_{vi[t]}|x_{vi[t-1]})$ entspricht. Also:

$$\begin{aligned}
\left\langle X_{vi[t]} = x_{vi[t]} \right\rangle |x_{vi[t-1]} &= \frac{\sum_{l'=1}^{m} l' \cdot \exp\left\{|l' - x_{vi[t-1]}| \cdot \eta_v - \beta_{il'}\right\}}{\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v - \beta_{il}\right\}} \\
&= \sum_{l'=1}^{m} \left[l' \cdot \frac{\exp\left\{|l' - x_{vi[t-1]}| \cdot \eta_v - \beta_{il'}\right\}}{\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v - \beta_{il}\right\}} \right] \qquad (4.10) \\
&= \sum_{l'=1}^{m} \left[l' \cdot p(X_{vi[t]} = l'|x_{vi[t-1]}) \right].
\end{aligned}$$

Tabelle 4.2.: Mögliche manifeste Reaktion $x_{vi[t]}$, absolute Differenzen $|x_{vi[t]} - x_{vi[t-1]}|$ und Wahrscheinlichkeiten der Wahl einer Kategorie $x_{vi[t]}$ bei zuletzt gewähltem Wert $x_{vi[t-1]} = 1$ und 4 Kategorien. ($\eta_v = -1.5$ und $\boldsymbol{\beta} = [-0.2, 0.3, 0.3, -0.4]$)

$x_{vi[t]}\|1$	$(\|x_{vi[t]} - 1\|)\|1$	$p(x_{vi[t]}\|1)$
1	$\|1-1\| = 0$	.69
2	$\|2-1\| = 1$	.25
3	$\|3-1\| = 2$	.06
4	$\|4-1\| = 3$	.01

Dieses Vorgehen ist vielleicht nicht unmittelbar zu verstehen, daher sei es näher erläutert. Tabelle 4.2 zeigt die möglichen manifesten Reaktionen ($x_{vi[t]}|1$), die möglichen absoluten Differenzen $|x_{vi[t]} - 1|$ und deren Wahrscheinlichkeiten bei zuletzt gewählter Kategorie $x_{vi[t-1]} = 1$ bei einem vier-kategoriellen Item. Nach einer Standardregel der Wahrscheinlichkeitstheorie gilt:

$$\langle f(x) \rangle = \sum_{l=1}^{m} f(x_l) \cdot p(x_l). \tag{4.11}$$

Also ergibt sich der Erwartungswert der absoluten Differenz bei gegebenem $x_{vi[t-1]}$ zu:

$$\left\langle |x_{vi[t]} - x_{vi[t-1]}| \right\rangle |1 = 0 \cdot 0.69 + 1 \cdot 0.25 + 2 \cdot 0.06 + 3 \cdot 0.01 = 0.4. \tag{4.12}$$

Das Vorgehen hier entspricht demjenigen, das durch Gleichung 4.8 ausgedrückt wird. Mit jeder einzelnen absoluten Differenz geht natürlich auch eine Ausprägung von $x_{vi[t]}$ einher, welche über die Scoring-Funktion mit den absoluten Differenzen verknüpft ist. In Tabelle geht ein $x_{vi[t]}$ von 1 mit einem Sprungbetrag von Null einher, ein $x_{vi[t]}$ von 2 geht mit einem Sprungbetrag von 1 einher, u.s.w. In der dritten Spalte der Tabelle 4.2 sehen wir die Wahrscheinlichkeiten der Wahl einer Kategorie unter dem Modell. Interessiert nun der Erwartungswert der manifesten Reaktion, und nicht derjenige der Scoring-Funktion so kann dieser Erwartungswert wie folgt berechnet werden:

$$\left\langle x_{vi[t]} \right\rangle |1 = 1 \cdot 0.69 + 2 \cdot 0.25 + 3 \cdot 0.06 + 4 \cdot 0.01 = 1.41. \tag{4.13}$$

Dieses Vorgehen entspricht der Bedeutung von Gleichung 4.10. Pro möglicher, zuletzt gewählter Kategorie $x_{vi[t-1]}$ existiert also je eine erwartete Antwort $\langle x_{vi[t]} \rangle | x_{vi[t-1]}$, da durch das Modell m bedingte Wahrscheinlichkeitsverteilungen modelliert werden, wobei m der Anzahl der wählbaren Kategorien entspricht. $x_{vi[t-1]}$ ist keine Variable, sondern ein fester Wert (Konstante), der in die Scoring-Funktion eingeht.

Die Gleichung 4.8 kann zur Parameterschätzung verwendet werden, da sie die erwartete absolute Differenz unter dem Modell abbildet. Wird diese erwartete absolute Differenz in Beziehung zu den beobachteten Differenzen gesetzt, so ergibt sich die Komponente des Gradienten der Likelihood-Funktion die mit der Schätzung des Parameters η_v korrespondiert. Hierauf wird bei der Herleitung der Likelihood des Modells genauer eingegangen.

Gleichung 4.10 erlaubt die Berechnung des erwarteten Wertes der manifesten Reaktion $x_{vi[t]}$. Die Gleichung kann zur Testung der Modellpassung genutzt werden, indem die Differenz der beobachteten, manifesten Reaktion $x_{vi[t]}$ und der unter dem Modell erwarteten manifesten Reaktion gebildet wird. Diese Differenz kann an der erwarteten Varianz der Reaktion bei Modellgeltung standardisiert werden, so dass sich z-Werte, bzw. standardisierte Residuen ergeben, über die sich die Modellpassung bewerten lässt.

Wird die logarithmierte Zustandssumme Z zwei mal partiell nach der latenten Variable η_v differenziert, so resultiert die Varianz der manifesten absoluten Differenzen bei Modellgeltung, bzw. die Varianz der Scoring-Funktion $f_1(x)$ unter dem Modell:

$$var(|x_{vi[t]} - x_{vi[t-1]}|)|x_{vi[t-1]} = \frac{\sum_{l'=1}^{m} |l' - x_{vi[t-1]}|^2 \cdot \exp\left\{|l' - x_{vi[t-1]}| \cdot \eta_v - \beta_{il'}\right\}}{\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v - \beta_{il}\right\}} - \left[\frac{\sum_{l'=1}^{m} |l' - x_{vi[t-1]}| \cdot \exp\left\{|l' - x_{vi[t-1]}| \cdot \eta_v - \beta_{il'}\right\}}{\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v - \beta_{il}\right\}}\right]^2. \tag{4.14}$$

Auch hier existieren wiederum m bedingte, erwartete Varianzen bei Modellgeltung, wie man sich leicht überzeugen kann, wenn der Ausdruck für jedes mögliche $x_{vi[t-1]}$ berechnet wird. Gleichung 4.14 ist mit der Definition der Varianz auf Basis von Erwartungswerten kompatibel:

$$\langle x^2 \rangle - \langle x \rangle^2 = var(x). \tag{4.15}$$

In Gleichung 4.14 entspricht der linke Term der Differenz dem Erwartungswert der quadrierten Scoring-Funktion $f_1(x)$, der rechte Term entspricht dem quadrierten Erwartungswert unter dem Modell. Die Varianz der *manifesten Reaktion* (nicht diejenige der absoluten Differenz) kann über folgenden Ausdruck errechnet werden:

$$var(x_{vi[t]})|x_{vi[t-1]} = \frac{\sum_{l'=1}^{m} l'^2 \cdot \exp\left\{|l' - x_{vi[t-1]}| \cdot \eta_v - \beta_{il'}\right\}}{\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v - \beta_{il}\right\}} - \left[\frac{\sum_{l=1}^{m} l' \cdot \exp\left\{|l' - x_{vi[t-1]}| \cdot \eta_v - \beta_{il'}\right\}}{\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v - \beta_{il}\right\}}\right]^2. \tag{4.16}$$

Bei diesen Berechnungen ist zu beachten, dass es sich bei $x_{vi[t-1]}$ um *keine* Variable handelt, sondern um eine bekannte Konstante in der Funktion $f_1(x)$. Die Zustandssumme läuft über alle möglichen Zustände von $x_{vi[t]}$ (1 bis m) und nicht über alle möglichen *gemeinsamen* Zustände von $x_{vi[t]}$ *und* $x_{vi[t-1]}$. Eine Wahl, die im Rahmen des Maximum-Entropie-Formalismus allerdings machbar wäre. Es würden in diesem Fall die gemeinsame Wahrscheinlichkeitsdichte (*joint density*) $p(x_{vi[t]}, x_{vi[t-1]})$ modelliert werden.

Die Erwartungswerte der manifesten Antworten unter dem Modell und die erwartete Varianzen können dazu genutzt werden, um standardisierte Residuen des Modells bezüglich der manifesten Antwort $x_{vi[t]}$ zu berechnen. Das bedingte, standardisierte Residuum $z_{vi[t]}|x_{vi[t-1]}$ einer manifesten Antwort $x_{vi[t]}|x_{vi[t-1]}$ ist:

$$z_{vi[t]}|x_{vi[t-1]} = \frac{x_{vi[t]}|x_{vi[t-1]} - \left\langle x_{vi[t]} \right\rangle |x_{vi[t-1]}}{\sqrt{\mathrm{var}(x_{vi[t]})|x_{vi[t-1]}}}. \tag{4.17}$$

4.7. Die Likelihood-Funktion und suffiziente Statistiken

Von der psychometrischen Bedeutung her bildet das Modell den Antwortprozess einer Person v auf ein Item i probabilistisch ab, wobei die Wahrscheinlichkeit der Wahl einer bestimmten Antwortkategorie einerseits von der latenten Variable η_v, dem zuletzt gewählten Wert $x_{vi[t-1]}$ und den Kategorien-Parametern β_{ix} abhängt. Die Parameter müssen aus beobachteten Daten geschätzt werden. Die Bedeutung der Parameter und den Bezug zum Konstrukt intraindividuelle Variabilität in Form der absoluten sukzessiven Differenz kann illustriert werden, indem die Likelihood-Funktion des Modells in klassischer Weise

dargestellt wird. Unter der Annahme der stochastischen Unabhängigkeit der manifesten Antworten gilt:

$$L = \prod_{v=1}^{N}\prod_{i=1}^{k} P(X_{vi[1]} = x_{vi[1]}) \prod_{v=1}^{N}\prod_{i=1}^{k}\prod_{t=2}^{T} p(X_{vi[t]} = x_{vi[t]}|x_{vi[t-1]}). \tag{4.18}$$

Der dreidimensionale Datenraum besteht aus Messzeitpunkten t, Personen v und Items i. Es werden pro Person und Item je eine Markov-Kette definiert, deren Übergangswahrscheinlichkeiten von den Modellparametern abhängen. Die Wahrscheinlichkeiten $P(X_{vi[1]} = x_{vi[1]})$ repräsentieren die Wahrscheinlichkeiten der Wahl einer Kategorie für Person v und Item i zu Messzeitpunkt 1. Allerdings werden durch das hier definierte Modell keine Wahrscheinlichkeiten für den ersten Messzeitpunkt geschätzt, vielmehr geht die erste Beobachtung $x_{vi[1]}$ als konstanter, beobachteter Wert in das Modell ein, auf dessen Basis die Wahrscheinlichkeiten zum folgenden Zeitpunkt $t = 2$ bei Kenntnis der Parameter geschätzt werden können. Von daher wird in den folgenden Rechnungen das erste Doppelprodukt in der Likelihood der Einfachheit der Darstellung wegen nicht berücksichtigt. Die weiter Produktbildung der Likelihood startet bezogen auf die Zeit bei $t = 2$. Die vereinfachte Likelihood, die nur die tatsächlich modellierten Daten berücksichtigt, ergibt sich zu:

$$L = \prod_{v=1}^{N}\prod_{i=1}^{k}\prod_{t=2}^{T} p(X_{vi[t]} = x_{vi[t]}|x_{vi[t-1]}). \tag{4.19}$$

Es sei angemerkt, dass potentiell unterschiedliche Längen von Beobachtungen pro Person kein Problem bei der Berechnung der Likelihood darstellen. Wird die logarithmierte Likelihood-Funktion gebildet, so erhalten wir:

$$\begin{aligned} logL &= \sum_{v=1}^{N}\sum_{i=1}^{k}\sum_{t=2}^{T} |x_{vi[t]} - x_{vi[t-1]}| \cdot \eta_v + \sum_{v=1}^{N}\sum_{i=1}^{k}\sum_{t=2}^{T} \beta_{ix} \qquad (4.20)\\ &\quad - \sum_{v=1}^{N}\sum_{i=1}^{k}\sum_{t=2}^{T} \log\left[\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v + \beta_{il}\right\}\right] \\ &= \sum_{v=1}^{N} \eta_v \left[\sum_{i=1}^{k}\sum_{t=2}^{T} |x_{vi[t]} - x_{vi[t-1]}|\right] + \sum_{i=1}^{k}\sum_{x}^{m} n_{ix} \cdot \beta_{ix} \qquad (4.21)\\ &\quad - \sum_{v=1}^{N}\sum_{i}^{k}\sum_{t=2}^{T} \log\left[\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v + \beta_{il}\right\}\right]. \end{aligned}$$

An dem Summand

$$\sum_{v=1}^{N} \eta_v \left[\sum_{i=1}^{k} \sum_{t=2}^{T} |x_{vi[t]} - x_{vi[t-1]}| \right] \tag{4.22}$$

zeigt sich, dass die Summe der absoluten Differenzen $|x_{vi[t]} - x_{vi[t-1]}|$ einer Person über alle Items k und Beobachtungszeitpunkte T hinweg eine suffiziente Statistik zur Schätzung des Parameters η_v darstellt. Wird die log-Likelihood partiell nach η_v differenziert, so erhalten wir:

$$\frac{\partial logL}{\partial \eta_v} = \sum_{i=1}^{k} \sum_{t=2}^{T} |x_{vi[t]} - x_{vi[t-1]}| - \sum_{i=1}^{k} \sum_{t=2}^{T} \frac{\partial}{\partial \eta_v} \log \left[\sum_{l=1}^{m} \exp \left\{ |l - x_{vi[t-1]}| \cdot \eta_v + \beta_{il} \right\} \right]. \tag{4.23}$$

Nullsetzen und umstellen ergibt:

$$\begin{aligned} \sum_{i=1}^{k} \sum_{t=2}^{T} |x_{vi[t]} - x_{vi[t-1]}| &= \sum_{i=1}^{k} \sum_{t=2}^{T} \frac{\partial}{\partial \eta_v} \log \left[\sum_{l=1}^{m} \exp \left\{ |l - x_{vi[t-1]}| \cdot \eta_v + \beta_{il} \right\} \right] \\ &= \sum_{i=1}^{k} \sum_{t=2}^{T} \frac{\partial}{\partial \eta_v} \log Z \end{aligned} \tag{4.24}$$

Gleichung 4.24 ist die Komponente des Gradienten der Likelihood, welcher die Schätzung von η_v gewährleistet. Ein interessanter Punkt an dieser Stelle ist, dass sich auf der rechten Seite der Gleichung Ableitungen der logarithmierten Zustandssumme Z des Modelles ergeben. Dies ist mit dem Maximum-Entropie-Formalismus kompatibel (vgl. Gleichung 3.52), da die partielle Ableitung der Zustandssumme nach einem bestimmten Parameter den Erwartungswert der korrespondierenden Funktion unter dem Modell ergibt. Bezüglich des Parameters η_v ist zu erkennen, dass die Summe der absoluten Differenzen in einer beobachteten Datenmatrix suffiziente Statistiken zur Parameterschätzung darstellen.

Inhaltlich bedeutet dies: zeigt eine Person relativ hohe Sprünge in den manifesten Itemantworten, so ist auch der aus den entsprechenden Daten geschätzte Parameter des Modells hoch. Zudem werden keine weiteren Informationen aus den Daten benötigt, als die Summe der absoluten Differenzen einer Person, um den Parameter η_v zu schätzen. Die Bedeutung der Kategorien-Parameter ist aus der Likelihood-Funktion nicht unmittelbar ersichtlich. Aber es ist möglich, sich deren Eigenschaften zu verdeutlichen, indem man sich eines Tricks bedient, den auch Andersen (1995a) bei der Darstellung der Likelihood-Funktion des Rasch-Modells verwendete.

Zur Ermittelung der suffizienten Statistiken zur Schätzung der Kategorien-Parameter β_{ix} kann ein Selektionsvektor s_{vitx} eingeführt werden (Andersen, 1995a, p. 275). Es gilt

$s_{vitx} = 1$, sofern Person v zum Zeitpunkt t Kategorie x auf Item i gewählt hat, ansonsten ist $s_{vitx} = 0$. Das Prinzip entspricht einer Dummy-Codierung, wie sie z.B. im Allgemeinen Linearen Modell zur Kodierung von Faktorstufen Anwendung findet. Also:

$$logL = \sum_{v=1}^{N}\sum_{i=1}^{k}\sum_{t=2}^{T}|x_{vi[t]} - x_{vi[t-1]}| \cdot \eta_v + \sum_{v=1}^{N}\sum_{i=1}^{k}\sum_{t=2}^{T}\sum_{x=1}^{m} s_{vitx}\beta_{ix} \tag{4.25}$$

$$- \sum_{v=1}^{N}\sum_{i=1}^{k}\sum_{t=2}^{T}\log\left[\sum_{l=1}^{m}\exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v - \beta_{il}\right\}\right]. \tag{4.26}$$

Die Doppelsumme $\sum_{v=1}^{N}\sum_{t=2}^{T} s_{vitx}$ entspricht der Häufigkeit der Wahl der Kategorie x auf Item i über alle Messzeitpunkte t und alle Personen v:

$$n_{ix} = \sum_{v=1}^{N}\sum_{t=2}^{T} s_{vitx}. \tag{4.27}$$

Einen ähnlichen Trick wendete auch Rost an (Rost, 2004, p.212). Durch Einsetzen und algebraisches Umformen folgt:

$$logL = \sum_{v=1}^{N}\eta_v\sum_{i=1}^{k}\sum_{t=2}^{T}|x_{vi[t]} - x_{vi[t-1]}| + \sum_{i=1}^{k}\sum_{x}^{m} n_{ix} \cdot \beta_{ix} \tag{4.28}$$

$$- \sum_{v=1}^{N}\sum_{i=1}^{k}\sum_{t=2}^{T}\log\left[\sum_{l=1}^{m}\exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v + \beta_{il}\right\}\right].$$

Die Häufigkeit n_{ix} der Wahl der Kategorie x auf Item i ist eine suffiziente Statistik zur Schätzung der Kategorien-Parameter β_{ix}. Dies bedeutet inhaltlich: je häufiger die Kategorie x eines Items i in einem Datensatz gewählt wurde, desto größer ist der Kategorien-Parameter β_{ix} des entsprechenden Items. Diese Parameter sind also als Kategorien-Leichtigkeiten, wie auch schon Mair und Hatzinger (Mair & Hatzinger, 2007) für das Partial-Credit-Modell feststellen.

Es ist zu vermuten, dass auch in diesem Fall sich durch das partielle Differenzieren der logarithmierten Zustandssumme des Modells nach den Kategorien-Parametern entsprechende Schätzgleichungen aus dem Maximum-Entropie-Formalismus replizieren lassen:

$$\frac{\partial logL}{\partial \beta_{ix}} = n_{ix} - \sum_{v=1}^{N}\sum_{t=2}^{T} \frac{\partial}{\partial \beta_{ix}} \log\left[\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v + \beta_{il}\right\}\right] \tag{4.29}$$

$$= n_{ix} - \sum_{v=1}^{N}\sum_{t=2}^{T} \frac{\exp\left\{|l' - x_{vi[t-1]}| \cdot \eta_v + \beta_{il'}\right\}}{\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v + \beta_{il}\right\}}, \tag{4.30}$$

wobei $l' = x$, da beim Differenzieren alle Summanden im Nenner entfallen, in denen die entsprechende Kategorie x nicht gewählt wurde. Aus der obigen Gleichung lässt sich schließen, dass die Häufigkeit der Wahl einer Kategorie eine suffiziente Statistik zur Schätzung der Kategorien-Parameter darstellt. Die rechte Seite der Differenz ist die erwartete Häufigkeit der Kategorienwahl auf einem Item über alle Messzeitpunkte und alle Personen unter dem Modell. Der Ausdruck

$$\frac{\exp\left\{|l' - x_{vi[t-1]}| \cdot \eta_v + \beta_{il'}\right\}}{\sum_{l=1}^{m} \exp\left\{|l - x_{vi[t-1]}| \cdot \eta_v + \beta_{il}\right\}} \tag{4.31}$$

entspricht der Wahrscheinlichkeit der Wahl der Kategorie l', bzw. x, wie leicht anhand der Modellgleichung überprüft werden kann. Die Notation, die sich besonders an Fischer & Molenaar (1995) orientiert, ist allerdings nicht einfach, da eine Unmenge von Summen anfallen und die implizite Struktur des Modells oder der Rasch-Modelle allgemein nicht unbedingt deutlich wird. Es muss für die Darstellung auf allerhand Kunstgriffe, wie z.B. die Einführung eines Selektionsvektors zurückgegriffen werden. Zudem ist die Koppelung des Laufindex für die Kategorien l mit der Scoring-Funktion im Nenner der Modelle nicht empfehlenswert, da diese Koppelung sich beim partiellen Differenzieren im Zähler repliziert. Von daher wäre es günstig, die Modelle unter Verwendung von Vektor und Matrix-Schreibweise umzuformen, um die Darstellung kompakter zu gestalten und auch die Rezeption zu erleichtern. Diese Umformulierung ist relativ problemlos möglich, wird hier allerdings nicht mehr vorgenommen werden.

Detaillierte Ausführungen zur Schätzung von Modellparametern bei Rasch-Modellen mit klassischen Methoden finden sich bei Fischer (1974) und Fischer & Molenaar (1995), wobei hier besonders die Kapitel 3 und 15 hervorzuheben sind. In der vorliegenden Arbeit wird auf die MCMC-Methode zur Parameterschätzung zurückgegriffen.

4.8. Der Zusammenhang zwischen dem Personen-Parameter und der manifesten Statistik

Hinsichtlich des Parameters η_v ist zu konstatieren, dass die Summe der absoluten Differenzen in einer Zeitreihe eine suffiziente Statistik zur Schätzung des Parameters darstellt. Dementsprechend ist zu vermuten, dass die Ausprägung des Parameters η_v in einem geordnet monotonen Verhältnis zur mittleren absoluten Differenz (MASD) in einer manifesten Zeitreihe steht. Es ist zweckmäßig, diese Annahme zu überprüfen. Dies geschieht hier auf simulative Art: Es werden Zeitreihen der Länge n bei gegebenen Parametern η_v und $\boldsymbol{\beta}$ simuliert, wobei η_v pro Zeitreihe über den Skalenbereich von $\eta_v = -3$ bis $\eta_v = 3$ erhöht wird. Zudem wird die mittlere absolute Differenz der erzeugten Zeitreihe berechnet und der Parameter η_v und die mittlere absolute Differenz werden grafisch in Beziehung gesetzt.

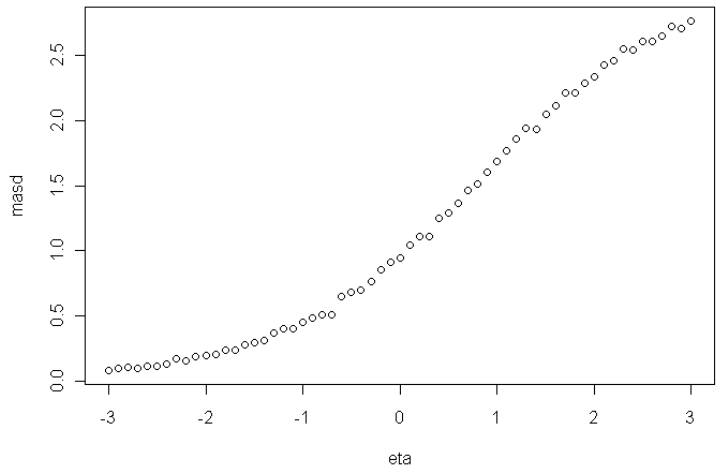

Abbildung 4.2.: Zusammenhang zwischen η_v und der MASD für ein Item mit $\boldsymbol{\beta} = [-0.5, 0.5, 0.5, -0,5]$ auf Basis von Zeitreihen der Länge $n = 1000$.

Abbildung 4.2 zeigt den Zusammenhang zwischen dem generierenden Parameter η_v und der mittleren absoluten Differenz der generierten, manifesten Zeitreihen der Länge

$n = 1000$. Auf der Abszisse sind die generierenden Parameter abgetragen, die Ordinate beschreibt die mittlere absolute Differenz der jeweiligen generierten Zeitreihe. Es zeigt sich der erwartete monotone Zusammenhang zwischen den generierenden Parametern η_v und den mittleren absoluten Differenzen der generierten Zeitreihen. Der Zusammenhang ist von sigmoider Gestalt. Der monotone Zusammenhang zeigt sich auch bei anderen Konfigurationen der Kategorien-Parameter $\boldsymbol{\beta}$, allerdings hängt die konkrete Form von der Ausprägung der Schwellenparameter ab.

4.9. Die Logits der Kategorien-Wahrscheinlichkeiten

Zur weiteren Verdeutlichung der Bedeutung des Modells ist es zweckmäßig, die Logits der benachbarten Kategorien-Wahrscheinlichkeiten näher zu untersuchen. Der Logit der Wahl der Kategorie x gegenüber der Kategorie $x+1$ ist nach der Modellgleichung wie folgt definiert:

$$\begin{aligned} \log\left(\frac{p(x_{vit}+1|x_{vi[t-1]})}{p(x_{vit}|x_{vi[t-1]})}\right) &= f(x_{vit}+1)\cdot\eta_v + \beta_{i[x+1]} - (f(x_{vit})\cdot\eta_v + \beta_{ix}) \quad (4.32)\\ &= (f(x_{vit}+1) - f(x_{vit}))\cdot\eta_v + \beta_{i[x+1]} - \beta_{ix}, \end{aligned}$$

wobei

$$f(x_{vit}+1) = |(x_{vit}+1) - x_{vi[t-1]}| \quad (4.33)$$

und

$$f(x_{vit}) = |x_{vit} - x_{vi[t-1]}|. \quad (4.34)$$

Es zeigt sich, dass die Logits der benachbarten Kategorien in der obigen Definition in einem linearen Zusammenhang mit dem Parameter η_v stehen. Die Richtung des Zusammenhangs hängt von der Differenz $f(x_{vit}+1) - f(x_{vit})$ ab. Diese Differenz kann immer nur 1 oder -1 sein. Das Vorzeichen hängt von den konkreten Ausprägungen von x_{vit} und $x_{vi[t-1]}$ ab. Die Bedeutung dessen sei an einem Beispiel verdeutlicht. Angenommen $x_{vi[t-1]}$ ist 1. Dann gilt für den Logit der Wahrscheinlichkeit der Wahl der Kategorie $x_{vit} = 1$ gegenüber der Kategorie $x_{vit}+1 = 2$:

$$\log\left(\frac{p(2|1)}{p(1|1)}\right) = \eta_v + \beta_{i[x+1]} - \beta_{ix}. \quad (4.35)$$

Dies bedeutet, dass der Logit eine lineare Funktion des Parameters η_v und der Schwellenparameter ist. Anders ausgedrückt, je höher die Variabilität einer Person, erfasst durch den Parameter η_v, desto höher die Wahrscheinlichkeit der Wahl der Kategorie 2 gegenüber der Kategorie 1, sofern die zuvor gewählte Kategorie $x_{vi[t-1]} = 1$ ist.

Nehmen wir nun an, der zuletzt gewählte Wert $x_{vi[t-1]}$ ist 4, so zeigt sich folgender Sachverhalt:

$$\log\left(\frac{p(2|4)}{p(1|4)}\right) = -\eta_v + \beta_{i[x+1]} - \beta_{ix}. \tag{4.36}$$

Dies bedeutet, dass die Wahrscheinlichkeit der Wahl der Kategorie 2 gegenüber der Kategorie 1 mit steigendem η_v *sinkt*, sofern die zuvor gewählte Kategorie $x_{vi[t-1]} = 4$ ist. Dies ist inhaltlich stimmig, da der Sprung von 4 nach 1 eine höhere Variabilität in Form der absoluten Differenz beinhaltet, als der Sprung von 4 nach 2. Die Differenz der Funktionen $\Delta = (f(x_{vit} + 1) - f(x_{vit}))$ adjustiert somit das Vorzeichen der Regression zur Berechnung der Logits in Abhängigkeit der zuvor gewählten Kategorie.

Gilt

$$\eta_v = \frac{-\beta_{i[x+1]} + \beta_{ix}}{f(x_{vit} + 1) - f(x_{vit})}, \tag{4.37}$$

so ist die Wahl der Kategorie x genauso wahrscheinlich, wie die Wahl der Kategorie $x + 1$, da an diesem Ort der Trait-Skala der Logit der benachbarten Kategorien-Wahrscheinlichkeiten Null ist und somit die Wahl jeder Kategorie gleich wahrscheinlich ist. Daher lassen sich aus Gleichung 4.38 die Schwellenparameter $\tau_{x[x+1]}$ des Modells berechnen:

$$\tau_{x[x+1]} = \frac{-\beta_{i[x+1]} + \beta_{ix}}{f(x_{vit} + 1) - f(x_{vit})}. \tag{4.38}$$

Die Schwellenparameter liegen an den Schnittpunkten der Kategorien-Funktionen der benachbarten Kategorien. Die Anwendung von Gleichung 4.38 sei an einem Beispiel verdeutlicht. Nehmen wir an, die Kategorien-Parameter eines vier-kategoriellen Items sind $\beta = [-0.5, 0.5, 0.5, -0.5]$. Nehmen wir ferner an es gilt $x_{vi[t-1]} = 2$, d.h. die zuletzt

gewählte Kategorie ist 2. Aus Gleichung 4.38 folgt:

$$\tau_{[12]} = \frac{-0.5 - 0.5}{|2-2| - |1-2|} = 1 \tag{4.39}$$

$$\tau_{[23]} = \frac{-0.5 + 0.5}{|3-2| - |2-2|} = 0 \tag{4.40}$$

$$\tau_{[34]} = \frac{0.5 + 0.5}{|4-2| - |3-2|} = 1. \tag{4.41}$$

Wie man sich anhand Diagramm B in Abbildung 4.1 leicht überzeugen kann, sind dies die Schnittpunkte der Kategorien-Funktionen. Beachtenswert ist, dass für den Fall $x_{vi[t]} = 2$ die Schnittpunkte nicht geordnet sind, was aber nicht bedeutet, dass dem Modell keine Messstruktur zugrunde liegt, da das Modell spezifisch objektive Vergleiche der Personen hinsichtlich der Variabilität auf einer Logit-Skala erlaubt, wie im Folgenden gezeigt wird.

Die Logit-Formulierung erlaubt ebenfalls die Untersuchung des Modells hinsichtlich der spezifischen Objektivität. Werden die Differenzen der Logits zweier Personen v und w gebildet und werden jeweils zwischen den Personen homologe Kategorien betrachtet ($x_{vit} = x_{wit}$ und $x_{vit} + 1 = x_{wit} + 1$), so zeigt sich folgender Sachverhalt:

$$\log\left(\frac{p(x_{vit}+1|x_{vi[t-1]})}{p(x_{vit}|x_{vi[t-1]})}\right) - \log\left(\frac{p(x_{wit}+1|x_{wi[t-1]})}{p(x_{wit}|x_{wi[t-1]})}\right) = \tag{4.42}$$

$$[f(x_{vit}+1) - f(x_{vit})] \cdot \eta_v + \beta_{i[x+1]} - \beta_{ix} - \left\{[f(x_{wit}+1) - f(x_{wit})] \cdot \eta_w + \beta_{i[x+1]} - \beta_{ix}\right\} =$$

$$[f(x_{vit}+1) - f(x_{vit})]\,(\eta_v - \eta_w),$$

da $x_{vit} = x_{wit}$ und $x_{vit} + 1 = x_{wit} + 1$ gilt. Daraus folgt:

$$\frac{\log\left(\frac{p(x_{vit}+1|x_{vi[t-1]})}{p(x_{vit}|x_{vi[t-1]})}\right) - \log\left(\frac{p(x_{wit}+1|x_{wi[t-1]})}{p(x_{wit}|x_{wi[t-1]})}\right)}{[f(x_{vit}+1) - f(x_{vit})]} = \eta_v - \eta_w. \tag{4.43}$$

Dieses Ergebnis bedeutet, dass die Differenz $\eta_v - \eta_w$ der Parameter der Differenz der Logits entspricht, wobei diese Differenz entweder ein positives oder ein negatives Vorzeichen nach Maßgabe der Funktionen im Nenner und damit der zuvor gewählten Kategorie erhält. Praktisch bedeutet dies, dass der Vergleich der Personen-Parameter hinsichtlich der Variabilität auf einer Differenzenskala erfolgt und unabhängig von den Kategorien-Parametern ist. Das Modell erlaubt also spezifisch objektive Vergleiche und ist somit ein

Rasch-Modell. Die Richtung des Vergleichs wird automatisch nach Maßgabe der zuletzt gewählten Kategorie durch die Differenz der Funktionen im Nenner gewährleistet.

Da die Bedeutung von Gleichung 4.43 nicht unbedingt einfach zu verstehen ist, sei diese anhand eines Beispiels erläutert. Nehmen wir an die Variabilität einer Person 1 wird durch den Parameter $\eta_v = -0.5$ und die Variabilität einer Person 2 wird durch den Parameter $\eta_w = -0.7$ beschrieben. Nehmen wir ferner an, dass die Personen auf ein Item mit den Kategorien-Parametern $\boldsymbol{\beta} = [-0.5, 0.5, 0.5, -0.5]$ Antworten und das die zuletzt gewählte Kategorie $x_{vi[t-1]} = x_{wi[t-1]} = 1$ war. Nach dem Modell sind die Kategorien-Wahrscheinlichkeiten $\boldsymbol{p}$ von Person v:

$$\boldsymbol{p}_v = [0.258, 0.426, 0.258, 0.058]. \tag{4.44}$$

Die Kategorien-Wahrscheinlichkeiten von Person w sind:

$$\boldsymbol{p}_w = [0.318, 0.430, 0.213, 0.039]. \tag{4.45}$$

Der Logit der Wahl der Kategorie 2 gegenüber Kategorie 1 sind für Person v nach Gleichung 4.32:

$$\log\left\{\frac{0.426}{0.258}\right\} = 0.5. \tag{4.46}$$

Der Logit der Wahl der Kategorie 2 gegenüber Kategorie 1 sind für Person w nach Gleichung 4.32:

$$\log\left\{\frac{0.430}{0.318}\right\} = 0.3. \tag{4.47}$$

Nach Gleichung 4.43 gilt:

$$1^{-1}\left[\log\left\{\frac{0.426}{0.258}\right\} - \log\left\{\frac{0.430}{0.318}\right\}\right] = \eta_v - \eta_w \tag{4.48}$$

$$0.5 - 0.3 = -0.5 - (-0.7) \tag{4.49}$$

$$0.2 = 0.2. \tag{4.50}$$

Diese Aussage ist offensichtlich wahr. Betrachten wir nun den interessanten Fall mit $x_{vi[t-1]} = 4$. Die Antwortwahrscheinlichkeiten von Person v sind:

$$\boldsymbol{p}_v = [0.058, 0.258, 0.426, 0.258]. \tag{4.51}$$

Die Antwortwahrscheinlichkeiten von Person w sind:

$$\boldsymbol{p}_w = [0.039, 0.213, 0.430, 0.318]. \tag{4.52}$$

Der Logit der Wahl der Kategorie 2 gegenüber Kategorie 1 sind für Person 1:

$$\log\left\{\frac{0.258}{0.058}\right\} = 1.5. \tag{4.53}$$

Der Logit der Wahl der Kategorie 2 gegenüber Kategorie 1 sind für Person 2:

$$\log\left\{\frac{0.213}{0.039}\right\} = 1.7. \tag{4.54}$$

Nach Gleichung 4.43 gilt:

$$\begin{aligned} -1^{-1}\left[\log\left\{\frac{0.258}{0.058}\right\} - \log\left\{\frac{0.213}{0.039}\right\}\right] &= \eta_v - \eta_w & (4.55)\\ -(1.5 - 1.7) &= -0.5 - (-0.7) & (4.56)\\ 0.2 &= 0.2. & (4.57) \end{aligned}$$

Diese Aussage ist offensichtlich ebenfalls wahr. Dies bedeutet, dass das Modell spezifisch objektive Vergleiche der Personen hinsichtlich ihrer Variabilität auf einer Logit-Skala erlaubt und eine sinnvolle Differenzenbildung der Parameter möglich ist. Hierbei wird nach Maßgabe der zuletzt gewählten Kategorie die Richtung des Vergleichs automatisch durch die Form des Modells adjustiert.

4.10. Simulative Evaluation des Modells auf Bias und Varianz der Schätzer

Nachdem die grundlegenden Eigenschaften des Modells bestimmt sind, ist die Überprüfung des Bias und der empirischen Varianz der Parameterschätzer von Interesse. Hier wird eine simulatives Vorgehen gewählt. Simulationsstudien bieten sich an, um formal abgeleitete Modelleigenschaften auf ihre Richtigkeit zu überprüfen. In Simulationsstudien zur Bestimmung des empirischen Bias und der empirischen Varianz der Parameterschätzer werden Daten bei bekannten Parametern aus dem Modell generiert und in einem zweiten

Schritt werden die Parameter aus den generierten Daten geschätzt. Die Bias der Parameterschätzer und deren Varianz lässt sich anhand der Verteilung der Parameterschätzer evaluieren. Zur Schätzung der Parameter kommen eine Reihe von Methoden in Frage, z.B. die Maximum-Likelihood-Methode (vgl. Skondral & Rabe-Hesketh, 2004, Kap.7 für einen Überblick). In der vorliegenden Arbeit wird die Monte-Carlo-Markov-Chain-Methode (MCMC-Methode) zur Bestimmung der Posterior-Verteilungen der Parameter verwendet, da diese relativ einfach zu implementieren ist und andererseits weiterer Vorteile, vor allem was die Bewertung der Modellpassung angeht, bietet. Da die MCMC-Methode in der Psychologie bisher wenig Anwendung findet, sei die Methode hier kurz dargestellt.

4.10.1. Parameterschätzung mit der MCMC-Methode

Die Anwendung der MCMC-Methode zur Bestimmung von sog. Posterior-Verteilungen von Modell-Parametern baut auf dem Theorem von Bayes auf und unterscheidet sich in mancherlei Hinsicht von dem klassischen Ansatz der Parameterschätzung nach der Maximum-Likelihood-Methode. Eine gute Einführung in die Materie bietet z.B. Lynch (2007). In diesem Buch wird auch die Maximum-Likelihood-Methode anschaulich dargestellt.

Während bei der Anwendung der Maximum-Likelihood-Methode die Likelihood-Funktion eines Modells in Abhängigkeit der Parameter maximiert wird, werden bei der Anwendung der MCMC-Methode zur Parameterschätzung die sog. Posterior-Verteilungen der Parameter auf Basis der Likelihood eines Modells und einer Prior-Verteilung der Parameter bestimmt. Der Begriff der Parameterschätzung ist in diesem Zusammenhang etwas irreführend, da in Rahmen eines Bayesianischen Ansatzes die Parameter nicht als fix und die Daten als zufällig aufgefasst werden, vielmehr verhält es sich umgekehrt, die Daten sind fix und den Parametern werden Verteilung unterstellt, die das „Wissen“ oder besser „Vorinformationen“ hinsichtlich der Lage der Parameter enkodieren. Auf die historischen Kontroversen, die mit diesem epistemischen Wahrscheinlichkeitsbegriff verbunden sind kann hier nicht eingegangen werden. Aus pragmatischen Gründen werden

die Mittelwerte der Posterior-Verteilungen der Parameter manchmal als frequentistische Punktschätzer und die Standardabweichungen der Posterior-Verteilungen als Standardfehler der Schätzer betrachtet.

Die MCMC-Methode der „Parameterschätzung“ basiert auf Bayes Theorem:

$$p(A|B) = p(B|A) \cdot \frac{p(A)}{p(B)}. \tag{4.58}$$

In dieser Gleichung stellen A und B Propositionen dar. Propositionen werden hier als Aussagen verstanden, wie z.B.:

- A: *Die Karte ist ein Ass.*
- B: *Die Karte ist rot.*

Nach Bayes Theorem lässt sich nun die Wahrscheinlichkeit $p(A|B)$, dass eine aus der Menge der roten Karten gezogene Karte ein Ass ist bestimmen, wenn folgende Wahrscheinlichkeiten bekannt sind:

- $p(B|A)$: Die Wahrscheinlichkeit, dass eine aus der Menge der Asse gezogene Karte rot ist.
- $p(A)$: Die Wahrscheinlichkeit, dass eine aus der Menge aller Karten gezogene Karte ein Ass ist.
- $p(B)$: Die Wahrscheinlichkeit, dass eine aus der Menge aller Karten gezogene Karte rot ist.

Der interessante Punkt, der den Kern der Bayesianischen Inferenz im Sinne der Bestimmung von Posterior-Verteilungen von Parametern eines Modells bei gegebenen Daten ausmacht ist derjenige, dass das Theorem verwendet werden kann, um die Posterior-Verteilungen von Modellparametern $\boldsymbol{\theta}$ angesichts beobachteter Daten $\mathbf{X}$ zu ermitteln. Die Posterior-Verteilungen enkodieren die verfügbare Information hinsichtlich der Lage von Parametern bei einem gegeben Modell und gegebenen Daten in Form von Verteilungen der Parameter, bzw. in Form von Dichten. Die folgenden Ausführungen lehnen sich

eng an Gill (2008) an.

Allgemein nimmt das Bayes-Theorem zur Bestimmung von Posterior-Verteilungen folgende Form an:

$$p(\boldsymbol{\theta}|\mathbf{X}) = L(\mathbf{X}|\boldsymbol{\theta}) \cdot \frac{p(\boldsymbol{\theta})}{p(\mathbf{X})}. \tag{4.59}$$

Das Theorem wird in diesem Fall nicht auf Einzelwahrscheinlichkeiten angewendet, sondern auf ganze Wahrscheinlichkeitsverteilungen, bzw. -dichten.

$p(\boldsymbol{\theta}|\mathbf{X})$ ist die Posterior-Verteilung der Parameter $\boldsymbol{\theta}$ angesichts beobachteter Daten $\mathbf{X}$, $p(\mathbf{X}|\boldsymbol{\theta})$ ist die Likelihood, welche von einem verwendeten Modell und den Daten abhängt, $p(\boldsymbol{\theta})$ ist die Prior-Verteilung der Parameter und $p(\mathbf{X})$ ist eine normalisierende Konstante, die gewährleistet, dass das Integral über den gesamten Wertebereich der Posterior-Verteilung 1 ergibt. Dementsprechend ist $p(\mathbf{X})$ wie folgt definiert:

$$p(\mathbf{X}) = \int_{\Theta} L(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \partial \boldsymbol{\theta}. \tag{4.60}$$

Die Posterior-Verteilung der Parameter ist also:

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\boldsymbol{\theta}) L(\mathbf{X}|\boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{\theta}) L(\mathbf{X}|\boldsymbol{\theta})}. \tag{4.61}$$

Zur Ermittelung der Posterior-Verteilung der Parameter ist es in obiger Gleichung notwendig, das Integral im Nenner zu bestimmen. Dies ist in vielen Fällen schwierig. Bei $p(\mathbf{X})$ handelt es sich um eine Konstante, die nicht von den Parametern abhängt. Daher kann zur Bestimmung der Posterior-Verteilung auf die normalisierende Konstante verzichtet werden, wodurch die Grundgleichung zur Bestimmung der Posterior-Verteilung der Parameter folgende, sehr viel einfachere Form annimmt:

$$p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\boldsymbol{\theta}) L(\mathbf{X}|\boldsymbol{\theta}). \tag{4.62}$$

Dies bedeutet, dass die Posterior-Wahrscheinlichkeit der Parameter proportional zum Produkt der Prior-Wahrscheinlichkeit und der Likelihood ist. Der klassische Maximum-Likelihood-Ansatz ist in dem Bayes'schen Formalismus enthalten, nämlich dann, wenn eine uninformative Prior-Verteilung verwendet wird. Im Bayes'schen Ansatz ist es darüber hinaus jedoch möglich, Vorinformationen über die Verteilung der Parameter durch

die Definition der Prior-Verteilung mit einzubeziehen.
Um die Posterior-Verteilungen der Parameter für ein konkretes Modell zu bestimmen ist es also nötig, Prior-Verteilungen der Parameter und eine Likelihood zu definieren. Die Likelihood baut auf der Modellgleichung und den beobachteten Daten $\mathbf{X}$ auf. Die Prior-Verteilungen werden so gewählt, dass sie entweder Vorinformationen hinsichtlich der Lage der Parameter enkodieren oder es werden diffuse Prior-Verteilungen verwendet, wenn keine Vorinformationen hinsichtlich der Lage der Parameter vorhanden sind. Wird z.B. für einen Parameter θ eine Prior-Verteilung aus der Klasse der Normalverteilungen definiert, deren Mittelwert 0 ist und deren Streuung sehr groß ist, so bringt diese Wahl zum Ausdruck, dass *a priori* wenig Informationen hinsichtlich der Lage der Parameter zur Verfügung stehen.

Um nun eine konkrete Posterior-Verteilung für einen Parameter zu erzeugen, ist es notwendig, Stichproben der Parameter aus der Posterior-Verteilungen zu ziehen, da die Posterior-Verteilungen nur in den wenigsten Fällen analytisch berechnet werden kann. Die Ziehung der Stichproben aus der Posterior-Verteilung erfolgt mit der MCMC-Methode. In der Regel kommen Gibb's Sampler oder der Metropolis-Hastings-Algorithmus (Metropolis et al., 1953) zum Einsatz. Die in dieser Arbeit verwendete Software WinBUGS (Lunn et al., 2000) verwendet Gibb's Sampler, von daher sei die Funktionsweise des Samplers kurz dargestellt. Eine detailliertere Darstellung der Methode findet sich unter anderem bei Gill (2008), p. 356; in Fox (2010), p. 64 und in Lynch (2007).

Um die Posterior-Verteilung der Parameter eines Modells zu bestimmen, wird ein Parametervektor $\boldsymbol{\theta}[0] = \theta_1^{[0]}, \theta_2^{[0]}, \ldots, \theta_k^{[0]}$ mit Startwerten für Gibb's Sampler erzeugt. Da die Likelihood auf Basis der beobachteten Daten und der Parameter und die Prior-Verteilung berechenbar sind, kann aus der Posterior-Verteilung ein Wert θ_1 gezogen werden, wobei für alle anderen Parameter die Startwerte eingesetzt werden.

Dieser gezogene Parameter θ_1 kann nun verwendet werden, um einen Parameter θ_2 aus der Posterior-Verteilung zu ziehen, wobei für die übrigen Parameter die Startwerte fungieren. Eine Ziehung für θ_3 erfolgt auf der Basis der zuvor gezogenen Werte θ_1 und θ_2, sowie der Startwerte und so fort. Nach einem ersten Zyklus sind die Startwerte vollständig

überschrieben und nun fungieren die jeweils zuletzt gezogenen Parameter als Werte der Posterior-Verteilung zur Ziehung eines Parameters $\theta_k^{[j]}$. Formal lässt sich dieser Prozess wie folgt darstellen:

$$
\begin{array}{lcllllll}
\theta_1^{[j]} & \sim & p(\theta_1| & \theta_2^{[j-1]}, & \theta_3^{[j-1]}, & \ldots, & \theta_{k-1}^{[j-1]}, & \theta_k^{[j-1]}, \quad \mathbf{X}\ ) \\
\theta_2^{[j]} & \sim & p(\theta_2| & \theta_1^{[j]}, & \theta_3^{[j-1]}, & \ldots, & \theta_{k-1}^{[j-1]}, & \theta_k^{[j-1]}, \quad \mathbf{X}\ ) \\
\theta_3^{[j]} & \sim & p(\theta_3| & \theta_1^{[j]}, & \theta_2^{[j]}, & \ldots, & \theta_{k-1}^{[j-1]}, & \theta_k^{[j-1]}, \quad \mathbf{X}\ ) \\
\vdots & & & & & & & \\
\theta_{k-1}^{[j]} & \sim & p(\theta_{k-1}| & \theta_1^{[j]}, & \theta_2^{[j]}, & \ldots, & \theta_{k-2}^{[j]}, & \theta_k^{[j-1]}, \quad \mathbf{X}\ ) \\
\theta_k^{[j]} & \sim & p(\theta_k| & \theta_1^{[j]}, & \theta_2^{[j]}, & \ldots, & \theta_{k-2}^{[j]}, & \theta_{k-1}^{[j]}, \quad \mathbf{X}\ ).
\end{array}
$$

In obiger Darstellung bezeichnet der Index k die Anzahl der Parameter eines Modells und der Index j die Anzahl der Zyklen. Das Grundprinzip von Gibb's Sampler besteht darin, dass iterativ Werte aus der Posterior-Verteilung bei Kenntnis der zuletzt gezogenen Parameter *gesampled* werden. Der Markov-Charakter des Prozesses zeigt sich darin, dass die Dichte, aus der ein Parameter gezogen wird, von den zuletzt gezogenen Parametern abhängt. Bei einer hinreichenden Anzahl von Zyklen konvergieren die Verteilungen der gezogenen Parameter gegen die Posterior-Verteilungen der Parameter, wobei keine Autokorrelationen in den jeweiligen so entstehenden Markov-Ketten vorliegen sollten. In der Praxis wird eine Anzahl von anfänglichen Iterationen des Algorithmus als Burn-In Phase bezeichnet. Die Werte der Ketten in dieser Phase werden zur Bestimmung der deskriptiven Statistiken der Posterior-Verteilung nicht verwendet, da die Ketten unter Umständen einige Zyklen benötigen, um gegen die Zielverteilung zu konvergieren. In der Regel werden pro Parameter mehrere Ketten erzeugt, um anhand deren Mischung die Konvergenz beurteilen zu können. Mischen sich die Ketten nicht, so operieren diese nicht auf einer gemeinsamen Zielverteilung, der Posterior-Verteilung, was wiederum darauf hindeutet, dass das Modell nicht korrekt spezifiziert ist.

Die Anwendung des MCMC-Verfahrens und die Zusammenfassung und Interpretation der Ergebnisse sei anhand eines konkreten Beispiels illustriert. In einem ersten Schritt wird eine manifeste Zeitreihen der Länge $n = 500$ aus dem entwickelten probabilistischen Modell bei gegebenen Parametern simuliert. Die Anzahl der Items ist 4, es wird ein

4-stufiges Antwortformat verwendet. Die datengenerierenden Parameter sind:

$$\eta = -0.8, \tag{4.63}$$

$$\boldsymbol{\beta}_1 = [-0.25, 0.25, 0.25, -0.25], \tag{4.64}$$

$$\boldsymbol{\beta}_2 = [-0.25, -0.10, 0.10, 0.25], \tag{4.65}$$

$$\boldsymbol{\beta}_3 = [0.25, -0.25, -0.25, 0.25], \tag{4.66}$$

$$\boldsymbol{\beta}_4 = [0, 0, 0, 0]. \tag{4.67}$$

Die Kategorien-Parameter wurden absichtlich relativ heterogen gewählt und zudem wurde die Summennormierung der Item-Parameter berücksichtigt. Die Posterior-Verteilungen werden aus der manifesten Zeitreihe mit Hilfe des MCMC-Verfahrens bestimmt. Zum Einsatz kommt die Software WinBUGS (Lunn et al., 2000) in Kombination mit dem R-Paket R2WinBUGS (Sturz, Ligges & Gelman, 2005).

Da in einem realen Anwendungsfall keine Vorinformationen über die Lage der Parameter vorhanden sind, werden als Prior-Verteilungen jeweils Normalverteilungen mit den Mittelwerten

$$\mu = 0 \tag{4.68}$$

und der Präzision

$$\frac{1}{\sigma^2} = 0.01 \tag{4.69}$$

gewählt. Zudem werden zwei Markov-Ketten pro Parameter eingesetzt, um die Mischung der Ketten bewerten zu können. Die Burn-In-Phase beträgt 200 Iterationen. Im realen Anwendungsfall würden die Ketten ohne die Burn-In-Phase die Länge 1000 oder mehr aufweisen, um die Posterior-Verteilungen mit einer hinreichenden Genauigkeit zu approximieren. Für das Beispiel wurden lediglich 200 Iterationen in der Burn-In Phase und insgesamt 400 Iterationen verwendet, da zwei Markov-Ketten zum Einsatz kommen.

Abbildung 4.3 zeigt exemplarisch die Markov-Ketten für den Parameter η. Eine Kette ist gestrichelt, die andere Kette ist mit einer durchgehenden Linie dargestellt. Aus der Abbildung geht hervor, dass die beiden Ketten konvergieren, was darauf hindeutet, dass

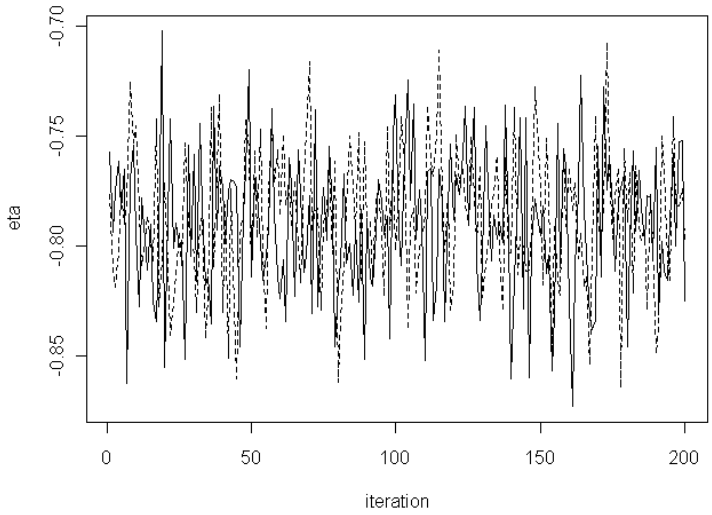

Abbildung 4.3.: Zwei Markov-Ketten für den Parameter η bei 200 Iterationen pro Kette.

beide Ketten in der selben Zielverteilung operieren, nämlich der Posterior-Verteilung von η. Neben der Mischung der Ketten ist gewünscht, dass die aufeinanderfolgenden Ziehungen voneinander unabhängig sind und keine Autokorrelationen in den Ketten vorliegen. Dies ist per optischer Inspektion in Abbildung 4.3 der Fall.

Neben der visuellen Überprüfung der Konvergenz der Markov-Ketten existiert die Gelman-Rubin-Statistik $\hat{R}$ (Gelman & Rubin, 1992), die es erlaubt die Konvergenz, bzw. Mischung von m Markov-Ketten der Länge n hinsichtlich eines Parameters zu bewerten. Um die Statistik zu bestimmen müssen nach (Gill, 2008, p. 478) folgende Kennwerte bestimmt werden:

1. Die Varianz innerhalb der Ketten W:

$$W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_{(j)}^{[i]} - \bar{\theta}_{(j)})^2.$$

2. Die Varianz zwischen den Ketten B:

$$B = \frac{n}{m-1} \sum_{j}^{m} (\bar{\theta}_{(j)} - \bar{\bar{\theta}})^2,$$

wobei $\overline{\overline{\theta}}$ den Gesamtmittelwert der Ketten bezeichnet.

3. Die geschätzte Varianz des Parameters $\widehat{Var(\theta)}$: $\widehat{Var(\theta)} = (1 - 1/n)W + (1/n)B$.

Mit diesen Werten ist

$$\hat{R} = \sqrt{\frac{\widehat{Var(\theta)}}{W}}. \tag{4.70}$$

Werte nahe an 1 deuten darauf hin, dass alle m spezifizierten Markov-Ketten innerhalb der selben Verteilung agieren und somit konvergieren. Höhere Werte zeigen an, dass sich die Markov-Ketten nur unzureichend mischen und die Modellparameter nicht eindeutig identifiziert werden können. Nach Gelman (1996) sind Werte von $\hat{R}$ kleiner als 1.02 akzeptabel.

Tabelle 4.3 zeigt einige deskriptiven Statistiken der Posterior-Verteilungen der Parameter auf Basis der aus dem Modell simulierten Daten. Es werden die Mittelwerte (M) und die Standardabweichungen (SD) der durch die Markov-Ketten erzeugten Posterior-Verteilungen angegeben. Aus einer pragmatischen Perspektive sind die Posterior-Mittelwerte Punktschätzer der Parameter, die Posterior-Standardabweichungen werden als Standardfehler betrachtet und die Grenzwerte der 2.5% und 97.5% Perzentile der Posterior-Verteilungen werden als Grenzwerte der 95% - Konfidenzintervalle der Parameter interpretiert. Technisch gesehen ist diese Betrachtungsweise nicht richtig, aus pragmatischen Gründen allerdings anscheinend üblich. Tatsächlich handelt es sich um ein sog. 95% - Kredibilitäts-Intervall, dass sich von der frequentistischen Definition eines Konfidenzintervalls unterscheidet. Es wird in der Bayes-Statistik nicht davon ausgegangen, dass bei einer theoretisch unendlichen Wiederholung eines Zufallsexperiments bei festen Parametern in 95% der Fälle das Konfidenzintervall den wahren Parameter überdeckt, vielmehr liegt ein epistemischer Wahrscheinlichkeitsbegriff zugrunde, d.h. die richtige Interpretation wäre tatsächlich die übliche Fehlinterpretation des Neyman-Pearson-Konfidenzintervalls: Der Parameter liegt bei gegebenem Modell mit einer epistemischen Wahrscheinlichkeit von 95% innerhalb dieses Intervalls. In der letzten Spalte der Tabelle 4.3 ist die Statistik $\hat{R}$ angegeben. Da sich alle Werte von $\hat{R}$ nahe 1 befinden, ist davon auszugehen, dass die zwei spezifizierten Markov-Ketten pro Parameter konvergieren und

Tabelle 4.3.: Ergebnisse der Bestimmung der Posterior-Verteilungen mit der MCMC-Methode: Mittelwerte, Standardabweichungen, Perzentile und $\hat{R}$

	M	SD	2.5 %	97.5 %	$\hat{R}$
$\beta_{1,1}$	-0.19	0.10	-0.38	0.00	1.00
$\beta_{1,2}$	0.29	0.07	0.15	0.43	1.00
$\beta_{1,3}$	0.26	0.07	0.12	0.41	1.00
$\beta_{1,4}$	-0.36	0.10	-0.54	-0.15	1.00
$\beta_{2,1}$	-0.14	0.11	-0.35	0.06	1.00
$\beta_{2,2}$	-0.22	0.09	-0.39	-0.05	1.00
$\beta_{2,3}$	0.14	0.08	-0.03	0.30	1.00
$\beta_{2,4}$	0.22	0.09	0.06	0.40	1.00
$\beta_{3,1}$	0.30	0.09	0.11	0.47	1.01
$\beta_{3,2}$	-0.16	0.09	-0.34	-0.01	1.01
$\beta_{3,3}$	-0.38	0.09	-0.57	-0.20	1.01
$\beta_{3,4}$	0.24	0.09	0.06	0.43	1.01
$\beta_{4,1}$	0.15	0.08	-0.02	0.31	1.00
$\beta_{4,2}$	0.00	0.08	-0.16	0.14	1.00
$\beta_{4,3}$	-0.07	0.08	- 0.23	0.09	1.00
$\beta_{4,4}$	-0.08	0.10	- 0.27	0.12	1.00
η	-0.79	0.03	-0.85	-0.73	1.00
deviance	4536.07	5.24	4528.00	4548.00	1.00
$pD = 13.3$ $DIC = 4549.3$					

Anmerkungen:

pD= Effektive Modellkomplexität; DIC= Deviance Information Criterion

auf der selben Posterior-Verteilung operieren. Im Beispiel zeigt sich für die meisten Parameter eine sehr gute Konvergenz der Ketten. Beachtenswert ist, dass der Parameter η relativ genau geschätzt wird. Für die Kategorien-Parameter zeigt sich eine leichte Verzerrung der Schätzung. Dies ist darauf zurückzuführen, dass lediglich *eine* Zeitreihe für *eine* Person generiert wurde und somit nur begrenzte Informationen über die Leichtigkeiten der Kategorien der Items zur Verfügung stehen. Es zeigt sich allerdings, dass alle generierenden Kategorien-Parameter innerhalb des entsprechenden 95% - Kredibilitäts-Intervalls liegen.

In Tabelle 4.3 sind ebenfalls die deskriptiven Statistiken für die Posterior-Verteilung der Devianz dargestellt. Die Devianz entspricht bekanntlich $-2\log L$. Die Anwendung der MCMC-Methode führt - neben den Posterior-Verteilungen für die Parameter - ebenfalls zu einer Posterior-Verteilung der Modell-Devianz, da für jede Iteration der Kette ein Devianz-Wert bestimmt werden kann. In der Bestimmung der Devianz geht auf diese Weise die Unsicherheit hinsichtlich der Lage der Parameter mit ein. Es lässt sich also - ebenso wie für die Parameter - ein Kredibilitäts-Intervall für die Devianz bestimmen.

Die Devianz ist eine Größe, die in die Bestimmung des *deviance information criterions* (DIC) (Spiegelhalter, Carlin, Carlin & Linde, 2002) eingeht. Das DIC ist ein Kennwert, der ähnlich wie das Bayes-Information-Criterion (BIC) (Schwarz, 1978) oder das Akaike-Information-Criterion (AIC) (Akaike, 1973) zu verstehen ist. Das DIC ist wie folgt definiert (Gill, 2008, p. 261):

$$DIC = \overline{D(\boldsymbol{\theta})} + \overline{D(\boldsymbol{\theta})} - D(\hat{\boldsymbol{\theta}}) \tag{4.71}$$

$$= \overline{D(\boldsymbol{\theta})} + p_D. \tag{4.72}$$

$\overline{D(\boldsymbol{\theta})}$ ist der Mittelwert der Posterior-Verteilung der Devianz (*posterior mean deviance*). $D(\hat{\boldsymbol{\theta}})$ ist die Devianz bei den Schätzern der Modellparameter θ. Die Differenz

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\hat{\boldsymbol{\theta}}) \tag{4.73}$$

ist die effektive Parameteranzahl des Modells. Je geringer die Devianz und je geringer die effektive Parameteranzahl des Modells, desto besser passt ein Modell auf einen gegebenen, festen Datensatz. Im Fall des Beispiels entspricht die effektive Parameteranzahl

$p_D = 13.3$, was relativ gut der tatsächlichen Parameteranzahl entspricht. Es werden für die Kategorien-Parameter wegen der Summennormierung 12 Parameter geschätzt. Ein Parameter wird für die Abbildung der Variabilität benötigt, somit müssen insgesamt 13 Parameter geschätzt werden, bzw. das Modell besitzt 13 Freiheitsgrade.

4.10.2. Empirischer Bias und Varianz von $\hat{\eta}_v$

Bias und Varianz der geschätzten Parameter sind klassische Gütemaße der Parameterschätzung (vgl. z.B. Edwards, 1972, Kap. 7.4. und 7.5. für eine sehr gute frequentistische Darstellung). Je geringer der Bias und je geringer die Varianz eines Schätzers, desto höher die Güte des Schätzers. Die simulative Bewertung des Bias und der Varianz eines Schätzers bietet sich an, um unterschiedliche Schätzprozeduren hinsichtlich der Güte zu vergleichen. So vergleichen z.B. Hojtink und Boomsma (1995) unterschiedliche Methoden der Schätzung der Personen-Parameter (MLE, WLE und BME) beim dichotomen Rasch-Modell hinsichtlich des Bias und der Varianz der Schätzer.

Aus einer frequentistischen Perspektive ist es von Interesse, inwiefern es möglich ist den „wahren“ Parameter η aus simulierten Daten zu extrahieren und mit welcher Effizienz dies geschehen kann.

Der empirische Bias eines Schätzers ist an Anlehnung an Hojtink und Boomtsma wie folgt definiert:

$$\text{Bias}(\hat{\eta}) = n^{-1} \sum_{i=1}^{n} (\hat{\eta}_i - \eta). \tag{4.74}$$

Der Index i läuft über n wiederholte Schätzungen des Parameters η aus simulierten Daten. Die Varianz des Schätzers ist

$$\text{Var}(\hat{\eta}) = n^{-1} \sum_{i=1}^{n} (\hat{\eta}_i - \bar{\hat{\eta}})^2. \tag{4.75}$$

Aus einer Bayesianischen Perspektive macht die Berechnung eines Bias keinen Sinn, da nicht davon ausgegangen wird, dass ein fixer Parameter die Daten generiert, sondern die Daten fix sind und die Parameter variieren.

Nichtsdestotrotz sei die Simulationsstudie hier durchgeführt, da die Bewertung des Bias vor einem frequentistischen Hintergrund durchaus von Interesse sein kann, um Hinweise zu erhalten, ob in bestimmten Skalenbereichen des Parameters η Schätzprobleme auftreten. In diesem Sinne werden die Mittelwerte der Posterior-Verteilungen als frequentistische Punktschätzer betrachtet.

Die Simulation konzentriert sich auf den Parameter η des probabilistischen Modells, da in diesem die psychometrische Information hinsichtlich der Variabilität einer intraindividuellen, multidimensionalen Zeitreihe erfasst wird. Der Wertebereich von η, der untersucht wird, ist $-2.5 < \eta < 2.5$, wobei der Wert des Parameters η in Schritten des Betrags 0.5 von -2.5 bis 2.5 erhöht wird. Pro Parameter wird eine multivariate, intraindividuelle Zeitreihe der Länge $N = 200$ auf der Basis von 4 Items mit den Kategorien-Parametern

$$\boldsymbol{\beta}_1 = [-0.25, 0.25, 0.25, -0.25], \tag{4.76}$$

$$\boldsymbol{\beta}_2 = [-0.25, -0.10, 0.10, 0.25], \tag{4.77}$$

$$\boldsymbol{\beta}_3 = [0.25, -0.25, -0.25, 0.25], \tag{4.78}$$

$$\boldsymbol{\beta}_4 = [0, 0, 0, 0]. \tag{4.79}$$

aus dem Modell simuliert. Die Kategorien-Parameter werden relativ heterogen gewählt. So zeigt z.B. Item 1 nach dem Modell eine eher umgekehrt U-förmige Verteilung der manifesten Werte der Zeitreihe, während Item 3 eher eine umgekehrt U-förmige Verteilung der Werte der manifesten Zeitreihe aufweist.

Auf die simulierten Zeitreihen wird die Markov-Chain-Monte-Carlo-Methode angewendet, um die Posterior-Verteilungen der Parameter zu bestimmten. Hierbei werden Normalverteilungen mit dem Mittelwert $\mu = 0$ und der Präzision $1/\sigma^2 = 0.01$ als Prior-Verteilungen der Parameter verwendet. Die Burn-In-Phase beträgt 500 Iterationen und die Kette, die zur Bestimmung der Posterior-Verteilungen verwendet wird, läuft über insgesamt $n = 500$ Iterationen. Die Markov-Kette eines jeden Parameters wird zur Schätzung des empirischen Bias und der empirischen Varianz verwenden. Die Berechnung der Statistiken erfolgt mit den Gleichungen 4.74 und 4.75, wobei der Index i über die Itera-

Tabelle 4.4.: Bias und Streuung von $\hat{\eta}$

η	$\hat{\eta}$	Bias($\hat{\eta}$)	SD($\hat{\eta}$)
-2.5	-2.48	0.02	0.11
-2.0	-1.99	0.01	0.08
-1.5	-1.49	0.01	0.07
-1.0	-1.00	0.00	0,05
-0.5	-0.49	0.01	0.04
0	0.01	0.01	0.04
0.5	0.51	0.01	0.04
1.0	1.03	0.02	0.07
1.5	1.66	0.16	0.23
2.0	2.5	0.50	0.59
2.5	4.57	2.07	1.67

tionen der Kette von 1 bis $n = 500$ läuft.

Tabelle 4.4 zeigt das Ergebnis der Simulation. Anstelle der Varianz der geschätzten Parameter wird die Streuung der geschätzten Parameter angegeben. Es zeigt sich, dass der empirische Bias der Schätzer für die Skalenbereiche für η von -2.5 bis 1.0 relativ gering ist, dann aber ab 1 sprunghaft ansteigt. Der Grund hierfür ist derjenige, dass ab einem η von etwa 1 die generierten Zeitreihen fast nur noch Extremwerte aufweisen. Zudem steigt auch die Streuung des Schätzers ab diesem Skalenwert stark an, bzw. es liegt relativ wenig Sicherheit hinsichtlich der Lage des generierenden Parameters vor. Praktisch hat dies zur Folge, dass Skalen-Werte von $\hat{\eta}$, die größer als 1 sind, bei der Verwendung von 4 Items mit Vorsicht zu interpretieren sind. Allerdings ist fraglich, ob solch extreme Werte empirisch in einer realen Untersuchung auftreten. Der Anwender ist in diesem Fall jedoch durch die hohen Standardfehler der Schätzer gewarnt. Für Skalenbereiche kleiner als 1 zeigt sich bei den gewählten Prior-Verteilungen ein sehr geringer Bias und auch eine recht geringe Standardabweichung von $\hat{\eta}$. Es ist zu vermuten, dass die Standardabweichung von $\hat{\eta}$ mit der Länge der multivariaten, beobachteten Zeitreihen sinkt, d.h. dass die Messgenauigkeit

von der Anzahl der Beobachtungen abhängt.

4.11. Die Überprüfung der Modellpassung mittels standardisierter Residuen

Sind die Parameter aus einem Datensatz geschätzt, so ist von Interesse, wie gut das Modell die beobachteten Daten beschreibt, bzw. wie gut das Modell passt. Es existiert eine Vielzahl von Möglichkeiten bei Rasch-Modellen, um die Modellgeltung zu überprüfen, auf die im theoretischen Hintergrund zu Rasch-Modellen kurz eingegangen wurde.

Der in der vorliegenden Arbeit angewendete Ausgangspunkt der Überprüfung der Modellpassung sind standardisierte Residuen (Wright & Stone, 1969; Davier von, 1996). Sind die Parameter eines Modells geschätzt, können die standardisierten Residuen wie folgt berechnet werden:

$$z_{vi[t]}|x_{vi[t-1]} = \frac{x_{vi[t]} - \left\langle x_{vi[t]} \right\rangle |x_{vi[t-1]}}{\sqrt{\mathrm{var}(x_{vi[t]})|x_{vi[t-1]}}}. \tag{4.80}$$

$z_{vi[t]|x_{vi[t-1]}}$ ist das standardisierte Residuum des Modells hinsichtlich der manifesten Reaktion $x_{vi[t]}$, $\left\langle x_{vi[t]} \right\rangle |x_{vi[t-1]}$ ist der Erwartungswert der Antwort unter dem Modell (vgl. Gleichung 4.10) und $\mathrm{var}(x_{vi[t]})|x_{vi[t-1]}$ ist die Varianz der Reaktion unter dem Modell (vgl. Gleichung 4.16).

Mittels der standardisierten Residuen ist es möglich, die Modellpassung anhand jeder einzelne Reaktion $x_{vi[t]}$ zu bewerten. Ferner können die Residuen quadriert und über die Items oder die Personen summativ aggregiert werden. Diese Summen sind theoretisch betrachtet χ^2-verteilte Test-Statistiken anhand derer die Passung der Items (Item-Fit) und die Passung der Personen (Person-Fit) bewertet werden kann. Für die Items gilt:

$$\sum_{v}^{N}\sum_{t}^{T} z^2_{vi[t]}|x_{vi[t-1]} \rightarrow \chi^2,\ df = n_i. \tag{4.81}$$

Für die Personen gilt:

$$\sum_{i}^{I}\sum_{t}^{T} z^2_{vi[t]}|x_{vi[t-1]} \rightarrow \chi^2,\ df = n_p. \tag{4.82}$$

Hier ist eine Anmerkung zu den Freiheitsgraden der Teststatistik angebracht. Da nach der Likelihood-Funktion die einzelnen Antworten der Personen auf den Items stochastisch unabhängig sind, entsprechen nach Ansicht des Autors die Freiheitsgrade der Anzahl der item-bezogenen bzw. der personenbezogenen Beobachtungen, bzw. Reaktionen. n_i ist die Anzahl aller Reaktionen auf ein Item, wobei die ersten Reaktionen der Personen auf ein Item nicht modelliert, sondern als gegeben betrachtet werden. Entsprechendes gilt für die Freiheitsgrade zur Bewertung der Passung der Personen. n_p ist die Anzahl der beobachteten Reaktionen einer Person, wobei die ersten Reaktionen auf allen Items einer Skala als gegeben betrachtet und nicht gezählt werden. Eine item- oder personenbezogene Fehlpassung des Modells im Sinne von zu hohen Residuen zeigt sich in einer χ^2-Teststatistik, die höher ist, als bei Modellgeltung erwartet würde. Dieser Sachverhalt wird auch als Underfit bezeichnet. Als Grenzwert zur Beurteilung der Fehlpassung könnte ein p-Wert kleiner als 0.05 als Richtlinie verwendet werden. Andererseits ist es auch denkbar, dass die Residuen kleiner sind, als dies unter dem Modell zu erwarten wäre (Overfit). Dies zeigt sich in einer kleineren χ^2-Teststatistik, als bei Modellgeltung erwartet. Ein p-Wert größer als .95 könnte als Kriterium verwendet werden.

Zur Bewertung des Overfits, bzw. Underfits, können die Mittelwerte der quadrierten z-Statistiken personenweise oder itemweise gebildet werden, um den sog. Outfit (OF) zu berechnen. Der Erwartungswert dieser Statistiken ist bei Modellgeltung 1, da die aggregierten Residuen zur Berechnung des Outfits an der Anzahl der Beobachtungen standardisiert werden und der Erwartungswert einer z^2-verteilten Statistik 1 ist. Sind die Residuen insgesamt größer, als bei Modellgeltung erwartet würde, so sind die Mittelwerte der quadrierten z-Statistiken größer als 1. Ist die Passung insgesamt besser, als bei Modellgeltung erwartet, so sind die Mittelwerte der quadrierten z-Statistiken kleiner als 1. Für die Items gilt:

$$OF_i = \frac{1}{n_i} \sum_{v}^{N} \sum_{t}^{T} z_{vit}^2 | x_{vi[t-1]}. \tag{4.83}$$

Für die Personen gilt:

$$OF_v = \frac{1}{n_p} \sum_{i}^{I} \sum_{t}^{T} z_{vit}^2 | x_{vi[t-1]}. \tag{4.84}$$

Die χ^2-Statistiken und die Outfit-Indices erlauben eine ziemlich detaillierte Bewertung der Modellpassung bis hinunter auf die Ebene einer einzelnen Item-Antwort. Zudem sind diese Statistiken intuitiv recht einleuchtend, leicht zu interpretieren und können flexibel aggregiert werden.

4.12. Bewertung der Messgenauigkeit

Im Rahmen der psychometrischen Messung kommt der Bewertung der Messgenauigkeit eines Instrumentes eine zentrale Bedeutung zu. Eine wesentliche Motivation der Verwendung von Messmodellen besteht darin, die Messgenauigkeit und die Homogenität eines Instrumentes prüfen zu können. Erfassen beispielsweise mehrere Indikatoren eines Instrumentes unterschiedliche latente Variablen, so wird die Reliabilität und damit die Homogenität des Instrumentes gering sein, was impliziert, dass das Instrument sich nur bedingt für eine Messung eignet. Im Rahmen probabilistischer Messmodelle wird die Homogenität über Modellgeltungstests überprüft. Zudem ist es möglich, die individuelle Messgenauigkeit über die Standardfehler der geschätzten Parameter und die globale Reliabilität zu schätzen. Ein Ansatz zur Schätzung der globalen Reliabilität, der schon im modelltheoretischen Hintergrund angesprochen wurde, ist derjenige der Andrich-Reliabilität (Andrich, 1988). Die Andrich-Reliabilität wird auch als Separabilität bezeichnet, da ein hoher Koeffizient darauf hindeutet, dass das verwendete Instrument die Personen in einer Stichprobe hinsichtlich der Merkmalsausprägung differenziert und damit misst. Dieser Index lehnt sich stark an die Klassische Testtheorie an und ist wie folgt definiert:

$$Rel_1 = \frac{\sigma^2(\theta)}{\sigma^2(\hat{\theta})} = \frac{\sigma^2(\hat{\theta}) - \sigma^2(e)}{\sigma^2(\hat{\theta})}. \tag{4.85}$$

Die Anlehnung an die KTT zeigt sich darin, dass die Reliabilität, bzw. Separabilität den Anteil der geschätzten „wahren“ Varianz der Personen-Parameter an der beobachteten Varianz der geschätzten Personen-Parameter darstellt. Die Varianz der Messfehler $\sigma^2(e)$ wird aus der mittleren Fehlervarianz über die Personen ermittelt. Dieses Vorgehen kann zur Berechnung der Reliabilität auf das vorliegende Modell übertragen werden, da die Verwendung der Andrich-Reliabilität auch bei mehrkategoriellen Rasch-Modellen nicht

unüblich ist und auch bei dem hier entwickelten Modell die Fehlervarianz der Schätzer der Merkmalsausprägung vorliegt. Die Bewertung der Messgenauigkeit der individuellen Merkmalsausprägung ist - wie üblich - über die Standardfehler der Schätzer der Merkmalsausprägung möglich.

4.13. Zusammenfassende Darstellung der Ergebnisse zur Modellentwicklung

Zu Fragestellung I.1.: Die deduktive Definition eines probabilistischen Modells unter Verwendung der absoluten sukzessiven Differenz (ASD) und der Berücksichtigung der kategorialen Natur der Ratings durch item- und kategorienspezifische Modellparameter führt unter Anwendung der Maximum-Entropie-Methode zu einem Modell, dass bis auf die Scoring-Funktion dem PCM von Masters gleicht. Zudem stellt das Modell bedingte Wahrscheinlichkeitsverteilungen da. Da in die Scoring-Funktion des definierten Modells zwei Größen eingehen, das Rating zum Zeitpunkt $[t]$ als Variable und das Rating zum Zeitpunkt $[t-1]$, definiert das Modell einen Markov-Prozess erster Ordnung, der neben den Kategorien-Parametern von den latenten Trait Parametern η_v abhängt. Pro Kategorie einer Rating-Skala existiert eine Wahrscheinlichkeitsverteilung der Ratings zum Zeitpunkt $[t]$ gegeben einer Reaktion zum Zeitpunkt $[t-1]$. Die Autokorrelation der Ratings, bzw. die Stabilität des Markov-Prozesses ist von dem latenten Parameter η_v abhängig. Niedrige Parameter η_v beschreiben einen relativ stabilen Prozess, hohe Parameter η_v beschreiben einen eher instabilen Prozess, d.h. hohe Sprünge in den manifesten Zeitreihen. Das Modell erfasst die Variabilität in manifesten, multivariaten Zeitreihen und führt diese auf einen latenten Parameter η_v zurück.

Zu Fragestellung I.2.: Die Übergangsmatrix des Modells und die Kategorien-Charakteristik Kurven sind kongruent mit dem Desideratum der latenten Skalierung von Variabilität. Hohe Personen-Parameter gehen mit Zeitreihen einher, die eine hohe Variabilität aufweisen. Je höher der Personen-Parameter, desto größer die Wahrscheinlichkeit der Wahl einer distalen Kategorie.

Zu Fragestellung I.3.: Die Varianzen und Erwartungswerte der manifesten Variablen ergeben sich aus dem Maximum-Entropie-Formalismus durch das ein- und zweimalige Differenzieren der logarithmierten Zustandssumme des Modells. Die Ergebnisse sind mit gängigen Erkenntnissen aus der Statistik hinsichtlich von Erwartungswerten und Varianzen von diskreten Zufallsvariablen kongruent.

Zu Fragestellung I.4.: Die generelle Maximum-Entropie-Verteilung für diskrete Daten gehört zur Exponentialfamilie. Die Herleitung der Likelihood-Funktion und das partielle Differenzieren nach den Modellparametern zeigt, dass suffiziente Statistiken für die Schätzung der Parameter existieren. Dies ist mit gängigen Erkenntnissen zur Exponentialfamilie kongruent.

Zu Fragestellung I.5.: Werden Zeitreihen aus dem Modell generiert, so zeigt sich, dass die mittleren, absoluten Differenzen dieser Zeitreihen in einem monotonen Zusammenhang mit den Personen-Parametern des Modells stehen. Der Zusammenhang ist von sigmoider Gestalt.

Zu Fragestellung I.6.: Die Frage der spezifisch objektiven Messung wurde durch die Bildung der Logits benachbarter Kategorien-Wahrscheinlichkeiten angegangen. Hierbei zeigt sich, dass die Logits der benachbarten Kategorien-Wahrscheinlichkeiten in einem linearen Verhältnis zu dem Personen-Parameter η_v stehen. Werden die Differenzen der Parameter η_v und η_w für zwei Personen gebildet, so verschwinden die Kategorien-Parameter aus der Gleichung verschwinden. Die Differenz der Personen-Parameter entspricht der Differenz der Logits der benachbarten Kategorien. Oder anders ausgedrückt, die Differenzen zweier Personen-Parameter sind linear auf einer Logit-Skala Eine Besonderheit besteht darin, dass die Richtung des Vergleichs in Abhängigkeit der Scoring-Funktion entweder ein positives oder negatives Vorzeichen erhält. Aus dem Sachverhalt, dass die Differenz der Personen-Parameter mit der Differenz der Logits in Beziehung steht und die Kategorien-Parameter für den Vergleich irrelevant sind, wird geschlossen, dass das Modell spezifisch objektive Vergleiche auf einer Differenzenskala ermöglicht.

Zu Fragestellung I.7.: Die Frage der Konvergenz der Parameterschätzung wurde simulativ angegangen. Es wurden manifeste Daten aus dem Modell simuliert und die Parameter

wurden mittels der MCMC-Methode geschätzt. Die Konvergenz der Schätzung wurde mit der Gelman-Rubin-Statistik und optisch bewertet. Es zeigt sich für alle Parameter, dass die Gelman-Rubin-Statistik den Wert von 1.01 nicht überschreitet, was bedeutet, dass zwei Markov-Ketten mit unterschiedlichen Startwerten gegen eine gemeinsame Zielverteilung konvergieren. Daraus lässt sich schließen, dass die Likelihood-Fläche des Modells ein absolutes Maximum besitzt.

Zu Fragestellung I.8.: Der empirische Bias und die empirische Varianz des Parameters η_v wurde simulativ für vier Items evaluiert. In den Skalenbereichen von -2 bis 0.5 zeigt sich ein relativ geringer Bias, die Parameter werden leicht unterschätzt. Ab einem Skalenwert von etwa 1 steigt der Bias sprunghaft an. Dies ist darauf zurückzuführen, dass ab einem Parameter von 1 durch das Modell fast nur noch extreme Antworten und maximale Sprünge produziert werden. Das Modell ist in diesen extrem positiven Skalen-Bereichen sozusagen gesättigt und manifest wird die extremste Sprungweite erreicht. Für die empirische Varianz der Schätzer zeigt sich ein von Rasch-Modellen bekannter umgekehrt U-förmiger Verlauf. In mittleren Skalenbereichen ist die Varianz an geringsten und steigt in den Extrembereichen an. Ausprägungen des Parameters in den Extrembereichen gehen mit extrem stabilen oder extrem sprunghaften Prozessen auf den manifesten Variablen einher.

Zu Fragestellung I.9.: Zur Fragestellung der Modellpassung wird in dieser Arbeit die Berechnung von standardisierten Residuen vorgeschlagen, die im Rahmen der Psychometrie zur Berechnung des sog. Outfits herangezogen werden. Der Grund für den Vorschlag besteht darin, dass die Residuen sich sehr einfach aus den Erwartungswerten und der Varianzen der manifesten Ratings bei Modellgeltung berechnen lassen. In Folge der angenommenen stochastischen Unabhängigkeit der Ratings können theoretisch die quadrierten, standardisierten Residuen flexibel z.B. über Personen oder Items kumuliert werden, was zu Teststatistiken führt, die bei Modellgeltung χ^2-verteilt sind.

Zu Fragestellung I.10.: Zur Berechnung der Reliabilität wird vorgeschlagen, das Vorgehen von Andrich (Andrich, 1988) zur Berechnung der Reliabilität bei probabilistischen Testmodellen auf das generierte Modell zu übertragen.

Zu Fragestellung I.11.: Die individuelle Messgenauigkeit der Merkmalsausprägung ist - wie auch bei den bekannten probabilistischen Testmodellen - anhand des Standardfehlers des Trait-Parameters, bzw. dem Streuung der entsprechenden Trait-Verteilung bewertbar.

5. Modellanwendung

Nachdem die Eigenschaften des Modells beschrieben wurden und Verfahren zur Parameterschätzung, der Bewertung der Reliabilität und der Modellpassung vorliegen, ist es angebracht, die Anwendung des Modells an realen Daten zu erproben. In Frage kommen massiv-längsschnittlich erhobene Daten, bei denen mehrere Konstrukte oder ein Konstrukt mit Hilfe von einem oder mehreren manifesten Indikatoren in Form von kategorialen Ratings auf Items erfasst werden. Anspruch des Modells ist es, die intraindividuelle Variabilität auf einem durch mehrere Items erfassten Konstrukt in dem Parameter η_v quantitativ abzubilden, um die Reliabilität der Messung und die Modellpassung zu bewerten. Zur Bewertung der praktischen Anwendbarkeit des Modells wird ein im Rahmen eines Ambulatory Assessments von Crayen, Eid, Lischetzke, Courvoisier und Vermunt (in Druck) erhobener Datensatz verwendet.

5.1. Beschreibung des Datensatzes

Die Beschreibung des Datensatzes orientiert sich an der Darstellung von Crayen et al. (in Druck). 165 Studierende der Freien Universität Berlin (89 Frauen) im Alter von 18 bis 35 Jahren, die über Aushänge an der Universität rekrutiert wurden, nahmen an einem zweiwöchigen Ambulatory Assessment im Zeitraum zwischen Oktober 2009 und Mai 2010 teil. Nach einer Laborsitzung und der Erhebung verschiedener Persönlichkeitsvariablen (u.a. Neurotizismus, Extraversion, Gewissenhaftigkeit), wurde jeder der Teilnehmenden mit einem mobilen Handheld-Computer (HP iPAQ rx1959) ausgestattet und im Gebrauch instruiert. Auf den Geräten war eine spezielle Software zur Durchführung von Ambulatory Assessments installiert (Izybuilder, IzyData Ltd., Fribourg, Schweiz). Die

Teilnehmer wurden gebeten, den auf dem Gerät implementierten Fragebogen zu beantworten, sobald das Gerät ein Signal aussendet. In der ersten Woche emittierte das Gerät automatisch 7 und in der zweiten Woche 8 Signale pro Tag, wobei der Abstand zwischen den Signalen zwischen 60 und 180 Minuten variieren konnte. Die Beantwortung des Fragebogens musste innerhalb von 30 Minuten geschehen, anderenfalls wurde die Sitzung als fehlend gewertet. Auf durchschnittlich 89 % Prozent der Signale wurde reagiert, so dass die Teilnahmebereitschaft als sehr hoch einzuschätzen ist. Crayen et al. (in Druck) erfassten zu jeder Messung externale Aspekte (gegenwärtiger Ort, anwesende Personen, derzeitige Aktivität), die Stimmungsregulation (stimmungsrelevante Aktivitäten, Intention zur Stimmungsregulation) und die momentane Stimmung. Für die vorliegende Arbeit ist die Erfassung der momentanen Stimmung relevant. Hierzu kam eine Kurzversion des Mehrdimensionalen Befindlichkeitsfragebogens (MDBF; Steyer et al., 1997) zum Einsatz. Die Dimension „gehobene vs. gedrückte Stimmung“ wurde mit vier bipolaren Items („unwohl-wohl“, „schlecht-gut“, „unzufrieden-zufrieden“, „unglücklich-glücklich“) erfasst. Die Dimension „Wachheit“ wurde mit den Items „müde-wach“, „schläfig-ausgeruht“ und die Dimension Anspannung-Entspannung mit den Items „angespannt-entspannt“, „unruhig-ruhig“ erfasst. Die Antwort erfolgte auf einer 4-stufigen Rating-Skala (z.B. „sehr unwohl“, „eher unwohl“, „eher wohl“, „sehr wohl“). Die Auswahl des Datensatzes zur Erprobung der Anwendung des Modells orientierte sich an folgenden Gesichtspunkten:

1. Ein oder mehrere psychologisch relevante Konstrukte sollten mit mehreren Indikatoren massiv-längsschnittlich bei mehreren Personen erhoben worden sein.

2. Die manifesten Indikatoren für ein latentes Konstrukt sollten Items, auf die kategoriale Ratings abgegeben wurden.

3. Die Rating-Skala sollte nicht zu viele Kategorien umfassen. Eine Vielzahl von Kategorien führt zu größeren Lücken in der Kategorien-Besetzung, was sich nachteilig auf die Parameterschätzung und die Testung der Modellgeltung auswirkt.

5.2. Fragestellungen zur Modellanwendung

Anhand des Datensatzes werden folgende Fragestellungen bearbeitet:

II.1. *Ist das Modell auch bei der Verwendung realer, nicht-simulierter Daten schätzbar?* Modellevaluation auf Basis von Simulationen ist ein günstiger Weg, um formale Eigenschaften von Modellen, wie z.B. die Konvergenz von Parameterschätzungen, die Abhängigkeit der Genauigkeit der Merkmalserfassung von der Stichprobengröße und verwandte Themen, wie z.B. Bias und Varianz genauer zu untersuchen. In der Regel werden Daten aus den Modellen simuliert, was bedeutet, dass die Modelleigenschaften unter Ideal-Bedingungen der Passung untersucht werden. In der Praxis ist es jedoch in der Regel der Fall, dass die Daten wesentlich komplexer sind, als unter einem Modell angenommen wird. Von daher ist es wichtig zu untersuchen, inwiefern ein Modell auf eine reale, empirisch im Feld oder dem Labor gewonnene Datenstruktur angewendet werden kann.

II.2. *Wie sind die Mittelwerte und Streuungen der latenten Trait-Verteilungen auf den jeweiligen latenten Variabilitäts-Skalen?* Diese Fragestellung bezieht sich vor allem auf die mit dem zu definierenden Modell ermittelte Trait-Verteilung in einer Stichprobe. Im Rahmen von probabilistischen Testmodellen sind die Mittelwerte und Streuungen von latenten Trait-Verteilungen schätzbar, sofern Techniken verwendet werden, die auch in der Multilevel-Analyse eine Anwendung finden. D.h. die latente Trait-Verteilung muss mit modelliert werden. Der Mittelwert der Trait-Verteilung gibt Aufschluss über die mittlere Merkmalsausprägung in der zugrundeliegenden Stichprobe und die Varianz der latenten Trait-Verteilung steht mit der Reliabilität und der Separation der Personen auf der latenten Dimension in Beziehung.

II.3. *Wie ist die Merkmalsausprägung der Personen und wie sind deren Standardfehler?* Dies Fragestellung bezieht sich auf die Erfassung der individuellen Merkmalsausprägung der Person und die Genauigkeit der individuellen Merkmalserfassung. Bei probabilistischen Testmodellen ist es üblich, das manifeste Antwortverhalten auf einen latenten Personen-Parameter zurückzuführen, dessen Lage geschätzt wer-

den muss. Zudem sollte die Genauigkeit der Parameterschätzung bestimmt werden, was es erlaubt, die Genauigkeit der individuellen Merkmalserfassung zu beurteilen.

II.4. *Wie sind die Reliabilitäten der Variabilitäts-Skalen?* Neben der Erfassung der individuellen Merkmalsausprägung ist es von Interesse, wie reliabel das Modell Personen hinsichtlich des Merkmals Variabilität differenziert. Diese Fragestellung soll anhand realer Daten überprüft werden.

II.5. *Wie hoch ist die Fehlpassung des Modells über die Gesamt-Daten einer Skala?* Um probabilistische Testmodelle sinnvoll anwenden zu können, ist es nötig, die Passung zu bewerten. Die Bewertung der Passung des zu generierenden Modells auf Basis standardisierter Residuen wird anhand realer Daten durchgeführt.

II.6. *Wie hoch ist die Fehlpassung des Modells bezogen auf einzelne Items einer Skala?* Passt ein probabilistisches Testmodell nicht, so sollte überprüft werden, warum das Modell nicht passt. Passen einzelne Items nicht, so kann dies darauf zurückzuführen sein, dass diese Items mit den anderen Items einer Skala zusammen keine latente Dimension erfassen und die Gründe mit einem Mischverteilungs-Modell exploriert werden müssen. Die im modelltheoretischen Teil der Arbeit vorgestellte, bekannte Methode der Bewertung der Passung anhand standardisierter Residuen wird im praktischen Teil angewendet.

II.7. *Wie ist der empirische Zusammenhang zwischen dem Modellparameter η_v und der beobachteten, mittleren absoluten Differenz?*. In der Modellentwicklung wurde gezeigt, dass der Variabilitäts-Parameter η_v in einem monotonen Verhältnis zur mittleren absoluten Differenz der manifesten Daten steht. Dieses Ergebnis beruht auf simulierten Daten und auf Betrachtung der Likelihood-Funktion des Modells. Es ist von Interesse, inwiefern sich dieser Zusammenhang auch zeigt, wenn die Modellparameter anhand empirischer Daten geschätzt werden.

II.8. *Wie ist der Zusammenhang zwischen den Variabilitäts-Parametern auf den drei Subskalen des MDBF?*. Auf Basis der Ergebnisse von Eid & Diener (1999) ist zu

vermuten, dass die Variabilitäten der Personen auf den drei Skalen hoch interkorrelieren. Dies wird anhand des vorliegenden Datensatzes überprüft.

II.9. *Wie ist der Zusammenhang zwischen den Variabilitäts-Parametern und den Skalen des NEO-FFI?* Um erste Hinweise auf die Korrelation der Variabilitäts-Parameter mit Persönlichkeitsvariablen zu erhalten, werden die Variabilitäten mit den drei im Datensatz vorhandenen Skalen des NEO-FFI (Borkenau & Ostendorf, 1993) (Neurotizismus, Extraversion und Gewissenhaftigkeit) korreliert. Es ist a priori zu vermuten, dass hohe Korrelationen der Variabilitäts-Parameter mit der Skala „Neurotizismus“ auftreten, da Neurotizismus auch als emotionale Labilität betrachtet wird.

5.3. Darstellung der Vorgehensweise zur Überprüfung der anwendungsorientierten Fragestellungen

Fragestellung II.1. wird überprüft, indem das Modell mit der MCMC-Methode an die jeweiligen Subskalen des MDBF angepasst wird, die Konvergenz der Markov-Ketten wird mittels der Gelman-Rubin-Statistik (Gelman, 1996) bewertet. Fragestellung II.2. wird bearbeitet, indem die Mittelwerte und Streuungen der latenten Trait-Verteilungen in der MCMC-Schätzfunktion explizit modelliert werden. Fragestellung II.3. wird Anhand der Posterior-Verteilungen der Variabilitäts-Parameter des Modells beantwortet. Fragestellung II.4. wird auf Basis der Andrich-Reliabilität (Andrich, 1988) angegangen. Die Andrich-Reliabilität setzt sich aus den mittleren Standardfehlern der individuellen Schätzer der Merkmalsausprägung und der Varianz der geschätzten Trait-Parameter zusammen. Die Fragestellungen II.5. und II.6. werden auf Basis der quadrierten, standardisierten Residuen bearbeitet, die sich einfach aus dem Maximum-Entropie-Formalismus berechnen lassen. Die Residuen werden für jedes einzelne Rating ermittelt und addiert, wodurch sich eine bei Modellgeltung χ^2-verteilte Test-Statistik ergibt. Werden die quadrierten, standardisierten Residuen gemittelt, ergibt sich der sog. Outfit-Index für die Gesamt-Skala. Werte nahe an 1 zeigen eine gute Modellpassung an. Ferner werden die

standardisierten Residuen des Gesamtmodells grafisch als Histogramm abgebildet und optisch inspiziert. Als Faustregel gelten standardisierte Residuen vom Betrag her größer als 2 als auffällig, da diese asymptotisch einer z-Verteilung folgen sollten. Die itembezogene Fehlpassung wird ermittelt, indem die Outfit-Statistik und die residuen-basierte χ^2-Teststatistik für die Items ermittelt werden. Signifikante Teststatistiken zeigen eine Fehlpassung an. Auffällig hohe p-Werte ($p<.95$) sprechen für eine bessere Passung, als unter dem Modell erwartet würde, d.h. das Antwortverhalten wäre in diesem Fall deterministischer als unter dem Modell angenommen. Das bedeutet, dass auf Items mit geringem Outfit weniger manifeste Variabilität vorliegt, als unter dem Modell erwartet würde. Eine zu gute Passung zeigt sich auch an einer Outfit-Statistik von OF$<$1. Outfit-Statistiken größer als 1 und p-Werte kleiner als $p = .05$ deuten auf eine Fehlpassung hin. Die Fragestellungen II.7. bis II.9. werden explorativ untersucht, indem die Interkorrelationsmatrix der Variabilitäten auf den Skalen des MDBF, den mittleren, absoluten sukzessiven Differenzen und den Skalen des NEO-FFI gebildet werden.

5.4. Die Bestimmung der Parameter mit der MCMC-Methode

Für die empirische Erprobung des Modells bieten sich die drei Subskalen des MDBF an. Im Vordergrund steht die Fragestellung, inwiefern das Modell in der Lage ist, Personen hinsichtlich der Variabilität auf den Subskalen des MDBF reliabel zu differenzieren. Zu diesem Zweck wird das Modell mittels der MCMC-Methode an die Daten angepasst. Als Prior-Verteilungen für die Kategorien-Parameter wird die Standardnormalverteilung mit $\mu = 0$ und der Präzision $1/\sigma^2 = 0.01$ verwendet. Die hohe Streuung bringt eine relative Unsicherheit hinsichtlich der Lage der Parameter zum Ausdruck. Die Wahl der Normalverteilung wird durch das Ergebnis begründet, dass ML-Schätzer asymptotisch normalverteilt sind. Zudem wird die latente Verteilung von η_v modelliert. Die Prior-Annahme über die latente Trait-Verteilung der Parameter η_v ist eine Normalverteilung der Parameter mit einem Mittelwert von μ_η und einer Streuung von σ_η. Der Mittelwert

und die Streuung werden aus den Daten mittels der MCMC-Methode geschätzt. Als Prior-Verteilung des Mittelwerts wird eine Normalverteilung mit $\mu_\eta = 0$ angenommen. Die Prior-Verteilung von σ_η entspricht der Gleichverteilung im Wertebereich zwischen 0 und 10, da eine negative Streuung nicht existiert und nicht damit gerechnet wird, dass die Streuung der latenten Trait-Verteilung einen Wert von 10 übersteigt. Um die Konvergenz der Markov-Ketten zu bewerten, kommen zwei Ketten zum Einsatz. Die Burn-In-Phase beträgt 500 Iterationen. Insgesamt werden 1000 Iterationen durchlaufen, so dass mit zwei Ketten die Posterior-Verteilungen der Parameter auf Basis von 1000 Samples evaluiert werden.

Tabelle 5.1 zeigt die Mittelwerte und die Streuungen der Kategorien-Parameter und die relative Häufigkeit der Kategorien-Wahlen in den Daten. Für alle Parameter und alle Skalen ist der Wert der Gelman-Rubin-Statistik kleiner als 1.02, was auf eine Konvergenz der Markov-Ketten hindeutet (siehe Anhang A.5.). Eine visuelle Inspektion der Markov-Ketten zeigt eine deutliche Mischung der zwei Markov-Ketten mit unterschiedlichen Startwerten. Bei den geschätzten Parametern handelt es sich *nicht* um geordnete Schwellenparameter, sondern um Kategorien-Leichtigkeiten, was in Tabelle 5.1 daran zu erkennen ist, dass die Kategorien-Parameter eines Items in einer ordinalen Beziehung zu der relativen Häufigkeit der Kategorien-Wahl stehen. Die Bedeutung der Kategorien-Parameter sei hier konkret an dem Item „wohl-unwohl“ der Skala „gute vs. schlechte Stimmung“ erläutert. Nach dem hier definierten Modell wird die manifeste Reaktion $x_{vi[t]}$ einerseits auf den Variabilitäts-Parameter einer Person η_v und die Leichtigkeiten der Kategorien zurückgeführt. Die Kategorien-Leichtigkeiten geben den Einfluss der Eigenschaften der Kategorien bezüglich der manifesten Reaktion $x_{vi[t]}$ unabhängig von der Variabilität einer Person an. Sind alle Kategorien-Leichtigkeiten gleich Null und ist auch der Variabilitäts-Parameter η_v gleich Null, so ist die Wahl einer jeden Kategorie gleich wahrscheinlich. Ist nun Information über die Kategorien auf Basis der beobachteten Daten vorhanden, so werden diese Informationen bei der Parameterschätzung in dem Modell abgebildet. Nehmen wir an, der Variabilitäts-Parameter einer Person ist $\eta_v = 0$ und die Kategorien-Leichtigkeiten entsprechen denen des Items „wohl-unwohl“ so ergeben sich

Tabelle 5.1.: Die Kategorien-Parameter der Items des MDBF, Standardfehler und beobachtete, relative Häufigkeiten der Kategorien-Wahl (kursiv gedruckt)

Item	β_{i1}	β_{i2}	β_{i3}	β_{i4}
wohl-unwohl	-1.51 (0.05)	0.04 (0.03)	1.18 (0.02)	0.29 (0.02)
relative Häufigkeit	*.02*	*.13*	*.68*	*.17*
gut-schlecht	-1.68 (0.06)	0.06 (0.03)	1.26 (0.02)	0.36 (0.03)
relative Häufigkeit	*.01*	*.13*	*.69*	*.17*
zufrieden-unzufrieden	-1.27 (0.04)	0.08 (0.02)	1.02 (0.02)	0.17 (0.02)
relative Häufigkeit	*.02*	*.17*	*.64*	*.17*
unglücklich-glücklich	-1.50 (0.05)	0.02 (0.03)	1.22 (0.02)	0.25 (0.03)
relative Häufigkeit	*.02*	*.13*	*.69*	*.16*
angespannt-entspannt	-1.42 (0.04)	0.18 (0.02)	1.02 (0.02)	0.21 (0.02)
relative Häufigkeit	*.02*	*.19*	*.61*	*.18*
unruhig-ruhig	-1.50 (0.05)	0.16 (0.02)	1.18 (0.02)	0.15 (0.02)
relative Häufigkeit	*.02*	*.16*	*.67*	*.15*
müde-wach	-0.70 (0.02)	0.17 (0.02)	0.74 (0.01)	-0.20 (0.02)
relative Häufigkeit	*.07*	*.26*	*.53*	*.14*
unausgeruht-ausgeruht	-0.68 (0.02)	0.24 (0.02)	0.78 (0.01)	-0.34 (0.02)
relative Häufigkeit	*.07*	*.27*	*.54*	*.11*

aus dem Modell unabhängig von der zuletzt gewählten Kategorie folgende erwarteten Wahrscheinlichkeiten der Wahl einer Kategorie x:

$$\boldsymbol{p} = [0.04, 0.18, 0.56, 0.23]. \tag{5.1}$$

Dies bedeutet praktisch, dass nach dem Modell in den beobachteten Daten die Kategorie 1 („sehr unwohl“) sehr selten gewählt wurde. Die relative Häufigkeit der Wahl der Kategorie 1 beträgt 0.02, wobei die Kategorie 3 („eher wohl“) am häufigsten gewählt wurde. Die relative Häufigkeit der Wahl beträgt 0.68. Dies ist mit dem theoretischen Ergebnis zur Schätzung der Kategorien-Leichtigkeiten kongruent (vgl. Sektion 4.7.). Je häufiger eine Kategorie in den Daten gewählt wird, desto höher ist der entsprechende Leichtigkeits-Parameter. Aus einer praktischen Perspektive könnte es durchaus gerechtfertigt sein, die Kategorien 1 und 2 zusammenzulegen. Die Standardfehler der Parameter zeigen an, dass die Kategorien-Leichtigkeiten alle relativ genau erfasst werden. Aus diesem Parametern lassen sich die Schwellenparameter der Kategorien-Response-Funktionen bei zuletzt gewählter Kategorie $x_{vi[t-1]}$ bestimmen (vgl. Sektion 4.9.). Dies sei hier kurz für das Item „wohl-unwohl“ und bei zuletzt gewählter Kategorie $x_{vi[t-1]} = 1$ demonstriert:

$$\tau_{[12]} = -\beta_{1[2]} + \beta_{1[1]} = -0.4 - 1.51 = -1.51 \tag{5.2}$$

$$\tau_{[23]} = -\beta_{1[3]} + \beta_{1[2]} = -1.18 + 0.04 = -1.14 \tag{5.3}$$

$$\tau_{[34]} = -\beta_{1[4]} + \beta_{1[3]} = -0.29 + 1.18 = 0.89. \tag{5.4}$$

Es zeigt sich, dass die klassischen Schwellenparameter bei einer zuletzt gewählten Kategorie von $x_{vi[t-1]} = 1$ in einem geordneten Verhältnis zueinander stehen. Gleiches gilt im übrigen auch für die anderen Items der Skalen des MDBF, wie sich der Leser mittels der im Anhang A 1.3. dokumentierten R-Funktion zur grafischen Darstellung der Kategorien-Funktionen auf Basis der Schwellenparameter leicht überzeugen kann.

Tabelle 5.2 zeigt einige wichtige Kennwerte der Skalen unter dem Modell. μ_η und σ_η sind die geschätzten Mittelwerte und Streuungen der latenten Trait-Verteilungen. Vergleichen wir die Mittelwerte μ_η der latenten Trait-Verteilungen miteinander, so zeigt sich, dass die Skala „gute Stimmung“ die höchste Stabilität der drei Skalen aufweist, wohingegen die Skala „Wachheit“ die höchste Variabilität der Skalen aufweist, da ja der Parameter

Tabelle 5.2.: Kennwerte der Skalen unter dem Modell

Skala	μ_η	σ_η	Rel	pD	DIC	n
gute Stimmung	-1.06 (0.04)	0.49 (0.03)	0.95	183.7	99342.9	64760
Ruhe	-0.86 (0.03)	0.42 (0.03)	0.89	179.5	55788.5	32380
Wachheit	-0.71 (0.03)	0.30 (0.02)	0.84	177.9	67033.4	32380

η_v die Variabilität der manifesten Zeitreihen abbildet. Die Skala „gute Stimmung“ zeigt die höchste Streuung der latenten Trait-Verteilung σ_η. Beachtenswert ist, dass für die Parameter μ_η und σ_η durch die MCMC-Schätzung Posterior-Verteilungen anfallen, die es erlauben, die Präzision dieser sog. Hyper-Parameter zu bewerten. Die Streuungen der Posterior-Verteilungen sind mit 0.03 und 0.02 relativ gering, was die Interpretation zulässt, dass die Parameter relativ genau geschätzt werden. Die Andrich-Reliabilitäten (Andrich, 1988) der Skalen rangiert zwischen 0.95 für die Skala „gute Stimmung“ bis 0.84 für die Skala „Wachheit“, so dass die Messgenauigkeit der drei Skalen als gut bis sehr gut bezeichnet werden kann. Die Reliabilitäten stehen in einem monotonen Verhältnis zur Streuung der jeweiligen latenten Trait-Verteilung. Je höher die Streuung der latenten Trait-Verteilung σ_η, desto höher die Separation der Personen, desto höher die interindividuelle Variabilität und desto höher die Reliabilität. Die Modellkomplexität pD liegt zwischen 183.7 und 177.9. Die höchste Modellkomplexität weist die Skala „gute Stimmung“ auf. pD entspricht in etwa der Anzahl der geschätzten Parameter. Für jede der 165 Personen wird ein Variabilitäts-Parameter geschätzt, pro Item werden vier Kategorien-Leichtigkeiten β_{ix} geschätzt, aus denen sich für jeden zuletzt gewählten Wert $x_{vi[t-1]}$ drei Schnittpunkte der bedingten Kategorien-Response-Funktionen τ_{ix} berechnen lassen. Die latente Trait-Verteilung wird durch 2 Parameter modelliert. Das *deviance information criterion* rangiert zwischen 99342.9 und 55799.5. n ist die Anzahl der beobachteten Daten und entspricht somit der Anzahl beobachteter Ratings auf den jeweiligen Items der Skalen. Inhaltlich bedeuten diese Ergebnisse, dass in der Stichprobe die Skala „gute Stimmung“ insgesamt die höchste Stabilität der multidimensionalen Zeitreihen aufweist und die Skala „Wachheit“ die höchste Variabilität. Die selbst eingeschätzte Stimmung ist also

insgesamt über alle Personen betrachtet stabiler, als die selbst eingeschätzte Wachheit. Die Skala „gute Stimmung“ zeigt die höchste Streuung der latenten Trait-Verteilung. Dies bedeutet, dass die Personen in der Stichprobe sich auf dieser Skala am stärksten hinsichtlich der Variabilität der Ratings unterscheiden, wohingegen die Trait-Verteilung der Skala „Wachheit“ lediglich eine Streuung von $\sigma_\eta = 0.30$ aufweist. Dies bedeutet, die Personen in der Stichprobe sind hinsichtlich der Variabilität auf der Skala „Wachheit“ homogener, als auf der Skala „gute Stimmung“. Mittels des Modells ist es also nicht nur möglich, einzelne Personen auf einem latenten Kontinuum anzuordnen, sondern auch einzelne Konstrukte hinsichtlich der Variabilität zwischen den Personen zu vergleichen. Insgesamt werden die Variabilitäten auf den Skalen relativ reliabel erfasst.

5.5. Bewertung der individuellen Messgenauigkeit und der Modellpassung

Die Diagramme in der linken Spalte von Abbildung 5.1 zeigen die Histogramme der standardisierten Residuen der drei Skalen über alle Beobachtungen n. Es zeigt sich für alle Skalen, dass die Mehrzahl der Residuen sich eng um den Wert 0 konzentrieren, was für eine relativ gute Passung der Modelle spricht. Da ein standardisiertes Residuum bei Modellpassung eine z-Verteilung aufweist, könnte eine interpretative Faustregel sein, dass alle Residuen vom Betrag her größer als 2 als auffällig anzusehen sind. Nur sehr wenige der Residuen überschreiten dieses Kriterium. Die Diagramme der rechten Spalte von Abbildung 5.1 zeigen die Skalenwerte der 165 Personen inklusive 95%- Kredibilitäts-Intervall. Neben einer globalen Schätzung der Reliabilität mit der Methode von Andrich fallen personenbezogene Kredibilitäts-Intervalle, bzw. Konfidenz-Intervalle an, die es erlauben, die Erfassung der Fähigkeit einer konkreten Person hinsichtlich der Genauigkeit zu beurteilen. Die optische Inspektion der Skalenwerte und der Kredibilitäts-Intervalle stützt das Ergebnis auf Basis der Andrich-Reliabilität: Die Fähigkeitsparameter sind relativ gut separiert. Ferner zeigt sich, dass Merkmalsausprägungen im unteren Bereich der latenten Skala, die mit einer hohen Stabilität der manifesten Ratings einhergehen, die

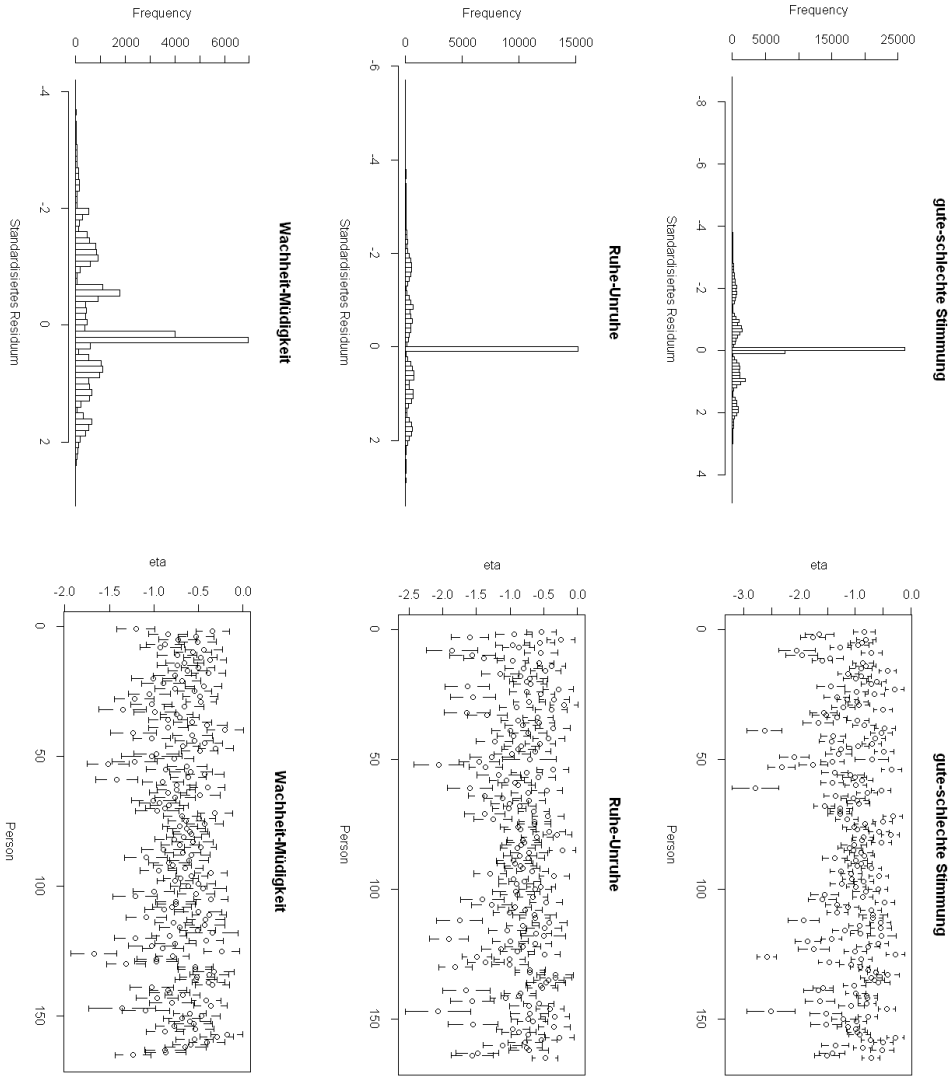

Abbildung 5.1.: Standardisierte Residuen der Modelle und Skalenwerte η_v der Personen

Tabelle 5.3.: Item-Fit und globaler Skalen-Fit

Item	Outfit	χ^2	*df*	*p*
wohl-unwohl	1.06	17217.63	16190	.00
gut-schlecht	1.00	16239.95	16190	.39
zufrieden-unzufrieden	0.99	16068.96	16190	.75
glücklich-unglücklich	0.88	14232.33	16190	1.00
angespannt-entspannt	1.01	16407.19	16190	.11
unruhig-ruhig	0.95	15326.77	16190	1.00
müde-wach	1.01	16308.87	16190	.25
unausgeruht-ausgeruht	0.99	16022.49	16190	.82
Skala				
gute-schlechte Stimmung	0.98	63758.87	64760	1
Ruhe-Unruhe	0.98	31733.96	32380	.99
Wachheit-Müdigkeit	1	32331.36	32380	.57

individuelle Merkmalsausprägung am ungenauesten geschätzt wird. Ein Ergebnis, dass mit der Simulation zur Bias und Varianz der Schätzer kongruent ist.

Zur statistischen Bewertung der Modellpassung wird der Outfit über die Gesamtskala und über die Items berechnet. Zudem werden entsprechende χ^2-Test durchgeführt.

Die Ergebnisse sind in Tabelle 5.5 dokumentiert. Hier ist eine Anmerkung zu den Freiheitsgraden der item-bezogenen Statistiken angebracht. Insgesamt sind in dem Originaldatensatz pro Item 16355 Reaktionen dokumentiert. Die ersten Reaktionen der 165 Personen auf ein Item werden nicht modelliert, somit entspricht die Anzahl der Freiheitsgrade pro Item 16355-165=16190. Die χ^2-Test-Statistiken unterstreichen das Ergebnis der optischen Inspektion der Residuen.

Insgesamt weisen alle Items relativ gute Outfit-Werte auf. Da ein Item mit einem sehr hohen Outfit ebenfalls sehr hohe Residuen aufweist, deutet dies darauf hin, dass das Modell das Antwortverhalten nicht hinreichend gut beschreibt. Für die Skala „gute Stimmung“ lässt sich feststellen, das das Item „wohl-unwohl“ im Mittel höherer Residuen aufweist, als unter dem Modell erwartet würde. Das Item „unglücklich-glücklich“ hingegen

passt besser, als unter dem Modell erwartet würde, was durch den geringen Outfit und dem hohen p-Wert angezeigt wird. Die Items „gut-schlecht“ und „zufrieden-unzufrieden“ hingegen zeigen eine sehr gute Modellpassung. Von den Items der Skala „Ruhe“ passt das Item „ruhig-unruhig“ sehr gut, wohingegen das Item „unruhig-ruhig“ besser passt, als unter dem Modell erwartet würde. Beide Items der Skala „Wachheit“ weisen eine sehr gute Modellpassung auf. Auch wenn einige der Items die kritischen Grenzen der p-Werte von .025 und .975 über-, bzw. unterschreiten, ist es angesichts der guten bis sehr guten Reliabilitäten der Skalen fraglich, ob einzelne Items aus der Analyse ausgeschlossen werden sollten. Zur Bewertung der Gesamt-Passung des Modells werden die quadrierten standardisierten Residuen (z-Werte) über die Gesamtdaten summativ aggregiert und ein entsprechender χ^2-Test wird durchgeführt. Die Skala „Wachheit“ zeigt insgesamt eine sehr gute Passung, wohin die Skalen „Ruhe“ und „Gute Stimmung“ eher besser passen, als unter dem probabilistischen Modell erwartet würde.

Hier stellt sich die Frage, wie eine „zu gute“ Modellpassung zu verstehen ist. Betrachten wir die Schätzer der Personen-Parameter η_v in Abbildung 5.1, so fällt auf, dass die Skalen „gute Stimmung“ und „Ruhe“ Parameterschätzer aufweisen, die stark am unteren Ende der Variabilitäts-Skala angesiedelt sind. Diese niedrigen Parameter korrespondieren mit sehr stabilen Ratings der Personen. Es ist zu vermuten, dass diese stabilen, deterministischen Zeitreihen, die fast keine Variabilität aufweisen, dazu beitragen, dass eine Vorhersage zu einfach ist und das Modell somit zu gut passt. Bezogen auf die Items bedeutet ein zu geringer Outfit, dass die manifesten Zeitreihen dieses Items stabiler sind, als unter dem Modell erwartet würde und nur wenig manifeste Variabilität auf diesen Items vorliegt.

5.6. Der Zusammenhang zwischen der MASD und $\hat{\eta}_v$

Um zu überprüfen, ob die Schätzung der Modellparameter auf Basis der absoluten sukzessiven Differenz zu plausiblen Ergebnissen gelangt, ist es zweckmäßig, die Beziehung zwischen den manifesten absoluten sukzessiven Differenzen (MASD) der multivariaten Zeitreihen und den per Modell extrahierten Parametern ($\hat{\eta}_v$)grafisch zu betrachten. Aus

den theoretischen Überlegungen ging hervor, dass der Zusammenhang zwischen den absoluten Differenzen und dem Variabilitäts-Parametern des Modells bei Modellgeltung von sigmoider Gestalt sein sollte (vgl. Abbildung 4.2) .

Abbildung 5.2 zeigt die Zusammenhänge der mit der MCMC-Methode geschätzten Parameter η_v und den jeweiligen mittleren absoluten Differenzen der Personen. Die absoluten Differenzen wurden für jede Person über alle Items einer Skala berechnet. Eine mittlere absolute Differenz von 0.7 auf der Skala „gute Stimmung“ dass die absoluten Differenzen der Ratings einer Person von $[t-1]$ zu t im Mittel über alle Items betrachtet 0.7 betragen. Abbildung A zeigt den Zusammenhang für die Skala „gute Stimmung“, Abbildung B zeigt den Zusammenhang für die Skala „Ruhe - Unruhe“ und Abbildung C zeigt den Zusammenhang für die Skala „Wachheit-Müdigkeit“ über alle Personen der Stichprobe. Wie aus den Abbildungen zu entnehmen, ist der Zusammenhang mit der theoretisch hergeleiteten Annahme der sigmoiden Gestalt des Zusammenhangs vereinbar. Allerdings sind empirisch keine Variabilitäten aufgetreten, die mit einem $\eta_v > 0$ einhergehen. Das Modell wäre also in der Lage, noch viel stärkere Variabilitäten abzubilden. Wie in einem vorherigen Abschnitt simulativ gezeigt wurde, tritt eine Verzerrung der Schätzer für η_v erst ab einem Bereich von $\eta_v = 1$ auf. Variabilitäten, die mit einem solch hohen Parameter einhergehen, wurden in dem Datensatz allerdings nicht beobachtet. Zudem ist zu verzeichnen, dass der empirische Zusammenhang von nicht ganz perfekt sigmoider Gestalt ist. Vielmehr ist zu beobachten, dass vereinzelt Variabilitäts-Parameter auftreten, die wesentlich größer sind, als bei perfekter Modellpassung erwartet würde. Insgesamt ist allerdings zu verzeichnen, dass die Parameter des Modells in einem monotonen Verhältnis zur mittleren absoluten Differenz stehen.

5.7. Korrelation der Variabilität mit ausgewählten Skalen des NEO-FFI

Tabelle 5.4 zeigt die Interkorrelationen der während der Labor-Sitzung erhobenen Persönlichkeitsmerkmale und der Variabilitäts-Skalen. Auf den ersten Blick ist auffällig, dass

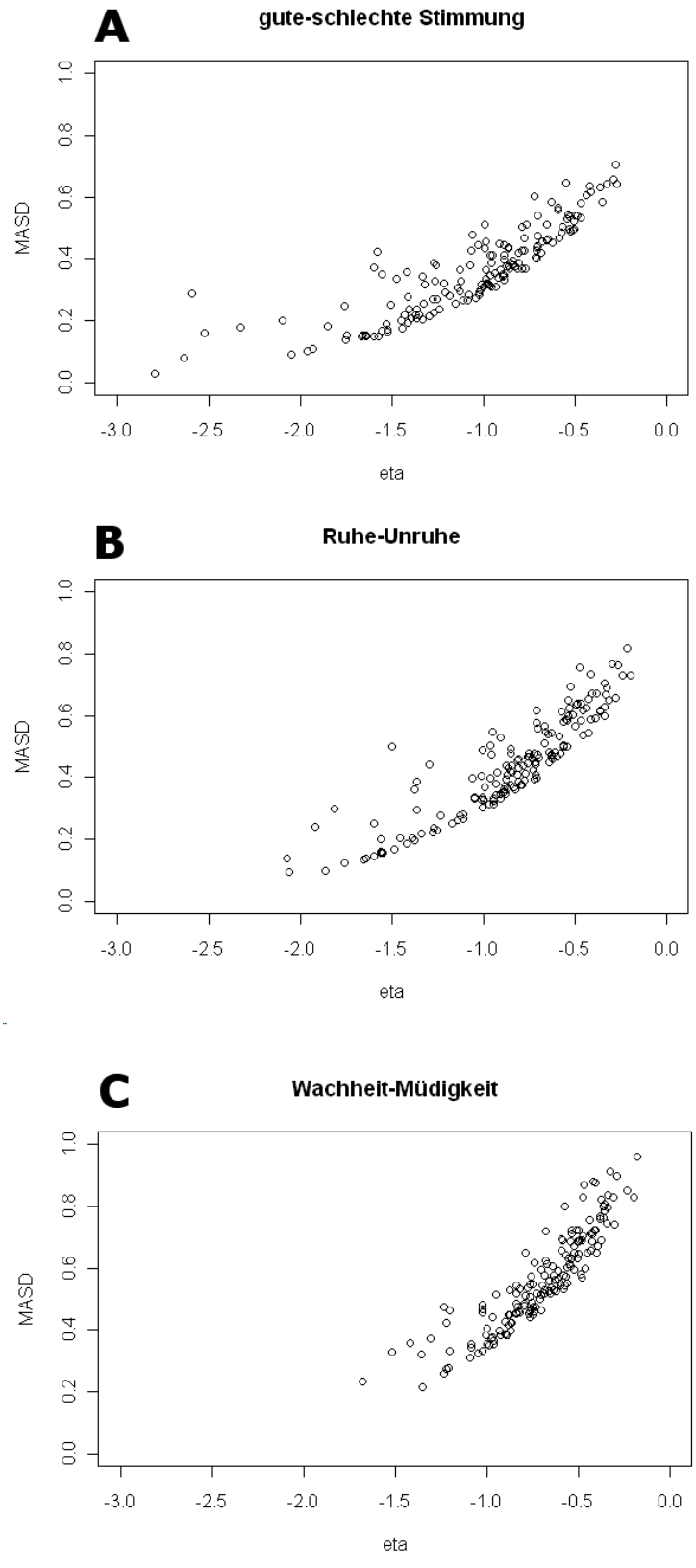

Abbildung 5.2.: Der Zusammenhang zwischen MASD und η_v für die drei Subskalen des MDBF

Tabelle 5.4.: Bivariate Korrelationen zwischen Persönlichkeitsmerkmalen, den Variabilitäts-Skalen und der MASD

	N	E	G	Stimmung	Ruhe	Wach	MASD Stimmung	MASD Ruhe
E	-0.42^{***}							
G	-0.30^{***}	0.16^{*}						
Stimmung	0.07	0.09	-0.23^{**}					
Ruhe	−0.09	0.07	−0.10	0.70^{***}				
Wach	−0.13	0.12	−0.14	0.58^{***}	0.59^{***}			
MASD Stimmung	0.09	0.10	-0.19^{*}	0.85^{***}	0.69^{***}	0.53^{***}		
MASD Ruhe	0.13	0.03	−0.11	0.64^{***}	0.90^{***}	0.48^{***}	0.77^{***}	
MASD Wach	−0.07	0.11	−0.14	0.58^{***}	0.63^{***}	0.89^{***}	0.64^{***}	0.65^{***}

Anmerkungen:

*: $p < .05$; **: $p < .01$, ***: $p < .001$

N: Neurotizismus; E: Extraversion; G: Gewissenhaftigkeit

Stimmung: Variabilitäts-Parameter gute Stimmung; Ruhe: Variabilitäts-Parameter Ruhe; Wach: Variabilitäts-Parameter Wachheit

MASD Stimmung: MASD Stimmung; Ruhe: MASD Ruhe; Wach: MASD Wach

die Persönlichkeitsmerkmale des NEO-FFI untereinander hoch korrelieren. So zeigen sich hohe negative Korrelationen der Skala Neurotizismus mit den Skalen Extraversion und Gewissenhaftigkeit. Zwischen der Skala Gewissenhaftigkeit und Extraversion zeigt sich eine mittlere, negative Korrelation. Die sehr hohen Interkorrelationen der Variabilitäts-Skalen deuten darauf hin, dass Personen, die eine hohe Variabilität auf einer der Skalen aufweisen, ebenfalls eine hohe Variabilität auf den anderen Skalen zeigen. Die einzige signifikante Korrelation zwischen den Persönlichkeitsmerkmalen und den Variabilitäts-Skalen findet sich zwischen den Variablen Gewissenhaftigkeit und „gute Stimmung“. Personen, die eine hohe Ausprägung auf der Variable Gewissenhaftigkeit aufweisen, zeigen ein eher stabileres Antwortverhalten bei der Beantwortung der Items der Skala „gute Stimmung“. Insgesamt deuten die Ergebnisse darauf hin, dass die Parameter des Modells etwas spezifisch anderes erfassen, als die in der Laborsitzung erhobenen Persönlichkeitsmerkmale. Ferner liegt die Interpretation nahe, dass die MASD-Variabilität der Zeitreihen selbst eher Trait-Charakter besitzt, d.h. eine Person, die eine hohe Variabilität auf einer der intraindividuellen Zeitreihen aufweist, zeigt auch tendenziell eine höhere Variabilität auf den anderen Zeitreihen. Bezüglich der Korrelationen der manifesten, mittleren absoluten sukzessiven Differenzen (MASD) mit den Persönlichkeitsmerkmalen zeigt sich ein zu den Variabilitäts-Skalen des Modells ähnliches Bild. Es zeigt sich nur eine negative Korrelation ($r = -.23^{**}$) der mittleren absoluten sukzessiven Differenz der Items der Skala „gute Stimmung“mit dem Persönlichkeitsmerkmal Neurotizismus. Auch die mittleren absoluten sukzessiven Differenzen sind untereinander hoch korreliert. Die Korrelationen der mittleren absoluten sukzessiven Differenzen mit den Variabilitäts-Skalen sind sehr hoch und rangieren im Bereich von .85 bis .95. Diese hohen Zusammenhänge sind mit den Ergebnissen der Modellentwicklung konform. Die mittleren absoluten sukzessiven Differenzen sollten a priori in einem monotonen Zusammenhang mit den Parametern des Modells stehen.

5.8. Zusammenfassende Darstellung der Ergebnisse der Modellanwendung

Zu Fragestellung II.1.: Die Anwendung der MCMC-Methode zur Schätzung der Modellparameter anhand des Datensatzes von Crayen et al. (in Druck) führt zu konvergenten Schätzern der Modellparameter, was sich an der Mischung der zwei eingesetzten Markov-Ketten zeigt. Die Gelman-Rubin-Statistik für alle Parameter weist einen Wert kleiner als 1.02 auf (vgl. Anhang A.5.1. - A.5.3.). Dies bedeutet, dass das Modell ein absolutes Maximum der Likelihood-Funktion besitzt. Die theoretisch abgeleitete Bedeutung der Kategorien-Parameter wird durch das empirische Ergebnis gestützt. Das Profil der itemspezifischen Kategorien-Leichtigkeiten folgt der manifesten Verteilung der beobachteten Kategorien-Häufigkeiten, was vor dem Hintergrund der Tatsache, dass die Häufigkeit der Wahl einer Kategorie auf einem Item eine suffiziente Statistik zur Schätzung des Parameters darstellt, nicht überrascht.

Zu Fragestellung II.2.: Die Modellierung der latenten Trait-Verteilungen erfolgt mit der MCMC-Methode problemlos und die entsprechenden Markov-Ketten konvergieren. Die geschätzten Streuungen der latenten Trait-Verteilungen deuteten darauf hin, dass eine interindividuelle Varianz der Personen hinsichtlich der Merkmalsausprägung vorliegt. Die geschätzten Mittelwerte der latenten Trait-Verteilungen zeigen, dass die vorhergesagten manifesten Verteilungen der Ratings eine eingipfelige Form aufweisen und im Mittel über alle Personen betrachtet eine Abhängigkeit des Ratings $x_{vi[t-1]}$ vom Rating $x_{vi[t]}$ vorliegt.

Zu Fragestellung II.3.: Die individuelle Merkmalsausprägung entspricht dem Parametern η_v. Die optische Inspektion der Posterior-Verteilungen der Variabilitäts-Parameter legt nahe, dass die individuelle Merkmalsausprägung in mittleren Bereichen der Variabilitäts-Skala genauer erfasst wird, als in unteren Extrembereichen der Skala. Extreme positive Variabilitäten ($\eta_v > 0$) traten in der Stichprobe nicht auf.

Zu Fragestellung II.4.: Die Reliabilität, basierend auf der Methode von Andrich (Andrich, 1988), ist für alle drei Skalen des MDBF als gut bis sehr gut zu bewerten. Auch die grafische Inspektion der Personen-Parameter und deren Standardfehler unterstreicht dieses

Ergebnis. Die vergleichsweise hohe Reliabilität der Skala „Wachheit“ könnte darauf zurückzuführen sein, dass diese Skala vier anstatt lediglich zwei Items aufweist.

Zu den Fragestellungen II.5. und II.6.: Die grafische Inspektion des Histogramms der standardisierten Residuen über alle Items zeigt einen stark ausgeprägten Gipfel der Häufigkeiten um Null, was darauf hindeutet, dass das Modell die Daten relativ gut beschreibt. Die Berechnung des Outfits der Items zeigt für alle Items relativ akzeptable Werte. Die χ^2-Statistiken der Items „wohl-unwohl“ und „glücklich-unglücklich“ der Skala „gehobene Stimmung“ weisen kritische Werte auf. Das Item „wohl-unwohl“ weist einen Underfit und das Item „glücklich-unglücklich“ weist einen Overfit auf. In der Skala „Ruhe“ zeigt das Item „ruhig-unruhig“ einen leichten Overfit, die Items der Skala „Wachheit“ hingegen passen beide sehr gut. Die Bewertung des Outfits über die Gesamt-Daten zeigt für die Skala „Wachheit“ eine gute Passung. Das Antwortverhalten auf der Skala „gehobene Stimmung“ ist insgesamt deterministischer als unter dem Modell erwartet würde. Gleiches gilt für die Skala „Wachheit“.

Zu den Fragestellungen II.7. bis II.9.: Die grafische Inspektion des Zusammenhanges zwischen den mittleren absoluten Differenzen der drei Skalen des MDBF mit den Modellparametern zeigt eine mit dem theoretisch abgeleiteten logistischen Zusammenhang konforme Form. Allerdings weichen einige Personen von dem theoretisch vermuteten Zusammenhang ab. Konform mit den Ergebnissen der Modellentwicklung zeigt sich ein starker Zusammenhang zwischen den Variabilitäts-Parametern des Modells und den manifesten, mittleren absoluten Differenzen der intraindividuellen Zeitreihen. Die Variabilitäts-Parameter des Modells und die mittleren absoluten Differenzen der manifesten Zeitreihen sind hoch interkorreliert. Zudem sind die Variabilitäten der Personen zwischen den Skalen des MDBF hoch korreliert. Allerdings zeigen sich keine eindeutig interpretierbaren Zusammenhänge zwischen den Persönlichkeitsvariablen des NEO-FFI und den Variabilitäts-Parametern Gleiches gilt für die manifesten, mittleren, absoluten Differenzen in den intraindividuellen Zeitreihen.

6. Diskussion

Zielsetzung der Diskussion ist es, die theoretische und praktische Signifikanz der Arbeit vor dem Korpus der bestehenden Forschung aufzuzeigen, hierbei auch auf die Grenzen und Schwierigkeiten der Vorgehensweise einzugehen und Perspektiven für die weitere Forschung aufzuzeigen.

6.1. Diskussion der modelltheoretischen Ergebnisse

Es hat sich gezeigt, dass die Anwendung der Maximum-Entropie-Methode sich dazu eignet, neue, probabilistische Testmodelle zu generieren. Das resultierende Modell ähnelt sehr stark dem PCM von Masters, lediglich die Scoring-Funktion unterscheidet sich. Die Ähnlichkeit mit Rasch-Modellen war a priori zu vermuten, da im modelltheoretischen Hintergrund versucht wurde zu zeigen, dass der globale Modellierungsansatz von Rasch mit dem Maximum-Entropie-Ansatz der Modellierung formal kompatibel ist. Wichtige Modelleigenschaften, die für die Anwendung des Modells zentral sind, wie die Erwartungswerte der manifesten Statistiken und die erwartete Varianz der manifesten Statistiken unter dem Modell, lassen sich vor dem Hintergrund des von Jaynes (2003) dargestellten Formalismus problemlos herleiten. Hierbei zeigt sich eine Kongruenz mit gängigen Ergebnissen in der Statistik und auch mit in der psychometrischen Literatur bekannten Ergebnissen zu Rasch-Modellen. Die Maximum-Entropie-Methode der Modellgenerierung scheint sich also zur Generierung neuer probabilistischer Testmodelle zu eignen. Zur Generierung eines Modells müssen lediglich die Scoring-Funktionen und die Zustandssumme definiert werden sowie der personenbezogene Parameter gegebenenfalls linear zerlegt werden. Je nach Wahl resultieren das dichotome Rasch-Modell, das Partial-Credit-Modell oder auch

neue, bisher unbekannte Modelle. Da die kanonische Maximum-Entropie-Verteilung zur Exponentialfamilie gehört, existieren automatisch suffiziente Statistiken zur Schätzung der Parameter in Form von (aggregierten) Erwartungswerten der jeweiligen Statistik. Psychometrisch ist dies insofern interessant, als dass es möglich ist, eine manifeste Statistik, wie z.B. die absolute sukzessive Differenz oder die intraindividuelle Standardabweichung als Scoring-Funktion zu definieren und somit ein probabilistisches Testmodell zu erzeugen, das Parameter enthält, für welche die jeweiligen Erwartungswerte der Scoring-Funktion suffiziente Statistiken darstellen. Durch die Wahl der Scoring-Funktionen und der Zustandssumme ist also eine relativ flexible Modellgenerierung möglich.

Ein weitere Stärke des Maximum-Entropie-Ansatzes liegt darin, dass die Formalismen sehr gut ausgearbeitet sind, einem kohärenten theoretischen Framework entspringen und die resultierenden Modelle transdisziplinär angewendet werden, d.h. nicht auf die Anwendung in einer Fachdisziplin begrenzt sind. So existiert ein breiter Literaturkorpus im Bereich des *machine learnings* und der *artificial intelligence* (AI) auf den der Autor erst relativ spät gestoßen ist. In dieser Literatur wird unter anderem der von Jaynes beschriebene Ansatz stärker aufgearbeitet und zu Lösung praktischer Fragestellung verwendet. Exemplarisch seien die Bücher *Probabilistic Graphical Models* von Koller und Friedman (Koller & Friedman, 2009) und *Artificial intelligence - a modern approach* (S. Russel & Norvig, 2010) genannt. Probabilistische Grafische Modelle (PGMs) sind eine Modellklasse, die aus der Fusion der mathematischen Graphentheorie und der Wahrscheinlichkeitstheorie hervorgehen. Das von Jaynes beschriebene Obermodell zählt zu der Klasse der *undirected graphical models* (vgl. Koller et al. 2009, Kap. 4). Wie unter anderem Koller und Friedman zeigen, lässt sich weite Reihe bekannter statistischer Modelle und Verteilungen, wie z.B. die multinomialen Logit-Modelle, die logistische Regression, das Perzeptron von Boltzmann, Hidden Markov Modelle, die Gibbs - Verteilung, Bayes Netzwerke und das Ising-Modell in dieses übergeordnete Framework einbetten. Im Rahmen der modernen AI werden PGMs zur Implementierung von *reasoning under uncertainty* eingesetzt, es existieren jedoch auch praktische Anwendungen in der Physik (siehe Gibb's Verteilung und das Ising-Modell), in den Neurowissenschaften (Schneidman et al., 2006)

und der komputationellen Linguistik (Berger et al., 1996). Wie in der Arbeit versucht wurde zu zeigen, bedient sich auch die quantitativ arbeitende Psychologie implizit in der Item-Response-Theorie ähnlicher Methoden, wie z.B. die statistische Mechanik (Jaynes, 1957a, 1957b) oder die Forschung im Bereich Künstlicher Intelligenz (S. Russel & Norvig, 2010), lediglich der Anwendungsbereich, bzw. die Forschungsgegenstände und auch die Terminologie unterscheiden sich. So wird z.B. im Bereich des *machine learnings* unter Parameterschätzung „maschinelles Lernen“ verstanden. Eine Implikation für die weitere methodische, modelltheoretische Forschung in der Psychologie könnte darin liegen, diese breit angelegte und besonders im Bereich des *machine learnings* gut ausgearbeitete Modellklasse zu erschließen und als eine Möglichkeit der Lösung komplexer Modellierungsprobleme zu begreifen. Ein erster Schritt zur Erschließung könnte darin bestehen, die bereits bestehenden Parallelen der in der Psychologie verwendeten Methoden explizit zu machen und deren Anwendbarkeit zu überprüfen. So zeigen zum Beispiel lineare Strukturgleichungsmodelle eine konzeptionell starke Ähnlichkeit mit Bayes-Netzwerken. Ferner bietet die Modellklasse eine Möglichkeit zum interdisziplinären Dialog zwischen Informatikern, Statistikern und Psychologen. Eine wesentliche Schwierigkeit beim Abfassen dieser Arbeit lag in notationellen und konzeptuellen Unterschieden zwischen den in der Bayes-Statistik verwurzelten Ansätzen von Jaynes und den frequentistisch orientierten Ansätzen in der Psychometrie. Bei weiteren Untersuchungen auf diesem Gebiet ist es ratsam, die konzeptionellen Ansätze von vornherein deutlich voneinander abzugrenzen und sich auf eine kohärente Notation festzulegen. Erste Ansätze zur Verwendung von Bayesianischen Methoden in der Item-Response-Theorie finden sich bei Fox (Fox, 2010).

6.2. Diskussion des resultierenden Testmodells

In der vorliegenden Arbeit wurde die Maximum-Entropie-Methode angewendet, um ein IRT-Modell zur Erfassung intraindividueller Variabilität auf Basis der absoluten sukzessiven Differenz (Ebner-Priemer et al., 2009) zu generieren. Das resultierende Modell gehört zur Exponentialfamilie und besitzt daher suffiziente Statistiken zur Schätzung der Parameter (Pitman, 1936; Koller & Friedman, 2009). Zudem ist das resultierende Modell

dem PCM von Masters (Masters, 1982) sehr ähnlich. Ein Unterschied liegt in der Scoring-Funktion, aus welcher ein Modell zur Beschreibung bedingter Wahrscheinlichkeiten resultiert. Der Variabilitäts-Parameter η steht in einer linearen Beziehung zu den Logits benachbarter Kategorien-Wahrscheinlichkeiten, was einen sinnvollen Vergleich der Parameter auf einer Differenzenskala ermöglicht. Die Kategorien-Parameter des Modells stehen in einer monotonen Beziehung zur relativen Häufigkeit der Wahl einer Kategorie. Das Modell selbst ist relativ einfach gehalten und die Variabilität auf allen personenspezifischen, diskreten, intraindividuellen Zeitreihen wird - neben den Kategorien-Leichtigkeiten - lediglich auf eine latente, personenspezifische Variable zurückgeführt. Relativ neu ist die Skalierung der Variabilität eines Markov-Prozesses auf einer Differenzenskala im Sinne eines Rasch-Modells. Intraindividuelle Variabilität ist in diesem Sinne nicht als Variabilität einer latenten Variable konzipiert, sondern eher als zeitstabile Disposition, die selbst parametrisch erfasst werden kann und die Variabilität auf manifesten Variablen abbildet . Gegenüber der in der Literatur dominierenden Verwendung manifester Indices bietet die latente Modellierung explizite Vorteile. Ein Messmodell liegt vor, dessen Passung überprüft werden kann. Nur wenn ein entsprechendes Messmodell passt, ist es sinnvoll, aggregierte Statistiken als Indikator für die latente Merkmalsausprägung zu verwenden. Neben der globalen Modellpassung lassen sich item- und personenspezifische Statistiken zur Bewertung der jeweiligen Passungen berechnen, zudem ist die Reliabilität der Gesamtmessung schätzbar und Homogenitätshypothesen sind potentiell prüfbar. Somit ist es mittels des Modells möglich zu prüfen, ob es psychometrisch sinnvoll ist, die mittlere absolute Differenz als Indikator für ein latentes Merkmal zu verwenden. Eine Beschränkung des hier generierten Modells liegen darin, dass in dem konstruierten Modell lediglich ein Markov-Prozess erster Ordnung beschrieben wird, da lediglich die absolute Differenz $|x_{vi[t]} - x_{vi[t-1]}|$ in die Modellgleichung eingeht. Es sollte allerdings möglich weitere Terme in die Modellgleichung einzubauen, um die Abhängigkeiten von weiteren vorhergehenden Ratings zu berücksichtigen (z.B. $|x_{vi[t]} - x_{vi[t-2]}|$). Der Einfachheit halber wurde in dieser Arbeit zunächst lediglich die mittlere absolute Differenz verwendet. In dem Modell werden auch folgende weitere, theoretisch mögliche Sachverhalte nicht abgebildet: situa-

tive Einflüsse auf die Variabilität und item-spezifische Einflüsse auf die Variabilität. Das Modell ermöglicht es leider nicht zu überprüfen, ob mit bestimmten Situationen oder Zeitpunkten eine höhere Variabilität der Ratings einhergeht oder nicht. Ebenso ist es theoretisch denkbar, dass bestimmte Items eine höhere Variabilität evozieren als andere. Da in dem Modell keine Parameter für diese Sachverhalte vorliegen, ist die Überprüfung von Fragestellungen solcher Art mit dem Modell auch nicht möglich. Es sollten jedoch Möglichkeiten der Modellerweiterung existieren, die diese Sachverhalte abbilden können, wie z.B. die Zerlegung der Variabilitäts-Parameter in item- und personenspezifische Komponenten. Eine weitere diagnostisch interessante Information - die mittlere Lage der intraindividuellen Zeitreihen - wird in dem hier generierten Modell ebenfalls nicht erfasst. Wie ist das Modell vor den im modelltheoretischen Hintergrund beschriebenen Modellen einzuordnen? Zunächst ist zu verzeichnen, dass das Modell intraindividuelle Variabilität als eine Trait-Variable beschreibt. Die Variabilität des Antwortverhaltens der Personen auf einem oder mehreren Items wird auf lediglich eine latente Variable zurückgeführt, die *zeitstabil* ist. Von daher unterscheidet sich das Modell von den Modellen der dynamischen Faktoranalyse, da diese Modelle davon ausgehen, dass sich der latente Trait selbst verändert und somit eine Auswirkung auf das manifeste Antwortverhalten besitzt. Zudem ist das vorliegende Modell im Vergleich zu den meisten dynamischen Faktormodellen nicht auf die Einzelfallanalyse beschränkt. Es ist allerdings zu verzeichnen, dass das Modell selbst ein dynamisches Modell ist, was sich man daran verdeutlichen kann, wie Daten aus dem Modell simuliert werden (vgl. Anhang A 1.5.): zur Simulation von Daten geht der jeweils zuletzt erzeugte Wert in die Simulation mit ein, so dass eine Wahrscheinlichkeitsverteilung für die nachfolgenden, möglichen Werte zum Zeitpunkt t durch das Modell erzeugt wird. Der simulierte Wert zum Zeitpunkt t selbst wiederum ist die Ausgangsbasis für die Erzeugung des Wertes zum Zeitpunkt $t + 1$ und so fort. Insofern besitzt das Modell eine gewisse rekursive Struktur. Im Vergleich zur Anwendung von Latent-State-Trait-Modellen auf intraindividuelle Standardabweichungen ist festzustellen, dass das vorliegende Modell es nicht erlaubt, mehrere Traits in einer Modellgleichung zu erfassen und es werden auch keine situationsabhängigen Effekte erfasst. Allerdings bietet

das vorliegende Modell den Vorteil, dass die Variabilität selbst parametrisch modelliert wird und nicht auf manifeste Standardabweichungen von Zeitreihen zurückgegriffen werden muss. Das hier generierte Modell zeigt von der Grundidee eine gewisse Ähnlichkeit zu dem Ansatz von Fleeson (Fleeson, 2001), der Traits als Verteilungen von States betrachtet. Fleeson wählt implizit als Prior-Verteilung für States eine Normalverteilung, wobei die mittlere Lage einer Zeitreihe in Form des Mittelwerts und die Variabilität in Form der Standardabweichung als Trait-Charakteristiken eine Person aufgefasst werden. Das hier entwickelte Modell berücksichtigt die kategorialen Natur des Antwortformates. Zudem wird in dem generierten Modell nicht die intraindividuelle Standardabweichung, sondern die absolute Differenzen als manifester Indikator der Variabilität aufgefasst, die latent und nicht manifest skaliert wird.

Die Parameterschätzung in der vorliegenden Arbeit erfolgte mittels der MCMC-Methode, allerdings ist deren Verwendung nicht zwingend notwendig. Es wäre auch möglich, die Modellparameter mit einer Variante der Maximum-Likelihood-Methode zu schätzen, die MCMC-Methode besitzt allerdings einige praktische Vorteile, wie z.B. die relative Einfachheit der Anwendung, die hohe Reichweite und die flexible Lösung des Problems fehlender Werten auf Basis von Posterior-Verteilungen von Missings. Allerdings wird bei der Verwendung dieser Methoden der Bayesianische Hintergrund der Statistik quasi „mitgekauft“. Da die in dieser Arbeit verwendete Methode der Modellgenerierung in Bayesianischem Gedankengut verwurzelt ist, lag es aufgrund einer gewissen Theorie-Homogenität nahe, ebenfalls die MCMC-Methode zur Bestimmung der Posterior-Verteilungen zu verwenden. Bezogen auf das Modell und die Verwendung von Bayesianischen Methoden könnte es von Interesse sein, die Performanz der MCMC-Methode mit anderen Möglichkeiten der Parameterschätzung zu vergleichen. Dies wurde in der vorliegenden Arbeit nicht durchgeführt.

Zwei weitere Punkte, die in der Arbeit nur oberflächlich angerissen wurden, sind die Berechnung der Reliabilität und die Bewertung der Modellpassung auf Basis standardisierter Residuen. Obwohl beide Verfahren einer gewissen internen Logik folgen und in der psychometrischen Literatur gut dokumentiert sind, wäre es sinnvoll, die Relia-

bilität von Andrich und die Bewertung der Modellpassung auf Basis standardisierter Residuen simulativ detaillierter zu untersuchen. So könnten z.B. Datensätze mit unterschiedlicher Varianz der latenten Trait-Verteilung aus dem Modell generiert werden. Die Varianz der Score-Verteilung kann graphisch in Beziehung mit der Andrich-Reliabilität gesetzt werden. Es ist auf Basis des theoretischen Hintergrundes zu vermuten, dass die Andrich-Reliabilität von der Streuung der latenten Trait-Verteilung abhängt: je hoher die Streuung, desto höher die Reliabilität.

6.3. Diskussion der Modellanwendung

In der Anwendung wird das generierte Modell an einem Ambulatory Assessment Datensatz von (Crayen et al., in Druck) praktisch erprobt. Die Anwendung der MCMC-Methode auf den Datensatz von (Crayen et al., in Druck) erfolgt problemlos und die Markov-Ketten konvergieren, was auf ein absolutes Maximum der Likelihood-Funktion des Modells hindeutet. Zudem sind die Personen-Parameter stark separiert, was sich an den hohen Andrich-Reliabilitäten und der Varianz der latenten Trait-Verteilungen zeigt. Die Messgenauigkeit der individuellen Parameter ist über den Standardfehler der Parameter, bzw. die Streuung der jeweiligen Posterior-Verteilung bewertbar. Die Profile der Kategorien-Leichtigkeiten entsprechen den Profilen der Häufigkeiten der Wahl der jeweiligen item-spezifischen Kategorien, was das theoretische Ergebnis hinsichtlich der suffizienten Statistiken dieser Parameter bestätigt. Das Modell passt relativ gut, was an der Verteilung der standardisierten Residuen deutlich wird. Insgesamt kann also gesagt werden, dass es möglich ist, die Personen hinsichtlich der absoluten sukzessiven Differenzen in den multivariaten Zeitreihen voneinander zu separieren. Aus einer mathematischen Perspektive funktioniert das Messmodell relativ gut. Was allerdings in dieser Arbeit nicht näher untersucht wurde ist die Validität der Messung der Stimmungsvariabilität mit dem MDBF, da der Hauptfokus darin bestand ein Modell zu konstruieren, dass Variabilität in Form der MASD mit einem IRT-Modell erfassbar macht. Um erste Hinweise auf die Validität zu bekommen wurden trotzdem die Variabilitäts-Parameter explorativ untereinander und mit den drei verfügbaren Skalen des NEO-FFI korreliert. Es zeigt sich, dass

die Variabilitäts-Parameter der Skalen des MDBF sehr hoch miteinander interkorrelieren. Dies ist ein Ergebnis, dass mit dem klassischen Ergebnis von Eid und Diener (1999) kongruent ist. Zudem sind die manifesten Variabilitäten mit den latenten Parametern korreliert, was für eine inhaltliche Validität der Merkmalserfassung spricht. Allerdings zeigen sich überraschender Weise keine Korrelationen der Variabilitäten mit der Variable Neurotizismus, d.h. in der Stichprobe der Studierenden werden keine Zusammenhänge zwischen der Stimmungsvariabilität und Neurotizismus gefunden. Dies gilt sowohl für die Variabilitäts-Parameter des Modells, als auch für die mittleren, absoluten Differenzen auf den manifesten, intraindividuellen Zeitreihen. Dieses Befund könnte einerseits darauf zurückzuführen sein, dass die studentische Stichprobe insgesamt zu homogen ist, um Theorien über den Zusammenhang zwischen intraindividueller Stimmungsvariabilität und Persönlichkeitsmerkmalen zu prüfen. Eine weiterer Grund könnte sein, dass es sich bei den manifesten Variabilitäten in den Zeitreihen um ein Response-Set handelt. Um diesen Einwand zu entkräften wäre es nötig mit eigenständigen, theoriegeleiteten Untersuchungen zu prüfen, inwiefern die Stimmungsvariabilität - erfasst mit dem MDBF - mit theoretisch relevanten Kriterien in einer geeigneten Population korreliert. So wäre es denkbar zu untersuchen, inwiefern die Stimmungsvariabilität, erfasst mit dem MDBF mit Diagnosen wie z.B. Borderline Persönlichkeitsstörung (BPS) oder bipolare Störung zusammenhängt. Der Fokus der vorliegenden Arbeit liegt allerdings darin, ein Messmodell zu entwickeln, dass es erlaubt, Fragestellungen dieser Art mit Hilfe eines Modells der Item-Response-Theorie im Rahmen von Ambulatory Assessments zu prüfen. Die Anwendung des Modells ist nicht auf die Skalen des MDBF beschränkt.

Bezüglich der Technik des Ambulatory Assessments ist zu verzeichnen, dass diese Methode der Datenerhebung potentiell nützliche, neue Möglichkeiten der psychologischen Diagnostik und auch der Intervention erschließt. Da sehr viele Beobachtungen im natürlichen Kontext anfallen, ist zu vermuten, dass sich die Aussagekraft und auch die Messgenauigkeit der Merkmalserfassung mit geeigneten Modellen erhöht werden kann. Zudem können Aussagen über Variabilität getroffen werden, die tatsächlich auf längsschnittlichen Beobachtungen fußen und nicht lediglich auf einem einmaligen Selbst- oder

Fremdbericht. Ferner ermöglicht die Beobachtung im natürlichen Setting die Prüfung von zeitlichen und situativen Effekten auf die subjektive Befindlichkeit oder Leistung. Zum Beispiel könnte die Technik sich - neben der Grundlagenforschung - dazu eignen, die Wirksamkeit psychologisch-therapeutischer Interventionen in einem natürlichen Setting zu evaluieren. Werden zusätzlich situative Faktoren erhoben, wäre es möglich, Kontingenzen zwischen Situation und Verhalten und Erleben aufzudecken.

6.4. Diskussion der Verwendung Bayesianischer Ansätze in der vorliegenden Arbeit

In den Methoden der Psychologie dominiert eher ein klassisch-frequentistischer Ansatz, welcher unter anderem auf die weitreichenden Arbeiten von R.A. Fischer zurückzuführen ist. Zu nennen wären hier die besonders einflussreichen und für die Akzeptanz der Sozialwissenschaften bahnbrechenden Werke *Statistical Methods for Research Workers* (Fisher, 1925) und *The Design of Experiments* (Fisher, 1935). Die Rolle von Fishers Ansätzen für die Akzeptanz der heutigen Sozialwissenschaften als Wissenschaften wird in Howie (2002) kurz diskutiert. In dieser Publikation werden die Entwicklungen des Wahrscheinlichkeitsbegriffs wissenschaftsgeschichtlich nachgezeichnet und die Unterschiede zwischen einem „subjektiven“ und „frequentistischen“ Wahrscheinlichkeitsbegriff werden anhand der historisch bedeutsamen Debatte zwischen dem Eugeniker R. A. Fischer und dem Geophysiker Harold Jeffreys erhellt. Howie kommt zu dem Schluss, dass die Unterschiede der beiden Betrachtungsweisen auf Wahrscheinlichkeiten und wissenschaftliche Inferenz vor den unterschiedlichen intellektuellen und wissenschaftlichen Hintergründen und Motivationen der Akteure zu verstehen sind. Während es Fisher hauptsächlich um Fragen der Genetik ging und die Annahme von hypothetischen, unendlich großen Populationen vor dem Hintergrund agrikulturellen Experimentierens Sinn macht, war für Jeffreys als Geophysiker z.B. die Annahme „unendlich vieler Erden“ auf deren Basis sich Stichprobenkennwerteverteilungen erzeugen lassen, schlichtweg absurd. Ferner führt Howie aus, dass Fisher ein geschickter Rhetoriker gewesen sein soll, was sich darin zeigen solle, dass

er seine eigenen Konzepte mit positiv konnotierten Adjektiven versah und konkurrierende Ansätze sprachlich brandmarkte. Man denke nur an die Antonyme subjektiv vs. objektiv, verzerrt vs. unverzerrt, effizient vs. ineffizient, etc. Howie argumentiert, dass die Ansätze von Jeffreys und Fisher in sich logisch kohärent sind, allerdings auf unterschiedlichen Annahmen, wissenschaftlichen Praxen und Zielen aufbauen. Während es Fischer eher um die Anwendbarkeit seiner Methoden in der agrikulturellen Praxis ging, um Fragen der Überlegenheit von z.B. einem bestimmten Dünger zu klären, ging es Jeffreys eher um Fragen der Beschaffenheit der Erde oder des Sonnensystems auf der Basis von unsicheren Messungen von z.B. seismologischen Messstationen. Von daher akzeptierte er vermutlich einen epistemischen Wahrscheinlichkeitsbegriff, der es erlaubt, dass Hypothesen eine Wahrscheinlichkeit angesichts beobachteter Daten besitzen, eine Annahme, die in einem frequentistischen Ansatz völlig ausgeschlossen ist. In jüngerer Zeit scheint sich der Bayesianische Ansatz einer gewissen Renaissance zu erfreuen, was nicht zuletzt an der Verfügbarkeit leistungsfähiger Prozessoren liegt, mit denen sich komputationelle Komplexität, die mit der Verwendung von Bayesianischen Ansätzen verbunden ist, lösen lässt. Das Lehrbuch von Gelman und Hill (Gelman & Hill, 2007) ist ein Beispiel für die Anwendung von Bayesianischen Schätzmethoden auf Multilevel-Modelle in der Politologie. Im Bereich des *machine learnings* sind Bayesianische Ansätze ebenfalls recht verbreitet (Koller & Friedman, 2009). Dies ist nicht überraschend, denn der epistemische Wahrscheinlichkeitsbegriff eignet sich auch aus einer theoretischen Perspektive dazu, das Wissenssystem artifizieller Agenten auf Basis beobachteter Daten und Lernen aus Erfahrung zu abzubilden. Auch der Kognitiven Psychologie, die eine gewissen Nähe zur künstlichen Intelligenz besitzt, scheint ein epistemischer Wahrscheinlichkeitsbegriff nicht fremd zu sein. Ein Anstoß zur Abfassung der vorliegenden Arbeit lag in der Beobachtung, dass die von Jaynes beschriebene Maximum-Entropie-Verteilung eine hohe formale Ähnlichkeit mit den in der Psychologie bekannten Rasch-Modellen besitzt. Eine nähere Exploration der Materie führte zu einer verstärkten Auseinandersetzung mit Bayes-Statistik und aus Gründen einer gewissen Theorie-Homogenität wurde die Verwendung der MCMC-Methode zur Bestimmung der Posterior-Verteilungen verwendet. Auch war

es für den Autor von Interesse, die Praktikabilität dieser Verfahren näher kennenzulernen. Praktische Probleme bei der Verwendung Bayesianischer Ansätze in der Psychologie sieht der Autor vor allem in folgenden Punkten:

1. Bayesianische Methoden, Konzepte, Notationen und Begriffe sind in der Psychologie so gut wie unbekannt. Es ist oft nicht klar, dass unterschiedliche Auffassungen über den Wahrscheinlichkeitsbegriff existieren und diese Unterschiede werden auch nicht im einzelnen erläutert, was zu Missverständnissen führen kann.

2. Die benötigte Mathematik zum Verständnis der Bayesianischen Ansätze ist relativ kompliziert, vor allem was die Verteilungstheorie angeht. MCMC-Ansätze gehen technisch betrachtet noch einen Schritt über klassische Maximum-Likelihood-Methoden durch die Verwendung von Prior-Verteilungen hinaus, was die Sache nicht vereinfacht.

3. Die Anwendung Bayesianischer Ansätze benötigt eine gewisse Expertise. Während für viele Standard-Analyse-Methode der Psychologie vorgefertigte Prozeduren in Softwarepaketen zur Verfügung stehen, müssen bei der Verwendung der MCMC-Methode die Likelihood, bzw. das verwendete Modell und die Prior-Verteilungen ggf. selbst spezifiziert werden, was eine gewissen programmiertechnische, statistische und modelltheoretische Erfahrung erfordert.

4. Die Verwendung von Prior-Verteilungen wird oft als kritisch betrachtet, da diese mit einem epistemischen Wahrscheinlichkeitsbegriff verbunden sind, welche einem Wissenschaftsverständnis von Wissenschaft als „objektiv“ widerspricht.

5. Die Diskussion um die Verwendung Bayesianischer vs. frequentistischer Ansätze kann aus historischen Gründen gegebenenfalls ideologisch und dogmatisch aufgeladen sein.

An dieser Stelle kann man sich durchaus fragen, warum diese Methoden überhaupt verwendet werden sollten. Nach Ansicht des Autors bietet die Verwendung der MCMC-

Methode in Kombination mit Bayesianischer Inferenz vermutlich potentiell folgende Vorteile:

1. Die Verwendung der MCMC-Methode befreit von der Abhängigkeit von tabellierten, asymptotischen Verteilungen von Test-Statistiken, ähnlich, wie dies auch beim parametrischen Bootstrapping der Fall ist, da Verteilungen von Parametern direkt über den Markov-Prozess generiert werden können, was Probleme mit klassischen Null-Hypothesen-Tests ausräumt. So besteht beispielsweise bei Multilevel-Analysen das Problem, dass die Freiheitsgrade der klassischen Signifikanz-Tests der festen Effekte einer gewissen Kontroverse unterliegen, bzw. nicht bekannt sind. Ein möglicher Lösungsansatz läge darin, Überschreitungswahrscheinlichkeiten auf Basis der MCMC-Methode zu generieren.

2. Das Problem fehlender Werte auf der abhängigen Variable, welches aus frequentistischer Perspektive unter anderem mit der Technik der multiplen Imputation angegangen wird, kann bei Verwendung der MCMC-Methode anders angegangen werden. Und zwar ließen sich die Posterior-Verteilungen von fehlenden Werten modellbasiert direkt über den MCMC-Prozess erzeugen. Eine mögliche Verzerrung von Test-Statistiken durch fehlende Werte muss vermutlich nicht berücksichtigt werden, da die Verteilungen der Modellparameter selbst auf Basis des MCMC-Prozesses induktiv erzeugt werden. Von daher ist kein Abgleich mit einer tabellierten Verteilung von Test-Statistiken, wie z.B. der F-Verteilung notwendig.

3. Kredibilitäts-Intervalle für standardisierte Effektstärke-Indices ließen sich theoretisch direkt aus der Anwendung des Markov-Prozess erzeugen, ohne auf nichtzentrale Verteilungen zurückgreifen zu müssen.

4. Modellgeltungs-Tests können über sogenannte posterior-prediktive Checks angegangen werden, bei der beobachtete Statistiken, wie z.B. Residuen, mit den Verteilungen dieser Statistiken bei Modellgeltung abgeglichen werden können.

5. Die Verwendung flexibler Programme wie WinBUGS (Lunn et al., 2000) ermöglicht

es, auch bisher unbekannte Modelle zu spezifizieren und die Posterior-Verteilungen der Parameter zu bestimmen.

Die angenommenen, potentiellen Vorteile der MCMC-Methode sind in der Psychologie noch wenig exploriert und es wäre interessant, deren Anwendbarkeit in der psychologisch-methodischen Forschung näher zu untersuchen. Ein Ziel hierbei kann es nicht sein, klassische Methoden zu ersetzen, allerdings ist eine deutlich sichtbare, konzeptuelle Trennung der frequentistischen und Bayesianischen Ansätze wünschenswert, um die Probleme, die sich aus der Mischung beider Methoden ergeben zu vermeiden und eine gewissen Theoriehomogenität zu gewährleisten.

Erste Ansatzpunkte für die weitere methodische Forschung in der Psychologie im Bezug auf Bayesianischen Methoden sieht der Autor in der Bestimmung von Posterior-Verteilungen von abgeleiteten Statistiken, wie z.B. Effektstärke-Indices, der modellbasierten Behandlung von fehlenden Werten, inklusive systematisch fehlender Werte und der Bewertung von Modell-Fehlpassung auf Basis von posterior-prediktiven Checks. Voraussetzung hierfür ist, dass der Bayesianische Ansatz didaktisch verständlich aufgearbeitet und klar von frequentistischen Methoden abgegrenzt wird. Oberflächlich betrachtet handelt es sich bei Bayesianischen Methoden lediglich um eine andere Methode der „Parameterschätzung“ die auch auf schon bestehende Modelle der Psychologie angewendet werden kann. Zu nennen wären hier insbesondere die multiple Regression (bzw. das Allgemeine Lineare Modell) und die Mehrebenenanalyse. Aber auch Item-Response-Modelle und (multilevel) Strukturgleichungsmodelle wären potentiell interessanten Kandidaten, anhand derer die Praktikabilität und die möglichen Vorteile und Nachteile der Methode näher herausgearbeitet und bewertet werden könnten.

6.5. Schlussbetrachtung und Ausblicke

Bezüglich des modelltheoretischen Teils der Arbeit ist zu verzeichnen, dass Rasch-Modelle als Maximum-Entropie-Modelle dargestellt werden können. Maximum-Entropie-Modelle wiederum lassen sich in den größeren Modellierungszusammenhang der Probabilistischen

Grafischen Modelle einbetten. Ein Vorteil der Einbettung von Rasch-Modellen in diese Zusammenhänge besteht darin, dass deutlich gemacht werden kann, dass diese Modelle der Item-Response-Theorie kein Exotikum darstellen, sondern mit statistischen Ansätzen kompatibel sind, die in anderen Wissenschaftszweigen angewendet werden. Ferner bietet die Darstellung der Rasch-Modelle aus diesen Modellierungszusammenhängen potentiell eine Lösung für das Problem der didaktischen Vermittlung der Modelle, da diese als Spezialfälle aus einem größeren Framework hergeleitet werden können, dessen Eigenschaften auf die Spezialfälle übertragbar sind. Zu nennen wäre beispielsweise die Herleitung der Likelihood aus der Differenzierung der Zustandssumme nach den Modellparametern und damit verbunden die Berechnung der Varianz und der Erwartungswerte der manifesten Variablen unter dem Modell. Ferner bietet die Maximum-Entropie-Methode die Möglichkeit, neue Modelle durch die Definition der Zustandssumme und die Wahl der Funktionen der Daten für - nicht nur - psychometrische Probleme zu generieren. Diese Techniken werden - wie bereits dargelegt - in mehreren Disziplinen bereits angewendet. Zu nennen wären z.B. die statistische Mechanik (Jaynes, 1957a, 1957b), die komputationelle Linguistik (Berger et al., 1996) und die Neurowissenschaften (Schneidman et al., 2006). Ein kurzes Beispiel für die Anwendung der Methode auf eine andere Fragestellung wäre z.B. das Affect-Grid (J. A. Russel, Weiss & Mendelsohn, 1989). Das Affect-Grid könnte sich dazu eignen, die Dimensionen Valenz und Arousal in einem längsschnittlichen Ambulatory Assessment zu erfassen. Im Unterschied zu den Skalen des MDBF wird nur ein einziges Item benötigt. Ferner liegt dem Affect-Grid die Annahme zugrunde, dass der Affektive Zustand einer Person innerhalb eines diskreten Koordinatensystems lokalisiert werden kann, dass durch die Dimensionen Valenz und Arousal gebildet wird. Die Response ist eine einfache Wahl in einem Koordinatensystem. Bisher liegt kein psychometrisches Messmodell für das Affect-Grid vor, dieses ließe sich allerdings einfach aus dem Maximum-Entropie-Formalismus erzeugen. Ein einfacher Modellkandidat wäre z.B.:

$$p(x_v, y_v) = \frac{\exp\{\lambda_{1v} \cdot x_v + \lambda_{2_v} \cdot y_v + \lambda_{3v} \cdot x_v \cdot y_v\}}{\sum_{x_v, y_v} \exp\{\lambda_{1v} \cdot x_v + \lambda_{2_v} \cdot y_v + \lambda_{3v} \cdot x_v \cdot y_v\}}.$$

Das bedeutet, dass die Wahrscheinlichkeit der Wahl des Kategorien-Paars (x_v, y_v) durch Person v auf die Merkmalsausprägung λ_{1v} der Person auf Dimension 1 und die Merk-

malsausprägung λ_{2v} der Person auf Dimension 2 zurückgeführt wird. Der Parameter λ_{3v} erfasst eventuell vorhandene Abhängigkeiten zwischen den Dimensionen. Psychometrisch bedeutet dies, dass die Wahl eines Kategorienpaares (x_v, y_v) auf die mittlere Merkmalsausprägung einer Person auf der Dimension Arousal und die mittlere Merkmalsausprägung einer Person auf der Dimension Valenz zurückgeführt wird. λ_3 erfasst die *intraindividuelle Koppelung* der Zustände innerhalb einer Person. Die Zustandssumme läuft über alle gemeinsamen Zustände der Variablen x_v und y_v, also über alle Felder des Affect-Grid. Das resultierende Modell ist kein bedingtes Modell, sondern es wird die gemeinsame Dichte der Variablen x_v und y_v modelliert, da die Reaktion der Person im Affect-Grid lediglich aus einem Kreuz besteht. Nach dem Maximum-Entropie-Formalismus sind die Mittelwerte der intraindividuellen Ratings auf den Skalen suffiziente Statistiken der Parameter λ_1 und λ_2. Die Summe von $x_v \cdot y_v$ ist eine suffiziente Statistik von λ_3, welches die intraindividuelle Koppelung der Prozesse erfasst. Ist λ_3 gleich Null, so sind Valenz und Arousal intraindividuell voneinander unabhängig.

Die Anwendung des Maximum-Entropie-Ansatzes ist innerhalb der Psychologie nicht auf die Psychometrie begrenzt. Ein weiteres interessantes Themengebiet, dass in Beziehung zum Maximum-Entropie-Formalismus steht, ist menschliches Entscheidungsverhalten. Es wäre interessant zu überprüfen, inwiefern die mittels der Maximum-Entropie-Methode generierten Modelle sich eignen, um menschliches Entscheidungsverhalten *under uncertainty* nicht-deterministisch, also probabilistisch zu modellieren. In der Literatur findet sich vereinzelt der Gedanke, dass die Natur das menschliche, kognitive System so gestaltet hat, dass z.B. das menschliche Entscheidungsverhalten mit dem Maximum-Entropie-Ansatz kompatibel ist (Jaynes, 1988). Schneidman et al. (2006) konnten anhand von Neuronenpopulationen in vitro zeigen, dass das Feuerungsverhalten von Neuronen mit einem Maximum-Entropie-Modell modellierbar ist. Die Hypothese, dass menschliches Entscheidungsverhalten (*human decision making*) mit dem Maximum-Entropie-Ansatz modelliert werden kann, ließe sich z.B. experimentell überprüfen, indem Studienteilnehmer gebeten werden den Ausgang von diskreten Zufallsexperimenten vorherzusagen. Diese Ergebnisse ließen sich mit den Vorhersagen von entsprechenden Maximum-Entropie-

Modellen vergleichen. Allerdings bedarf die Klärung dieser Fragestellungen zunächst einer Aufbereitung der bereits bestehenden Literatur zur Thematik *decision making* und einer Aufarbeitung der Theorie, was etwas von einer rein methodischen Forschung wegführen würde.

Während die Umsetzung der hier dargestellten Maximum-Entropie-Modelle als Computerprogramm bei einfachen Modellen relativ problemlos ist, ist die Darstellung in der Literatur und auch in der vorliegenden Arbeit häufig schwierig. Es wäre günstig, die Darstellung der Methode in Absprache mit Fachkollegen zu verbessern und notationell zu vereinfachen. Ein Ansatzpunkt besteht darin, Vektor-Matrix-Notation zu verwendet und eine Matrix einzuführen, die den möglichen Zustandsraum eines diskreten Systems abbildet. Das Einführen einer Zustands-Matrix $\boldsymbol{S}$ vereinfacht das Verständnis der Modelle, da die Zeilen der Matrix den möglichen Zuständen entsprechen und die Spalten den entsprechenden Ausprägungen der Funktionen $f_j(x_j)$, denen wiederum Parameter zugeordnet sind, um die Erwartungswerte dieser Funktionen unter dem Modell entsprechend der suffizienten, beobachteten Statistiken zu adjustieren. Aus dem globalen Ansatz würden stringent die bekannten Rasch-Modelle, als auch neue Modelle, z.B. zur Anwendung in Ambulatory Assessments, generierbar sein

Letztlich sei zum modelltheoretischen Teil der Arbeit angemerkt, dass sich Maximum-Entropie-Modelle vermutlich potentiell dazu eignen, um Link-Inversen für Multilevel Modelle zu erzeugen. Dies sei hier kurz an einem Beispiel für drei-kategorielle, diskrete Reaktionen ($k = 3$) skizziert. Wie in der Arbeit dargelegt, folgt das Rating-Scale-Modell aus dem Maximum-Entropie-Formalismus. Wird das Rating-Scale-Modell für lediglich eine Reaktion $y_i \in \{1, \dots 3\}$ dargestellt, so folgt aus der Anwendung der Methode:

$$p(y_i) = \frac{\exp\left(\lambda_i \cdot y_i + \tau_y\right)}{\sum_{l=1}^{k} \exp\left\{l \cdot \lambda_i + \tau_l\right\}}. \tag{6.1}$$

Bei Gleichung 6.1 handelt es sich um eine Link-Inverse für kategoriale Reaktionen. λ_i ist der lineare Prädiktor und die Parameter τ_y sind die Kategorien-Leichtigkeiten. Zur Identifikation muss eine Summennormierung über diese Parameter durchgeführt werden. Der springende Punkt ist, dass die Link-Inverse den linearen Prädiktor λ_i in Beziehung zu dem Erwartungswert der Reaktion y_i unter dem Modell setzt. Wird λ_i nach Maßgabe

eines gemischten Modells linear zerlegt, so resultiert daraus ein Multilevel-Modell für kategoriale Reaktionen, wobei impliziert wird, dass die Leichtigkeits-Parameter für jede Kategorie über alle möglichen Reaktionen identisch sind:

$$\boldsymbol{\lambda} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}. \tag{6.2}$$

Zur Notation vergleiche z.B. DeBoeck & Wilson, (2004, p. 22). $\boldsymbol{\lambda}$ ist der Vektor der linearen Prädiktoren für eine Reaktion y_i, $\mathbf{X}$ ist die Design-Matrix der festen Effekte, $\boldsymbol{\beta}$ ist der Zeilen-Vektor der festen Parameter, $\mathbf{Z}$ ist die Matrix der zufälligen Effekte und $\mathbf{b}$ ist der Zeilen-Vektor der Parameter der zufälligen Effekte. Die Parameter $\boldsymbol{\beta}$ folgen einer multivariaten Normalverteilung. Das resultierende Modell ähnelt dem Linear Logistischen Testmodell von Fischer (Fischer, 1995b) und ist ein gemischtes, multinomiales Logit-Modell. Zum Zusammenhang zwischen Multilevel-Modellen und IRT-Modellen siehe DeBoeck & Wilson (2004). Der Themenbereich um diese Angelegenheit wäre näher zu explorieren, zumal kategoriale Reaktionen in der Psychologie relativ häufig vorkommen und Multilevel-Modelle zur Modellierung derselben selten angewendet werden.

Ein weiteres Forschungsfeld, das mit den in dieser Arbeit angewendeten Methoden zusammenhängt, ist dasjenige der Bayes-Statistik und der Methoden der Bestimmung von Posterior-Verteilungen mit der MCMC-Methode, wie schon angesprochen wurde. Die MCMC-Methode bietet potentielle Vorteile gegenüber den klassisch frequentistischen Methoden. Zunächst ist die Methode nicht an asymptotische Voraussetzungen der Verteilung einer Test-Statistik gebunden. Vielmehr werden Posterior-Verteilungen der Parameter auf Basis der Modellgleichung (Likelihood) und von Prior-Verteilungen der Parameter erzeugt. Dies macht es möglich, parametrische Methoden auch auf kleine Stichproben anzuwenden. Ferner lassen sich Verteilungen von manifesten, abgeleiteten Größen, wie z.B. der Reliabilität unter dem Modell erzeugen. Die Passung eines Modells lässt sich anhand eines Vergleichs der empirischen Größe mit der unter dem Modell erwarteten Verteilung dieser Größe bewerten. Diese Methode der Bewertung der Modellpassung ist auch als Posterior-Prediktiver-Check (*posterior predictive check*) bekannt (vgl. z.B. Lynch, 2007) und ähnelt von den Grundzügen einem parametrischen Bootsrap (Efron & Tibshirani, 1993; Davier von, 1996), nur dass zusätzlich die Unsicherheit hinsichtlich der Lage der

Parameter mit berücksichtigt wird. Potentiell bietet diese Methode einen Ansatzpunkt, um die in der Arbeit angesprochenen Probleme zu vermeiden, die mit frequentistischen Modellgeltungstests und der Nullhypothese als „Wunschhypothese“ verbunden sind. Das Generieren von Verteilungen abgeleiteter Größen bietet zudem potentiell die Möglichkeit, Verteilungen von Effektstärke-Indices aus dem MCMC-Prozess zu erzeugen. Diese Themenkomplexe wären für die in der Psychologie verwendeten Modelle näher zu prüfen. Software zur Implementierung der MCMC-Methode steht zur Verfügung (Lunn et al., 2000) mit der Modelle sehr flexibel spezifiziert werden können. Allerdings kann dies auch als Nachteil betrachtet werden, da die Modellspezifikation eine gewisse Expertise erfordert und der Benutzer nicht durch eine Menüstruktur „geführt“ wird. Vielmehr muss die mathematische Form des Modells verstanden worden sein, um die Modellspezifikation vornehmen zu können. Die weiteren potentiellen Nachteile der Verwendung der Methode wurden bereits angesprochen.

Bezüglich des in der Arbeit generierten Testmodells ist zu verzeichnen, dass die hier vorgeschlagenen Methoden der Bewertung der Reliabilität und der Modellpassung näher simulativ untersucht werden sollten. Zwar folgend die Ableitungen der Verfahren einer gewissen mathematischen Kongruenz und sind in der Literatur verankert, allerdings ist es sicherer, diese Ableitungen zusätzlich einer simulativen Prüfung zu unterziehen.

Im Bezug auf das Konstrukt intraindividuelle Variabilität wäre es interessant, das Modell an weitere geeignete Datensätze anzupassen, um zu überprüfen, ob sich schon bestehende Ergebnisse hinsichtlich des Zusammenhangs der Merkmalsausprägung Variabilität und externen, konvergenten Konstrukten mit dem Modell replizieren lassen. Auf Basis der Literatur (Ebner-Priemer et al., 2009; J. Russel et al., 2007) ist zu vermuten, dass affektive, intraindividuelle Variabilität vor allem mit den Diagnose Borderline-Persönlichkeitsstörung (BPS) und bipolaren Störungen einhergeht.

Zusammenfassend ist zu sagen, dass das Modell immer dann verwendet werden kann, wenn die Variabilität auf multivariaten Zeitreihen psychometrisch mit einem Modell der Item-Response-Theorie skaliert werden soll und die absoluten Differenzen der Zeitreihen als manifeste Ausprägung der Variabilität interessieren. Die Verwendung des Modells

bietet gegenüber der Verwendung von manifesten Indices die Vorteile, dass es möglich ist Homogenitätshypothesen hinsichtlich der verwendeten Indikatoren zu prüfen sowie die globale Reliabilität und die individuelle Messgenauigkeit zu bewerten. Zudem wird die kategoriale Art der Reaktionen berücksichtigt und es muss kein Intervallskalenniveau der manifesten Reaktionen unterstellt werden.

Literatur

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Hrsg.), *Proceedings of the second international symposium on information theory* (S. 610-624). Budapest: Akademiai Kiado.

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

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42* (1), 69-81.

Andersen, E. B. (1995a). Polytomous Rasch models and their estimation. In G. Fischer & I. Molenaar (Hrsg.), *Rasch models: foundations, recent developments and applications.* New York: Springer.

Andersen, E. B. (1995b). What Georg Rasch would have thought about this book? In G. Fischer & I. Molenaar (Hrsg.), *Rasch models: foundations, recent developments and applications.* New York: Springer.

Andrich, D. (1978a). Application of a psychometric rating model to ordered response categories which are scored with successive integers. *Applied Psychological Measurement* (2), 581-594.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika*, *43*.

Andrich, D. (1978c). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement* (38), 665-680.

Andrich, D. (1988). *Rasch models for measurement.* Newbury Park: Sage.

Berger, A., Della Pietra, S. & Della Pietra, V. (1996). A maximum entropy approach to

natural language processing. *Computational Linguistics*, *20* (1).

Boker, S. (2001). Differential structural equation models of intraindividual variability. In L. M. Collins & A. G. Sayer (Hrsg.), *New methods for the analysis of change.* Washington, DC: American Psychological Association.

Borkenau, P. & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI) nach Costa und McCrae. Handanweisung.* Göttingen: Hogrefe.

Box, G. E. P. & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control.* San Francisco, CA: Holden-Day.

Brendan, M. B., Kimdy, L. & Lucas, R. E. (2006). On the nature of intraindividual personality variability: reliability, validity and associations with well-being. *Journal of Personality and Social Psychology*, *90* (3), 512-527.

Cattell, R. B., Cattell, A. K. S. & Rhymer, R. M. (1947). P-technique demonstrated in determining psychophysical source traits in a normal individual. *Psychometrika*, *12*, 267-288.

Chow, S. M., Ram, N., Boker, S. M., Fujita, F. & Clore, G. (2005). Capturing weekly fluctuations in emotion using a latent differential structural approach. *Emotion*, *5*, 208-225.

Costa, P. T. & McCrae, R. R. (1992). *Revised neo personality inventory (neo pi-r) and neo five factor inventory. professional manual.* Odessa, FL: Psychological Assessment Ressources.

Crayen, C., Eid, M., Lischetzke, T., Courvoisier, D. & Vermunt, J. (in Druck). Exploring dynamics in mood regulation - mixture latent markov modeling of ambulatory assessment data. *Psychosomatic Medicine*.

Davier von, M. (1996). *Methoden zur Prüfung probabilistischer Testmodelle.* Kiel: Institut für die Pädagogik der Naturwissenschaften an der Universität Kiel.

Davier von, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: results of a monte carlo study. *Methods of Psychological Research*, *2* (2).

Davier von, M. (2000). Winmira - a program system for the analyses with the Rasch model, with the lantent class analysis und with the mixed rasch model. [Software-

Handbuch]. Programma.

De Boeck, P. & Wilson, P. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer.

Ebner-Priemer, U. W., Eid, M., Kleindienst, N., Stabenow, S. & Trull, T. J. (2009). Analytic strategies for understanding affective (in)stability and other dynamic processes in psychopathology. *Journal of Abnormal Psychology*, *118* (1), 195-202.

Edwards, A. W. F. (1972). *Likelihood - an account of the statistical concept of likelihood and its application in scientific inference.* Cambridge: Cambridge University Press.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Eid, M. & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity and personality correlates. *Journal of Personality and Social Psychology*, *76*, 662-676.

Elliott, R. J., Aggoun, L. & Moore, J. B. (1995). *Hidden markov models.* New York, NY: Springer.

Fahrenberg, J. & Myrtek, M. (1996). *Ambulatory assessment: Computer-assisted psychological and psychophysiological methods in monitoring and field studies.* Göttingen: Hogrefe.

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

Fischer, G. H. (1995a). Derivations of the Rasch model. In G. Fischer & I. Molenaar (Hrsg.), *Rasch models: foundations, recent developments and applications.* New York: Springer.

Fischer, G. H. (1995b). The linear logistic test model. In G. Fischer & I. Molenaar (Hrsg.), *Rasch models: foundations, recent developments and applications.* New York: Springer.

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

Fischer, G. H. & Ponocny, I. (1995). Extended rating scale and partial credit models for assessing change. In G. Fischer & I. Molenaar (Hrsg.), *Rasch models: foundations,*

recent developments and applications. New York: Springer.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver and Boyd.

Fiske, D. & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin*, *52*, 217-250.

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*, 1011-1027.

Fox, J.-P. (2010). *Bayesian item response modeling.* New York: Springer.

Gelman, A. (1996). Inference and monitoring convergence. In W. Gilks, S. Richardson & D. Spiegelhalter (Hrsg.), *Markov chain monte carlo in practice.* New York: Chapman & Hall.

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierachical models.* Cambridge: Cambridge University Press.

Gelman, A. & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-472.

Gill, J. (2008). *Bayesian methods - a social and behavioral sciences approach.* Boca Raton: Chapman & Hall.

Glas, G. A. W. & Verhelst, N. D. (1995). Testing the Rasch model. In G. Fischer & I. Molenaar (Hrsg.), *Rasch models: foundations, recent developments and applications.* New York: Springer.

Gottman, J., Murray, J., Swanson, C., Tyson, R. & Swanson, K. (2002). *The mathematics of marriage: Dynamic nonlinear models.* Cambridge, MA: MIT Press.

Hoijtink, H. & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G. Fischer & I. Molenaar (Hrsg.), *Rasch models: foundations, recent developments and applications.* New York: Springer.

Howie, D. (2002). *Interpreting probability - controversies and developments in the early twentieth century.* Cambridge: Cambridge University Press.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical Review*, *106* (4), 620 -630.

Jaynes, E. T. (1957b). Information theory and statistical mechanics ii. *Physical Review*, *108* (2), 171 - 190.

Jaynes, E. T. (1988). How does the brain do plausible reasoning. In G. J. Erickson & C. R. Smith (Hrsg.), *Maximum-entropy and bayesian methods in science and engineering* (Bd. 1). New York: Springer.

Jaynes, E. T. (2003). *Probability theory: the logic of science.* Cambridge: Cambridge University Press.

Jenkins, G. M. & Watts, D. G. (1968). *Spectral analysis and its applications.* San Francisco, CA: Holden Day.

Jones, C. J. & Nesselroade, J. R. (1990). Multivariate, replicated, single-subject designs and P-technique factor analysis: A selective review of the literature. *Experimental Aging Research* (16), 171-183.

Kettunen, J. & Ravaja, N. (2000). A comparison of different time series techniques to analyse phasic coupling: A case study of cardiac and electrodermal activity. *Psychophysiology*, *37*, 395-408.

Keynes, J. M. (1921). *A treatise on probability.* London: McMillian and Co.

Koller, D. & Friedman, N. (2009). *Probabilistic graphical models - principles and techniques.* Cambridge, MA: MIT Press.

Koopman, B. (1936). On distributions permitting sufficient statistics. *Transactions of the American Mathematical Society*, *39*, 399-409.

Larsen, R. J. & Diener, E. (1992). Promises and problems with the circumplex model of emotions. *Review of Personality ad Social Psychology*, *13*, 25-59.

Luborsky, L. (1995). The first trial of the P-technique in psychotherapy research - a still-lively legacy. *Journal of Consulting and Clinical Psychology*, *63* (1), 6-14.

Lunn, D., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS - a bayesian modelling framework: concepts, structure and extensibility. *Statistics and Computing*, *10*, 325-331.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists.* Heidelberg: Springer.

Mair, P. & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20* (9), 1-20.

Martin-Loef, P. (1973). *Statistical models. Note from seminars 1969-70 by Rolf Sundberg.* Stockholm.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *49* (4), 529-544.

Metropolis, N., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1092.

Mintz, J. & Luborsky, L. (1970). P-technique factor analysis in psychotherapy research: an illustration of a method. *Psychotherapy: Theory, Research and Practice*, *7* (1), 13-18.

Molenaar, E. (1995). Estimation of item parameters. In G. Fischer & I. Molenaar (Hrsg.), *Rasch models: foundations, recent developments and applications.* New York: Springer.

Molenaar, P. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*, 181-202.

Molenaar, P. (1987). Dynamic factor analysis in the frequency domain: Causal modeling of multivariate psychophysiological time series. *Multivariate Behavioral Research*, *22*, 329-353.

Moskowitz, D. S. & Zuroff, D. C. (2004). Flux, pulse and spin: Dynamic additions to the personality lexicon. *Journal of Personality and Social Psychology*, *86*, 880-893.

Nesselroade, J. R. (1991). The warp and woof of the developmental fabric. In A. E. Rabin, R. A. Zucker, R. A. Emmons & S. Frank (Hrsg.), *Visons of development, the environment and aesthetics: The legacy of Joachim F. Wohlwill* (S. 213-240). Nillsdale, N. J.: Erlbaum.

Nesselroade, J. R., McArdle, J. J., Aggens, S. H. & Meyers, J. M. (2002). Dynamic factor

analysis models for representing process in multivariate time-series. In *Modeling intraindividual variability with repeated measures data: Advances and techniques* (S. 235-265). Mahawah, N. J.: Lawrence Erlbaum Associates, Inc.

Nesselroade, J. R. & Ram, N. (2004). Studying intraindividual variability: What we have learned that will help us understand lives in context. *Research in Human Development*, *1*, 9-29.

Nunally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Pitman, E. (1936). Sufficient statistics and intrinsic accuracy. *Proceedings of the Cambridge Philosophical Society*, *32*, 567-579.

R Development Core Team. (2010). R: A language and environment for statistical computing [Software-Handbuch]. Vienna, Austria. Verfügbar unter `http://www.r-project.org`

Ram, N., Chow, S. M., Bowles, R., Wang, L., Grimm, K., Fujita, F. et al. (2005). Examining interindividual differences in cyclicity of pleasant and unpleasant affect using spectral analysis and item response modeling. *Psychometrika*, *70* (4), 773-790.

Ram, N. & Gerstorf, D. (2009). Time-structured and net intraindividual variability: tools for examining the development of characteristics and processes. *Psychology and Aging*.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, *37*, 81-89.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology* (Bd. 4). Berkeley: University of California Press.

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

Rost, J. & Spada, H. (1983). Die Quantifizierung von Lerneffekten anhand von Testdaten. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *4*, 29-42.

Russel, J., Moskowitz, D., Zuroff, D., Sookman, D. & Paris, J. (2007). Stability and va-

riability of affective experience and interpersonal behavior in borderline personality disorder. *Journal of Abnormal Psychology*, *116* (3).

Russel, J. A., Weiss, A. & Mendelsohn, G. A. (1989). Affect grid: a single item scale for pleasure and arousal. *Journal of Personality and Social Psychology*, *57* (3), 493-502.

Russel, S. & Norvig, P. (2010). *Artificial intelligence - a modern approach*. Upper Saddle River, NJ: Pearson.

Russell, J. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*, 145-172.

Schneidman, E., Berry, M. J., Ronen, S. & Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, *440*, 1007-1012.

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics* (6), 461-464.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell systems technical journal*, *27*, 379-623.

Shifrin, K., Hooker, K. A., Wood, P. K. & Nesselroade, J. R. (1997). The structure and variation in mood in individuals with Parkinson's disease: A dynamic factor analysis. *Psychology and Aging*, *12*, 328-339.

Skondral, A. & Rabe-Hesketh, S. R. (2004). *Generalized latent variable modeling - multilevel, longitudinal and structural equation models*. Boca-Raton: Chapman & Hall.

Spiegelhalter, D. J., Carlin, B., Carlin, B. P. & Linde, A. van der. (2002). Bayesian measures of complexity and fit (with discussion). *Journal of the Royal Statistiscal Society B*.

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

Steyer, R., Schmitt, M. & Eid, M. (1999). Latent state-trait theory and research in

personality and individual differences. *European Journal of Personality*, *13* (5), 389-408.

Steyer, R., Schwenkmezger, P. & Eid, M. (1997). *Der Mehrdimensionale Befindlichkeitsfragebogen (MDBF)*. Göttingen: Hogrefe.

Sturz, S., Ligges, U. & Gelman, A. (2005). R2WinBUGS: a package for running WinBUGS from R. *Journal of Statistical Software*, *12* (3).

Wright, B. & Stone, M. (1969). A procedure for sample free item analysis. *Journal of Educational and Psychological Measurement*, *29*, 23-48.

Zhang, Z. (2007). Bayesian estimation of categorical dynamic factor models. *Multivariate Behavioral Research*, *42* (4), 729-756.

A. Appendix

Um die in dieser Arbeit verfolgte Argumentation nicht nur mathematisch, sondern auch praktisch nachvollziehen zu können, ist des zweckmäßig, einige der bei der Arbeit entstandenen Skripte zur Verfügung zu stellen. Hauptsächlich kam die Software R (R Development Core Team, 2010) und WinBUGS (Lunn et al., 2000) zum Einsatz. Der Anhang gliedert sich in fünf Teile. In Teil A.1. werden R-Funktionen zur Verfügung gestellt, um das Modell komputationell, d.h. mit dem Rechner nachzuvollziehen und simulativ untersuchen zu können. Teil A.2. behandelt die Schätzung der Parameter mit der MCMC-Methode und WinBUGS (Lunn et al., 2000). Teil A.3. liefert Code zur Überprüfung des Modells per Simulation, Teil A.4. beinhaltet ein Skript zur Überprüfung der Modellgeltung und in Teil A.5. ist die Ausgabe von WinBUGS für den Anwendungsteil dargestellt. Die Anhänge A.6. bis A.9. beinhalten einen kurzgefassten Lebenslauf, die Zusammenfassung und eine Erklärung über die verwendeten Hilfsmittel.

A.1. R-Funktionen zum Modell

In diesem Abschnitt finden sich einige Funktionen, die Hilfreich zur Verdeutlichung der Eigenschaften des in dieser Arbeit vorgestellten Modells sind. Die Funktionen bauen teilweise aufeinander auf.

A.1.1. Berechnung der Kategorien-Wahrscheinlichkeiten

Beschreibung

Die Funktion gibt die Kategorien-Wahrscheinlichkeiten bei gegebenem Personen-Parameter, Kategorien-Parametern, Scoring-Funktion und zuletzt gewähltem Wert unter dem Modell aus.

Funktion

```
expected.prob<-function(eta, beta, cat_funct, last)
{
 m=length(cat_funct)
 Z<-0;
 d<-exp(sqrt((cat_funct-last)^2)*eta+beta)
 for (i in 1:m)
 {
  Z[i]=exp(sqrt((cat_funct[i]-last)^2)*eta+beta[i])
 }
 result<-d/sum(Z)
}
```

Akzeptierte Argumente

eta: Personen-Parameter η

beta: Vektor der Kategorien-Parameter β_{ix} für ein Item

cat_funct: Vektor mit Werten der Scoring-Funktion

last: Zuletzt gewählter Wert $x_{vi[t-1]}$

Beispiel

```
> # Personen-Parameter
> eta=-1.5
> # Kategorien-Parameter
> beta=c(-0.5,0.5,0.5,-0.5)
> # Kategorien-Funktion
> cat_funct=c(1:4)
> # Zuletzt gewählter Wert
> last=2
>
> expected.prob(eta, beta, cat_funct, last)
[1] 0.06201971 0.75555476 0.16858706 0.01383847
```

```
>
```

A.1.2. Berechnung der Übergangsmatrix

Beschreibung

Die Funktion gibt die Übergangsmatrix unter dem Modell bei gegebenem Personen-Parameter, Kategorien-Parametern und gegebener Scoring-Funktion aus.

Funktion

```
transition_mat<-function(eta, beta, cat_funct)
{
 k=length(cat_funct)
 result<-expected.prob(eta, beta, cat_funct, 1)
 for (i in 2:k)
 {
  result_tmp<-expected.prob(eta, beta, cat_funct, i)
  result<-rbind(result, result_tmp)
 }
 dimnames(result)<-list(c(1:k), c(1:k))
 result
}
```

Akzeptierte Argumente

eta Personen-Parameter η

beta Vektor der Kategorien-Parameter β_{ix} für ein Item

cat_funct Vektor mit Werten der Scoring-Funktion

Beispiel

```
> # Personen-Parameter
> eta=-1.5
> # Kategorien-Parameter
> beta=c(-0.5,0.5,0.5,-0.5)
> # Kategorien-Funktion
> cat_funct=c(1:4)
> transition_mat(eta,beta,cat_funct)
           1 2 3 4
 1 0.570458811 0.3460008 0.0772032 0.006337225
 2 0.062019712 0.7555548 0.1685871 0.013838468
```

```
 3 0.013838468 0.1685871 0.7555548 0.062019712
 4 0.006337225 0.0772032 0.3460008 0.570458811
>
```

A.1.3. Darstellung der Kategorien-Funktionen

Beschreibung

Die Funktion stellt die Kategorien-Funktionen bei gegebenem, zuletzt gewählten Wert und gegebener Kategorien-Funktion grafisch dar. Der Skalen-Bereich für η beträgt -3 bis 3.

Darstellung der Funktion

```
plot_cat_funct<-function(beta, cat_funct, last)
{
 eta=-3
 j=1
 a<-expected.prob(eta, beta, cat_funct, last)

  for (i in seq(-2.9, 3, by = 0.1))
  {
   a<-rbind(a, expected.prob(i, beta ,cat_funct, last))
   j=j+1
   eta[j]<-i
  }
 matplot(eta, a, ylab="p(x_i[t])", ylim=c(0,1), xlab="eta")
}
```

Akzeptierte Argumente

beta: Vektor der Kategorien-Parameter β_{ix} für ein Item

cat_funct: Vektor mit Werten der Scoring-Funktion

last: Zuletzt gewählter Wert $x_{vi[t-1]}$

Beispiel

```
> # Kategorien-Parameter
> beta=c(-0.5,0.5,0.5,-0.5)
> # Kategorien-Funktion
> cat_funct=c(1:4)
```

```
> last=2
> plot_cat_funct(beta, cat_funct, 2)
```

A.1.4. Simulation einer Antwort aus dem Modell

Beschreibung

Die Funktion erlaubt die probabilistische Simulation einer Reaktionen für ein Item aus dem Modell bei gegebenem Personen-Parameter η, gegebenen Kategorien-Parametern β_{ix}, gegebener Scoring-Funktion und letzter Reaktion.

Darstellung der Funktion

```
sim_response<-function(eta, beta, cat_funct, last)
{
 k=length(cat_funct)
 prob<-expected.prob(eta, beta, cat_funct, last)

 #compute cumulative probabilities
 p<-c(0)
 cum_prob<-c(0)
 for (i in 1:k)
 {
   p=p+prob[i]
   cum_prob[i]=p
 }
 #generate value probabilistically
 result=sum(cum_prob<runif(1))+1
 result
}
```

Akzeptierte Argumente

eta Personen-Parameter η

beta: Vektor der Kategorien-Parameter β_{ix} für ein Item

cat_funct: Vektor mit Werten der Scoring-Funktion

last: Zuletzt gewählter Wert $x_{vi[t-1]}$

Beispiel

```
> # Personen-Parameter
> eta=-1.5
> # Kategorien-Parameter
> beta=c(-0.5,0.5,0.5,-0.5)
> # Kategorien-Funktion
> cat_funct=c(1:4)
> # Zuletzt gewählter Wert
> last=2
>
> sim_response(eta, beta, cat_funct, last)
[1] 2
>
```

A.1.5. Simulation von n Reaktionen aus dem Modell

Beschreibung

Die Funktion erlaubt die probabilistische Simulation von n Reaktionen für ein Item aus dem Modell bei gegebenem Personen-Parameter η, gegebenen Kategorien-Parametern β_{ix}, gegebener Scoring-Funktion und letzter Reaktion.

Darstellung der Funktion

```
sim_n_response<-function(eta, beta, cat_funct, last, n)
{
 result<-c(0)
 for (i in 1:n)
 {
  result[i]<-sim_response(eta, beta, cat_funct, last)
  last=result[i]
 }
 result
}
```

Akzeptierte Argumente

eta Personen-Parameter η

beta: Vektor der Kategorien-Parameter β_{ix} für ein Item

cat_funct: Vektor mit Werten der Scoring-Funktion

last: Zuletzt gewählter Wert $x_{vi[t-1]}$

n: Anzahl der zu simulierenden Reaktionen

Beispiel

```
> # Personen-Parameter
> eta=-1.5
> # Kategorien-Parameter
> beta=c(-0.5,0.5,0.5,-0.5)
> # Kategorienfunktion
> cat_funct=c(1:4)
> # Zuletzt gewählter Wert
> last=2
>
> (sim_n_response(eta, beta, cat_funct, last, 10))
 [1] 2 2 2 2 4 4 3 2 2 2
>
```

A.2. Parameterschätzung mit der MCMC-Methode

Während im vorherigen Abschnitt Funktionen im Vordergrund stehen, die es erlauben, die Eigenschaften des in dieser Arbeit generierten Modells nachzuvollziehen, geht es in diesem Abschnitt um die Schätzung der Modellparameter. Die Parameterschätzung kann mit der Maximum-Likelihood-Methode und der Markov-Chain-Monte-Carlo-Methode geschehen. In der Arbeit wird die Markov-Chain-Monte-Carlo-Methode verwendet. Zum Einsatz kommt die Software WinBUGS (Lunn et al., 2000) und das R-Paket R2WinBUGS (Sturz et al., 2005), welches es ermöglicht, WinBUGS aus der R-Umgebung anzusprechen.

A.2.1. Darstellung der Eingangsdaten

Im folgenden wird das Rohdaten-Format dargestellt, das benötigt wird, um die intraindividuelle Variabilität mittels R und WinBUGS zu untersuchen. Es handelt sich um eine einfache ASCII-Datei.

```
"y" "id" "item" "last"
"1" 3 1 1 NA
```

```
"2" 2 1 2 NA
"3" 2 1 3 NA
"4" 3 1 4 NA
"5" 2 1 5 NA
"6" 2 1 6 NA
"7" 2 1 7 NA
"8" 3 1 8 NA
"9" 3 1 1 3
"10" 2 1 2 2
"11" 2 1 3 2
"12" 3 1 4 3
"13" 2 1 5 2
"14" 4 1 6 2
"15" 1 1 7 2
"16" 4 1 8 3
"17" 2 1 1 3
"18" 2 1 2 2
"19" 2 1 3 2
"20" 2 1 4 3
"21" 2 1 5 2
"22" 3 1 6 4
"23" 3 1 7 1
"24" 3 1 8 4
...
"130825" 3 165 1 3
"130826" 3 165 2 3
"130827" 3 165 3 1
"130828" 4 165 4 3
"130829" 2 165 5 3
"130830" 3 165 6 2
"130831" 3 165 7 2
"130832" 3 165 8 2
"130833" 4 165 1 3
```

```
"130834" 2 165 2 3
"130835" 3 165 3 3
"130836" 3 165 4 4
"130837" 2 165 5 2
"130838" 3 165 6 3
"130839" 3 165 7 3
"130840" 3 165 8 3
```

Die Kopfzeile enthält die Variablen-Namen, wobei zu beachten ist, dass für den Laufindex der Zeilen kein Name definiert ist. Die erste Spalte der folgenden Zeilen bezeichnet den Laufindex der Beobachtungen. Insgesamt liegen 130840 Beobachtungen vor. In der zweiten Spalte steht die Reaktion eines Individuums, welche als `"y"` bezeichnet ist. Die dritte Spalte bezeichnet die Identifikationsnummer (`"id"`) der betreffenden Person und die vierte Spalte (`"item"`) beinhaltet den Item-Identifikator. Die letzte Spalte (`"last"`) beinhaltet die letzte registrierte Reaktion auf dem entsprechenden Item. Der Wert von `"last"` für die ersten Reaktionen eines Individuums ist als „missing“ deklariert, da keine entsprechenden Beobachtungen vorliegen. Im Prinzip ist es hinreichend, die Variablen, `"y"`, `"id"`, `"item"` und `last"` in einem Datenframe zu speichern. Diejenigen Items, die zu einer zu analysierenden Subskala gehören, können separat in einer Liste gespeichert und WinBUGS übergeben werden.

A.2.2. WinBUGS-Code zur Schätzung der Modellparameter

```
# estimate_mcmc.bugs
model state
{

# Definition der Likelihood

 for (i in 1:n)
 {
  Z1[i]<-exp((eta[id[i]])*sqrt(pow((1-last[i]),2))+tau[item[i],1])
  Z2[i]<-exp((eta[id[i]])*sqrt(pow((2-last[i]),2))+tau[item[i],2])
  Z3[i]<-exp((eta[id[i]])*sqrt(pow((3-last[i]),2))+tau[item[i],3])
  Z4[i]<-exp((eta[id[i]])*sqrt(pow((4-last[i]),2))+tau[item[i],4])

  Z_ges[i]<-Z1[i]+Z2[i]+Z3[i]+Z4[i]

  for (k in 1:4)
```

```
  {
   p[i,k]<-exp((eta[id[i]])*sqrt(pow((k-last[i]),2))+tau[item[i],k])/Z_ges[i]
  }

 y[i]~dcat(p[i,])
}

# Prior-Verteilungen und Summennormierung

for (i in 1:4)
{
  tau[1,i]~dnorm(0,0.01)
  tau[2,i]~dnorm(0,0.01)
  tau[3,i]~dnorm(0,0.01)
  tau[4,i]~dnorm(0,0.01)
  beta[1,i] <- tau[1,i] - mean(tau[1,])
  beta[2,i] <- tau[2,i] - mean(tau[2,])
  beta[3,i] <- tau[3,i] - mean(tau[3,])
  beta[4,i] <- tau[4,i] - mean(tau[4,])
}

# Prior-Verteilung der Personen-Parameter

for (i in 1:N)
{
 eta[i]~dnorm(mu.eta, tau.eta)
}

# Hyper-Parameter
mu.eta~dnorm(0,0.01)
tau.eta<-pow(sigma.eta, -2)
sigma.eta~dunif(0,10)

}
```

Das Listing zeigt den Code zur Spezifikation des Modells in WinBUGS. Im ersten Abschnitt wird die Likelihood des Modells definiert. Der Code ist für 4-kategorielle Items ausgelegt ist. Bei Bedarf muss der Code also entsprechend geändert werden. Der Index i läuft in diesem Abschnitt von 1 bis n, wobei n die Anzahl der beobachteten Ratings darstellt. Der Index k läuft über die Anzahl der Kategorien. Im zweiten Abschnitt werden die Prior-Verteilungen der Kategorien-Parameter definiert. Diese sind Normalverteilungen mit $\mu = 0$ und $1/\sigma^2 = 0.01$. Durch die breite Streuung der Prior-Verteilungen wird zum Ausdruck gebracht, das so gut wie keine Information hinsichtlich der Lage der Parameter zur Verfügung steht. In Abschnitt drei werden die Kategorien-Parameter itemweise summennormiert. In Abschnitt vier wird die Prior-Verteilung der Personen-Parameter, bzw. der latenten Trait-Verteilung als eine Normalverteilung definiert. Die Kennwerte

dieser Verteilung (μ_η) und (σ_η) werden aus den Daten selbst geschätzt. Diese Schätzung wird in Abschnitt 5 durch die Definition von sog. Hyper-Parametern weiter spezifiziert. Dem Mittelwert der latenten Trait-Verteilung wird eine Prior-Verteilung von $\mu_\eta = 0$ und eine Präzison von $1/\sigma_\eta^2 = 0.01$ zugewiesen. Dies bedeutet, dass angenommen wird, dass der Mittelwert der latenten Trait-Verteilung Null ist, was inhaltlich mit der Annahme korrespondiert, dass *keine* Abhängigkeit des Wertes $x_{vi[t]}$ vom Wert $x_{vi[t-1]}$ besteht. Diese Prior-Annahme soll anhand der Beobachtungen und der Likelihood „aktualisiert" (*updated*) werden. Der Streuung der latenten Trait-Verteilung wird eine Gleichverteilung (auch als uniforme Verteilung bekannt) im Wertebereich von 0 bis 10 zugewiesen. Dies bedeutet inhaltlich, dass angenommen wird, dass die Streuung der latenten Trait-Verteilung sich im Bereich von 0 bis 10 bewegt, wobei keiner der Streuungen in diesem Intervall ein Vorzug gegeben wird. Es sei angemerkt, dass das Modell auch anders spezifiziert werden kann, z.B. über kumulative Logits.

A.2.3. R-Skript zum Ansteuern von WinBUGS

```
# extract.R
library(arm)

# Einlesen der Daten
data<-read.table("input.dat", head=T)

# Spezifikation der Anzahl der Personen
# (Kann auch dynamisch geschehen)

N=165

# Zuweisung der Daten des Daten-Frames zu den Variablen
id<-data$id
item<-data$item
last<-data$last
y<-data$y

# Auszählung der Anzahl der Beobachtungen
n=length(id)

# Erstellen einer Liste der Variablen
# für die Übergabe an WinBUGS
dat<-list("y", "id","item", "last", "n", "N")

# Spezifikation der zu bestimmenden Parameter
parameters<-c("beta", "tau", "eta", "mu.eta", "sigma.eta")
```

```
# Übergabe an WinBUGS zur Bestimmung der Posterior-Verteilungen
simlist<-bugs(dat, inits=NULL, parameters, "estimate_mcmc_bugs.bug", n.chains=2,
\ n.burnin=500, n.iter=1000, debug=TRUE, bugs.directory="c:/WinBUGS14/")

# Grafische Darstellung der Ergebnisse
plot(simlist)

# Ausgabe der Ergebnisse auf der Konsole
print(simlist)

# Speicherung Markov-Ketten in eine Datei zur weiteren Analyse
attach.bugs(simlist)
dput(simlist, file="simlist.dat")
```

Die Funktion `extract.R` dient dem Einlesen der Daten aus einer tabulator-getrennten `ASCII`-Datei (`input.dat`). Die Daten müssen im `long`-Format vorliegen. `y` bezeichnet die manifeste Reaktion, `id` ist der Personen-Identifikator, `item` ist der Item-Identifikator und `last` ist der zuletzt gewählte Wert. Der Personen- und der Item-Identifikator laufen von jeweils `1` bis `N`, bzw. `k`. Die Daten werden in eine Liste `dat`überführt, die zusammen mit der Definition der zu schätzenden Parameter (`parameters`) der Funktion `bugs()` übergeben wird. Über die Argumente `n.burnin` und `n.iter` wird die Länge der Burn-In-Phase und die Anzahl der Iterationen zur Bestimmung der Posterior-Verteilungen definiert. Die Funktion gibt das Objekt `simlist` zurück, welches die Ergebnisse der Analyse mit WinBUGS enthält. Diese Ergebnisse werden grafisch und auf der Konsole dargestellt und schließlich wird das Objekt zur weiteren Verwendung (z.B. für die Testung der Modellgeltung) gespeichert.

A.3. Simulation

Die beiden Listings in dieser Sektion können genutzt oder modifiziert werden, um die Modelleigenschaften simulativ zu untersuchen.

```
model state
{

 for (i in 1:n)
 {
   Z1[i]<-exp(eta*sqrt(pow((1-last[i]),2))+tau[item[i],1])
   Z2[i]<-exp(eta*sqrt(pow((2-last[i]),2))+tau[item[i],2])
   Z3[i]<-exp(eta*sqrt(pow((3-last[i]),2))+tau[item[i],3])
   Z4[i]<-exp(eta*sqrt(pow((4-last[i]),2))+tau[item[i],4])
```

```
  Z_ges[i]<-Z1[i]+Z2[i]+Z3[i]+Z4[i]

  for (k in 1:4)
  {
    p[i,k]<-exp(eta*sqrt(pow((k-last[i]),2))+tau[item[i],k])/Z_ges[i]
  }

  y[i]~dcat(p[i,])
 }

# Itemweise Summennormierung und Prior-Verteilung

 for (i in 1:4)
 {
  tau[1,i]~dnorm(0,0.01)
  tau[2,i]~dnorm(0,0.01)
  tau[3,i]~dnorm(0,0.01)
  tau[4,i]~dnorm(0,0.01)
  beta[1,i] <- tau[1,i] - mean(tau[1,])
  beta[2,i] <- tau[2,i] - mean(tau[2,])
  beta[3,i] <- tau[3,i] - mean(tau[3,])
  beta[4,i] <- tau[4,i] - mean(tau[4,])
 }

 eta~dnorm(0,0.01)

}
```

Listing A.3 repräsentiert WinBUGS-Code zum Zwecke der Simulation. Hier wird lediglich ein Parameter aus einer simulierten Zeitreihe geschätzt.

```
# Funktion zur simulation von manifesten Zeitreihen und zur Schätzung der
# Modell-Parameter mit der MCMC-Methode
#
# eta ist ein Funktions-Argument

estim_param_sim<-function(eta)
{

# Parameter
tau_1<-c(-0.25, 0.25, 0.25,-0.25)
tau_2<-c(-0.25, -0.10, 0.10, 0.25)
tau_3<-c(0.25, -0.25, -0.25, 0.25)
tau_4<-c(0, 0, 0, 0)

# Kategorien-Funktion
cat_funct<-c(1:4)

# Länge der Zeitreihe
n=500

# Letzter Wert
last=2
```

```
# Simulation einer multivariaten Zeitreihe
t1<-sim_n_response(eta, tau_1, cat_funct, last, n+1)
t2<-sim_n_response(eta, tau_2, cat_funct, last, n+1)
t3<-sim_n_response(eta, tau_3, cat_funct, last, n+1)
t4<-sim_n_response(eta, tau_4, cat_funct, last, n+1)

# Vektor der zuletzt gewählten Werte
l1<-t1[1:n]
l2<-t2[1:n]
l3<-t3[1:n]
l4<-t4[1:n]

# Vektor der AV
y1<-t1[2:(n+1)]
y2<-t2[2:(n+1)]
y3<-t3[2:(n+1)]
y4<-t4[2:(n+1)]

# Generierung von Item-Indikatoren
i1<-rep(1,n)
i2<-rep(2,n)
i3<-rep(3,n)
i4<-rep(4,n)

# Zusammenfügung der Vektoren
y<-c(y1, y2, y3, y4)
last<-c(l1, l2, l3, l4)
item<-c(i1, i2, i3, i4)

return<-list(y=y, last=last, item=item)

n=length(y)

# Aufruf an WinBUGS
library(arm)

dat <- list ("y", "item", "last", "n")
parameters<-c("beta", "eta")


simlist<-bugs(dat, inits=NULL, parameters, "simulation.bugs", \
n.chains=2, n.burnin=500, n.iter=1000, debug=TRUE)

result=simlist
}

# Initialisierung der Matrix der Posterior-Mittelwerte
mean_matrix<-matrix(0, 100, 21)

# Initialisierung der Matrix für Posterior-SD
sd_matrix<-matrix(0, 100, 21)


simlist<-estim_param_sim(-2.1)
i=1
for  (k in seq(-2.5, 2.5, by = 0.25))
```

```
      {
            for (j in 1:100)
            {
                  simlist<-estim_param_sim(k)
                  attach.bugs(simlist)
                  mean_matrix[j,i]=mean(eta)
                  sd_matrix[j,i]=sd(eta)
                  rm(simlist)
            }
      i=i+1
      }
dput(mean_matrix, file="mean_matrix.dat")
dput(sd_matrix, file="sd_matrix.dat")
```

Der WinBUGS-Code wird von R aus mit dem Paket `R2WinBUGS` angesprochen. Im obigen Code werden Zeitreihen mit unterschiedlichen Parametern aus dem Modell simuliert und anschließend werden die Posterior-Verteilungen der Parameter geschätzt. Die Posterior-Verteilungen werden zur weiteren Analyse und zur grafischen Darstellung gespeichert.

A.4. Modellgeltung

```
library(arm)

# Einlesen der Daten
data<-read.table("input.dat", head=T)

# Einlesen der Prior-Verteilungen
params<-dget("simlist.dat")
attach.bugs(params)

n=dim(data)[1]
n_rep=dim(eta)[1]
cat_funct=c(1:4)

# Initialisierung von benötigten Variablen
resp<-c(0)
expected<-c(0)
info<-c(0)
std_res<-c(0)
std_res<-c(0)
mean_res_g<-c(0)
mean_res_g_rep<-c(0)
mean_res_it<-c(0)
mean_res_it_rep<-c(0)
diff_g<-c(0)
diff_it<-c(0)
```

```
# Berechnung der erwarteten Reaktion unter dem Modell
# und Berechnung der Streuung für jede manifeste
# Reaktion

for (i in 1:n)
{
            eta_rep=mean(eta[,data$id[i]]);
            beta1_rep=mean(beta[,,1] [,data$item[i]]);
            beta2_rep=mean(beta[,,2] [,data$item[i]]);
            beta3_rep=mean(beta[,,3] [,data$item[i]]);
            beta4_rep=mean(beta[,,4] [,data$item[i]]);
            beta_rep<-c(beta1_rep, beta2_rep, beta3_rep, beta4_rep)
            prob=expected.prob(eta_rep, beta_rep, cat_funct, data$last[i])

            # Das ist der Erwertungswert von $x_vi[t]|x_vi[t-1]|$
            # unter dem Modell
            expected[i]<-sum(prob*cat_funct)

            # Anmerkung: Das ist die erwartete Varianz von $x_vi[t]|x_vi[t-1]|$
            # unter dem Modell
            info[i]<- sum((abs(cat_funct)^2*prob))-sum((abs(cat_funct)*prob))^2
}

# Berechnung der standardisierten Residuen
std.res<-(data$y-expected)^2/info

# Aggregierung der Residuen über Item und Person
sum.res.it<-tapply(std.res, data$item, sum)
sum.res.id<-tapply(std.res, data$id, sum)

# Berechnung des Outfits
outfit.it<-tapply(std.res, data$item, mean)
outfit.id<-tapply(std.res, data$id, mean)

# Berechnung des Infits
infit.it.enum<-tapply((data$y-expected)^2, data$item, mean)
infit.it.denom<-tapply(info, data$item, mean)
infit.it<-infit.it.enum/infit.it.denom

infit.id.enum<-tapply((data$y-expected)^2, data$id, mean)
infit.id.denom<-tapply(info, data$id, mean)
infit.id<-infit.id.enum/infit.id.denom


# Berechnung der Chi-Quadrat-Statistiken
chisq.it<-as.vector(sum.res.it)
chisq.id<-as.vector(sum.res.id)

# Bestimmung der Freiheitsgrade
df.it=as.vector(table(data$item))
df.id=as.vector(table(data$id))

# Berechnung der Überschreitungswahrscheinlichkeiten
p.it<-round(1-pchisq(chisq.it, df.it),2)
p.id<-round(1-pchisq(chisq.id, df.id),2)
```

```
# Berechnung von Outfit und Infit
# über die Gesamt-Daten
outfit.ges<-mean(std.res)
infit.ges<-mean((data$y-expected)^2/info)
chisq.ges<-sum(std.res)
df.ges<-length(data$y)
p.ges<-round(1-pchisq(chisq.ges, df.ges),2)

# Überführung der Fit-Statistiken in jeweils ein Objekt
fit.stat.id<-data.frame(outfit.id, infit.id, chisq.id, df.id, p.id)
fit.stat.it<-data.frame(outfit.it, infit.it, chisq.it, df.it, p.it)
fit.stat.ges<-data.frame(outfit.ges, infit.ges, chisq.ges, df.ges, p.ges)

# Berechnung der Reliabilität
var_error<-mean(apply(eta,2,var))
var_eta_hat<-var(apply(eta,2,mean))

# Andrich Reliabilität
Rel_1=(var_eta_hat-var_error)/(var_eta_hat)

# Reliabilität auf Basis der latenten Score-Verteilung
# Experimentell, nicht empfohlen
Rel_2=var_eta_hat/mean(sigma.eta)^2

# Plots
eta_mean<-(apply(eta,2,mean))
eta_sd<-(apply(eta,2,sd))

# Plot der Personen-Parameter mit Konfidenzintervall
plotCI(x=eta_mean, uiw=2*eta_sd, lwd=1, sfrac=0.005, xlab="Person", ylab="eta", main="Ruhe-Unruhe")

#Histogramm der standardisierten Residuen
hist((data$y-expected)/sqrt(info), main="Ruhe-Unruhe", xlab="Standardisiertes Residuum", 100)
```

Die Modelltestung basiert auf auf standardisierten Residuen. Zunächst werden die unstandardisierten Residuen berechnet und mittels der erwarteten Streuung der manifesten Variable unter dem Modell standardisiert. Somit liegt für jede einzelne Reaktion einer Person ein quadrierter z-Wert vor. Die Summen von quadrierten, unabhängigen z-Werten sind χ^2-verteilt. Daher bietet es sich an, χ^2-Test-Statistiken zu verwenden, um die Passung des Modells bezüglich der Personen und der Items zu evalieren, indem die quadrierten z-Werte über die jeweiligen relevanten Partitionen der Daten aggregiert werden. Diese resultierenden Statistiken können in die bekannten Outfit- und Infit-Indices der Item-Response-Theorie überführt werden. Ferner wird die Reliabilität nach der Methode von Andrich (Andrich, 1988) berechnet.

A.5. Ausgabe der MCMC-Schätzung über die Skalen des MDBF

Hier werden die Ergebnisse der Parameterschätzung über die Skalen des MDBF dargestellt.

A.5.1. Subskala „gehobene Stimmung“

```
Inference for Bugs model at "tau_each_item.bug", fit using WinBUGS,
 2 chains, each with 1000 iterations (first 500 discarded)
 n.sims = 1000 iterations saved
```

	mean	sd	2.5%	25%	50%	75%	97.5%	Rhat	n.eff
beta[1,1]	-1.51	0.05	-1.62	-1.54	-1.51	-1.47	-1.40	1.00	1000
beta[1,2]	0.04	0.03	-0.01	0.02	0.04	0.05	0.08	1.00	1000
beta[1,3]	1.18	0.02	1.14	1.17	1.18	1.20	1.23	1.00	1000
beta[1,4]	0.29	0.02	0.24	0.27	0.29	0.30	0.34	1.00	1000
beta[2,1]	-1.68	0.06	-1.80	-1.72	-1.68	-1.64	-1.57	1.01	260
beta[2,2]	0.06	0.03	0.01	0.04	0.06	0.08	0.11	1.00	1000
beta[2,3]	1.26	0.02	1.22	1.25	1.26	1.28	1.31	1.00	520
beta[2,4]	0.36	0.03	0.31	0.34	0.36	0.38	0.41	1.01	220
beta[3,1]	-1.27	0.04	-1.36	-1.30	-1.27	-1.24	-1.18	1.00	390
beta[3,2]	0.08	0.02	0.04	0.07	0.08	0.10	0.13	1.01	300
beta[3,3]	1.02	0.02	0.98	1.01	1.02	1.03	1.06	1.00	1000
beta[3,4]	0.17	0.02	0.12	0.15	0.17	0.18	0.21	1.00	560
beta[4,1]	-1.50	0.05	-1.61	-1.53	-1.50	-1.46	-1.39	1.00	1000
beta[4,2]	0.02	0.03	-0.03	0.00	0.02	0.04	0.07	1.00	1000
beta[4,3]	1.22	0.02	1.18	1.21	1.22	1.24	1.27	1.00	1000
beta[4,4]	0.25	0.03	0.20	0.24	0.25	0.27	0.30	1.00	620
eta[1]	-0.84	0.09	-1.03	-0.90	-0.83	-0.77	-0.64	1.01	300
eta[2]	-1.64	0.13	-1.90	-1.72	-1.64	-1.55	-1.39	1.00	1000
eta[3]	-1.76	0.11	-1.97	-1.83	-1.76	-1.68	-1.54	1.00	1000
eta[4]	-0.81	0.09	-0.98	-0.87	-0.81	-0.75	-0.64	1.00	1000
eta[5]	-0.89	0.10	-1.08	-0.96	-0.88	-0.82	-0.69	1.00	1000
eta[6]	-0.95	0.10	-1.13	-1.01	-0.95	-0.88	-0.76	1.00	1000
eta[7]	-1.26	0.10	-1.45	-1.32	-1.27	-1.20	-1.07	1.00	820
eta[8]	-2.05	0.16	-2.39	-2.15	-2.04	-1.94	-1.74	1.01	360
eta[9]	-0.70	0.09	-0.89	-0.77	-0.70	-0.64	-0.52	1.00	1000
eta[10]	-1.96	0.14	-2.24	-2.06	-1.96	-1.86	-1.69	1.00	1000
eta[11]	-1.45	0.12	-1.69	-1.53	-1.44	-1.37	-1.24	1.00	770
eta[12]	-1.60	0.09	-1.79	-1.65	-1.59	-1.53	-1.42	1.00	780
eta[13]	-0.86	0.10	-1.05	-0.93	-0.86	-0.80	-0.68	1.01	1000
eta[14]	-0.81	0.09	-1.00	-0.87	-0.81	-0.75	-0.64	1.00	1000
eta[15]	-0.89	0.10	-1.11	-0.96	-0.89	-0.83	-0.69	1.00	1000
eta[16]	-0.42	0.08	-0.56	-0.47	-0.42	-0.36	-0.28	1.00	1000
eta[17]	-1.12	0.10	-1.30	-1.19	-1.12	-1.06	-0.94	1.00	1000
eta[18]	-0.89	0.09	-1.07	-0.95	-0.89	-0.83	-0.70	1.00	410
eta[19]	-0.99	0.11	-1.20	-1.07	-0.99	-0.91	-0.78	1.00	1000
eta[20]	-0.62	0.09	-0.79	-0.68	-0.62	-0.56	-0.44	1.00	1000
eta[21]	-0.71	0.09	-0.88	-0.77	-0.71	-0.65	-0.51	1.00	1000
eta[22]	-1.44	0.12	-1.69	-1.53	-1.44	-1.36	-1.20	1.00	1000
eta[23]	-0.29	0.08	-0.45	-0.35	-0.29	-0.23	-0.12	1.00	1000
eta[24]	-0.97	0.11	-1.18	-1.04	-0.97	-0.91	-0.77	1.00	490

eta[25]	-0.65	0.09	-0.82	-0.71	-0.65	-0.59	-0.48	1.00	420
eta[26]	-1.33	0.11	-1.55	-1.41	-1.33	-1.25	-1.12	1.00	1000
eta[27]	-1.14	0.10	-1.32	-1.20	-1.14	-1.07	-0.96	1.00	1000
eta[28]	-0.97	0.10	-1.17	-1.03	-0.97	-0.90	-0.78	1.00	1000
eta[29]	-0.93	0.10	-1.12	-1.00	-0.93	-0.86	-0.74	1.00	1000
eta[30]	-1.21	0.11	-1.44	-1.28	-1.21	-1.15	-1.02	1.00	1000
eta[31]	-0.50	0.09	-0.67	-0.56	-0.50	-0.44	-0.34	1.00	780
eta[32]	-1.57	0.12	-1.84	-1.65	-1.57	-1.48	-1.34	1.00	1000
eta[33]	-1.53	0.12	-1.76	-1.61	-1.52	-1.45	-1.30	1.00	950
eta[34]	-1.33	0.09	-1.51	-1.39	-1.33	-1.27	-1.15	1.00	960
eta[35]	-0.96	0.10	-1.16	-1.02	-0.96	-0.89	-0.76	1.00	1000
eta[36]	-1.67	0.13	-1.92	-1.76	-1.66	-1.58	-1.41	1.00	1000
eta[37]	-0.84	0.10	-1.04	-0.91	-0.84	-0.77	-0.64	1.00	1000
eta[38]	-0.47	0.10	-0.65	-0.53	-0.46	-0.40	-0.28	1.00	540
eta[39]	-2.63	0.16	-2.96	-2.75	-2.63	-2.52	-2.35	1.00	1000
eta[40]	-0.53	0.10	-0.71	-0.60	-0.53	-0.47	-0.35	1.00	560
eta[41]	-1.39	0.11	-1.62	-1.47	-1.39	-1.31	-1.19	1.00	1000
eta[42]	-0.65	0.09	-0.83	-0.71	-0.65	-0.58	-0.47	1.00	1000
eta[43]	-1.43	0.11	-1.65	-1.50	-1.43	-1.36	-1.21	1.00	480
eta[44]	-0.70	0.09	-0.87	-0.76	-0.70	-0.64	-0.54	1.00	1000
eta[45]	-0.95	0.10	-1.15	-1.02	-0.95	-0.89	-0.75	1.00	830
eta[46]	-1.25	0.09	-1.44	-1.32	-1.25	-1.19	-1.07	1.00	620
eta[47]	-0.48	0.09	-0.66	-0.54	-0.48	-0.42	-0.32	1.00	900
eta[48]	-1.29	0.12	-1.51	-1.37	-1.30	-1.22	-1.06	1.00	1000
eta[49]	-2.10	0.13	-2.36	-2.18	-2.10	-2.02	-1.84	1.00	1000
eta[50]	-0.70	0.15	-0.99	-0.81	-0.70	-0.60	-0.41	1.00	910
eta[51]	-1.40	0.10	-1.61	-1.48	-1.40	-1.34	-1.20	1.00	1000
eta[52]	-1.75	0.13	-2.02	-1.84	-1.75	-1.66	-1.50	1.00	1000
eta[53]	-2.32	0.12	-2.56	-2.41	-2.32	-2.24	-2.10	1.00	610
eta[54]	-0.35	0.09	-0.54	-0.41	-0.35	-0.29	-0.18	1.00	1000
eta[55]	-1.37	0.12	-1.60	-1.45	-1.37	-1.29	-1.15	1.00	1000
eta[56]	-1.08	0.12	-1.33	-1.16	-1.08	-1.00	-0.85	1.00	1000
eta[57]	-1.01	0.11	-1.22	-1.08	-1.01	-0.94	-0.80	1.00	520
eta[58]	-0.88	0.10	-1.08	-0.95	-0.88	-0.82	-0.69	1.00	1000
eta[59]	-1.41	0.10	-1.60	-1.48	-1.41	-1.34	-1.22	1.01	290
eta[60]	-0.99	0.10	-1.17	-1.05	-0.99	-0.92	-0.79	1.00	1000
eta[61]	-2.80	0.21	-3.22	-2.94	-2.79	-2.65	-2.41	1.00	1000
eta[62]	-0.54	0.08	-0.70	-0.59	-0.54	-0.48	-0.37	1.00	1000
eta[63]	-0.79	0.09	-0.95	-0.85	-0.79	-0.72	-0.60	1.00	1000
eta[64]	-1.35	0.11	-1.55	-1.43	-1.35	-1.28	-1.13	1.00	460
eta[65]	-1.04	0.11	-1.25	-1.11	-1.04	-0.96	-0.84	1.00	1000
eta[66]	-0.96	0.10	-1.17	-1.02	-0.96	-0.89	-0.77	1.00	600
eta[67]	-0.94	0.10	-1.14	-1.01	-0.94	-0.87	-0.74	1.00	1000
eta[68]	-1.56	0.09	-1.74	-1.62	-1.56	-1.49	-1.38	1.00	1000
eta[69]	-1.27	0.11	-1.49	-1.34	-1.27	-1.20	-1.05	1.00	1000
eta[70]	-1.27	0.09	-1.44	-1.33	-1.27	-1.21	-1.10	1.00	1000
eta[71]	-1.28	0.11	-1.49	-1.35	-1.27	-1.21	-1.07	1.00	1000
eta[72]	-0.32	0.09	-0.50	-0.38	-0.32	-0.26	-0.16	1.00	580
eta[73]	-1.15	0.10	-1.36	-1.22	-1.15	-1.08	-0.95	1.00	630
eta[74]	-0.50	0.09	-0.67	-0.56	-0.51	-0.44	-0.34	1.01	290
eta[75]	-0.83	0.09	-1.01	-0.89	-0.83	-0.76	-0.66	1.00	320
eta[76]	-1.02	0.10	-1.24	-1.09	-1.02	-0.96	-0.82	1.00	1000
eta[77]	-0.96	0.10	-1.15	-1.03	-0.96	-0.89	-0.78	1.00	1000
eta[78]	-0.57	0.09	-0.75	-0.63	-0.56	-0.50	-0.40	1.00	1000
eta[79]	-0.37	0.08	-0.53	-0.42	-0.37	-0.31	-0.20	1.01	240
eta[80]	-0.86	0.09	-1.05	-0.92	-0.85	-0.79	-0.68	1.00	1000

eta[81]	-0.87	0.10	-1.07	-0.94	-0.87	-0.81	-0.69	1.00	1000
eta[82]	-0.52	0.09	-0.71	-0.58	-0.52	-0.47	-0.34	1.01	170
eta[83]	-1.03	0.11	-1.25	-1.10	-1.03	-0.94	-0.81	1.00	1000
eta[84]	-1.11	0.10	-1.32	-1.18	-1.11	-1.03	-0.90	1.00	1000
eta[85]	-1.03	0.09	-1.20	-1.09	-1.03	-0.97	-0.86	1.00	1000
eta[86]	-0.89	0.11	-1.11	-0.96	-0.89	-0.83	-0.68	1.00	940
eta[87]	-0.84	0.09	-1.01	-0.90	-0.83	-0.77	-0.66	1.00	1000
eta[88]	-1.36	0.11	-1.59	-1.44	-1.36	-1.29	-1.16	1.00	1000
eta[89]	-0.98	0.10	-1.18	-1.05	-0.98	-0.91	-0.79	1.00	1000
eta[90]	-0.89	0.10	-1.08	-0.96	-0.89	-0.83	-0.70	1.00	1000
eta[91]	-0.96	0.09	-1.14	-1.01	-0.96	-0.90	-0.78	1.00	1000
eta[92]	-0.71	0.09	-0.89	-0.77	-0.71	-0.64	-0.53	1.00	1000
eta[93]	-1.25	0.10	-1.45	-1.32	-1.25	-1.18	-1.05	1.00	540
eta[94]	-1.07	0.09	-1.25	-1.13	-1.07	-1.00	-0.89	1.00	500
eta[95]	-0.57	0.09	-0.76	-0.63	-0.57	-0.51	-0.41	1.00	1000
eta[96]	-1.06	0.08	-1.22	-1.12	-1.06	-1.00	-0.90	1.00	1000
eta[97]	-0.85	0.09	-1.02	-0.91	-0.85	-0.79	-0.67	1.00	1000
eta[98]	-1.24	0.11	-1.45	-1.31	-1.24	-1.16	-1.05	1.00	1000
eta[99]	-0.98	0.09	-1.16	-1.04	-0.98	-0.92	-0.82	1.00	1000
eta[100]	-0.59	0.09	-0.76	-0.64	-0.58	-0.53	-0.42	1.00	1000
eta[101]	-0.80	0.12	-1.05	-0.89	-0.80	-0.72	-0.58	1.00	730
eta[102]	-1.55	0.12	-1.80	-1.63	-1.55	-1.46	-1.31	1.00	670
eta[103]	-0.86	0.09	-1.05	-0.93	-0.86	-0.80	-0.68	1.00	1000
eta[104]	-1.60	0.13	-1.85	-1.68	-1.60	-1.51	-1.35	1.00	1000
eta[105]	-0.47	0.09	-0.64	-0.53	-0.47	-0.41	-0.31	1.00	1000
eta[106]	-1.33	0.11	-1.54	-1.40	-1.33	-1.25	-1.12	1.00	1000
eta[107]	-1.13	0.10	-1.33	-1.20	-1.13	-1.06	-0.95	1.00	730
eta[108]	-0.71	0.10	-0.89	-0.78	-0.71	-0.65	-0.53	1.00	1000
eta[109]	-1.32	0.10	-1.50	-1.38	-1.32	-1.25	-1.13	1.00	1000
eta[110]	-0.68	0.13	-0.93	-0.76	-0.68	-0.60	-0.43	1.00	1000
eta[111]	-0.69	0.09	-0.85	-0.74	-0.68	-0.62	-0.51	1.00	840
eta[112]	-1.93	0.14	-2.22	-2.03	-1.93	-1.84	-1.66	1.00	1000
eta[113]	-0.55	0.10	-0.74	-0.61	-0.55	-0.49	-0.35	1.00	1000
eta[114]	-0.92	0.10	-1.12	-0.98	-0.92	-0.85	-0.72	1.00	420
eta[115]	-0.54	0.09	-0.73	-0.60	-0.54	-0.48	-0.35	1.00	1000
eta[116]	-1.18	0.11	-1.40	-1.26	-1.19	-1.10	-0.97	1.00	1000
eta[117]	-0.91	0.09	-1.09	-0.97	-0.90	-0.84	-0.74	1.00	1000
eta[118]	-0.53	0.13	-0.78	-0.61	-0.53	-0.44	-0.28	1.01	240
eta[119]	-1.42	0.09	-1.60	-1.48	-1.42	-1.36	-1.25	1.00	1000
eta[120]	-1.85	0.12	-2.07	-1.93	-1.86	-1.77	-1.63	1.00	1000
eta[121]	-0.58	0.09	-0.76	-0.64	-0.58	-0.52	-0.42	1.00	440
eta[122]	-0.77	0.10	-0.98	-0.83	-0.77	-0.70	-0.57	1.00	840
eta[123]	-1.75	0.14	-2.02	-1.83	-1.75	-1.65	-1.50	1.00	1000
eta[124]	-1.00	0.10	-1.19	-1.06	-1.00	-0.93	-0.81	1.00	1000
eta[125]	-0.28	0.08	-0.44	-0.33	-0.28	-0.22	-0.11	1.00	1000
eta[126]	-2.59	0.09	-2.77	-2.65	-2.59	-2.53	-2.42	1.01	1000
eta[127]	-0.89	0.09	-1.06	-0.95	-0.89	-0.83	-0.71	1.00	1000
eta[128]	-1.47	0.10	-1.67	-1.53	-1.47	-1.41	-1.29	1.00	1000
eta[129]	-1.08	0.10	-1.28	-1.15	-1.09	-1.01	-0.87	1.00	1000
eta[130]	-0.91	0.09	-1.08	-0.97	-0.91	-0.85	-0.74	1.00	1000
eta[131]	-0.77	0.08	-0.93	-0.82	-0.76	-0.71	-0.60	1.00	1000
eta[132]	-0.63	0.08	-0.78	-0.68	-0.63	-0.58	-0.48	1.00	1000
eta[133]	-0.41	0.10	-0.61	-0.48	-0.41	-0.35	-0.21	1.00	950
eta[134]	-0.71	0.10	-0.90	-0.78	-0.71	-0.64	-0.51	1.00	670
eta[135]	-0.65	0.09	-0.82	-0.72	-0.66	-0.59	-0.47	1.00	640
eta[136]	-0.59	0.11	-0.82	-0.66	-0.59	-0.51	-0.37	1.00	1000

```
eta[137]     -1.01  0.10    -1.21    -1.08    -1.01    -0.94    -0.82 1.00  1000
eta[138]     -1.58  0.09    -1.75    -1.64    -1.57    -1.51    -1.40 1.00  1000
eta[139]     -1.66  0.13    -1.93    -1.74    -1.65    -1.57    -1.41 1.00   780
eta[140]     -0.99  0.09    -1.18    -1.05    -1.00    -0.93    -0.83 1.00   420
eta[141]     -0.79  0.10    -0.99    -0.85    -0.78    -0.73    -0.61 1.00  1000
eta[142]     -0.77  0.09    -0.94    -0.83    -0.78    -0.72    -0.60 1.00   430
eta[143]     -1.64  0.13    -1.90    -1.73    -1.64    -1.55    -1.39 1.00  1000
eta[144]     -0.79  0.08    -0.93    -0.84    -0.79    -0.74    -0.63 1.00  1000
eta[145]     -1.07  0.11    -1.29    -1.14    -1.07    -0.99    -0.85 1.01   270
eta[146]     -0.43  0.08    -0.58    -0.49    -0.43    -0.38    -0.29 1.00  1000
eta[147]     -2.52  0.22    -2.98    -2.66    -2.52    -2.38    -2.10 1.00  1000
eta[148]     -1.52  0.14    -1.82    -1.62    -1.52    -1.43    -1.26 1.00  1000
eta[149]     -0.77  0.09    -0.96    -0.83    -0.77    -0.71    -0.59 1.00  1000
eta[150]     -1.21  0.11    -1.41    -1.28    -1.21    -1.14    -1.00 1.00  1000
eta[151]     -0.98  0.10    -1.17    -1.04    -0.98    -0.92    -0.80 1.00   550
eta[152]     -1.52  0.14    -1.79    -1.62    -1.52    -1.42    -1.27 1.00  1000
eta[153]     -1.13  0.10    -1.34    -1.19    -1.13    -1.06    -0.93 1.00  1000
eta[154]     -0.99  0.09    -1.16    -1.05    -0.99    -0.93    -0.83 1.00  1000
eta[155]     -0.96  0.09    -1.14    -1.02    -0.96    -0.90    -0.79 1.00   970
eta[156]     -0.92  0.10    -1.12    -0.98    -0.92    -0.86    -0.73 1.00  1000
eta[157]     -0.27  0.08    -0.42    -0.32    -0.27    -0.22    -0.12 1.00  1000
eta[158]     -0.72  0.09    -0.89    -0.77    -0.72    -0.66    -0.55 1.00   890
eta[159]     -0.50  0.09    -0.67    -0.56    -0.50    -0.44    -0.33 1.00  1000
eta[160]     -1.36  0.12    -1.60    -1.45    -1.36    -1.28    -1.12 1.00  1000
eta[161]     -0.86  0.09    -1.03    -0.92    -0.86    -0.80    -0.69 1.00  1000
eta[162]     -0.51  0.10    -0.73    -0.58    -0.51    -0.44    -0.31 1.00  1000
eta[163]     -1.41  0.12    -1.65    -1.49    -1.42    -1.34    -1.18 1.00  1000
eta[164]     -1.51  0.11    -1.73    -1.58    -1.50    -1.43    -1.29 1.00  1000
eta[165]     -0.72  0.08    -0.87    -0.77    -0.72    -0.67    -0.56 1.01  1000
mu.eta       -1.06  0.04    -1.14    -1.08    -1.05    -1.03    -0.98 1.01   150
sigma.eta     0.49  0.03     0.44     0.47     0.49     0.51     0.55 1.00  1000
deviance  99159.12 19.18 99120.00 99150.00 99160.00 99170.00 99200.00 1.00   550

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 183.7 and DIC = 99342.9
DIC is an estimate of expected predictive error (lower deviance is better).
```

A.5.2. Subskala „Ruhe“

```
Inference for Bugs model at "tau_each_item_2.bug", fit using WinBUGS,
 2 chains, each with 1000 iterations (first 500 discarded)
 n.sims = 1000 iterations saved
              mean    sd     2.5%      25%      50%      75%    97.5% Rhat n.eff
beta[1,1]    -1.42  0.04    -1.51    -1.45    -1.42    -1.39    -1.34 1.00  1000
beta[1,2]     0.18  0.02     0.14     0.17     0.18     0.20     0.23 1.00  1000
beta[1,3]     1.02  0.02     0.99     1.01     1.02     1.03     1.06 1.00  1000
beta[1,4]     0.21  0.02     0.17     0.20     0.21     0.23     0.25 1.00  1000
beta[2,1]    -1.50  0.05    -1.59    -1.53    -1.50    -1.47    -1.41 1.00  1000
beta[2,2]     0.16  0.02     0.12     0.15     0.16     0.18     0.20 1.00  1000
beta[2,3]     1.18  0.02     1.14     1.17     1.18     1.20     1.22 1.00  1000
beta[2,4]     0.15  0.02     0.11     0.14     0.15     0.17     0.20 1.00  1000
eta[1]       -0.53  0.11    -0.75    -0.61    -0.53    -0.45    -0.32 1.00  1000
eta[2]       -0.94  0.13    -1.20    -1.03    -0.94    -0.85    -0.67 1.01   230
```

eta[3]	-1.60	0.14	-1.87	-1.69	-1.59	-1.50	-1.32	1.00	1000
eta[4]	-0.24	0.10	-0.45	-0.31	-0.24	-0.17	-0.04	1.01	230
eta[5]	-0.56	0.11	-0.78	-0.63	-0.56	-0.48	-0.33	1.00	650
eta[6]	-0.96	0.13	-1.21	-1.05	-0.96	-0.87	-0.69	1.00	1000
eta[7]	-0.87	0.12	-1.08	-0.95	-0.87	-0.79	-0.64	1.00	1000
eta[8]	-1.86	0.19	-2.24	-1.99	-1.85	-1.73	-1.50	1.00	1000
eta[9]	-0.54	0.12	-0.77	-0.62	-0.53	-0.46	-0.30	1.00	1000
eta[10]	-1.56	0.17	-1.94	-1.67	-1.56	-1.45	-1.24	1.00	460
eta[11]	-1.39	0.15	-1.69	-1.49	-1.39	-1.28	-1.10	1.00	1000
eta[12]	-0.96	0.12	-1.19	-1.04	-0.96	-0.88	-0.73	1.00	1000
eta[13]	-0.56	0.11	-0.77	-0.63	-0.56	-0.48	-0.35	1.00	1000
eta[14]	-0.38	0.11	-0.59	-0.46	-0.38	-0.30	-0.18	1.00	1000
eta[15]	-0.80	0.14	-1.07	-0.89	-0.80	-0.70	-0.53	1.00	1000
eta[16]	-0.47	0.11	-0.69	-0.54	-0.47	-0.39	-0.25	1.00	1000
eta[17]	-1.14	0.14	-1.43	-1.24	-1.14	-1.04	-0.89	1.01	260
eta[18]	-0.84	0.12	-1.08	-0.92	-0.84	-0.76	-0.60	1.00	1000
eta[19]	-0.70	0.13	-0.95	-0.79	-0.70	-0.61	-0.46	1.00	480
eta[20]	-0.75	0.13	-1.00	-0.83	-0.75	-0.67	-0.52	1.01	200
eta[21]	-0.70	0.12	-0.94	-0.78	-0.70	-0.61	-0.46	1.00	990
eta[22]	-1.64	0.17	-1.98	-1.75	-1.63	-1.52	-1.33	1.00	1000
eta[23]	-0.28	0.11	-0.49	-0.36	-0.28	-0.20	-0.04	1.01	280
eta[24]	-0.76	0.13	-1.00	-0.84	-0.75	-0.67	-0.51	1.00	1000
eta[25]	-0.54	0.11	-0.75	-0.61	-0.54	-0.47	-0.33	1.00	1000
eta[26]	-1.55	0.17	-1.90	-1.67	-1.55	-1.44	-1.24	1.01	290
eta[27]	-0.34	0.11	-0.55	-0.41	-0.34	-0.27	-0.11	1.00	1000
eta[28]	-0.71	0.12	-0.95	-0.79	-0.71	-0.64	-0.47	1.01	1000
eta[29]	-0.20	0.11	-0.40	-0.27	-0.19	-0.13	0.01	1.00	1000
eta[30]	-0.91	0.12	-1.16	-0.98	-0.91	-0.83	-0.66	1.00	1000
eta[31]	-0.39	0.12	-0.62	-0.46	-0.39	-0.31	-0.16	1.00	1000
eta[32]	-1.64	0.17	-1.97	-1.76	-1.64	-1.52	-1.33	1.00	1000
eta[33]	-1.34	0.15	-1.62	-1.44	-1.34	-1.23	-1.06	1.00	1000
eta[34]	-0.59	0.11	-0.81	-0.66	-0.59	-0.51	-0.36	1.00	1000
eta[35]	-0.88	0.13	-1.14	-0.96	-0.88	-0.80	-0.64	1.00	1000
eta[36]	-0.56	0.12	-0.79	-0.64	-0.56	-0.48	-0.33	1.00	1000
eta[37]	-0.81	0.13	-1.07	-0.90	-0.81	-0.73	-0.54	1.00	1000
eta[38]	-0.36	0.13	-0.61	-0.45	-0.36	-0.28	-0.10	1.00	880
eta[39]	-1.00	0.14	-1.28	-1.09	-1.00	-0.90	-0.74	1.00	1000
eta[40]	-0.73	0.14	-1.01	-0.82	-0.72	-0.64	-0.47	1.00	1000
eta[41]	-0.94	0.13	-1.19	-1.03	-0.94	-0.86	-0.69	1.00	1000
eta[42]	-0.42	0.12	-0.66	-0.51	-0.42	-0.34	-0.19	1.00	1000
eta[43]	-1.23	0.14	-1.50	-1.32	-1.23	-1.13	-0.97	1.00	780
eta[44]	-0.55	0.11	-0.78	-0.63	-0.55	-0.47	-0.34	1.00	1000
eta[45]	-0.83	0.13	-1.09	-0.91	-0.83	-0.75	-0.57	1.00	580
eta[46]	-0.87	0.13	-1.12	-0.96	-0.87	-0.79	-0.63	1.00	1000
eta[47]	-0.63	0.12	-0.86	-0.71	-0.63	-0.55	-0.40	1.00	770
eta[48]	-0.97	0.14	-1.24	-1.06	-0.97	-0.87	-0.70	1.00	1000
eta[49]	-1.27	0.16	-1.60	-1.38	-1.27	-1.15	-0.98	1.00	1000
eta[50]	-0.71	0.20	-1.09	-0.84	-0.71	-0.57	-0.30	1.00	1000
eta[51]	-1.46	0.15	-1.75	-1.56	-1.45	-1.35	-1.17	1.00	1000
eta[52]	-2.06	0.19	-2.43	-2.18	-2.05	-1.93	-1.72	1.00	1000
eta[53]	-1.36	0.12	-1.62	-1.44	-1.36	-1.28	-1.13	1.00	1000
eta[54]	-0.36	0.12	-0.62	-0.43	-0.36	-0.28	-0.14	1.01	140
eta[55]	-0.75	0.12	-0.98	-0.82	-0.75	-0.67	-0.53	1.00	1000
eta[56]	-1.17	0.16	-1.52	-1.28	-1.17	-1.06	-0.86	1.00	860
eta[57]	-0.58	0.13	-0.83	-0.66	-0.58	-0.49	-0.31	1.00	1000
eta[58]	-1.05	0.13	-1.30	-1.14	-1.05	-0.96	-0.79	1.00	1000

eta[59]	-0.93	0.12	-1.17	-1.01	-0.93	-0.85	-0.68	1.00	420
eta[60]	-0.87	0.12	-1.11	-0.96	-0.87	-0.79	-0.62	1.00	1000
eta[61]	-1.60	0.17	-1.93	-1.71	-1.59	-1.48	-1.29	1.00	1000
eta[62]	-0.45	0.11	-0.66	-0.53	-0.46	-0.37	-0.23	1.00	1000
eta[63]	-0.73	0.13	-0.97	-0.81	-0.74	-0.65	-0.48	1.00	1000
eta[64]	-1.37	0.13	-1.64	-1.46	-1.37	-1.29	-1.14	1.00	1000
eta[65]	-1.11	0.14	-1.39	-1.20	-1.10	-1.01	-0.83	1.00	740
eta[66]	-0.72	0.12	-0.95	-0.80	-0.72	-0.63	-0.48	1.01	380
eta[67]	-1.00	0.14	-1.26	-1.10	-1.00	-0.91	-0.72	1.00	1000
eta[68]	-0.85	0.12	-1.10	-0.94	-0.85	-0.77	-0.60	1.00	1000
eta[69]	-1.01	0.13	-1.26	-1.10	-1.01	-0.93	-0.76	1.00	760
eta[70]	-0.64	0.11	-0.87	-0.72	-0.64	-0.57	-0.42	1.00	1000
eta[71]	-1.38	0.15	-1.66	-1.48	-1.37	-1.27	-1.09	1.01	340
eta[72]	-0.44	0.12	-0.67	-0.51	-0.43	-0.36	-0.21	1.00	1000
eta[73]	-1.25	0.14	-1.52	-1.35	-1.25	-1.15	-0.97	1.00	1000
eta[74]	-0.63	0.12	-0.87	-0.71	-0.63	-0.55	-0.40	1.00	1000
eta[75]	-0.64	0.12	-0.88	-0.72	-0.63	-0.56	-0.40	1.00	1000
eta[76]	-0.88	0.13	-1.15	-0.97	-0.88	-0.79	-0.62	1.00	1000
eta[77]	-0.85	0.12	-1.07	-0.93	-0.85	-0.77	-0.60	1.00	1000
eta[78]	-0.41	0.12	-0.65	-0.50	-0.41	-0.33	-0.18	1.00	600
eta[79]	-0.29	0.11	-0.50	-0.37	-0.29	-0.22	-0.09	1.00	1000
eta[80]	-0.60	0.12	-0.84	-0.68	-0.59	-0.52	-0.38	1.00	1000
eta[81]	-0.81	0.12	-1.04	-0.89	-0.81	-0.72	-0.57	1.00	1000
eta[82]	-0.79	0.13	-1.04	-0.88	-0.79	-0.70	-0.52	1.00	1000
eta[83]	-0.83	0.15	-1.11	-0.92	-0.83	-0.73	-0.54	1.01	320
eta[84]	-0.89	0.13	-1.12	-0.98	-0.89	-0.80	-0.64	1.00	1000
eta[85]	-0.22	0.10	-0.42	-0.29	-0.22	-0.15	-0.03	1.00	1000
eta[86]	-0.72	0.14	-1.00	-0.81	-0.72	-0.62	-0.45	1.00	1000
eta[87]	-0.85	0.12	-1.09	-0.93	-0.85	-0.77	-0.62	1.01	320
eta[88]	-0.73	0.12	-0.99	-0.81	-0.73	-0.65	-0.49	1.00	1000
eta[89]	-0.94	0.13	-1.20	-1.03	-0.94	-0.85	-0.69	1.00	1000
eta[90]	-0.93	0.13	-1.16	-1.02	-0.93	-0.83	-0.67	1.00	1000
eta[91]	-0.84	0.12	-1.09	-0.92	-0.85	-0.76	-0.60	1.00	1000
eta[92]	-0.81	0.13	-1.05	-0.90	-0.81	-0.72	-0.54	1.02	120
eta[93]	-0.67	0.12	-0.91	-0.75	-0.66	-0.58	-0.45	1.00	1000
eta[94]	-1.30	0.12	-1.54	-1.38	-1.29	-1.21	-1.06	1.00	1000
eta[95]	-0.34	0.11	-0.56	-0.42	-0.34	-0.26	-0.13	1.00	1000
eta[96]	-0.95	0.11	-1.17	-1.03	-0.95	-0.88	-0.73	1.00	1000
eta[97]	-0.78	0.12	-1.01	-0.86	-0.78	-0.70	-0.54	1.00	620
eta[98]	-0.91	0.13	-1.18	-0.99	-0.90	-0.82	-0.66	1.01	220
eta[99]	-0.54	0.11	-0.76	-0.61	-0.53	-0.46	-0.32	1.00	1000
eta[100]	-0.63	0.12	-0.87	-0.71	-0.62	-0.54	-0.39	1.00	1000
eta[101]	-0.91	0.18	-1.26	-1.03	-0.91	-0.79	-0.57	1.00	1000
eta[102]	-0.89	0.13	-1.15	-0.97	-0.88	-0.80	-0.65	1.01	340
eta[103]	-0.66	0.11	-0.89	-0.74	-0.67	-0.59	-0.45	1.00	1000
eta[104]	-1.42	0.16	-1.74	-1.52	-1.41	-1.30	-1.12	1.00	1000
eta[105]	-0.46	0.12	-0.69	-0.54	-0.46	-0.38	-0.21	1.00	1000
eta[106]	-1.28	0.15	-1.57	-1.38	-1.28	-1.18	-0.97	1.00	1000
eta[107]	-0.93	0.13	-1.20	-1.01	-0.93	-0.84	-0.69	1.00	1000
eta[108]	-0.77	0.12	-1.01	-0.85	-0.77	-0.69	-0.53	1.00	1000
eta[109]	-1.01	0.12	-1.23	-1.09	-1.01	-0.93	-0.78	1.00	1000
eta[110]	-0.64	0.16	-0.97	-0.73	-0.63	-0.54	-0.32	1.01	270
eta[111]	-0.63	0.12	-0.86	-0.71	-0.63	-0.55	-0.41	1.00	1000
eta[112]	-1.75	0.17	-2.10	-1.87	-1.75	-1.65	-1.42	1.00	1000
eta[113]	-0.41	0.13	-0.66	-0.50	-0.41	-0.33	-0.15	1.01	260
eta[114]	-0.81	0.12	-1.05	-0.89	-0.81	-0.73	-0.57	1.00	1000

```
eta[115]     -0.49  0.13    -0.74    -0.57    -0.49    -0.40    -0.23 1.00   480
eta[116]     -1.05  0.14    -1.31    -1.14    -1.05    -0.95    -0.76 1.00  1000
eta[117]     -0.61  0.11    -0.85    -0.68    -0.61    -0.52    -0.39 1.00  1000
eta[118]     -0.50  0.17    -0.83    -0.61    -0.50    -0.38    -0.18 1.00  1000
eta[119]     -1.92  0.15    -2.20    -2.01    -1.91    -1.82    -1.63 1.00  1000
eta[120]     -0.99  0.13    -1.25    -1.08    -0.99    -0.91    -0.75 1.00   400
eta[121]     -0.62  0.12    -0.86    -0.71    -0.62    -0.53    -0.39 1.00  1000
eta[122]     -0.81  0.13    -1.08    -0.90    -0.81    -0.71    -0.56 1.00  1000
eta[123]     -1.13  0.15    -1.42    -1.22    -1.13    -1.03    -0.83 1.00  1000
eta[124]     -0.89  0.13    -1.16    -0.97    -0.88    -0.80    -0.65 1.01   210
eta[125]     -0.49  0.12    -0.72    -0.56    -0.48    -0.40    -0.26 1.00   830
eta[126]     -1.50  0.11    -1.72    -1.57    -1.50    -1.42    -1.28 1.00  1000
eta[127]     -1.01  0.11    -1.23    -1.08    -1.00    -0.93    -0.79 1.00  1000
eta[128]     -1.36  0.14    -1.62    -1.46    -1.36    -1.27    -1.08 1.00  1000
eta[129]     -1.00  0.14    -1.27    -1.09    -1.00    -0.91    -0.75 1.00  1000
eta[130]     -1.81  0.13    -2.06    -1.90    -1.81    -1.73    -1.57 1.00   510
eta[131]     -0.70  0.11    -0.93    -0.77    -0.70    -0.62    -0.48 1.00  1000
eta[132]     -0.71  0.11    -0.94    -0.78    -0.70    -0.63    -0.50 1.00  1000
eta[133]     -0.33  0.13    -0.59    -0.41    -0.32    -0.24    -0.07 1.00  1000
eta[134]     -0.32  0.12    -0.56    -0.39    -0.32    -0.24    -0.09 1.00  1000
eta[135]     -0.42  0.11    -0.64    -0.49    -0.42    -0.35    -0.20 1.00  1000
eta[136]     -0.47  0.14    -0.74    -0.57    -0.46    -0.37    -0.19 1.00   650
eta[137]     -0.54  0.11    -0.76    -0.62    -0.54    -0.46    -0.32 1.01   220
eta[138]     -0.51  0.12    -0.73    -0.59    -0.52    -0.44    -0.28 1.00  1000
eta[139]     -1.65  0.18    -2.02    -1.76    -1.64    -1.54    -1.33 1.00  1000
eta[140]     -0.85  0.12    -1.07    -0.93    -0.85    -0.77    -0.63 1.01   140
eta[141]     -0.87  0.13    -1.13    -0.96    -0.87    -0.78    -0.62 1.00   980
eta[142]     -1.06  0.12    -1.29    -1.14    -1.06    -0.98    -0.84 1.00   350
eta[143]     -1.56  0.16    -1.87    -1.68    -1.56    -1.46    -1.24 1.00  1000
eta[144]     -0.57  0.11    -0.80    -0.64    -0.57    -0.49    -0.37 1.00   750
eta[145]     -0.72  0.14    -1.00    -0.81    -0.72    -0.62    -0.45 1.01   270
eta[146]     -0.40  0.11    -0.60    -0.47    -0.40    -0.34    -0.19 1.00  1000
eta[147]     -2.07  0.24    -2.56    -2.24    -2.06    -1.90    -1.62 1.00  1000
eta[148]     -0.71  0.14    -1.00    -0.80    -0.71    -0.61    -0.43 1.00  1000
eta[149]     -0.33  0.11    -0.54    -0.40    -0.33    -0.26    -0.12 1.00   740
eta[150]     -0.73  0.13    -0.97    -0.82    -0.72    -0.64    -0.48 1.00  1000
eta[151]     -0.66  0.11    -0.88    -0.73    -0.66    -0.59    -0.44 1.00  1000
eta[152]     -1.55  0.19    -1.88    -1.68    -1.55    -1.42    -1.19 1.00  1000
eta[153]     -0.34  0.12    -0.56    -0.43    -0.34    -0.26    -0.11 1.00  1000
eta[154]     -0.96  0.12    -1.22    -1.04    -0.95    -0.87    -0.72 1.00  1000
eta[155]     -0.86  0.13    -1.11    -0.95    -0.86    -0.77    -0.62 1.00   710
eta[156]     -0.73  0.12    -0.97    -0.82    -0.73    -0.65    -0.49 1.00  1000
eta[157]     -0.26  0.11    -0.47    -0.33    -0.26    -0.19    -0.05 1.00  1000
eta[158]     -0.53  0.11    -0.75    -0.60    -0.53    -0.45    -0.30 1.00  1000
eta[159]     -0.78  0.13    -1.03    -0.86    -0.77    -0.69    -0.53 1.00   610
eta[160]     -1.11  0.15    -1.38    -1.21    -1.11    -1.01    -0.83 1.00  1000
eta[161]     -0.75  0.12    -1.00    -0.83    -0.75    -0.67    -0.53 1.00  1000
eta[162]     -0.72  0.15    -1.02    -0.81    -0.72    -0.62    -0.43 1.00  1000
eta[163]     -1.49  0.16    -1.81    -1.59    -1.48    -1.37    -1.19 1.00  1000
eta[164]     -1.56  0.16    -1.89    -1.67    -1.56    -1.45    -1.28 1.01   330
eta[165]     -0.48  0.10    -0.66    -0.55    -0.47    -0.41    -0.29 1.01   920
mu.eta       -0.86  0.03    -0.93    -0.88    -0.86    -0.84    -0.79 1.00  1000
sigma.eta     0.42  0.03     0.38     0.40     0.42     0.44     0.48 1.00  1000
deviance  55608.99 18.94 55570.00 55600.00 55610.00 55620.00 55640.00 1.00  1000

For each parameter, n.eff is a crude measure of effective sample size,
```

```
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 179.5 and DIC = 55788.5
DIC is an estimate of expected predictive error (lower deviance is better).
```

A.5.3. Subskala „Wachheit“

```
Inference for Bugs model at "tau_each_item_2.bug", fit using WinBUGS,
 2 chains, each with 1000 iterations (first 500 discarded)
 n.sims = 1000 iterations saved
```

	mean	sd	2.5%	25%	50%	75%	97.5%	Rhat	n.eff
beta[1,1]	-0.70	0.02	-0.75	-0.72	-0.70	-0.69	-0.65	1.01	330
beta[1,2]	0.17	0.02	0.14	0.16	0.17	0.18	0.20	1.00	920
beta[1,3]	0.74	0.01	0.71	0.73	0.74	0.75	0.77	1.00	990
beta[1,4]	-0.20	0.02	-0.24	-0.22	-0.20	-0.19	-0.16	1.00	1000
beta[2,1]	-0.68	0.02	-0.73	-0.70	-0.68	-0.67	-0.64	1.00	1000
beta[2,2]	0.24	0.02	0.21	0.23	0.24	0.25	0.28	1.00	1000
beta[2,3]	0.78	0.01	0.76	0.77	0.78	0.79	0.81	1.00	1000
beta[2,4]	-0.34	0.02	-0.39	-0.36	-0.34	-0.33	-0.30	1.00	1000
eta[1]	-1.20	0.11	-1.41	-1.27	-1.20	-1.13	-1.00	1.00	340
eta[2]	-0.34	0.09	-0.51	-0.40	-0.34	-0.28	-0.15	1.00	1000
eta[3]	-0.84	0.11	-1.05	-0.91	-0.84	-0.77	-0.63	1.00	470
eta[4]	-0.53	0.10	-0.72	-0.59	-0.53	-0.45	-0.33	1.00	1000
eta[5]	-0.73	0.11	-0.94	-0.80	-0.73	-0.66	-0.53	1.00	1000
eta[6]	-0.50	0.10	-0.71	-0.57	-0.51	-0.43	-0.31	1.00	550
eta[7]	-0.87	0.11	-1.08	-0.95	-0.87	-0.80	-0.67	1.00	1000
eta[8]	-0.92	0.12	-1.15	-1.00	-0.92	-0.85	-0.70	1.00	1000
eta[9]	-0.43	0.11	-0.63	-0.50	-0.43	-0.36	-0.23	1.01	410
eta[10]	-0.56	0.10	-0.76	-0.63	-0.56	-0.49	-0.35	1.00	1000
eta[11]	-0.74	0.11	-0.96	-0.81	-0.73	-0.66	-0.52	1.00	1000
eta[12]	-0.47	0.09	-0.65	-0.54	-0.47	-0.41	-0.29	1.01	290
eta[13]	-0.37	0.09	-0.55	-0.43	-0.36	-0.30	-0.18	1.00	1000
eta[14]	-0.65	0.11	-0.88	-0.72	-0.65	-0.58	-0.45	1.00	1000
eta[15]	-0.73	0.12	-0.95	-0.81	-0.73	-0.66	-0.49	1.00	1000
eta[16]	-0.50	0.09	-0.69	-0.56	-0.50	-0.43	-0.32	1.00	960
eta[17]	-0.62	0.11	-0.84	-0.69	-0.62	-0.54	-0.41	1.00	360
eta[18]	-0.38	0.10	-0.56	-0.45	-0.39	-0.32	-0.19	1.00	1000
eta[19]	-0.76	0.12	-1.00	-0.85	-0.76	-0.68	-0.51	1.00	1000
eta[20]	-1.00	0.13	-1.24	-1.09	-0.99	-0.91	-0.77	1.00	1000
eta[21]	-0.68	0.11	-0.89	-0.76	-0.68	-0.61	-0.47	1.00	1000
eta[22]	-0.89	0.12	-1.12	-0.97	-0.89	-0.81	-0.65	1.00	1000
eta[23]	-0.44	0.10	-0.65	-0.50	-0.43	-0.37	-0.24	1.00	1000
eta[24]	-0.76	0.11	-0.98	-0.84	-0.76	-0.69	-0.56	1.00	690
eta[25]	-0.54	0.10	-0.73	-0.60	-0.54	-0.47	-0.34	1.00	1000
eta[26]	-1.05	0.12	-1.29	-1.13	-1.05	-0.97	-0.83	1.00	1000
eta[27]	-0.48	0.10	-0.67	-0.54	-0.48	-0.41	-0.30	1.00	1000
eta[28]	-1.22	0.11	-1.44	-1.30	-1.22	-1.14	-1.00	1.00	1000
eta[29]	-0.47	0.09	-0.64	-0.53	-0.47	-0.41	-0.30	1.00	500
eta[30]	-1.02	0.12	-1.25	-1.11	-1.02	-0.94	-0.80	1.00	1000
eta[31]	-0.66	0.11	-0.88	-0.73	-0.66	-0.58	-0.44	1.00	550
eta[32]	-1.35	0.13	-1.62	-1.44	-1.35	-1.26	-1.10	1.00	1000
eta[33]	-0.99	0.12	-1.22	-1.07	-0.99	-0.91	-0.75	1.00	1000
eta[34]	-0.74	0.09	-0.92	-0.81	-0.74	-0.68	-0.56	1.01	950
eta[35]	-0.70	0.11	-0.92	-0.78	-0.71	-0.63	-0.49	1.00	1000
eta[36]	-0.84	0.11	-1.05	-0.91	-0.84	-0.77	-0.63	1.00	540

eta[37]	-0.57	0.11	-0.77	-0.64	-0.57	-0.49	-0.37	1.01	200
eta[38]	-0.40	0.11	-0.61	-0.48	-0.41	-0.33	-0.18	1.00	1000
eta[39]	-0.84	0.11	-1.06	-0.91	-0.84	-0.76	-0.63	1.00	1000
eta[40]	-0.20	0.10	-0.40	-0.27	-0.20	-0.12	0.00	1.01	200
eta[41]	-1.23	0.13	-1.48	-1.31	-1.23	-1.14	-0.99	1.00	1000
eta[42]	-0.56	0.11	-0.79	-0.64	-0.56	-0.49	-0.34	1.00	740
eta[43]	-1.02	0.11	-1.22	-1.10	-1.02	-0.94	-0.81	1.00	1000
eta[44]	-0.55	0.11	-0.77	-0.62	-0.54	-0.47	-0.34	1.00	1000
eta[45]	-0.43	0.10	-0.62	-0.50	-0.43	-0.36	-0.23	1.00	1000
eta[46]	-0.67	0.10	-0.87	-0.74	-0.67	-0.61	-0.49	1.00	1000
eta[47]	-0.30	0.10	-0.50	-0.37	-0.30	-0.23	-0.12	1.00	750
eta[48]	-0.48	0.10	-0.68	-0.56	-0.48	-0.41	-0.28	1.00	1000
eta[49]	-0.97	0.13	-1.23	-1.06	-0.97	-0.87	-0.73	1.00	390
eta[50]	-1.02	0.17	-1.37	-1.14	-1.02	-0.91	-0.69	1.00	1000
eta[51]	-0.77	0.11	-0.98	-0.84	-0.77	-0.70	-0.57	1.00	1000
eta[52]	-1.22	0.13	-1.47	-1.30	-1.22	-1.14	-0.97	1.00	1000
eta[53]	-1.51	0.12	-1.75	-1.60	-1.52	-1.44	-1.28	1.00	410
eta[54]	-0.63	0.11	-0.85	-0.70	-0.62	-0.55	-0.41	1.00	740
eta[55]	-0.87	0.11	-1.08	-0.94	-0.87	-0.80	-0.65	1.00	1000
eta[56]	-0.58	0.12	-0.81	-0.66	-0.58	-0.50	-0.34	1.00	1000
eta[57]	-0.42	0.11	-0.63	-0.50	-0.42	-0.35	-0.20	1.01	160
eta[58]	-0.61	0.10	-0.81	-0.68	-0.61	-0.54	-0.42	1.00	1000
eta[59]	-1.42	0.12	-1.64	-1.50	-1.41	-1.34	-1.18	1.00	1000
eta[60]	-0.90	0.12	-1.14	-0.97	-0.89	-0.83	-0.67	1.00	1000
eta[61]	-0.75	0.11	-0.97	-0.83	-0.75	-0.68	-0.52	1.00	1000
eta[62]	-0.40	0.10	-0.61	-0.46	-0.39	-0.33	-0.21	1.00	840
eta[63]	-0.75	0.11	-0.95	-0.82	-0.75	-0.67	-0.53	1.00	1000
eta[64]	-0.84	0.10	-1.04	-0.91	-0.84	-0.77	-0.62	1.00	1000
eta[65]	-0.48	0.11	-0.71	-0.55	-0.48	-0.40	-0.26	1.00	570
eta[66]	-0.76	0.11	-0.99	-0.84	-0.76	-0.69	-0.55	1.00	1000
eta[67]	-1.00	0.13	-1.25	-1.09	-1.00	-0.92	-0.75	1.00	1000
eta[68]	-0.93	0.11	-1.16	-1.00	-0.93	-0.86	-0.71	1.00	1000
eta[69]	-0.83	0.11	-1.03	-0.90	-0.83	-0.76	-0.62	1.00	1000
eta[70]	-0.88	0.11	-1.09	-0.95	-0.88	-0.81	-0.66	1.00	1000
eta[71]	-0.96	0.12	-1.21	-1.05	-0.96	-0.88	-0.72	1.00	1000
eta[72]	-0.31	0.10	-0.50	-0.37	-0.31	-0.24	-0.12	1.00	1000
eta[73]	-0.68	0.10	-0.88	-0.75	-0.68	-0.61	-0.48	1.00	1000
eta[74]	-0.49	0.10	-0.68	-0.56	-0.49	-0.43	-0.30	1.00	1000
eta[75]	-0.64	0.10	-0.83	-0.71	-0.64	-0.57	-0.44	1.01	800
eta[76]	-0.42	0.10	-0.64	-0.49	-0.42	-0.35	-0.21	1.00	740
eta[77]	-0.70	0.10	-0.91	-0.77	-0.70	-0.63	-0.51	1.00	840
eta[78]	-0.63	0.11	-0.85	-0.70	-0.63	-0.56	-0.43	1.00	630
eta[79]	-0.59	0.10	-0.79	-0.66	-0.59	-0.52	-0.40	1.00	710
eta[80]	-0.58	0.10	-0.77	-0.64	-0.58	-0.51	-0.38	1.00	1000
eta[81]	-0.66	0.10	-0.85	-0.73	-0.66	-0.59	-0.47	1.00	1000
eta[82]	-0.77	0.11	-0.98	-0.84	-0.76	-0.69	-0.56	1.00	1000
eta[83]	-0.56	0.12	-0.78	-0.64	-0.56	-0.49	-0.33	1.00	970
eta[84]	-0.70	0.11	-0.92	-0.78	-0.70	-0.63	-0.49	1.00	1000
eta[85]	-0.47	0.09	-0.65	-0.53	-0.47	-0.42	-0.29	1.00	1000
eta[86]	-0.89	0.12	-1.13	-0.98	-0.89	-0.81	-0.66	1.00	1000
eta[87]	-0.67	0.10	-0.87	-0.74	-0.67	-0.61	-0.48	1.00	1000
eta[88]	-0.57	0.10	-0.77	-0.64	-0.57	-0.50	-0.38	1.00	1000
eta[89]	-1.08	0.12	-1.32	-1.17	-1.08	-1.00	-0.85	1.00	1000
eta[90]	-0.60	0.11	-0.81	-0.67	-0.60	-0.53	-0.38	1.00	1000
eta[91]	-0.83	0.11	-1.05	-0.91	-0.83	-0.75	-0.60	1.00	1000
eta[92]	-0.79	0.11	-1.01	-0.87	-0.79	-0.71	-0.61	1.00	1000

eta[93]	-0.64	0.10	-0.84	-0.71	-0.64	-0.57	-0.45	1.00	1000
eta[94]	-0.95	0.10	-1.16	-1.02	-0.95	-0.88	-0.75	1.01	770
eta[95]	-0.36	0.09	-0.55	-0.42	-0.36	-0.29	-0.18	1.00	1000
eta[96]	-0.57	0.09	-0.75	-0.64	-0.57	-0.51	-0.40	1.00	1000
eta[97]	-0.70	0.10	-0.89	-0.76	-0.69	-0.63	-0.50	1.00	630
eta[98]	-0.76	0.10	-0.95	-0.83	-0.76	-0.69	-0.57	1.00	910
eta[99]	-0.50	0.10	-0.69	-0.57	-0.50	-0.43	-0.29	1.00	1000
eta[100]	-0.61	0.10	-0.81	-0.68	-0.61	-0.54	-0.42	1.00	1000
eta[101]	-0.45	0.13	-0.72	-0.53	-0.45	-0.36	-0.19	1.01	1000
eta[102]	-1.00	0.11	-1.22	-1.08	-1.00	-0.92	-0.77	1.01	550
eta[103]	-0.52	0.10	-0.74	-0.59	-0.52	-0.46	-0.33	1.01	1000
eta[104]	-1.21	0.13	-1.47	-1.30	-1.20	-1.12	-0.99	1.00	900
eta[105]	-0.36	0.10	-0.54	-0.42	-0.36	-0.29	-0.17	1.00	1000
eta[106]	-0.75	0.11	-0.96	-0.82	-0.74	-0.67	-0.54	1.00	790
eta[107]	-0.75	0.11	-0.96	-0.82	-0.75	-0.67	-0.54	1.00	1000
eta[108]	-0.80	0.11	-1.01	-0.87	-0.80	-0.72	-0.58	1.00	1000
eta[109]	-0.88	0.11	-1.10	-0.95	-0.88	-0.81	-0.68	1.00	1000
eta[110]	-0.50	0.13	-0.76	-0.59	-0.50	-0.41	-0.24	1.00	1000
eta[111]	-0.38	0.10	-0.56	-0.44	-0.38	-0.31	-0.19	1.00	990
eta[112]	-1.08	0.11	-1.31	-1.16	-1.08	-1.00	-0.86	1.00	1000
eta[113]	-0.42	0.11	-0.62	-0.49	-0.41	-0.35	-0.20	1.00	590
eta[114]	-0.77	0.10	-0.98	-0.84	-0.78	-0.71	-0.56	1.00	380
eta[115]	-0.42	0.11	-0.65	-0.50	-0.42	-0.35	-0.22	1.00	430
eta[116]	-0.71	0.12	-0.93	-0.79	-0.71	-0.63	-0.49	1.00	1000
eta[117]	-0.51	0.10	-0.71	-0.57	-0.50	-0.44	-0.31	1.00	1000
eta[118]	-0.34	0.15	-0.64	-0.44	-0.34	-0.24	-0.06	1.00	730
eta[119]	-0.82	0.10	-1.02	-0.89	-0.82	-0.75	-0.64	1.01	260
eta[120]	-1.20	0.12	-1.43	-1.29	-1.20	-1.12	-0.99	1.00	1000
eta[121]	-0.41	0.10	-0.59	-0.48	-0.41	-0.34	-0.22	1.00	1000
eta[122]	-0.76	0.12	-0.98	-0.83	-0.75	-0.68	-0.53	1.00	1000
eta[123]	-1.02	0.13	-1.29	-1.11	-1.02	-0.93	-0.78	1.00	350
eta[124]	-0.90	0.11	-1.12	-0.98	-0.90	-0.82	-0.68	1.01	210
eta[125]	-0.23	0.10	-0.44	-0.30	-0.23	-0.17	-0.05	1.00	340
eta[126]	-1.68	0.13	-1.92	-1.76	-1.68	-1.59	-1.42	1.00	500
eta[127]	-0.79	0.09	-0.98	-0.85	-0.79	-0.72	-0.62	1.00	890
eta[128]	-0.97	0.11	-1.18	-1.05	-0.97	-0.89	-0.75	1.01	190
eta[129]	-0.97	0.12	-1.22	-1.05	-0.97	-0.89	-0.75	1.00	1000
eta[130]	-1.31	0.11	-1.52	-1.39	-1.31	-1.24	-1.09	1.00	1000
eta[131]	-0.53	0.10	-0.72	-0.60	-0.53	-0.47	-0.34	1.00	1000
eta[132]	-0.54	0.10	-0.74	-0.60	-0.54	-0.47	-0.35	1.00	1000
eta[133]	-0.32	0.12	-0.55	-0.41	-0.32	-0.25	-0.09	1.00	800
eta[134]	-0.37	0.11	-0.59	-0.45	-0.37	-0.30	-0.17	1.00	580
eta[135]	-0.51	0.10	-0.72	-0.58	-0.51	-0.44	-0.31	1.00	1000
eta[136]	-0.52	0.12	-0.75	-0.61	-0.52	-0.45	-0.28	1.00	920
eta[137]	-0.38	0.09	-0.55	-0.45	-0.38	-0.32	-0.21	1.00	1000
eta[138]	-0.34	0.10	-0.52	-0.40	-0.34	-0.27	-0.16	1.00	690
eta[139]	-1.02	0.12	-1.27	-1.11	-1.02	-0.94	-0.78	1.00	1000
eta[140]	-0.84	0.11	-1.05	-0.91	-0.83	-0.76	-0.63	1.00	840
eta[141]	-0.82	0.11	-1.05	-0.89	-0.82	-0.74	-0.60	1.00	1000
eta[142]	-0.68	0.10	-0.88	-0.74	-0.67	-0.61	-0.48	1.00	1000
eta[143]	-0.96	0.12	-1.19	-1.04	-0.96	-0.88	-0.73	1.00	1000
eta[144]	-0.41	0.08	-0.56	-0.46	-0.41	-0.35	-0.24	1.00	1000
eta[145]	-0.79	0.13	-1.03	-0.88	-0.80	-0.71	-0.55	1.00	1000
eta[146]	-0.36	0.09	-0.54	-0.43	-0.36	-0.30	-0.17	1.00	1000
eta[147]	-1.35	0.19	-1.75	-1.47	-1.35	-1.22	-1.01	1.00	1000
eta[148]	-1.09	0.13	-1.36	-1.18	-1.09	-1.00	-0.84	1.00	400

```
eta[149]     -0.59  0.09    -0.77    -0.66    -0.59    -0.53    -0.41 1.00   980
eta[150]     -0.46  0.10    -0.66    -0.54    -0.47    -0.39    -0.25 1.00  1000
eta[151]     -0.69  0.10    -0.89    -0.76    -0.68    -0.62    -0.49 1.00  1000
eta[152]     -0.59  0.12    -0.83    -0.67    -0.59    -0.52    -0.37 1.00  1000
eta[153]     -0.55  0.11    -0.77    -0.62    -0.55    -0.48    -0.35 1.00   760
eta[154]     -0.77  0.11    -0.98    -0.83    -0.77    -0.69    -0.55 1.00   700
eta[155]     -0.53  0.11    -0.74    -0.60    -0.53    -0.46    -0.32 1.00   320
eta[156]     -0.87  0.11    -1.09    -0.95    -0.87    -0.80    -0.67 1.00  1000
eta[157]     -0.18  0.09    -0.35    -0.24    -0.18    -0.12     0.00 1.00  1000
eta[158]     -0.29  0.09    -0.48    -0.35    -0.29    -0.22    -0.11 1.00  1000
eta[159]     -0.54  0.10    -0.74    -0.61    -0.54    -0.47    -0.34 1.00  1000
eta[160]     -0.40  0.10    -0.61    -0.47    -0.40    -0.34    -0.20 1.00  1000
eta[161]     -0.59  0.10    -0.78    -0.65    -0.59    -0.52    -0.39 1.00  1000
eta[162]     -0.64  0.12    -0.88    -0.72    -0.64    -0.57    -0.43 1.00   850
eta[163]     -0.87  0.12    -1.11    -0.95    -0.87    -0.80    -0.65 1.00  1000
eta[164]     -0.86  0.11    -1.08    -0.94    -0.86    -0.79    -0.64 1.00  1000
eta[165]     -1.23  0.10    -1.44    -1.30    -1.23    -1.16    -1.04 1.00  1000
mu.eta       -0.71  0.03    -0.76    -0.73    -0.71    -0.69    -0.66 1.00  1000
sigma.eta     0.30  0.02     0.27     0.29     0.30     0.31     0.34 1.00   330
deviance  66855.54 18.89 66820.00 66840.00 66860.00 66870.00 66890.00 1.01   250

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 177.9 and DIC = 67033.4
DIC is an estimate of expected predictive error (lower deviance is better).
```

A.6. Zusammenfassung

In der vorliegenden Arbeit wird ein probabilistisches Item-Response-Modell zur Erfassung von intraindividueller Variabilität in multivariaten, diskreten Zeitreihen generiert. Manifester Indikator der Variabilität ist die mittlere absolute Differenz auf den Items (MASD). Potentielles Anwendungsgebiet ist die Skalierung von intraindividueller Variabilität im Rahmen von Ambulatory Assessments. Im Theorieteil werden einige Theorien zur intraindividuellen Variabilität dargelegt und der Modellierungshintergrund von Rasch zur Erzeugung von probabilistischen Testmodellen und die Maximum-Entropie-Methode zur Modellgenerierung werden gegenübergestellt. Es wird gezeigt, wie gängige Testmodelle, wie das dichotome Rasch-Modell, das Partial-Credit-Modell und das bedingte Rasch-Modell aus der Anwendung der Methode resultieren. Im Teil zur Modellentwicklung wird die Maximum-Entropie-Methode angewendet, um ein neues IRT-Modell zur Erfassung intraindividueller Variabilität auf Basis sukzessiver, absoluter Differenzen in Zeitreihen zu

generieren. Es resultiert ein Modell, dass einen Markov-Prozess erster Ordnung abbildet, wobei die manifeste Variabilität durch einen personenbezogenen, latenten Parameter bestimmt wird. Die Eigenschaften des Modells werden untersucht, indem die vorhergesagten Wahrscheinlichkeiten, die Item-Response-Kurven und die Likelihood-Funktion dargestellt werden. Anhand der Likelihood-Funktion zeigt sich, dass das Modell suffiziente Statistiken zur Schätzung der Parameter besitzt. Um die Bewertung der Modellpassung auf der Basis von standardisierten Residuen zu ermöglichen, werden die Erwartungswerte und die Varianz der manifesten Variable unter dem Modell hergeleitet. Die Eigenschaften der latenten Trait-Skala des Modells werden anhand der Bildung der Logits der Kategorien-Wahrscheinlichkeiten untersucht. Es zeigt sich, dass das Modell auf einer Differenzen-Skala misst. Der Zusammenhang zwischen der manifesten Variable und der Trait-Skala sowie der empirische Bias und die empirische Varianz der MCMC-Parameterschätzer werden simulativ untersucht. Es zeigt sich, dass die latente Trait-Skala in einem monotonen Verhältnis zu der Variabilität in den manifesten Zeitreihen steht. In nicht-extremen Regionen der Trait-Skala ist der Bias relativ gering und die Varianz der Schätzer steigt in Extrembereichen der Trait-Skala. Die Ergebnisse der simulativen Untersuchung des Bias und der Varianz sind mit gängigen Ergebnissen zu Rasch-Modellen kompatibel. Auf Basis der Ergebnisse der Modellentwicklung wird geschlossen, dass es sich bei dem generierten Modell um ein Rasch-Modell handelt und die Maximum-Entropie-Methode verwendet werden kann, um neue, probabilistische, psychometrische Modelle zu generieren. Zur Bewertung der Reliabilität wird die Andrich-Reliabilität vorgeschlagen. Im Anwendungteil der Arbeit wird das Modell auf Daten angewendet, die im Rahmen eines Ambulatory Assessments zur Affektregulation von Crayen et al. (in Druck) mit einer Kurzform des Mehrdimensionalen Befindlichkeitsfragebogens (Steyer et al., 1997) angefallen sind. Es wird überprüft, ob das Modell zur Bewertung intraindividueller Variabilität auf reale Daten angewendet werden kann. Anhand der standardisierten Residuen des Modells zeig sich, dass das Modell relativ gut passt. Die Andrich-Reliabilität für alle drei Skalen ist relativ hoch, was darauf hindeutet, dass die intraindividuelle Variabilität auf den drei Skalen des MDBF reliabel erfasst wird. Die individuellen Variabilitäts-Parameter der

drei Skalen sind hoch miteinander korrelieren, ein Ergebnis, dass mit demjenigen von Eid und Diener (Eid & Diener, 1999) zur Affekt-Variabilität kompatibel ist. Die Korrelationen der Variabilitäts-Parameter mit der manifesten Variabilität ist hoch, was darauf hindeutet, dass die Parameter die manifeste Variabilität in den Zeitreihen abbilden. Die Korrelation der Variabilitäts-Skalen und der manifesten Indikatoren der Variabilität mit drei Skalen des NEO-FFI (Neurotizismus, Extraversion und Gewissenhaftigkeit) zeigte jedoch keine deutlichen, signifikanten Zusammenhänge, was darauf hindeutet, dass die Variabilität der Skalen des MDBF in der Stichprobe etwas spezifisch anderes erfasst als die drei Skalen des NEO-FFI.

A.7. Summary

The present thesis focuses on the development of an item response model for the assessment of intraindividual variability in multivariate, discrete time series based on the mean absolute successive difference of the time-series. A potential application of the model is the scaling of intraindividual variability in ambulatory assessments. In the theoretical part some current theories with regard to intraindividual variability are presented and the IRT-modeling framework by Rasch and the maximum entropy method for model construction are discussed. It is shown that well known IRT-models, such as the dichotomous Rasch model, the partial credit model and the conditional Rasch model can be obtained by applying the maximum entropy modeling framework. Then the maximum entropy modeling framework is applied to the problem of generating a new psychometric model for the assessment of intraindividual variability based on absolute successive differences in multivariate, discrete time series. The resulting model describes a first order markov process that is governed by a latent trait parameter. The properties of the resulting model are described by examining the predicted category response probabilities, the item characteristic curves and the likelihood function. An detailed examination of the likelihood function reveals that the model features sufficient statistics for parameter estimation. To allow for the assessment of model fit by standardized residuals, the expectations and the variance of the expected responses under the model are derived. To clarify the measure-

ment structure of the model, the logits of adjacent response categories are examined. It is found that the model captures intraindividual variability on a difference scale. The relationship between the latent trait scale and the manifest variability of the responses are examined by simulation. The simulation showed a monotonous relationship between the variability of the manifest variables and the latent trait scale. A simulative examination of the MCMC-estimators' empirical bias and variance indicated that bias is relatively small in non extreme regions of the trait scale. The variance of the estimators increases in extreme regions of the trait scale. These results are in accordance with known results for Rasch models. It is concluded that the generated model is a Rasch model and that the maximum entropy framework can be used to generate new probabilistic psychometric models. It is proposed to assess measurement reliability by Andrich's method and to examine model fit based on standardized residuals. The model's practical applicability is examined based on a longitudinal ambulatory assessment data set for the study of affect regulation by Crayen et al. (in press). The generated model is applied to a short form of the multidimensional mood questionnaire (MDBF) (Steyer et al., 1997) with the aim to check if the model could be used to measure mood variability in ambulatory assessment studies. The model shows overall good fit based on standardized residuals and good reliability of measurement based on Andrich's reliability coefficient. It is concluded that the model is capable of reliably differentiating individuals with regard to intraindividual variability on a latent scale. The three sub-scales of the MDBF were highly intercorrelated. To get first hints with regard to validity, the variability parameters were correlated with three scales of the NEO-FFI (neuroticism, extraversion and conscientiousness). The model's latent variability parameters were highly intercorrelated with the manifest variability in the time series. However, there were no pronounced correlations between the variability scales and the three avaliable scales of the NEO-FFI. The same counts for the manifest variability in the time series.

A.8. Kurzgefasster Lebenslauf

Der Lebenslauf ist aus Gründen des Datenschutzes in der Online-Version nicht enthalten.

Der Lebenslauf ist aus Gründen des Datenschutzes in der Online-Version nicht enthalten.

Der Lebenslauf ist aus Gründen des Datenschutzes in der Online-Version nicht enthalten.

A.9. Erklärung über die verwendeten Hilfsmittel

Hiermit erkläre ich, die Dissertation

Ein probabilistisches Testmodell zur Erfassung intraindividueller Variabilität

vollständig selbständig angefertigt zu haben. Sämtliche zu ihrer Erstellung verwendeten Hilfsmittel und Hilfen sind angegeben. Die Arbeit wurde in noch keinem früheren Promotionsverfahren angenommen oder abgelehnt.

Georg Hosoya