- Ludwig-Maximilians-Universität München -
Institut für Statistik

Methoden der Veränderungsmessung basierend auf dem rasch-Modell

Bachelorarbeit

zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)

Vorgelegt von: Silke Janitza

Erstgutachter: Prof. Dr. Helmut Küchenhoff

München, den 14. Juli 2010

Abstract

In der vorliegenden Arbeit werden verschiedene Modelle zur Messung gruppenspezifischer und individuenspezifischer Veränderungen in der Psychometrie vorgestellt. Alle vorgestellten Modelle basieren auf dem dichotomen Rasch-Modell, welches die objektive Messung latenter Eigenschaften ermöglicht. Ausgehend von diesem Modell werden das Linear Logistic Test Model (LLTM), das Linear Logistic Test Model with Relaxed Assumptions (LLRA), das Hybrid LLRA (HLLRA) und das Multidimensional Rasch Model for Learning and Change (MRMLC) vorgestellt, die durch Hinzunahme einer oder mehrerer Parameter Veränderung beschreiben. Die ersten drei genannten Modelle ermöglichen die Schätzung von gruppenspezifischer Veränderung. Mittels Likelihood-Quotiententest ist es möglich, Hypothesen über die Veränderung einzelner Gruppen zu testen und damit Aussagen über die homogene bzw. heterogene Entwicklung von Gruppen zu machen. Das MRMLC unterscheidet sich von diesen Modellen in der Hinsicht, dass die Veränderung individuenspezifisch erfasst wird. Dies kann von Bedeutung sein, wenn die Veränderung einzelner Personen von Interesse ist. Anschließend werden Möglichkeiten aufgezeigt, verschiedene Veränderungsmuster innerhalb einer Gruppe von Personen anhand von Mixed Rasch Models (MRM) zu identifizieren. Dabei werden latente Gruppen ausfindig gemacht, innerhalb derer die Veränderung homogen ist, zwischen den Gruppen allerdings Unterschiede bezüglich der Veränderung bestehen. Abschließend werden weitere Ansätze aufgezeigt, wie Veränderung erfasst werden kann, die jedoch aufgrund ihrer komplexen Schätzung und begrenzten Hypothesentestung in der Praxis kaum Verwendung finden.

Inhaltsverzeichnis

Begriffserklärung

Bezeichnung	**Bedeutung**
Laufindizes:	
$v = 1, \cdots, N$	Personen
$i = 1, \cdots, I$	Aufgaben/Items
$t = 1, \cdots, T$	Zeitpunkte
$g = 1, \cdots, G$	Gruppen
Modellparameter:	
θ	Personenparameter
β	Aufgaben-/Itemparameter
δ	Veränderungsparameter
τ	Trend
$x_{vit} \in (1, 0)$	Antwort der Person v auf Item i zu Zeitpunkt t
s_{vt}	Gesamtpunktzahl der Person v zu Zeitpunkt t
π_g	Anteil der Subgruppe g in der Stichprobe

Bezeichnung	**Bedeutung**	**Modelle**
θ_v	Eigenschaft der Person v	RM, LLTM
θ_{vi}	Eigenschaft der Person v bzgl. latenter Eigenschaft i	LLRA, HLLRA
θ_{vt} θ_{v1} $\theta_{v2}, \cdots, \theta_{vT}$ $\sum_{m=1}^{t} \theta_{vm}$	Fähigkeiten der Person v: Anfangsfähigkeit der Person v Modifizierbarkeiten der Person v Effektive Fähigkeit der Person v zu Zeitpunkt t	MRMLC
$\theta_{v\|g}$	Eigenschaft der Person v aus Gruppe g	MRM
β_i	Schwierigkeit des Items i	RM, LLTM, MRMLC
β_{it}	Schwierigkeit eines Hybrid-Items it	HLLRA
$\beta_{i\|g}$	Schwierigkeit des Items i in Gruppe g	MRM
δ_g	Veränderung in Gruppe g	LLRA
δ_{gt}	Veränderung in Gruppe g zum Zeitpunkt t	LLTM, HLLRA
$\delta_{t\|g}$	Veränderung zum Zeitpunkt t in Gruppe g	MRM

1 Einleitung

Die Psychometrie beschäftigt sich als Teildisziplin der Psychologie mit der Theorie und Methode des psychologischen Messens. Ziel der Psychometrie ist die Entwicklung neuer Lösungswege sowie die Verbesserung bestehender Ansätze. In vielen Bereichen der Psychometrie wird versucht die Veränderung bestimmter Fähigkeiten oder Eigenschaften zu erfassen. So kann beispielsweise die Effektivität einer neuen Unterrichtsmethodik bei Schülern von Interesse sein oder auch die Wirksamkeit einer neu entwickelten Therapie.

Zumeist handelt es sich in der Psychometrie um latente Eigenschaften, die der Veränderung unterliegen. Diese sind nicht direkt messbar. So ist es beispielsweise nicht möglich, die schulische Leistungsfähigkeit direkt zu erfassen. Ebenso wenig können Symptome eines Patienten gemessen werden, die sich äußerlich nicht ausdrücken, wie Schmerz und körperliche Beschwerden. Somit ist auch die direkte Messung der Veränderung in der latenten Eigenschaft, die beispielsweise durch eine Therapie oder auch durch eine Unterrichtsmethodik erfolgt, nicht möglich.
Latente Eigenschaften werden hauptsächlich durch Fragebögen oder Tests erfasst und sind mit einem auf individueller Ebene unvermeidlichen Messfehler behaftet. Der Gesamtpunktwert, der in einem Test erreicht wird, kann als Schätzwert der latenten Fähigkeit dienen. Dieser Ansatz ist jedoch problematisch, unter anderem deshalb, weil eine lineare Beziehung zwischen der latenten Fähigkeit und der Gesamtpunktzahl unterstellt wird. Dies impliziert die Annahme, dass latente Fähigkeiten nach oben und nach unten hin begrenzt sind.

Ein Modell, das speziell für die Erfassung latenter Eigenschaften verwendet wird, ist das Rasch-Modell. Es beschreibt einen log-linearen Zusammenhang zwischen latenter Fähigkeit und Punktwert, der sich für die Messung von latenten Fähigkeiten als geeigneter erweist (Abbildung 2). Latente Fähigkeiten werden in diesem Modell nicht durch eine obere und eine untere Grenze eingeschränkt und daraus resultierende Ceiling- und Bottom-Effekte können vermieden werden.
Zudem ermöglicht das Modell durch Einbeziehung von Aufgabenschwierigkeiten eine objektive Messung der latenten Fähigkeit, die unabhängig vom Messinstrument ist. Eine Vergleichbarkeit zwischen Personen, denen unterschiedliche Aufgaben vorgelegt werden, ist damit möglich.

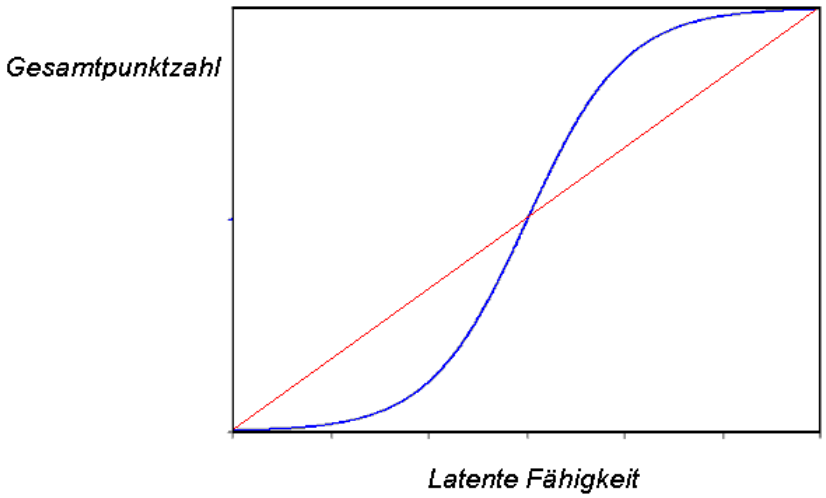

Abbildung 1: linearer (rot) und log-linearer (blau) Zusammenhang zwischen Punktwert und Fähigkeit

Diese Arbeit beschäftigt sich mit Modellen zur Messung von Veränderung, die auf dem Rasch-Modell aufbauen. Dabei wird die Veränderung zwischen zwei oder mehreren Zeitpunkten für Untersuchungseinheiten gemessen. Untersuchungseinheiten können sowohl einzelne Individuen als auch Gruppen von Personen sein.
Wird beispielsweise eine Therapie hinsichtlich ihrer Wirksamkeit bewertet, so ist die Veränderung in der gesamten Versuchsgruppe, die die Therapie erhält, von Interesse. Indem den Personen zu mindestens zwei Zeitpunkten ein Fragebogen mit bestimmten Symptomen vorgelegt wird, kann eine mögliche Verbesserung durch die Therapie ausgemacht werden. Bei der Erfassung einer 'globalen' Veränderung für die Versuchsgruppe könnte es passieren, dass sich entgegengesetzte Auswirkungen der Therapie in der Versuchsgruppe überlagern. Dies könnte im Extremfall sogar zu einer Aufhebung des Therapieeffektes führen. Es werden ebenfalls Methoden vorgestellt, mit denen es möglich ist solche heterogene Veränderungen innerhalb einer Gruppe von Personen zu identifizieren.

Bei der Entwicklung einer neuen Unterrichtsmethodik könnten Leistungstests zu zwei Testzeitpunkten Aufschluss über den Effekt der neuen Methode geben. Falls es sich um dieselben Aufgaben zu den beiden Testzeitpunkten handelt, ist allerdings zu bedenken, dass Erinnerungseffekte auftreten könnten. Auch dieser Fall findet in einigen Modellen Berücksichtigung, die im Rahmen dieser Arbeit behandelt werden sollen.

Bevor verschiedene Modelle zur Messung von Veränderung aufgezeigt werden, soll im folgenden Unterkapitel kurz das Rasch-Modell vorgestellt werden, auf dem die Modelle für die Veränderungsmessung basieren. Neben den wesentlichen Eigenschaften des Rasch-Modells werden drei gängige Methoden zur Parameterschätzung angeführt.

1.1 Das Rasch-Modell

Das Rasch-Modell wurde zwar nicht zur Messung von Veränderungen konzipiert, bildet jedoch die Basis vieler Modelle, durch die Veränderung erfasst werden kann. Aufgrund seiner nützlichen Eigenschaften gewinnt es in der Praxis bei der Messung latenter Eigenschaften immer mehr an Bedeutung. Einer der bedeutendsten Vorteile dieses Modells ist, dass es objektive Messungen einer latenten Eigenschaft erlaubt, indem sowohl die Fähigkeit einer Person als auch die Schwierigkeit einer Aufgabe Berücksichtigung finden. Eine anschauliche Einführung zum Rasch-Modell ist in der Literatur von Strobl (2010) zu finden.

Die Modellgleichung für das Rasch-Modell lautet:

$$P(X_{vi} = 1|v, i) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)} \tag{1}$$

Die latente Eigenschaft einer Person v wird dabei mit dem Parameter θ_v gemessen, der Parameter β_i misst die Schwierigkeit für die Aufgabe i. Der Ausdruck $P(X_{vi} = 1)$ gibt die Wahrscheinlichkeit für eine Person v an, Aufgabe i richtig zu lösen.

Im Folgenden sollen weitere bedeutende Eigenschaften des Rasch-Modells neben der Trennung von item- und personenspezifischer Parameter sowie die Schätzung der Parameter, aufgeführt werden.

1.1.1 Eigenschaften

Die fünf grundlegenden Eigenschaften des Rasch-Modells sollen nun kurz erläutert werden.

Suffiziente Statistiken
Die Gesamtanzahl positiv beantworteter Fragen ist eine suffiziente Statistik für den Personenparameter. Die Schätzung der Personenfähigkeit hängt also nur davon ab, wie viele

Aufgaben die Person in einem Test positiv beantwortet hat, nicht jedoch, welche.

Lokale Stochastische Unabhängigkeit
Die lokale stochastische Unabhängigkeit besagt, dass sich die gemeinsame Wahrscheinlichkeit mehrere Aufgaben positiv zu beantworten bei fester Personenfähigkeit aus dem Produkt der Einzelwahrscheinlichkeiten ergibt. Das bedeutet inhaltlich, dass es nicht zulässig ist, dass sich die Wahrscheinlichkeit eine Aufgabe positiv zu beantworten verändert, je nachdem ob eine andere Aufgabe positiv oder negativ beantwortet wurde. Die lokale stochastische Unabhängigkeit ist zum Beispiel verletzt, wenn Aufgaben aufeinander aufbauen. Diese Eigenschaft ist für die Parameterschätzung mittels der Maximum-Likelihood-Methode unabdingbar, auf die später eingegangen werden soll.

Spezifische Objektivität
Nach dem Grundsatz der spezifischen Objektivität ist die Schätzung personenspezifischer Parameter unabhängig von den itemspezifischen Parametern und vice versa. Das bedeutet, dass es für den Vergleich der Fähigkeiten von Personen gleichgültig ist, welche Aufgabe die Personen beantwortet haben. Umgekehrt können Aufgaben hinsichtlich ihrer Schwierigkeit miteinander verglichen werden unabhängig davon, welche Person die Aufgaben gelöst hat.

Eindimensionalität
Die geschätzte latente Fähigkeit und die Aufgabenschwierigkeit können auf einer gemeinsamen Skala abgebildet werden. Inhaltlich bedeutet dies, dass der Test nur den Bereich umfasst, der gemessen werden soll. Soll beispielsweise die mathematische Fähigkeit deutscher Schüler untersucht werden, wäre ein Mathematik-Test, der in englischer Sprache gestellt wird, nicht Rasch-konform, da eine weitere Fähigkeit - die englischen Sprachkenntnisse der Schüler - im Test erfasst wird.

Intervallskalenniveau
Item- und personenspezifische Parameter sind intervallskaliert. Damit ist es möglich Vergleiche zwischen Personen und auch zwischen Aufgaben aufzustellen. Eine Absolutskalenmetrik für die Parameter kann allerdings kaum gerechtfertigt werden, da *latente* Eigenschaften gemessen werden sollen, die nicht unmittelbar messbar sind.

Das Rasch-Modell gilt als ein recht strenges Modell, da es relativ hohe Anforderungen an die Beschaffenheit eines Tests stellt. Diese Anforderungen sind in der Praxis jedoch meist eher schwer zu verwirklichen. Ein bedeutender Aspekt ist, dass Testaufgaben auf Rasch-Konformität hin überprüft werden und so ungeeignete Testaufgaben eliminiert werden können. Dies trägt wesentlich zur Vermeidung von Fehlinterpretationen bei. Eine Übersicht über verschiedene Testverfahren findet sich in der Literatur von Strobl (2010) (Kapitel 4).

1.1.2 Parameterschätzung

Es gibt verschiedene Ansätze zur Schätzung der personenspezifischen und itemspezifischen Parameter. Im Folgenden werden drei Schätzverfahren vorgestellt, die zu den bekanntesten gehören und auch für die Schätzung bei Modellen der Veränderungsmessung angewendet werden.

Gemeinsame Maximum-Likelihood-Schätzung (UML)
Dieses Verfahren wird aufgrund resultierender inkonsistenter Schätzer in der Praxis eher seltener verwendetet. Von den gängigen Schätzverfahren ist es vermutlich das anschaulichste. Hierbei wird die gemeinsame Likelihood über alle Aufgaben und alle Personen gebildet. Aufgrund der Annahme der lokalen stochastischen Unabhängigkeit kann die gemeinsame Likelihood über das Produkt der einzelnen Likelihoods gemäß (1) über alle I Aufgaben und über alle N Personen gebildet werden:

$$L(\theta, \beta) = \prod_{v=1}^{N} \prod_{i=1}^{I} \frac{\exp\left((\theta_v - \beta_i) \cdot x_{vi}\right)}{1 + \exp(\theta_v - \beta_i)} \tag{2}$$

Der Parameter θ_v bezeichnet dabei die Personenfähigkeit für Person v, β_i die Aufgabenschwierigkeit für Aufgabe i. x_{vi} nimmt den Wert 1 an, falls Person v Aufgabe i richtig gelöst hat, ansonsten den Wert 0.

Durch Maximierung der Likelihood erhält man gleichzeitig Schätzungen für personen- und für itemspezifische Parameter. Diese Schätzer sind allerdings nicht konsistent, da sich bei Erhöhung der Stichprobenanzahl N die Anzahl zusätzlich zu schätzender Parameter ebenfalls erhöht.

Bedingte Maximum-Likelihood-Schätzung (CML)

Als bessere Schätzmethode erweist sich die bedingte Maximum-Likelihood-Schätzung. Hier werden personen- und itemspezifische Parameter unabhängig voneinander geschätzt. Unter Bedingung auf die Summe aller positiv beantworteten Items einer Person als suffiziente Statistik für den personenspezifischen Parameter, kann die bedingte Likelihood aufgestellt werden, die nur noch von den itemspezifischen Parametern abhängt.

Die Parameterschätzung erfolgt dann in zwei Schritten:
In einem ersten Schritt werden die itemspezifischen Parameter iterativ geschätzt, indem die bedingte Likelihood abgeleitet und Null gesetzt wird. Meist erfolgt die iterative Schätzung mittels *Newton-Raphson*-Verfahren. Durch die Eliminierung der personenspezifischen Parameter aus der (bedingten) Likelihood ist es möglich die itemspezifischen Parameter unabhängig von den personenspezifischen Parametern zu schätzen.
Im zweiten Schritt werden diese Schätzungen in die 'gewöhnliche' Likelihood (2) eingesetzt um die personenspezifischen Parameter zu schätzen.

Marginale Maximum-Likelihood-Schätzung (MML)

Auch bei der marginalen Maximum-Likelihood-Schätzung erfolgt die Schätzung von personen- und itemspezifischen Parametern in zwei Schritten. Für die Schätzung mittels MML müssen jedoch Angaben über die Verteilung der personenspezifischen Parameter vorliegen. Bei Fehlspezifikation der Verteilung kann die Schätzung fehlerhaft sein.

Es werden wie bei der CML zuerst die itemspezifischen Parameter geschätzt, und zwar indem über die personenspezifischen Parameter integriert wird, sodass die marginale Likelihood nur noch von den itemspezifischen Parametern abhängt. Für die marginale Likelihood L_v einer Person v mit Antwortvektor $x_v = (x_{v1}, \cdots, x_{vI})'$ ergibt sich:

$$L_v(\beta) = \int P(x_v|\theta, \beta) f(\theta) \; \partial\theta \tag{3}$$

$f(\theta)$ bezeichnet die Verteilung der Personenfähigkeiten. Häufig wird von einer Normalverteilung der Fähigkeiten ausgegangen. Die marginale Likelihood $L(\beta)$ für alle Personen ergibt sich aus dem Produkt der marginalen Likelihoods aus (3) über alle $v = 1, \cdots, N$ Personen. Wie gewöhnlich erhält man nun die itemspezifischen Parameter indem man die gemeinsame marginale Likelihood $L(\beta)$ durch Ableiten und Null setzen maximiert.
In einem zweiten Schritt werden die Schätzungen für die itemspezifischen Parameter in die ursprüngliche Likelihood (2) eingesetzt um die personenspezifischen Parameter zu schätzen.

1.1.3 Das Rasch-Modell in der Veränderungsmessung

Es stellt sich die Frage, wie nun das Rasch-Modell zur Messung von Veränderung verwendet werden kann. Um eine Messung von Veränderung zu ermöglichen, müssen Daten zu mindestens zwei Testzeitpunkten vorliegen. Die Datenmatrix, wie sie für die Messung von Veränderung notwendig ist, nimmt also eine dreidimensionale Struktur an (Abbildung 2).

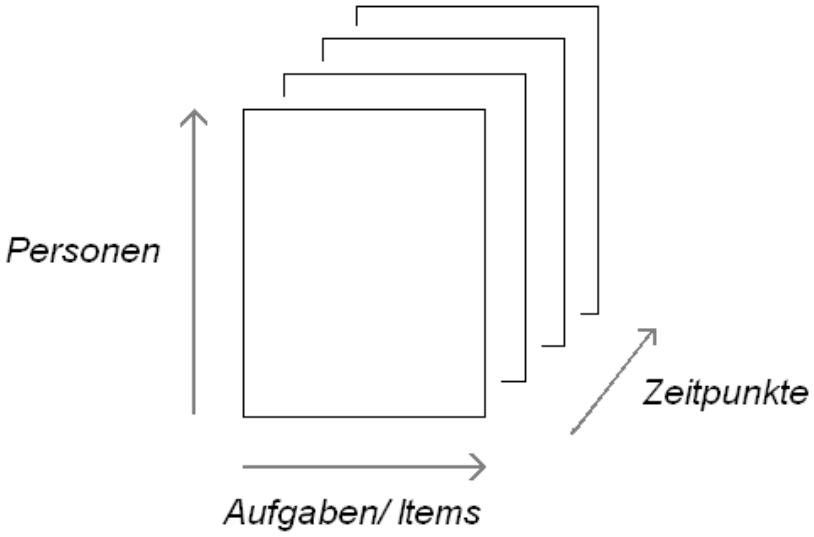

Abbildung 2: Dreidimensionale Datenstruktur

Das Rasch-Modell hingegen geht von einer zweidimensionalen Datenmatrix aus: Eine Dimension für die Personen und eine weitere für die Aufgaben, die den Personen gestellt werden. Die zusätzliche Dimension der Zeitpunkte, die die Messung von Veränderung erst ermöglicht, ist nicht Bestandteil des Konzepts des Rasch-Modells.

Natürlich wäre es möglich, für eine Person die latente Fähigkeit zu jedem Testzeitpunkt zu schätzen und die Veränderung der Person durch die Differenz ihrer latenten Fähigkeiten zu den Testzeitpunkten zu berechnen. Dieses Vorgehen hat allerdings den Nachteil, dass die Veränderung als Parameter nicht Bestandteil im Rasch-Modell ist. Eine Schätzung der Veränderung ist damit nur über den Umweg der Differenzbildung möglich.

Die Modelle zur Veränderungsmessung, die im Rahmen dieser Arbeit vorgestellt werden, haben den Vorteil, dass die Veränderung Bestandteil des Modells ist. Damit sind die direkte Schätzung der Veränderung sowie Tests auf Signifikanz des Veränderungsparameters möglich. Diese Modelle sind durch Einführung eines zusätzlichen Parameters für die Veränderung auf die dreidimensionale Datenstruktur angepasst und stellen somit eine bessere Möglichkeit dar, Veränderung zu messen.

In Kapitel 2 werden vier bekannte Modelle zur Messung von Veränderung vorgestellt. Die ersten drei Modelle, die hauptsächlich auf den Überlegungen von Fischer (1976, 1977, 1983, 1989, 1995) basieren, können zur Messung von gruppenspezifischer Veränderung herangezogen werden. Als viertes Modell wird ein Modell von Embretson (1991a,b) angeführt, welches Aussagen über Veränderung auf individuenspezifischer Ebene ermöglicht. In Kapitel 3 wird die Anwendung von Mixed Rasch Models (MRM) beschrieben, die die Möglichkeit zur explorativen Suche nach Gruppen mit unterschiedlichen Veränderungen bietet. Zugleich kann dieses auch zur Überprüfung von Modellannahmen im Linear Logistic Test Model (LLTM) dienen. Abschließend werden in Kapitel 4 weitere Modellansätze zur Messung von Veränderung vorgestellt.

2 Modelle zur Messung von Veränderung

Das Rasch-Modell selbst ist weniger dazu geeignet Veränderung zu messen. In diesem Kapitel sollen Modelle zur Veränderungsmessung vorgestellt werden, die auf dem Konzept des Rasch-Modells aufbauen und die Vorteile nutzen, die dieses mit sich bringt. Neben allgemeinen Erläuterungen dieser Modelle sollen Vor- und Nachteile diskutiert sowie situationsbedingte Anwendungsmöglichkeiten aufgezeigt werden.

Zu Beginn wird das Linear Logistic Test Model (LLTM) vorgestellt, das zwar nicht primär für die Messung von Veränderung konzipiert wurde, sich jedoch als durchaus geeignet dafür erweist. Im darauf folgenden Unterkapitel wird das Linear Logistic Test Modell with Relaxed Assumptions (LLRA) aufgeführt, welches sich direkt von dem LLTM ableiten lässt. Das Hybrid LLRA (HLLRA), das im Anschluss daran aufgezeigt wird, stellt eine Optimierung des LLRA dar. Schließlich wird mit dem Multidimensional Rasch Model for Learning and Change (MRMLC) ein Modell aufgezeigt, welches eine individuenspezifische Messung von Veränderung ermöglicht.

2.1 Das Linear Logistic Test Model (LLTM)

Das LLTM ist eine multivariate Erweiterung des Rasch-Modells. Folgende Erläuterungen zum LLTM basieren auf der Literatur von Fischer (1983, 1995). Die Veränderung wird durch Einführung von einem zusätzlichen (Veränderungs-)Parameter in das Rasch-Modell in (1) beschrieben. Dabei werden jeder Gruppe und jedem Zeitpunkt ein eigener Aufgabenschwierigkeitsparameter zugrunde gelegt. Die Modellformel des LLTM zur Messung von Veränderung nimmt dann folgende Form an:

$$P(X_{vit} = 1|v, i, t) = \frac{\exp(\theta_v - \beta_i + \delta_{gt})}{1 + \exp(\theta_v - \beta_i + \delta_{gt})} \tag{4}$$

Die Fähigkeit einer Person wird, wie beim Rasch-Modell, mit θ_v bezeichnet. Diese wird über alle Zeitpunkte hinweg als konstant angenommen. Die beiden Parameter β_i und δ_{gt} werden als Schwierigkeit bzw. als Schwierigkeitszuwachs einer Aufgabe angesehen. Mit der Restriktion $\delta_{g1} = 0$ ist zum Zeitpunkt des ersten Tests die Schwierigkeit einer Aufgabe für

alle Gruppen dieselbe[1]. Der gruppenspezifische Schwierigkeitszuwachs ist gleichbedeutend mit der Veränderung in der Gruppe. Ist eine Aufgabe für eine Gruppe zu einem späteren Zeitpunkt einfacher, kann man von einem Lernzuwachs der Gruppe ausgehen. δ_{gt} kann in beliebig viele Effektparameter und deren Wechselwirkungen zerlegt werden.

Modellannahmen

Bevor das LLTM angewendet werden kann, müssen einige Annahmen getroffen werden. Da das LLTM eine Verallgemeinerung des zweiparametrigen Rasch-Modells ist, werden im LLTM dieselben Annahmen wie im Rasch-Modell getroffen. Das schließt die Annahme der lokalen stochastischen Unabhängigkeit ein, sowie die Annahme, dass alle Aufgaben dieselbe latente Fähigkeit messen. Es muss also Rasch-homogenes Testmaterial vorliegen.
Darüber hinaus müssen weitere Annahmen getroffen werden, die sich aus der Hinzunahme des Veränderungsparameters in das Modell ergeben. Der Veränderungsparameter ist nicht individuenspezifisch, sondern gruppenspezifisch. Das bedeutet, dass im LLTM die Annahme getroffen wird, dass alle Personen in einer Gruppe einer gleich starken Veränderung unterliegen. Zudem ist der Veränderungsparameter nicht aufgabenspezifisch. Das LLTM nimmt also an, dass alle Aufgaben gleich sensibel in Bezug auf Veränderung sind. Durch Aufstellung eines quasisaturierten Modells ist es möglich diese Annahme auf ihre Gültigkeit hin zu überprüfen (siehe Seite 11 Hypothesentestung).

Parameterschätzung

Die Parameter im LLTM werden mittels der bedingten Maximum-Likelihood (CML) geschätzt. Ausgangspunkt der Schätzung ist die Modellformel in (4). Um eine konsistente Schätzung der gruppen- bzw. zeitpunktspezifischen Veränderungsparameter δ_{gt} zu ermöglichen, ist es notwendig die personenspezifischen Parameter θ_v aus der Likelihood zu eliminieren. Dabei wird wie beim Rasch-Modell ausgenutzt, dass der Gesamtpunktwert einer Person über alle Aufgaben und Testzeitpunkte suffizient für ihren Fähigkeitsparameter ist und somit der Fähigkeitsparameter aus der Likelihood herausgerechnet werden kann. Zur Bildung der bedingten Likelihood L_{C_v} für eine Person v wird die Annahme der Unabhängigkeit der Lösungswahrscheinlichkeiten der Aufgaben innerhalb sowie zwischen den Testzeitpunkten ausgenutzt.

[1] Die Voraussetzung von Rasch-homogenem Testmaterial im LLTM impliziert, dass es zum Anfangszeitpunkt keine Schwierigkeitsunterschiede zwischen den Gruppen gibt.

Die gemeinsame bedingte Likelihood L_C ergibt sich nun unter der Annahme, dass die Lösungswahrscheinlichkeiten der Aufgaben für die Personen unabhängig voneinander sind, aus dem Produkt der L_{C_v} über alle Personen $v = 1, \cdots, N$. Diese enthält sowohl die Aufgabenparameter β_i mit $i = 1, \cdots, I$ als auch die gruppen- und zeitpunktspezifischen Veränderungsparameter δ_{gt} mit $g = 1, \cdots, G$ und $t = 1, \cdots, T$. Durch iterative Verfahren (beispielsweise *Newton-Raphson*) werden die Aufgabenparameter und die Veränderungsparameter so gewählt, dass die bedingte Likelihood L_C maximiert wird.

Für eine effiziente Schätzung der gruppen- und zeitpunktspezifischen Veränderungsparameter ist ein ausreichend großer Stichprobenumfang in jeder Gruppe erforderlich. Das LLTM sieht demnach keine Messung von Veränderung für einzelne Individuen vor.

Hypothesentestung

In vielen Anwendungen ist von Interesse, inwiefern Veränderungen einer latenten Eigenschaft lediglich durch Zufallsstreuung in der Stichprobe zustande kommen. So würde es keinen Sinn machen, eine neu entwickelte Therapie einzuführen, wenn die Stichprobe der Probanden so beschaffen ist, dass die Personen sehr gut auf die Therapie ansprechen, die Therapie in der Bevölkerung aber eher weniger gut anschlägt. Es ist daher notwendig die Effekte auf Signifikanz hin zu überprüfen.

Die Hypothesentestung kann mittels bedingtem Likelihood-Quotiententest (LRT) erfolgen (Andersen (1973)). Dabei bezeichnet L_{C_0} die bedingte Likelihood unter der Nullhypothese, L_{C_1} bezeichnet die bedingte Likelihood unter der Alternativhypothese. Die Anzahl der zu überprüfenden Parameter ergibt sich aus der Differenz der Parameteranzahl p_1 des vollen Modells und der Parameteranzahl p_0 des restringierten Modells. Die Teststatistik ist asymptotisch χ^2-verteilt mit $p_1 - p_0$ Freiheitsgraden.

$$T = -2(\ln L_0 - \ln L_1) \overset{as.}{\sim} \chi^2_{(p_1 - p_0)}$$

Neben der Überprüfung auf eine signifikante Abweichung einzelner Effekte von Null können weitere Hypothesen aufgestellt und geprüft werden. Folgende Tabelle zeigt einige typische Hypothesen, die mittels LRT getestet werden können.

Hypothese	Bedeutung am Beispiel für zwei Effekte δ_{1t}, δ_{2t} und Trendeffekt τ_t zum Zeitpunkt t
keine Veränderung	$\delta_{1t} = \delta_{2t} = \tau_t = 0$
kein Trend	$\tau_t = 0$
keine Interaktionseffekte	$\rho_t = 0$
ineffektive Effekte	$\delta_{1t} = 0$ und/oder $\delta_{2t} = 0$
Gleichheit von Effekten	$\delta_{1t} = \delta_{2t}$

Tabelle 1: Typische Hypothesen

Auch ist es möglich die Annahme der itemunabhängigen Veränderung im LLTM mittels LRT zu überprüfen. Dies geschieht durch Aufstellung eines quasisaturierten Modells. Das Wort 'quasi' wird vorangestellt, da man zur Anwendung des Modells bereits relativ strenge Annahmen (siehe Seite 10 Modellannahmen) trifft, die nur teilweise überprüfbar sind, und die es daher kritisch zu hinterfragen gilt. Das quasisaturierte Modell enthält die maximale Anzahl der zu schätzenden Parameter. Das bedeutet, dass es zu jedem Item für jeden Zeitpunkt und für jede Gruppe einen eigenen Schwierigkeitsparameter gibt.
Neben dem quasisaturierten Modell wird ein Modell gemäß dem LLTM aufgestellt, in dem die Veränderung zwar gruppen- und zeitpunktspezifisch, nicht aber itemspezifisch ist. Mittels Likelihood-Quotiententest (LRT) kann nun geprüft werden, inwieweit die gruppenspezifische Veränderung über alle Items hinweg dieselbe ist.

Anwendungsbeispiel

Im Folgenden soll eine neu entwickelten Therapie hinsichtlich ihrer Wirksamkeit evaluiert werden. Den Probanden wird zu zwei Zeitpunkten ein Fragebogen mit denselben klinischen Symptomen vorgelegt. Die Probanden sollen zu beiden Zeitpunkten angeben, ob sie diese Symptome haben oder nicht. Damit kann der Effekt der Therapie geschätzt werden.

Der Therapieeffekt sei im Folgenden mit δ bezeichnet. Um diesen von einem möglichen Trendeffekt, der sich unabhängig von der Therapie bei allen Personen ergibt, unterscheiden zu können, wird ein zusätzlicher Parameter für den Trend eingeführt, der mit τ bezeichnet wird. Um sowohl den Effekt für die Behandlung als auch den Effekt für den Trend schätzen zu können, wird eine Kontrollgruppe in den Versuchsplan mit aufgenommen, die die Therapie nicht erhält. Ansonsten wären Behandlungs- und Trendeffekt miteinander konfundiert,

also nicht eindeutig bestimmbar.

Es ergeben sich gemäß der Modellformel des LLTM in (4) folgende vier Modelle für Kontrollgruppe bzw. Versuchsgruppe zu zwei Zeitpunkten T_1 und T_2:

$$\text{Kontrollgruppe zu } T_1: \quad P(X_{vi1}=1|v,i,T_1)=\frac{\exp(\theta_v-\beta_i)}{1+\exp(\theta_v-\beta_i)} \tag{5}$$
$$(v=1,\cdots,n)$$

$$\text{Versuchsgruppe zu } T_1: \quad P(X_{vi1}=1|v,i,T_1)=\frac{\exp(\theta_v-\beta_i)}{1+\exp(\theta_v-\beta_i)} \tag{6}$$
$$(v=n+1,\cdots,N)$$

$$\text{Kontrollgruppe zu } T_2: \quad P(X_{vi2}=1|v,i,T_2)=\frac{\exp(\theta_v-\beta_i+\tau)}{1+\exp(\theta_v-\beta_i+\tau)} \tag{7}$$
$$(v=1,\cdots,n)$$

$$\text{Versuchsgruppe zu } T_2: \quad P(X_{vi2}=1|v,i,T_2)=\frac{\exp(\theta_v-\beta_i+\delta+\tau)}{1+\exp(\theta_v-\beta_i+\delta+\tau)} \tag{8}$$
$$(v=n+1,\cdots,N)$$

Mittels CML können sowohl der Therapieeffekt δ als auch der Trendeffekt τ geschätzt werden. Ob die Behandlung zu einer signifikanten Verbesserung führt, kann mittels LRT geprüft werden.
Bei der Überprüfung des Therapieeffektes wird ein Modell aufgestellt, in dem der Therapieeffekt auf Null gesetzt wird. Somit ist nach diesem Modell die Veränderung für Versuchs- und Kontrollgruppe gleich - nämlich lediglich bestimmt durch den Trend. Weist der LRT beim Vergleich dieses restringierten mit dem unrestringierten Modell ein signifikantes Ergebnis auf, so ist ein Effekt durch die Therapie nachweisbar. Ist das Ergebnis nicht signifikant, so kann davon ausgegangen werden, dass die Therapie keine Auswirkung hat.
Insbesondere dann ist es empfehlenswert mittels MRM (siehe Kapitel 3) abzusichern, dass es in der Versuchsgruppe keine heterogene Entwicklung gibt. Anderenfalls wäre es denkbar, dass die Therapie bei einem Teil der Versuchsgruppe sehr gut wirkt, bei dem anderen allerdings gegenteilige Effekte hervorruft und sich die Wirkungen somit ausgleichen.
Neben der Prüfung auf die Wirksamkeit der Therapie kann ebenfalls die Signifikanz des Trendeffektes mittels LRT geprüft werden. Dieser Test gibt Aufschluss darüber, ob eine signifikante Veränderung ohne Therapie stattfindet.

Es ist zu empfehlen, die Annahme der itemunabhängigen Veränderung mittels eines quasi-saturierten Modells zu überprüfen. Es kann beispielsweise vorkommen, dass in einem Fragebogen nach ein oder mehreren klinischen Symptomen gefragt wird, bei denen die Therapie keine Wirkung zeigt, bei anderen Symptomen allerdings zu einer sehr starken Verbesserung führt. Im LLTM könnte daher die Wirksamkeit der Therapie erheblich unterschätzt werden, da hier die Veränderung über alle Symptome gemeinsam hinweg betrachtet wird.

Dieses simple Beispiel kann um beliebig viele Effekte erweitert werden. Beispielsweise könnte sich die Frage stellen, ob die Therapie bei Männern anders anschlägt als bei Frauen. Oder etwa, ob eine zusätzliche Medikation zu einem größeren Erfolg führt und ob es Wechselwirkungen zwischen der Medikation und der Therapie gibt. Dieser Fall soll im Folgenden genauer betrachtet werden.

Um die Effekte von Therapie und Medikation, sowie deren Wechselwirkung eindeutig schätzen zu können, müssen Versuchs- und Kontrollgruppe nochmals unterteilt werden nach Medikation. Der Effekt der Therapie sei wie vorhin mit τ bezeichnet, der des Medikamentes mit γ. Die Wechselwirkung zwischen Therapie und Medikation wird mit ρ bezeichnet. Es ergeben sich gemäß (4) folgende Modelle:

Alle Gruppen zu T_1 : $(v = 1, \cdots, N)$	$P(X_{vi1} = 1 \mid v, i, T_1) = \dfrac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}$
Kontrollgruppe zu T_2 : $(v = 1, \cdots, n_1)$	$P(X_{vi2} = 1 \mid v, i, T_2) = \dfrac{\exp(\theta_v - \beta_i + \tau)}{1 + \exp(\theta_v - \beta_i + \tau)}$
Therapie-Gruppe zu T_2 : $(v = n_1 + 1, \cdots, n_2)$	$P(X_{vi2} = 1 \mid v, i, T_2) = \dfrac{\exp(\theta_v - \beta_i + \delta + \tau)}{1 + \exp(\theta_v - \beta_i + \delta + \tau)}$
Medikations-Gruppe zu T_2 : $(v = n_2 + 1, \cdots, n_3)$	$P(X_{vi2} = 1 \mid v, i, T_2) = \dfrac{\exp(\theta_v - \beta_i + \gamma + \tau)}{1 + \exp(\theta_v - \beta_i + \gamma + \tau)}$

Therapie-Medikations-Gruppe zu T_2 : $P(X_{vi2} = 1|v, i, T_2) = \dfrac{\exp(\theta_v - \beta_i + \delta + \gamma + \rho + \tau)}{1 + \exp(\theta_v - \beta_i + \delta + \gamma + \rho + \tau)}$

$(v = n_3 + 1, \cdots, N)$

Die Schätzung von Trend-, Therapie- und Medikationseffekt sowie die Wechselwirkung von Therapie und Medikation erfolgt wieder durch CML.
Daraufhin kann mittels LRT auf Signifikanz der Effekte geprüft werden. So kann unter anderem herausgefunden werden...

...ob Therapie bzw. Medikation zu einem Erfolg führt ($H_0 : \delta = 0$ bzw. $H_0 : \gamma = 0$)
...ob ihr Effekt gleich groß ist ($H_0 : \delta = \gamma$)
...und ob sich Therapie und Medikation wechselseitig bedingen ($H_0 : \rho = 0$).

Diskussion

Ein bedeutender Aspekt, der bei der Anwendung des LLTM unbedingt betrachtet werden sollte, sind die Modellannahmen. Das LLTM trifft recht strenge Annahmen, die möglichst auf ihre Gültigkeit hin untersucht werden sollten. Bei einigen der Annahmen ist eine direkte Überprüfung möglich:

- Mixed Rasch Models (MRM) ermöglichen es zu überprüfen, ob die Veränderung für alle Personen einer Gruppe gleich ist, wie es im LLTM angenommen wird (siehe Kapitel 3). Gibt es in einer Gruppe heterogene Entwicklungen, beispielsweise durch eine unterschiedliche Verträglichkeit einer Medikation, so wird diese Heterogenität durch MRM entdeckt. Anhand dieser Ergebnisse kann eine neue Gruppenaufteilung vorgenommen werden, die eine homogene Entwicklung innerhalb der Gruppen garantiert. Es empfiehlt sich also bei Anwendung des LLTM in jedem Fall eine Prüfung der Homogenität innerhalb der Gruppen mittels MRM vorzunehmen.

- Die Annahme, dass die Veränderung für alle Items in gleichem Ausmaß erfolgt, kann durch einen Vergleich mit einem quasisaturierten Modell mittels Likelihood-Quotiententest überprüft werden (siehe Seite 11 Hypothesentestung).

- Die Homogenität der Items, die im LLTM gefordert wird, kann durch verschiedene Methoden überprüft werden. Der bekannteste Test zur Prüfung auf Itemhomogenität ist der Martin-Löf-Test (beispielsweise in Fischer and Molenaar (1995)). Items, die nicht Rasch-konform sind, können dann aus der Analyse ausgeschlossen werden.

Es gibt jedoch auch Annahmen des LLTM, die nicht überprüft werden können und die in besonderen Anwendungssituationen eher zweifelhaft sind. Dies gilt beispielsweise für die Annahme der lokalen stochastischen Unabhängigkeit. Im LLTM werden einer Person zu jedem Zeitpunkt dieselben Items vorgelegt, was bei Erinnerungseffekten problematisch sein könnte. In Anwendungssituationen im medizinischen Bereich, wenn der Fragebogen beispielsweise Symptome auflistet, die der Patient entweder hat oder auch nicht, ist dies eher unproblematisch. Soll allerdings eine neue Unterrichtsmethodik evaluiert werden und den Schülern werden zu mehreren Zeitpunkten dieselben Aufgaben gestellt, so könnte es passieren, dass sich die Schüler an die Aufgaben erinnern und damit die Lösungswahrscheinlichkeiten zwischen den Testzeitpunkten nicht unabhängig voneinander sind.
Meist wird ein Trendeffekt in das Modell mit aufgenommen, welcher neben der 'globalen' Veränderung, die in allen Gruppen erfolgt, auch den Erinnerungseffekt erfasst. Dies ändert jedoch nichts an der Tatsache, dass die lokale stochastische Unabhängigkeit verletzt ist und damit die Schätzung fehlerhaft sein könnte.

Die Forderung nach Rasch-homogenem Testmaterial im LLTM stellt eine Einschränkung für die praktische Anwendung dar. In manchen Anwendungen kann es von Interesse sein eine Veränderung in mehreren latenten Eigenschaften zu messen, beispielsweise in medizinischen Bereichen, wenn die Wirkung einer Therapie oder einer Medikation untersucht werden soll. Oftmals wirken sich eine Therapie oder eine Medikation auf mehrere latente Bereiche aus. So kann sich ein Medikament neben dem physischen ebenfalls auf das psychische Wohlbefinden des Patienten auswirken. Durch die strenge Forderung der Eindimensionalität im LLTM ist die Befragung des Patienten nur auf einen latenten Bereich eingeschränkt.

Nicht nur für die Praxis stellt die Anforderung von homogenem Testmaterial ein Problem dar: Die Anforderung der Rasch-Homogenität ergibt sich aus der Modellformel, in der personen- und itemspezifische Parameter auf einer gemeinsamen Skala abgebildet werden. So vermerkt Fischer (1995), dass das LLTM nicht für die Messung von Veränderung konzipiert wurde und daher sowohl personenspezifische Parameter als auch itemspezifische Parameter als sogenannte *Nuisance*-Parameter in das Modell mit eingehen, da sie für die Veränderungsmessung nicht von primärem Interesse sind. Das sei insofern problematisch, als itemspezifische Parameter mit in die bedingte Likelihood mit eingehen und so die Schätzung des Veränderungsparameters beeinflussen, wenn nicht sogar verzerren.

Um eine mögliche Verzerrung durch die itemspezifischen Parameter zu vermeiden, schlägt Fischer (1983, 1995) ein von ihm für die Messung von Veränderung konzipiertes Modell vor. In diesem werden personen- und itemspezifische Parameter als ein gemeinsamer Parameter aufgefasst und bei der Parameterschätzung aus der Likelihood eliminiert, sodass die Veränderung unabhängig von personen- und itemspezifischen Parametern geschätzt werden kann. Dieses Modell, das einen sehr engen Bezug zum LLTM aufweist, wird nach Fischer Linear Logistic Test Model with Relaxed Assumptions (LLRA) genannt und im folgenden Kapitel vorgestellt.

2.2 Das Linear Logistic Test Model with Relaxed Assumptions (LLRA)

Wie im vorigen Kapitel dargelegt, besitzt das LLTM einige Nachteile, wie die Existenz von *Nuisance*-Parametern und der damit verbundenen Gefahr von verzerrten Veränderungsparameter-Schätzungen sowie der strengen Voraussetzung von homogenem Testmaterial. Fischer (1976, 1977, 1983, 1995) konzipierte daher das Linear Logistic Test Model with Relaxed Assumptions (LLRA) für zwei Testzeitpunkte, welches item- und personenspezifischen Parameter zu einem Parameter zusammenfasst. Dieses Modell ermöglicht eine unabhängige Schätzung des Veränderungsparameters von personen- und von itemspezifischen Parametern. Die Zusammenfassung der beiden Parameter hat ebenso zur Folge, dass nun die strikte Voraussetzung der Eindimensionalität verworfen werden kann, da item- und personenspezifischer Parameter nicht mehr auf einer gemeinsamen Skala abgebildet werden. Es ist somit möglich, die Veränderung mehrerer latenter Eigenschaften zu messen.

Die Modellformel des LLRA ergibt sich durch Ersetzen von $\theta_v - \beta_i$ in der Modellgleichung des LLTM in (4) durch den Parameter θ_{vi}, der die Fähigkeit der Person v in Bezug auf die latente Eigenschaft, die Item i misst, beschreibt.

$$P(X_{vi1} = 1|v, i, T_1) = \frac{\exp(\theta_{vi})}{1 + \exp(\theta_{vi})} \tag{9}$$

$$P(X_{vi2} = 1|v, i, T_2) = \frac{\exp(\theta_{vi} + \delta_g)}{1 + \exp(\theta_{vi} + \delta_g)} \tag{10}$$

Die gruppenspezifische Veränderung wird mit δ_g bezeichnet. Diese ist im Gegensatz zur Veränderung im LLTM nicht zeitpunktspezifisch, da die Veränderungsmessung im LLRA nur zwischen zwei Testzeitpunkten vorgesehen ist.
Es ist zu beachten, dass durch die Zusammenfassung von personen- und itemspezifischem Parameter in Kauf genommen werden muss, dass keine Schätzung der Personenfähigkeiten und der Itemschwierigkeiten mehr möglich ist.

Modellannahmen

Die Struktur des LLRA unterscheidet sich von der des LLTM lediglich darin, dass es anstatt zwei getrennter Parameter für Person und Item einen Parameter gibt, der diese beiden Komponenten in einem Wert zusammenfasst. Die Veränderung wird analog zum LLTM durch einen bzw. mehrere gruppenspezifische Parameter beschrieben. Im LLRA werden also dieselben Annahmen in Bezug auf die Veränderung getroffen wie im LLTM. Das bedeutet also, dass die Annahme getroffen wird, dass die Veränderung für alle Personen innerhalb einer Gruppe gleich ist.
Ebenso wird angenommen, dass die gruppenspezifische Veränderung unabhängig von den Items ist, d.h. diese wird über alle Items hinweg als konstant angenommen. In Bezug auf die Items bedeutet dies, dass alle Items gleich sensibel in Bezug auf Veränderung sind. Dies impliziert die Annahme, dass die Veränderung in allen abgefragten latenten Eigenschaften erfolgt und dies sogar in gleichem Ausmaß.
Auch das Konzept der lokalen stochastischen Unabhängigkeit wird im LLRA übernommen, wovon bei der Schätzung der Veränderungsparameter Gebrauch gemacht wird.

Parameterschätzung

Der Parameter δ_g in (10), der die Veränderung beschreibt, wird wie auch im LLTM durch die bedingte Maximum-Likelihood-Methode (CML) geschätzt. Durch Umformung des LLRA in ein LLTM kann die bedingte Likelihood analog zum LLTM hergeleitet werden (Fischer (1983)) und nimmt folgende Form an:

$$L(X_1, X_2|X_1 + X_2) = \prod_{v=1}^{N} \prod_{i=1}^{I} \left\{ \frac{\exp(x_{vi2}\delta_g)}{1 + \exp(\delta_g)} \right\}^{(x_{vi1} - x_{vi2})^2} \tag{11}$$

Dabei bezeichnet $X_1 = ((x_{vi1}))$ die Datenmatrix mit den Antworten zum Zeitpunkt des ersten Tests und $X_2 = ((x_{vi2}))$ die Datenmatrix zum Zeitpunkt des zweiten Tests. Ein Eintrag der Datenmatrix nimmt den Wert 1 bei einer positiven Beantwortung eines Items an, ansonsten den Wert 0.

Die suffiziente Statistik für die Personenfähigkeit θ_{vi} (in Bezug auf die durch das Item gemessene latente Eigenschaft) ist die Summe der positiven Antworten der Person auf dieses Item zu den beiden Zeitpunkten. Auf diese Summe wird in (11) bedingt.
Im Gegensatz zum LLTM ist die Summe der richtigen Antworten einer Person auf alle Items keine suffiziente Statistik, da die Fähigkeit einer Person aufgrund der Zulässigkeit von mehrdimensional messendem Testmaterial in Bezug auf das Item gesehen werden muss. Jedes Item bildet sozusagen für sich genommen einen eigenen (homogenen) Test über die zwei Zeitpunkte.
Durch Bedingung auf die suffiziente Statistik ist θ_{vi} in der Likelihood in (11) nicht mehr enthalten und es ist möglich die Veränderung δ_g, die primär von Interesse ist, unabhängig von item- und personenspezifischem Parameter zu schätzen. Hier nimmt das LLRA das Konzept der spezifischen Objektivität auf, welches auch im Rasch-Modell postuliert wird. Die spezifische Objektivität schließt Verzerrungen, die durch die gemeinsame Schätzung von Parametern auftreten können, aus.

Wie man in (11) erkennen kann, nimmt die Likelihood für Beobachtungen mit $x_{vi1} = x_{vi2}$ den Wert Eins an. Das bedeutet, dass identische Antworten einer Person auf ein Item zu den beiden Zeitpunkten keinen Informationsgehalt tragen und somit aus der Schätzung ausgeschlossen werden.
Existieren in einer Gruppe nur Beobachtungen, die in dieselbe Richtung gehen (d.h. entweder nur Beobachtungen mit $x_{vi1} = 0$ und $x_{vi2} = 1$ oder nur Beobachtungen mit $x_{vi1} = 1$ und $x_{vi2} = 0$), so ist die Schätzung der Veränderung für diese Gruppe nicht möglich (Fischer (1983)), da die Veränderung einen unendlich großen/kleinen Schätzwert und damit keinen endlichen Wert liefert. Für eine vernünftige Schätzung aller gruppenspezifischen Veränderungsparameter müssen also in jeder Gruppe sowohl Beobachtungen mit $x_{vi1} = 0$ und $x_{vi2} = 1$ als auch mit $x_{vi1} = 1$ und $x_{vi2} = 0$ existieren.

Hypothesentestung

Grundsätzlich können im LLRA dieselben Hypothesen über die Veränderungsparameter getestet werden, wie im LLTM (siehe Seite 12). Auch hier ist es möglich mit Hilfe eines quasisaturierten Modells zu überprüfen, ob die Veränderung itemunspezifisch erfolgt. So könnten Items, deren latente Eigenschaft überhaupt keiner Veränderung unterliegen, und die damit im LLRA eine Unterschätzung der Veränderung bewirken könnten, ausfindig gemacht und gegebenenfalls entfernt werden.
Hypothesentests über itemspezifische Parameter sowie über latente Fähigkeiten von Personen, wie sie im LLTM durchgeführt werden könnten, sind im LLRA nicht möglich.

Diskussion

Auch im LLRA sind die Modellannahmen noch recht streng, wenn auch - wie der Name schon vermuten lässt - lockerer als im LLTM. Der bedeutende Vorteil gegenüber dem LLTM insbesondere für die praktische Anwendung ist die Zulässigkeit von mehrdimensional messendem Testmaterial. Zum einen wird dadurch die Konstruktion von Tests erheblich erleichtert, zum anderen kann die Messung von Veränderung in mehreren latenten Eigenschaften vorgenommen werden, was in manchen Anwendungsfällen wünschenswert sein kann. Dies erklärt, warum das LLRA dem LLTM oftmals in der Praxis vorgezogen wird.

Allerdings muss durch die Zulässigkeit von mehrdimensional messendem Testmaterial in Kauf genommen werden, dass die Schätzung von personen- und itemspezifischen Parametern aufgrund der Zusammenfassung der beiden Parameter nicht mehr möglich ist. Möchte man also neben der Veränderung auch Informationen über die Personen bzw. über die Items gewinnen, so empfiehlt sich die Anwendung des LLTM (vorausgesetzt es ist Rasch-homogenes Testmaterial verfügbar).
Auch die Annahme, dass die Veränderung nicht itemspezifisch ist, ist im LLRA kritischer zu betrachten als im LLTM. Die Einbeziehung mehrerer latenter Eigenschaften könnte dazu führen, dass es ein oder mehrere Items gibt, die eine latente Eigenschaft messen, die nicht von der Veränderung betroffen ist. Somit würde der Veränderungsparameter unterschätzt werden. Die Zulässigkeit von mehrdimensional messendem Testmaterial verschärft somit die Annahme der itemunabhängigen Veränderung. Es ist daher empfehlenswert, den Vergleich mit dem quasisaturierten Modell mittels Likelihood-Quotiententest durchzuführen. Werden dadurch Items ausfindig gemacht, die keiner Veränderung unterliegen, könnten die-

se aus dem Test ausgeschlossen werden um mögliche Verzerrungen der Veränderungsparameters zu vermeiden. Solche Entscheidungen sollten allerdings situationsbedingt getroffen werden.

Im Gegensatz zum LLTM ist die Anwendung von MRM aufgrund der Zulässigkeit von mehrdimensional messendem Testmaterial kaum möglich. Denn im MRM wird die Rasch-Homogenität innerhalb von latenten Subgruppen vorausgesetzt. Diese ist im LLRA nicht erfüllt. Damit ist die Überprüfung der Gruppeneinteilung nicht durchführbar und mögliche Heterogenitäten in einer Gruppe können nicht entdeckt werden. Die Einteilung der Gruppen sollte also bei Anwendung des LLRA sorgfältig überdacht sein um mögliche Fehlschlüsse zu vermeiden.

Ein weiterer Nachteil bei der Anwendung des LLRA ist der relativ hohe Datenverlust, der dazu führen kann, dass keine bzw. nur eine instabile Parameterschätzung möglich ist. Daten, die nicht auswertbar sind, umschließen alle diejenigen Beobachtungen, bei denen ein Item zu beiden Zeitpunkten gleich beantwortet wurde. Aufgrund der Wiederholung derselben Items zu beiden Testzeitpunkten liegt es nahe, dass die Anzahl identischer Antworten relativ hoch sein wird und damit sehr viele Beobachtungen nicht auswertbar sind. Dies liegt zum einen daran, dass der Schwierigkeitsgrad einer Aufgabe zu den beiden Zeitpunkten gleich ist und zum anderen daran, dass gewisse Erinnerungseffekte auftreten können. Beides tritt insbesondere in Bereichen der pädagogischen Psychologie auf, wenn es darum geht Lernfähigkeiten zu erfassen.
Angenommen in einem Intelligenztest gibt es eine sehr leichte Aufgabe. So wird die Wahrscheinlichkeit die Aufgabe zu beiden Testzeitpunkten zu lösen für (durchschnittlich) intelligente Personen sehr hoch sein. Insbesondere äußerst schwere bzw. leichte Testaufgaben führen zu identischen Antworten. Zudem besteht die Möglichkeit, dass sich die Personen an die Aufgaben erinnern und sie daher gleich lösen. Solche Erinnerungseffekte können außerdem zu einer Verletzung der lokalen stochastischen Unabhängigkeit führen.

Ein weiteres Problem ergibt sich, wenn die Veränderung aller Personen einer Gruppe nur in eine Richtung erfolgt. Dann ist eine Schätzung der Veränderung für die Gruppe nicht möglich. Eine Veränderung in eine Richtung tritt beispielsweise bei (starken) Lerneffekten auf, wie es bei Kindern der Fall ist. Es kommt eher selten vor, dass ein Kind eine Aufgabe zu einem Zeitpunkt richtig löst und zu einem späteren Zeitpunkt nicht mehr lösen kann.

Einen Lösungsansatz zu diesen Problemen bietet das Hybrid LLRA durch Einführung von paarweise homogenen Testitems mit unterschiedlichen Schwierigkeitsstufen. Dieses wird im folgenden Kapitel genauer erläutert.

2.3 Das Hybrid LLRA (HLLRA)

Das Hybrid LLRA wurde von Fischer (1989) für die Erfassung von Veränderung über mehrere Zeitpunkte hinweg konzipiert und kann als eine Optimierung des LLRA angesehen werden. Durch die Vermeidung von Wiederholung derselben Items zu den Testzeitpunkten wird sowohl Veränderung in nur eine Richtung unwahrscheinlicher gemacht als auch die Anzahl an identischen Antworten auf ein Item zu den Zeitpunkten reduziert.

Beides wird dadurch realisiert, dass den Personen zu jedem Testzeitpunkt andere Items mit unterschiedlichen Schwierigkeitsgraden vorgelegt werden. Dabei müssen zu jedem Zeitpunkt Items existieren, die dieselbe latente Fähigkeit messen, sodass diese Items einander zugeordnet werden können. Im einfachsten Fall von nur zwei Zeitpunkten gibt es paarweise homogene Items (siehe Tabelle 2). Die einander zugeordneten homogenen Items ('Hybrid-Items') bilden dann über die Zeitpunkte hinweg eindimensionale Tests mit unterschiedlichen Schwierigkeiten.

Hybrid-Itempaar mit latenter Fähigkeit	T_1	T_2
1	$Item_{11}$	$Item_{12}$
2	$Item_{21}$	$Item_{22}$
⋮	⋮	⋮
I	$Item_{I1}$	$Item_{I2}$

Tabelle 2: Hybrid-Items für zwei Zeitpunkte

Je nachdem wie die Annahmen bezüglich der Veränderung sind, werden die einander zugeordneten Items in aufsteigendem, absteigendem oder auch zufälligem Schwierigkeitsgrad präsentiert (Fischer (1989), Formann and Spiel (1989)).

Wird beispielsweise angenommen, dass eine starke Verbesserung der Fähigkeit stattfindet, so werden die schwierigeren Items zu späteren Zeitpunkten gestellt. Geht man von der Annahme einer geringen Veränderung aus, so können die einander zugeordneten homogenen Items zufällig auf die Testzeitpunkte verteilt werden.

Die Modellformel für das HLLRA ergibt aus der des LLRA ((9) und (10)) unter Berücksichtigung von Schwierigkeitsunterschieden der Hybrid-Items und auf mehrere Testzeitpunkte erweitert. Diese nimmt dann folgende Form an:

$$P(X_{vit} = 1|v, i, t) = \frac{\exp(\theta_{vi} - \beta_{it} + \delta_{gt})}{1 + \exp(\theta_{vi} - \beta_{it} + \delta_{gt})} \tag{12}$$

θ_{vi} bezeichnet die Personenfähigkeit bezüglich der latenten Eigenschaft, die die einander zugeordneten Items $it, t = 1, \cdots, T$ messen. Da sich diese Items in ihren Schwierigkeiten voneinander unterscheiden, werden sie durch die Parameter β_{it} in der Modellformel mit berücksichtigt.
Die Veränderung wird, wie im LLTM und im LLRA, durch einen Veränderungsparameter erfasst. Dieser wird mit δ_{gt} bezeichnet. δ_{gt} ist gruppen- und zeitpunktspezifisch und wird zum Zeitpunkt des ersten Tests auf Null gesetzt, da bis dahin noch keine Veränderung stattgefunden hat. Der Veränderungsparameter δ_{gt} kann analog zu den zuvor vorgestellten Modellen durch beliebig viele Effektparameter und deren Wechselwirkungen ausgedrückt werden.

Wird ein Trendeffekt τ mit in das Modell aufgenommen, so ist dieser jedoch nicht eindeutig bestimmbar, da er mit der Schwierigkeitsdifferenz der einander zugeordneten Items konfundiert ist. Das bedeutet, dass die Schätzung einen Wert $\tau_i = \tau - d_i$ liefert (hier vereinfacht dargestellt für nur zwei Testzeitpunkte), der sowohl den Trend τ als auch die Schwierigkeitsdifferenz d_i zweier einander zugeordneter Items beinhaltet. Die Vermischung des Trendeffektes mit den Schwierigkeitsdifferenzen macht eine sinnvolle Interpretation des Effektes kaum möglich. Falls der Trendeffekt von Bedeutung ist, sollten daher Voruntersuchungen angestellt werden, in denen die Itemschwierigkeiten geschätzt werden können. Anhand dieser Schätzungen können die Schwierigkeitsdifferenzen der Items und damit auch der Trendeffekt eindeutig bestimmt werden.

Spezialfall:
Das LLRA ergibt sich als Spezialfall aus dem HLLRA. Dies tritt genau dann ein, wenn es nur zwei Testzeitpunkte gibt und dieselben Items zu den zwei Zeitpunkten vorliegen.

Modellannahmen

Eine sehr wichtige Annahme, die im HLLRA gelten muss, ist die Homogenität der einander zugeordneten Items. Diese müssen also dieselbe latente Fähigkeit messen. Zudem ist es sinnvoll, die einander zugeordneten Items so zu wählen, dass sie sich in ihren Schwierigkeiten voneinander unterscheiden um die Anzahl an identischen Antworten zu verringern. Neben der Rasch-Homogenität einander zugeordneter Items wird die Annahme der lokalen stochastischen Unabhängigkeit getroffen.
Das HLLRA besitzt wie das LLRA einen gruppenspezifischen Parameter, der die Veränderung beschreibt. Auch hier wird also die Annahme getroffen, dass die Veränderung für die Personen einer Gruppe gleich ist. Zudem ist der Veränderungsparameter nicht itemspezifisch, das heißt, dass angenommen wird, die Veränderung erfolgt in allen latenten Eigenschaften und dies in gleichem Ausmaß.
Vor Anwendung des HLLRA sollte klar sein, ob der Trend ebenfalls von Interesse ist. Soll der Trend untersucht werden, ist es notwendig, dass die itemspezifischen Parameter bekannt sind. Nur so kann der Trendeffekt eindeutig bestimmt werden. Ansonsten ist er mit der Schwierigkeitsdifferenz d_i der einander zugeordneten Items konfundiert und damit nicht interpretierbar. In diesem Falle wäre der konfundierte Trendeffekt $\tau_i = \tau - d_i$, den die Schätzung liefert, itemabhängig. Die Veränderung δ_{gt}, die den Trendeffekt beinhaltet, ist dann nur unter der Annahme, dass alle Schwierigkeitsdifferenzen d_i gleich sind, itemunspezifisch.

Parameterschätzung

Fischer (1989) nutzt bei der Parameterschätzung aus, dass das HLLRA auf die Form eines LLTM gebracht werden kann. Mittels CML ist es möglich, die gruppen- und zeitpunktspezifische Veränderung zu schätzen. Die suffiziente Statistik für den Personenparameter θ_{vi} ist dabei die Summe der durch die Person v richtig beantworteten Items, welche die latente Eigenschaft i messen. Damit kann unter Annahme der Unabhängigkeit der Lösungswahrscheinlichkeiten der Items zwischen den Testzeitpunkten die bedingte Like-

lihood $L_{C_{vi}}$ aufgestellt werden, die nun nur noch von dem aufgabenspezifischen Parameter β_{it} und dem Veränderungsparameter δ_{gt} abhängt.
Die gemeinsam bedingte Likelihood L_C wird über das Produkt aller Hybrid-Items, über alle Personen einer Gruppe und schließlich über alle Gruppen gebildet (Fischer (1989)):

$$L_C = \prod_{g=1}^{G} \prod_{Person_v \in Gruppe_g} \prod_{i=1}^{I} L_{C_{vi}} \tag{13}$$

Die bedingte Likelihood wird in Abhängigkeit der unbekannten Parameter mittels iterativer Verfahren wie dem *Newton-Raphson* Algorithmus maximiert. Die gruppen- bzw. zeitpunktspezifische Veränderung mit Ausnahme des Trends ist eindeutig schätzbar. Sind die itemspezifischen Parameter bekannt (beispielsweise durch vorhergehende Untersuchungen), so können diese zuvor in die Schätzgleichung eingesetzt werden, und der Trendeffekt wäre damit ebenfalls eindeutig bestimmbar. Falls allerdings itemspezifische Parameter nicht bekannt sind, so sind Trendeffekt und Differenz der itemspezifischen Parameter miteinander konfundiert.

Hypothesentestung

Im HLLRA können analog zum LLTM und im LLRA Hypothesen bezüglich der Effektparameter aufgestellt und mittels bedingtem Likelihood-Quotiententest getestet werden. Mögliche Hypothesen über die Effekte - wie Gleichheit der Effekte bei bestimmten Gruppen, keine Interaktionseffekte, überhaupt keine Effekte - (siehe Seite 12) können getestet werden. Bei einem Trendeffekt ist eine Hypothesentestung nur sinnvoll, wenn dieser nicht mit den Schwierigkeitsdifferenzen der einander zugeordneten Items konfundiert ist. Ist der Trendeffekt aufgrund von Vorkenntnissen über die itemspezifischen Parameter eindeutig bestimmbar, so können auch in diesem Modell Hypothesen über den Trend analog wie im LLTM und im LLRA aufgestellt werden.

Anwendungsbeispiel

Im Folgenden soll eine Anwendung des HLLRA vorgestellt werden, die von Formann and Spiel (1989) beschrieben wird:

In einer Untersuchung sollten verschiedene Übungsmethoden zum Textverständnis evaluiert werden. Dabei wurde Schülern ein Text vorgelegt, zu dem sie vier Aufgaben zum Text-

verständnis bearbeiten sollten. Nach Bearbeitung der Aufgaben wurden die Schüler zufällig in vier Gruppen eingeteilt. Drei der Gruppen erhielten unterschiedliche Übungsmethoden, von denen angenommen wurde, dass sie das Textverständnis der Schüler verbessern. Eine weitere Gruppe erhielt als Kontrollgruppe keine Übungen. Zu einem zweiten Testzeitpunkt wurde den Schülern ein anderer Text vorgelegt, der von Experten als schwieriger als der erste bewertet wurde. Die Schüler sollten dieselben Aufgaben an diesem Text bearbeiten.

Dieses Beispiel verdeutlicht das Konzept des HLLRA: Indem den Schülern dieselben Aufgaben zu beiden Zeitpunkten vorgelegt wurden, wurde die paarweise Homogenität der Aufgaben zu den zwei Testzeitpunkten sichergestellt. Der für das HLLRA bedeutende Aspekt zeigt sich in der Verschiedenheit der Texte zu den beiden Zeitpunkten. Dadurch werden Erinnerungen der Kinder an den Lösungsweg vermieden.
Die den Schülern vorgelegten Texte waren zudem von unterschiedlicher Schwierigkeit. Da man von einem Lernzuwachs der Schüler ausging, war der Schwierigkeitsgrad des zweiten Textes höher. Dadurch sollten Veränderungen in nur eine Richtung vermieden werden. Veränderung in nur eine Richtung bedeutet hier, dass es nur Fälle gibt, in denen ein Kind die Aufgaben zum ersten Zeitpunkt nicht lösen kann, zum zweiten dagegen schon.
Anhand der Testergebnisse zu den zwei Zeitpunkten konnte die Effektivität der verschiedenen Übungsmethoden bereinigt von jeglichen Erinnerungseffekten geschätzt und auf Signifikanz hin überprüft werden.

Diskussion

In Bereichen der pädagogischen Psychologie, in denen beispielsweise die Lernfähigkeit von Personengruppen erfasst werden soll, ist das HLLRA das wohl geeignetste der in dieser Arbeit vorgestellten Modellen. Der Unterschied dieses Modells zu dem LLRA besteht darin, dass zu jedem Testzeitpunkt andere Aufgaben gestellt werden. Dies bringt einige Vorteile mit sich:

Zum einen ist die stochastische Unabhängigkeit der Lösungswahrscheinlichkeiten zwischen den Testzeitpunkten im HLLRA gegeben, die im LLTM und im LLRA aufgrund von Erinnerungseffekten verletzt sein könnte. Dadurch werden Verzerrungen der Parameterschätzungen vermieden.

Zum anderen ist es möglich effizientere Schätzer als im LLRA zu erhalten. Dies hat mehrere Gründe:

- Die Wahrscheinlichkeit für identische Antworten[2] verringert sich mit zunehmender Anzahl von Zeitpunkten. Da das LLRA nur für zwei Zeitpunkte konstruiert wurde, besitzt dieses erwartungsgemäß die höchste Anzahl an identischen Antworten. Der Datenverlust ist also bei nur zwei Testzeitpunkten am höchsten.

- Bei voneinander abhängigen Antwortwahrscheinlichkeiten ist erwartungsgemäß die Anzahl gleicher Antwortmuster größer als bei voneinander unabhängigen Antwortwahrscheinlichkeiten. Im LLRA besteht die Gefahr, dass die Antwortwahrscheinlichkeiten aufgrund von Erinnerungseffekten voneinander abhängig sind. Die Erinnerungseffekte treten im HLLRA nicht auf, da zu jedem Zeitpunkt andere Aufgaben gestellt werden.

- Bei starken Lerneffekten kann es im LLRA vorkommen, dass in einer Gruppe wenige oder sogar keine Beobachtungen mit $x_{vi} = (1, 0)'$ vorliegen. Eine zu geringe Anzahl an solchen Beobachtungen führt zu ineffizienten Schätzern. Liegen in einer Gruppe keine solche Beobachtungen vor, so ist der gruppenspezifische Effekt nicht schätzbar. Im HLLRA kann die Anzahl an Beobachtungen mit $x_{vi} = (1, 0)'$ dadurch erhöht werden, dass zu dem späteren Zeitpunkt eine schwierigere Aufgabe derselben latenten Eigenschaft gestellt wird. Dies ist im LLRA nicht möglich, da zum zweiten Zeitpunkt dieselbe Aufgabe vom ersten Testzeitpunkt wiederholt wird.

Das HLLRA besitzt jedoch nicht nur Vorteile gegenüber dem LLRA. Ein wesentlicher Schwachpunkt des HLLRA ist die Konfundierung des Trendeffektes mit den Schwierigkeitsdifferenzen der einander zugeordneten Items. Fischer (1989) schlägt daher zwei Möglichkeiten vor:

1. Die erste Möglichkeit ist die, Aufgaben zu verwenden, deren Schwierigkeiten beispielsweise durch Voruntersuchungen bekannt sind. Diese werden dann als Konstanten in die Schätzgleichung eingesetzt, und es können gruppenspezifische Effekte, Interakti-

[2]Identische Antworten zu den Zeitpunkten, d.h. $x_{vi} = (0, 0, 0, \cdots, 0)'$ bzw. $x_{vi} = (1, 1, 1, \cdots, 1)'$ tragen keinen Informationsgehalt und werden aus der Parameterschätzung ausgeschlossen.

onseffekte und Trendeffekt eindeutig geschätzt werden.Im Gegensatz zum LLTM und zum LLRA beinhaltet der Trendeffekt dann keinen Erinnerungseffekt, da die Aufgaben nicht wiederholt werden und kann so als ein 'globaler Lernzuwachs' interpretiert werden.

2. Die zweite Möglichkeit ist, dieselben Aufgaben zu den Testzeitpunkten zu verwenden. Die Schwierigkeitsdifferenz nimmt damit den Wert Null an, und der Trendeffekt ist eindeutig bestimmbar. Bei Anwendung dieser aufgrund ihrer Einfachheit attraktiv erscheinenden Lösung muss in Kauf genommen werden, dass auch hier Erinnerungseffekte auftreten können, und dass die Einbringung von Schwierigkeitsunterschieden zur Erhöhung seltener Beobachtungen nicht möglich ist.

Ist der Trendeffekt für den Anwender überhaupt nicht von Interesse, so bietet das HLLRA eine geeignete Möglichkeit um gruppenspezifische Veränderung zu messen.

Wie beim LLRA müssen durch die Zulässigkeit von mehrdimensional messendem Testmaterial einige Nachteile in Kauf genommen werden:
Eine Schätzung von personen- und itemspezifischen Parametern ist nicht möglich.
Zudem kann keine Überprüfung auf heterogene Veränderungen innerhalb der vordefinierten Gruppen mittels MRM vorgenommen werden. Die Gruppeneinteilung muss also gut überdacht sein um Fehlinterpretationen aufgrund von heterogenen Entwicklungen innerhalb der Gruppen zu vermeiden.

2.4 Das Multidimensional Rasch Model for Learning and Change (MRMLC)

Das Multidimensional Rasch Model for Learning and Change (MRMLC) wurde von Embretson (1991a,b) primär für die Messung von Lernfähigkeit entwickelt, kann aber auch auf andere Bereiche angewendet werden. Das MRMLC basiert auf der Struktur des LLTM, wobei die Veränderung auch in diesem Modell durch einen bzw. mehrere Parameter beschrieben wird. Dieses Modell unterscheidet sich zu den zuvor vorgestellten Modellen darin, dass sich die Veränderung nicht in einer Veränderung der Schwierigkeit der Aufgabe ausdrückt, sondern vielmehr in einer Veränderung der latenten Fähigkeit der Person. Dadurch ist die Messung von Veränderung auf individueller Ebene möglich.

Die Modellformel des MRMLC nimmt folgende Form an:

$$P(X_{vit} = 1|v, i, t) = \frac{\exp(\sum_{m=1}^{t} \theta_{vm} - \beta_i)}{1 + \exp((\sum_{m=1}^{t} \theta_{vm} - \beta_i)} \tag{14}$$

wobei die Schwierigkeit eines Items i mit β_i bezeichnet und über alle Zeitpunkte hinweg als konstant angenommen wird. Die individuelle Veränderung wird erfasst durch eine Veränderung in der Personenfähigkeit.

Embretson unterscheidet drei Arten von Fähigkeiten:

- Zum einen die Anfangsfähigkeit θ_{v1}, die die Fähigkeit einer Person zum Zeitpunkt des ersten Tests beschreibt, zu dem noch keine Veränderung statt gefunden hat.
- Eine weitere Fähigkeit ist die Lernfähigkeit oder auch Modifizierbarkeit. Diese entspricht der Veränderung von einem Zeitpunkt zum nächsten. Somit gibt es für jede Person $T - 1$ zeitpunktabhängige Modifizierbarkeiten $\theta_{v2}, \cdots, \theta_{vT}$. Die Veränderung über einen größeren Zeitraum hinweg ergibt sich dann aus der Summe der entsprechenden Modifizierbarkeiten.
- Als letzte Fähigkeit nennt Embretson die effektive Fähigkeit. Diese beschreibt die latente Fähigkeit einer Person zu einem bestimmten Zeitpunkt und setzt sich demnach aus der Anfangsfähigkeit und den bis dahin stattgefundenen Lernzuwächsen (Modifizierbarkeiten) zusammen.

Die Zusammenhänge zwischen den einzelnen Fähigkeiten werden nochmals in Abbildung 3 für drei Testzeitpunkte veranschaulicht.

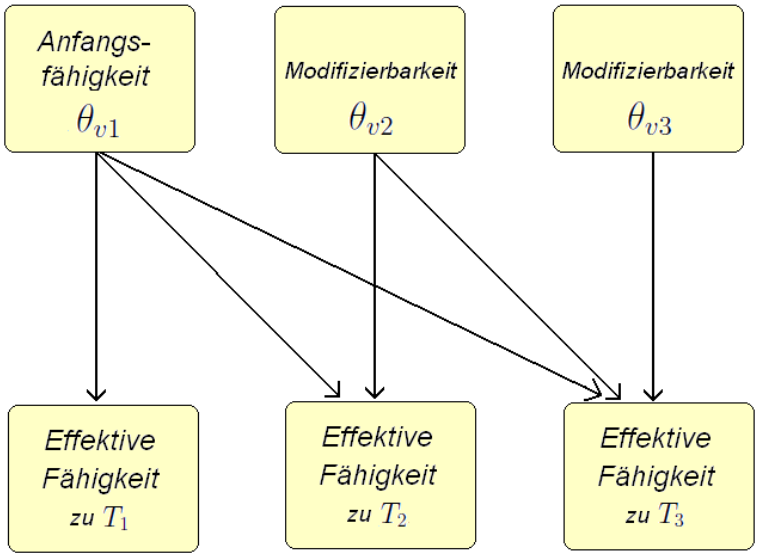

Abbildung 3: Zusammenhang zwischen den Fähigkeiten für drei Zeitpunkte

Studiendesign

Bei der Messung von Lernfähigkeiten ist es sinnvoll, einer Person zu jedem Zeitpunkt andere Aufgaben zu stellen, um Erinnerungseffekte zu vermeiden. Embretson (1991a) schlägt daher ein balanciertes Design vor. Die Idee bei diesem Design ist, dass jede Aufgabe zu jedem Zeitpunkt gestellt wird, jedoch nicht derselben Person vorgelegt wird um Erinnerungseffekte auszuschließen. Dabei werden die Personen zufällig in Gruppen eingeteilt. Die Aufgabenblöcke werden den Gruppen in einer anderen Reihenfolge vorgelegt (siehe Tabelle 3). Dieses Vorgehen ermöglicht eine Schätzung aller Aufgabenschwierigkeiten. Da die Aufgabenschwierigkeiten lediglich auf einer Intervallskala messbar sind, (fehlender natürlicher Nullpunkt), sind diese zwischen den Gruppen bis auf einen konstanten Wert nicht eindeutig definiert. Um die Aufgabenschwierigkeiten zwischen den Gruppen vergleichbar zu machen, müssen daher einige Aufgaben, die sogenannten *Linking-Items*, allen Gruppen zum gleichen Zeitpunkt gestellt werden.

Personengruppe	T_1	T_2	T_3	Linking-Items
1	Block 1	Block 2	Block 3	Block 4
2	Block 2	Block 3	Block 1	Block 4
3	Block 3	Block 1	Block 2	Block 4

Tabelle 3: Latin-Square-Design mit vier Aufgaben-Blöcken für drei Testzeitpunkte

Modellannahmen

Das MRMLC setzt Rasch-homogenes Testmaterial voraus, da die latente Fähigkeit einer Person und die Schwierigkeit einer Aufgabe auf derselben Skala abgebildet werden.
Zudem geht das Modell von der lokalen stochastischen Unabhängigkeit aus. Insbesondere bei Lernfähigkeitstests könnte diese durch Erinnerungseffekte verletzt werden. Um die lokale stochastischen Unabhängigkeit zu gewährleisten wird ein balanciertes Design verwendet, bei dem einer Person dieselbe Aufgabe nur zu einem Zeitpunkt vorgelegt wird.

Parameterschätzung

Die Parameterschätzung teilt sich in zwei Schritte auf:

Im ersten Schritt werden die itemspezifischen Parameter unter Bedingung auf die effektiven Fähigkeiten geschätzt. Dabei wird ausgenutzt, dass die Gesamtpunktzahl einer Person zu einem Testzeitpunkt eine suffiziente Statistik für ihre effektive Fähigkeit darstellt. Somit ist die Gesamtpunktzahl s_{v1} einer Person v zum ersten Testzeitpunkt suffizient für die Anfangsfähigkeit der Person v. Die Gesamtpunktzahl s_{vm} für den $m-$ten Testzeitpunkt ist suffizient für die effektive Fähigkeit $\theta_{v1} + \theta_{v2} + \cdots + \theta_{vm}$ der Person v zum Zeitpunkt m. Die bedingte Likelihood hängt dann nur noch von den itemspezifischen Parametern ab. Sie wird in Abhängigkeit dieser Parameter mittels iterativer Verfahren wie beispielsweise der *Newton-Raphson* Methode maximiert.

In einem zweiten Schritt werden die geschätzten itemspezifischen Parameter in die gemeinsame Likelihood eingesetzt um die Anfangsfähigkeit und Modifizierbarkeiten einer Person zu schätzen.

Eine detaillierte Beschreibung der einzelnen Schritte zur Schätzung von item- und personenspezifischen Parametern findet sich in Embretson (1991a).

Neben der Schätzung von Aufgabenschwierigkeiten, Personenfähigkeiten und individuellem Lernzuwachs können der Informationsmatrix, die sich als das Negative der zweiten Ableitung ergibt, Varianzen bzw. Kovarianzen von Anfangsfähigkeit und Modifizierbarkeiten entnommen werden (siehe Embretson (1991a)). Anhand von Varianzen könnte geprüft werden, inwieweit die Veränderung von Personen zufällig erfolgt.

Diskussion

Das MRMLC wurde speziell für die Messung von Lernfähigkeit konzipiert. Ein wesentlicher Aspekt ist, dass einer Person zu jedem Testzeitpunkt andere Aufgaben gestellt werden um Erinnerungseffekte zu vermeiden. Dieser Aspekt wird auch im HLLRA berücksichtigt. Im MRMLC können dadurch, dass jede Aufgabe zu jedem Testzeitpunkt gestellt wird, die Schwierigkeiten aller Aufgaben geschätzt werden. Im Vergleich zum HLLRA ist damit eine nicht-konfundierte Schätzung der gesamten Veränderung möglich. Zur Anwendung des MRMLC muss allerdings die Rasch-Homogenität für alle Items gelten, und nicht etwa nur für die Items, die einander zugeordnet werden, wie im HLLRA.

Im Gegensatz zum HLLRA erfolgt die Messung von Veränderung im MRMLC individuenspezifisch. Dies kann für manche Anwendungssituationen von Vorteil sein. Kann man beispielsweise keine feste Gruppeneinteilung der Stichprobe aufgrund von sachlogischen Überlegungen vornehmen, oder zweifelt man an einer bestehenden Gruppeneinteilung, so bietet es sich an, individuelle Veränderung von Personen zu messen. Mit Hilfe dieser individuellen Veränderung können dann weitere Analysen durchgeführt werden. Die Gefahr falsche Schlüsse über die Veränderung aufgrund von Heterogenitäten innerhalb einer oder mehrerer Gruppen zu ziehen, wie sie im HLLRA sowie im LLTM und LLRA besteht, wird im MRMLC ausgeschlossen.
Eine Alternative zur Vermeidung von falschen Schlüssen aufgrund einer fehlspezifizierten Gruppenaufteilung bieten Mixed Rasch Models, die im nächsten Kapitel vorgestellt werden.

2.5 Überblick

Nachdem die bekanntesten Modelle der Veränderungsmessung vorgestellt wurden, sollen im Folgenden die wichtigsten Aspekte der Modelle gegenübergestellt werden. Tabelle 4 gibt einen groben Überblick über die charakteristischen Eigenschaften dieser Modelle. Anhand der aufgeführten Aspekte kann für eine bestimmte Anwendungssituation das geeignetste Modell gewählt werden.

	LLTM	LLRA	HLLRA	MRMLC
Veränderungs-parameter	gruppen-spezifisch	gruppen-spezifisch	gruppen-spezifisch	individuen-spezifisch
Anforderung an Testmaterial	Rasch-homogen	keine	Rasch-homogene Hybrid-Items	Rasch-homogen
Wiederholung der Items	ja	ja	nein	nein
schätzbare Parameter	Veränderung Trend Personen Aufgaben	Veränderung Trend - -	Veränderung - - -	Veränderung * Personen Aufgaben
Bemerkungen	Gruppen-zugehörigkeit mittels MRM überprüfbar	nur für zwei Testzeitpunkte	Trend ist mit Schwierigkeits-differenzen konfundiert	Schätzung von Varianzen und Kovarianzen möglich

Tabelle 4: Überblick über die wichtigsten Eigenschaften der vier Modelle

* Ein Trend wie beim LLTM, LLRA und HLLRA ist aufgrund der individuenspezifischen Messung nicht vorhanden.

3 Mixed Rasch Models (MRM)

Mixed Rasch Models (MRM) verbinden die latente Klassenanalyse mit dem Rasch-Modell, indem sie die Stichprobe in Subgruppen unterteilen, für die jeweils das Rasch-Modell gilt. Innerhalb dieser Subgruppen ist dann die Richtung und das Ausmaß der Veränderung homogen, während es zwischen den Gruppen Unterschiede bezüglich der Veränderung gibt. MRM ermöglichen damit die Identifizierung von verschiedenen Entwicklungsmustern in einer Stichprobe.

Bei dem MRM wird für jede Subgruppe ein eigenes Modell mit einem globalen Veränderungsparameter aufgestellt. Dieses Modell wird in Form eines LLTM bedingt auf die Subgruppe g beschrieben (vgl. Meiser (2007)):

$$P(X_{vit} = 1|g, v, i, t) = \frac{\exp(\theta_{v|g} - \beta_{i|g} + \delta_{t|g})}{1 + \exp(\theta_{v|g} - \beta_{i|g} + \delta_{t|g})} \tag{15}$$

Die Parameter in (15) für die Subgruppe g werden analog wie im LLTM interpretiert:
Die zeitpunktspezifische Veränderung wird durch den Parameter $\delta_{t|g}$ erfasst und ist für alle Personen in der Subgruppe gleich. Da zum ersten Testzeitpunkt noch keine Veränderung stattgefunden hat, wird $\delta_{1|g}$ auf Null gesetzt.
Während die Veränderung $\delta_{t|g}$ für eine Subgruppe g über die Zeitpunkte hinweg variiert, werden die Personenparameter $\theta_{v|g}$ und die itemspezifischen Parameter $\beta_{i|g}$ über die Zeit hinweg als konstant angenommen. Aussagen über die Fähigkeit einer Person innerhalb ihrer Subgruppe sind also über die Schätzung von Personenfähigkeiten möglich.

Wie eben verdeutlicht, ist die Veränderung innerhalb einer Subgruppe homogen. Zwischen den Gruppen liegt allerdings eine heterogene Entwicklung vor. Qualitative Unterschiede in der Entwicklung von Personen zeigen sich also durch die Ungleichheit ihrer Klassenzugehörigkeit. Die Zuordnung einer Person in eine latente Klasse erfolgt dabei nicht deterministisch, sondern mit einer bestimmten Wahrscheinlichkeit. So wird eine Person der Subgruppe g zugeordnet, für die die Wahrscheinlichkeit, dass sie dieser Gruppe angehört unter Hinzuziehung ihrer Testergebnisse, maximal ist. Die Zuordnungswahrscheinlichkeit einer Person mit Antwortvektor x für die Subgruppe g wird folgendermaßen bestimmt (vgl. Rost and Von Davier (1995)):

$$P(g|X = x) = \frac{\pi_g P(X = x|g)}{P(X = x)} \tag{16}$$

Hierbei entspricht π_g dem Anteil der Subgruppe g. $P(X = x|g)$ berechnet sich aus (15) [3] und entspricht der Wahrscheinlichkeit für das Testergebnis x unter der Bedingung, dass die Person aus Subgruppe g stammt.
Für die Berechnung der Zuordnungswahrscheinlichkeit muss schließlich noch die marginale Verteilung $P(X = x)$ bekannt sein. Sie entspricht der Wahrscheinlichkeit für den Antwortvektor x unabhängig von der Gruppenzugehörigkeit, die gegeben ist durch

$$P(X = x) = \sum_{g=1}^{G} \pi_g P(X = x|g) \tag{17}$$

Die Schätzung der Parameter erfolgt mittels *Expectation-Maximization* (EM) Algorithmus und wird in Rost (1990, 1991) sowie in Von Davier and Rost (1995) beschrieben.

Die Anzahl der Gruppen wird nicht etwa geschätzt, sondern im Modell a priori spezifiziert. In der Regel werden mehrere Modelle mit unterschiedlich vielen Gruppen gefittet und daraufhin miteinander verglichen. Ein Vergleich mittels Likelihood-Quotiententest ist allerdings nicht möglich, da es sich bei den Modellen nicht um genestete Modelle handelt.

Glück and Spiel (1997) nennen zwei Vergleichsmöglichkeiten:

- Zum einen können zum Vergleich der Modelle Informationskriterien wie das Akaike Information Criterion (AIC), das Best Information Criterion (BIC) oder das Consistent Akaike Information Criterion (CAIC) herangezogen werden. Diese berücksichtigen sowohl die Likelihood als auch die Anzahl der zu schätzenden Parameter. Dabei wird das Modell bevorzugt, für welches das Informationskriterium den geringsten Wert annimmt.

- Eine andere Möglichkeit ist die Betrachtung der mittleren Klassenzuordnungswahrscheinlichkeiten der Modelle. Die Personen einer Stichprobe werden mit einer bestimmten Wahrscheinlichkeit den latenten Klassen zugeordnet. Das Modell ist umso geeigneter, je eindeutiger die Zuordnung der Personen in die latenten Klassen erfolgt. Die Eindeutigkeit der Zuordnung drückt sich durch eine hohe mittlere Zuordnungswahrscheinlichkeit aus. Es werden demnach diejenigen Modelle gewählt, deren mittlere Zuordnungswahrscheinlichkeit maximal ist.

Nachdem das allgemeine Konzept von MRM vorgestellt wurde, sollen im Folgenden Möglichkeiten aufgezeigt werden, wie MRM zur Messung von Veränderung herangezogen werden können.

[3] Für die Berechnung sei auf die Literatur von Rost and Von Davier (1995) verwiesen.

3.1 Ansätze zur Aufdeckung von Veränderungsstrukturen

Im Folgenden sollen unter Hinzuziehung von Beispielen zwei verschiedene Ansätze vorgestellt werden, wie man Veränderung mit Hilfe von MRM untersuchen kann. Bei dem ersten Ansatz, der beispielsweise in Meiser et al. (1995) beschrieben wird, wird zunächst die Datenstruktur so angepasst, dass sich ein 'großer Gesamttest' ergibt. Die Dimension der Zeitpunkte wird aufgelöst, wodurch sogenannte *virtuelle* Aufgaben entstehen (Abbildung 4).

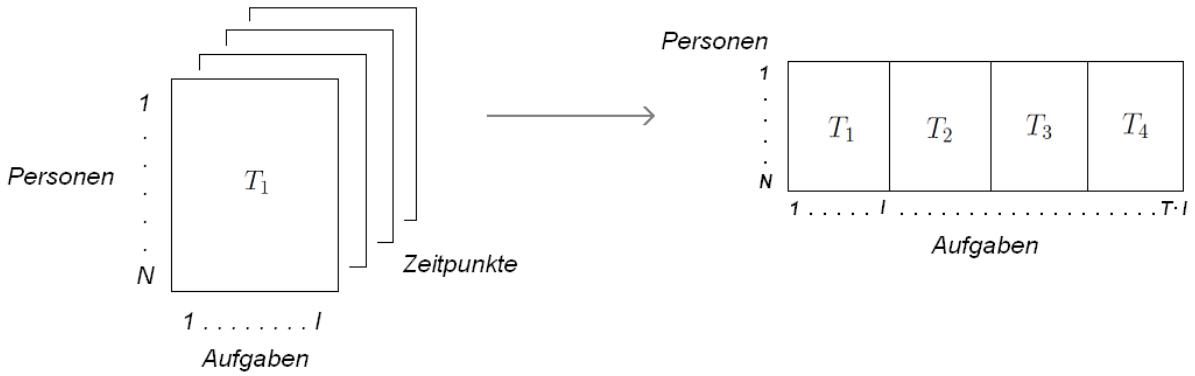

Abbildung 4: Datenstruktur

Gibt es Veränderungen von unterschiedlichem Ausmaß in der Stichprobe, so geht dies mit unterschiedlichen Schwierigkeiten für die Aufgaben zu späteren Zeitpunkten einher. Diese Schwierigkeitsunterschiede werden durch MRM entdeckt. Zur Veranschaulichung soll ein kurzes Beispiel gegeben werden, das sich auf nur zwei Testzeitpunkte beschränkt:

Bekommt ein Teil der Stichprobe nach dem ersten Testzeitpunkt Übungen, so werden den geübten Personen vermutlich die Aufgaben zu einem zweiten Testzeitpunkt leichter fallen als den Personen, die keine Übungen erhielten. Dieser Schwierigkeitsunterschied für geübte und ungeübte Personen zu dem späteren Zeitpunkt wird mittels MRM erkannt (siehe Abbildung 5). Es werden zwei latente Klassen gefunden, für die jeweils ein eigenes Modell mit unterschiedlichen Aufgabenschwierigkeitsparametern angepasst wird. Die eine latente Klasse umfasst (im Idealfall) die Personen, die Übungen erhalten haben, während der anderen latenten Klasse die Personen zugeordnet werden, die keine Übungen erhielten.

Die Schwierigkeitsdifferenzen $\beta_{i|1} - \beta_{2\cdot i|1}$[4] der Aufgaben zum ersten und zweiten Testzeitpunkt sind im Modell für die geübten Personen dann vermutlich größer, da der Lerneffekt bei diesen vermutlich stärker ist als bei den Personen, die keine Übungen erhielten.

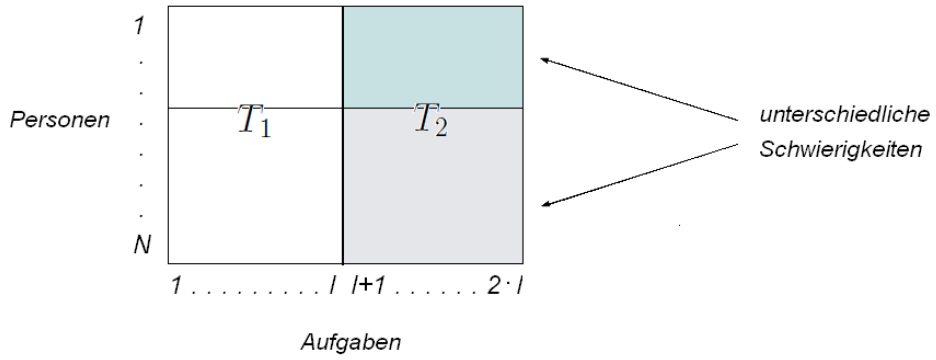

Abbildung 5: verschiedene Schwierigkeiten zum zweiten Testzeitpunkt für geübte und ungeübte Personen

Ein Spezialfall von MRM, der nach diesem Ansatz in der Praxis bereits häufig angewendet wurde, ist das Mover-Stayer Mixed Rasch-Model, welches nun kurz vorgestellt werden soll.

Mover-Stayer Mixed Rasch Models

Das Mover-Stayer Mixed Rasch Model ist ein Spezialfall der MRM mit der Annahme, dass es zwei Klassen gibt, wobei die eine Klasse keiner Veränderung unterliegt ('*Stayer*'), die andere Klasse dagegen schon ('*Mover*'). Es werden demnach unter der Annahme der Existenz von zwei latenten Klassen folgende zwei Restriktionen bezüglich der Veränderungsparameter in (15) für die beiden Klassen eingeführt:

$$\delta_{t|1} = 0 \ \forall t = 1, \cdots, T$$

$$\delta_{t|2} \neq 0 \ \forall t = 1, \cdots, T$$

Die Veränderung $\delta_{t|1}$ in der ersten Klasse nimmt für alle Zeitpunkte den Wert Null an. Es erfolgt

[4] $\beta_{i|1} - \beta_{2\cdot i|1} = \beta_{i|1} - (\beta_{i|1} + \delta_{|1}) = -\delta_{|1}$ entspricht dem Veränderungsparameter

also weder eine Verbesserung noch eine Verschlechterung. Bei dieser Klasse handelt es sich um die 'Stayer'. Die Veränderung $\delta_{t|2}$ in der zweiten Klasse nimmt hingegen einen Wert verschieden von Null an. In der zweiten Klasse der 'Movers' findet also eine Verbesserung bzw. eine Verschlechterung statt.

Anwendungsbeispiele für das Mover-Stayer Mixed Rasch Model finden sich in Meiser et al. (1998) sowie in Meiser and Rudinger (1997).

Ein weiterer Ansatz zur Messung von Veränderung mittels MRM wird von Rost (2004) vorgeschlagen. Bei diesem werden Personen zu allen Zeitpunkten anhand ihrer Testergebnisse den latenten Klassen zugeordnet, für die ihre Zuordnungswahrscheinlichkeit am größten ist. Daraufhin kann die qualitative Entwicklung von Personen anhand eines Wechsels in eine andere latente Klasse ausgemacht werden. Dieser Ansatz soll nun anhand eines Beispiels erläutert werden.

Veränderung bei Klassenwechsel

In dem Ansatz, der von Glück and Spiel (1997) an einem Anwendungsbeispiel beschrieben ist, wird qualitative Veränderung durch einen Wechsel der Klassenzugehörigkeit ausgedrückt. Zunächst sollte geklärt werden, was unter einer qualitativen Veränderung genau zu verstehen ist: Häufig trifft man die Annahme, dass Personen latenten Klassen zugeordnet werden können. Zwischen diesen Klassen gibt es Unterschiede bezüglich der Rangreihe der Aufgabenschwierigkeiten.
So kann beispielsweise für eine latente Klasse Aufgabe 1 als schwieriger empfunden werden als Aufgabe 2, für eine andere latente Klasse wird hingegen Aufgabe 2 als schwieriger empfunden. Wechselt eine Person, von Zeitpunkt T_1 zu Zeitpunkt T_2 in die jeweils andere latente Klasse, so spricht man von einer qualitativen Veränderung oder auch Entwicklung (Abbildung 6).

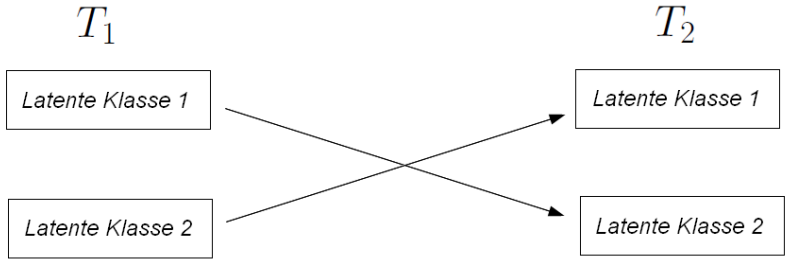

Abbildung 6: Existenz einer qualitativen Veränderung bei einem Klassenwechsel (vereinfachte Darstellung für zwei latente Klassen zu zwei Zeitpunkten)

Die qualitative Veränderung von Fähigkeiten drückt sich also durch eine Veränderung der Rangreihe der Aufgabenschwierigkeiten oder durch die Veränderung der Schwierigkeitsdifferenzen aus[5].

Um qualitative Veränderung durch Klassenwechsel mittels MRM zu untersuchen, muss die dreidimensionale Datenstruktur in eine zweidimensionale überführt werden. Dabei werden die Daten für die insgesamt T Testzeitpunkte so angeordnet, als ob ein Test zu nur einem Zeitpunkt bei $T \cdot N$ verschiedenen Personen durchgeführt wurde. Dies ist in Abbildung 7 veranschaulicht.

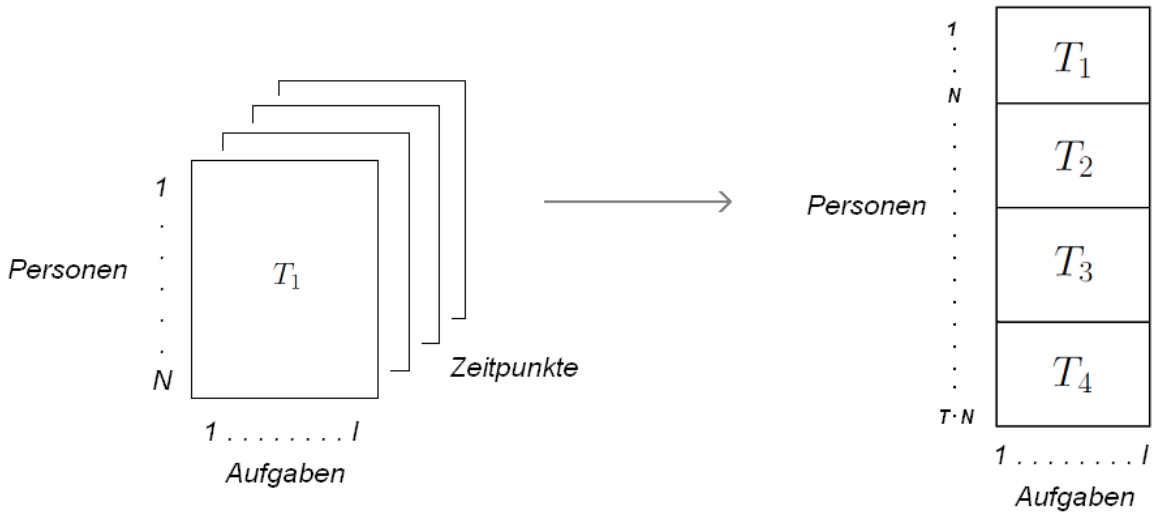

Abbildung 7: Datenstruktur

In einer Untersuchung, die von Glück and Spiel (1997) beschrieben ist, wurden 94 Kindergartenkinder Aufgaben gestellt, anhand derer die Visumotorik der Kinder getestet werden sollte. Diese Aufgaben wurden zu einem späteren Testzeitpunkt wiederholt um die Entwicklung der Visumotorik in Kindesalter schätzen zu können. Mittels MRM sollte herausgefunden werden, ob es qualitative Entwicklungsunterschiede zwischen den Kindern gibt.

Es zeigte sich bei Anwendung von MRM, dass es zwei latente Klassen in der Datenstruktur mit den insgesamt $T \cdot N = 2 \cdot 94$ Beobachtungen gibt, für die die Aufgabenschwierigkeiten differieren. Abbildung 8 stellt die Aufteilung der $2 \cdot 94$ Beobachtungen in die zwei latenten Klassen exemplarisch dar.

[5]Verändern sich die Schwierigkeiten aller Aufgaben in gleichem Ausmaß, so handelt es sich um eine quantitative Veränderung. Diese tritt bei Personen ein, die keinen Klassenwechsel vollziehen.

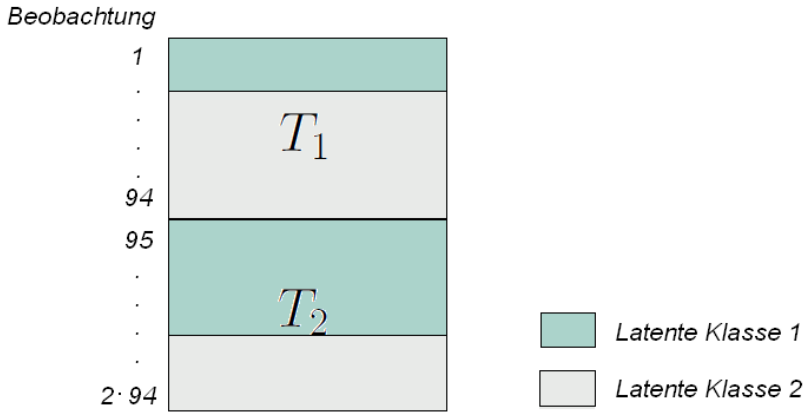

Abbildung 8: Identifizierung von zwei latenten Klassen mit verschiedenen Aufgabenschwierigkeiten in der Untersuchung von Glück and Spiel (1997)

Qualitative Unterschiede in der Entwicklung der Kinder zeigen sich durch einen Klassenwechsel vom ersten zum zweiten Testzeitpunkt in die jeweils andere latente Klasse. Für eine Interpretation der qualitativen Entwicklung ist es nützlich die latenten Klassen genauer zu untersuchen. So stellten Glück and Spiel (1997) fest, dass der erwartete Gesamtpunktwert in der ersten latenten Klasse höher ist als der in der zweiten. Die erste latente Klasse wurde daher als leistungsstärker eingeschätzt. Ein Wechsel von der ersten in die zweite Klasse würde demnach einen Rückschritt in der Entwicklung bedeuten. Umgekehrt kennzeichnet ein Wechsel von der zweiten in die erste Klasse eine Verbesserung der Visumotorik. In folgender Tabelle sind die Ergebnisse der Untersuchung zu den Entwicklungsmustern in der Stichprobe der Kindergartenkinder zusammengefasst.

T_1 \ T_2	Klasse 1	Klasse 2
Klasse 1	Leistungsstark (30)	Rückschritt (1)
Klasse 2	Verbesserung (27)	Leistungsschwach (36)

Tabelle 5: Klassenzugehörigkeiten zu beiden Zeitpunkten und Bedeutung; Anzahl der Kinder in Klammern (Glück and Spiel (1997))

Diskussion

Mixed Rasch Models (MRM) haben sich in mehrerer Hinsicht als sehr hilfreich für die Veränderungsmessung erwiesen:
Mittels MRM können latente Klassen in einer Stichprobe ausfindig gemacht werden, innerhalb derer das Rasch-Modell gilt. Damit gewährleisten sie im Gegensatz zu Modellen, in denen die Gruppeneinteilung a priori aufgrund sachlogischer Überlegungen vorgenommen wird, eine homogene Veränderung innerhalb von Gruppen. Fehlinterpretationen, die aufgrund von Heterogenität innerhalb einer oder mehrerer Gruppen zustande kommen, werden durch MRM ausgeschlossen.
Darüber hinaus können die Ergebnisse von MRM auch für den Vergleich zu einer vordefinierten Gruppeneinteilung dienen (Glück and Spiel (1997)). Stimmen die latenten Gruppen, die mittels MRM ausfindig gemacht wurden, mit den vordefinierten überein, so ist die Homogenität innerhalb der Gruppen abgesichert. Weitere Analysen sowie Hypothesentests können dann durchgeführt werden.
Eine andere Möglichkeit ist es, Untersuchungen mittels MRM innerhalb vordefinierter Gruppen vorzunehmen. So ist es beispielsweise empfehlenswert die Homogenität einer Versuchsgruppe zu überprüfen. Medikationen können für eine Person sehr hilfreich sein, einer anderen Person dagegen Schaden zufügen. Während es bei Anwendung von Modellen mit einem gruppenspezifischem Veränderungsparameter passieren kann, dass der Medikationseffekt durch entgegengesetzte Wirkungen in der Versuchsgruppe überlagert bzw. gänzlich überdeckt wird, können im MRM die unterschiedlichen Reaktionen in der Versuchsgruppe entdeckt und positive und negative Wirkungen durch Aufteilung der Versuchsgruppe getrennt voneinander betrachtet werden.
MRM können also auch zur Überprüfung von Modellannahmen der in Kapitel 2 vorgestellten Modelle herangezogen werden. Jedoch muss beachtet werden, dass die Voraussetzung der Rasch-Konformität innerhalb der latenten Gruppen erfüllt sein muss. Die Rasch-Homogenität innerhalb von Subgruppen ist beispielsweise für das LLRA und das HLLRA bei Anwendung von mehrdimensional messendem Testmaterial nicht gewährleistet. So kann die Überprüfung auf Homogenität innerhalb vordefinierter Gruppen mittels MRM im LLRA und im HLLRA nicht erfolgen, wenn mehrere latente Fähigkeiten untersucht werden sollen. Im LLTM dagegen muss die Forderung von homogenem Testmaterial erfüllt sein, und somit ist es in jedem Fall empfehlenswert, die Homogenität innerhalb vordefinierter Gruppen durch MRM abzusichern.

Allerdings gibt es auch Kritikpunkte bei der Anwendung von MRM. Glück and Spiel (1997) führen an, dass die Entscheidung für eine optimale Anzahl latenter Klassen von dem gewählten Informationskriterium abhängen kann. So könnte die Entscheidung für eine Anzahl latenter Klassen anders ausfallen, je nachdem welches Kriterium gewählt wird.

4 Weitere Modellansätze

In dieser Arbeit wurden die bekanntesten Modelle zur Messung von Veränderung dargestellt, die alle denkbar möglichen Situationen abdecken. Mit dem LLTM, LLRA und dem HLLRA ist es möglich, gruppenspezifische Veränderung zu erfassen und Hypothesentests durchzuführen. Zweifelt man an einer homogenen Entwicklung innerhalb einer oder mehrerer Gruppen oder möchte man latente Subgruppen in einer Stichprobe ausfindig machen, so bietet sich die Anwendung von MRM an. Ist man hingegen an der Veränderung einzelner Personen interessiert, so kann eine individuenspezifische Veränderung mit MRMLC geschätzt werden. Alle bisher vorgestellten Modelle wurden in der Praxis bereits erfolgreich angewendet.
Im Folgenden sollen weitere Modellansätze zur Erfassung von Veränderung angeführt werden, die wegen ihrer Komplexität und Begrenztheit in der Praxis seltener verwendet werden.

Im LLTM sowie im LLRA und im HLLRA wird die Veränderung der Personenfähigkeit durch eine Veränderung in der Aufgabenschwierigkeit erfasst. Die Personenfähigkeiten werden dabei als über die Zeitpunkte hinweg konstant angenommen. Im Gegensatz dazu wird in den Ansätzen von Andersen (1985, 1980), Mislevy (1985) und Glas (1992) die Veränderung durch eine Veränderung der Personenfähigkeiten gemessen. Die drei Autoren gehen dabei davon aus, dass die Fähigkeiten einer Person, die zu verschiedenen Zeitpunkten gemessen wurden, miteinander korreliert sind und einer multivariaten Normalverteilung folgen. Unter Modellierung einer gemeinsamen Verteilung der (korrelierten) Personenfähigkeiten ist es möglich Veränderung durch Differenzbildung der Erwartungswerte von Personenfähigkeiten zu den Zeitpunkten zu messen.

Der einfachste Fall ergibt sich bei zwei Testzeitpunkten mit Wiederholung derselben Aufgaben zu beiden Zeitpunkten. Auf diesen von Andersen (1985, 1980) beschriebenen Ansatz soll im Folgenden näher eingegangen werden.

Andersen setzt voraus, dass alle aufgabenspezifischen Parameter β_i bereits aus Voruntersuchungen bekannt sind. Die Wahrscheinlichkeit, Aufgabe i zum ersten bzw. zweiten Testzeitpunkt richtig zu beantworten, ergibt sich gemäß dem Rasch-Modell unter Bedingung auf die Personenfähigkeit θ_1 bzw. θ_2 (vgl. Andersen (1985)):

$$P(X_{vi1} = 1|\theta_1) = \frac{\exp(\theta_1 + \beta_i)}{1 + \exp(\theta_1 + \beta_i)} \tag{18}$$

$$P(X_{vi2} = 1|\theta_2) = \frac{\exp(\theta_2 + \beta_i)}{1 + \exp(\theta_2 + \beta_i)} \tag{19}$$

Die Aufgabenparameter werden als konstante Werte in (18) und in (19) eingesetzt. Von den Personenfähigkeiten θ_1 zum ersten Testzeitpunkt und θ_2 zum zweiten Testzeitpunkt wird angenommen, dass sie einer bivariaten Normalverteilung folgen. In der marginalen Likelihood ist die Verteilung der Personenfähigkeiten mit den Erwartungswerten μ_1 und μ_2 für den ersten bzw. zweiten Testzeitpunkt und mit der Kovarianz-Matrix Σ enthalten[6], wobei

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \varrho(\theta_1, \theta_2) \\ \varrho(\theta_1, \theta_2) & \sigma_2^2 \end{pmatrix}$$

Mittels iterativer Verfahren, beispielsweise dem EM-Algorithmus (siehe Dempster et al. (1977)), wird die marginale Likelihood (bei gegebenen Aufgabenschwierigkeitsparametern) in Abhängigkeit der unbekannten Verteilungsparameter μ_1, μ_2, σ_1^2, σ_2^2 und $\varrho(\theta_1, \theta_2)$ maximiert.
Die mittlere Veränderung der Personenfähigkeiten, die sich zwischen den beiden Testzeitpunkten ereignet, wird demnach nicht explizit modelliert, sondern ergibt sich aus der Differenz der geschätzten Erwartungswerte $\hat{\mu}_1 - \hat{\mu}_2$.

Andersen geht in seinem Konzept von einer Wiederholung derselben Aufgaben zu zwei Zeitpunkten aus. Dabei müssen die aufgabenspezifischen Parameter aus Voruntersuchungen bekannt sein. Die Wiederholung derselben Aufgaben könnte im Falle von Erinnerungseffekten problematisch sein, da dadurch die lokale stochastische Unabhängigkeit verletzt wird. Andersens Konzept wurde von Glas (1992) bezüglich folgender Aspekte erweitert (Verhelst (1995)):

- Die Aufgaben müssen nicht zu den Testzeitpunkten wiederholt werden.
- Aufgabenspezifische Parameter müssen nicht vorab bekannt sein, sondern können gleichzeitig mit den Verteilungsparametern geschätzt werden.
- Veränderung kann auch für drei Zeitpunkte unter Annahme einer dreidimensionalen Normalverteilung der Personenparameter geschätzt werden.

Es gibt allerdings bedeutende Nachteile[7], die schließlich auch die Tatsache erklären, dass dieser Ansatz zur Messung von Veränderung in der Praxis kaum Verwendung findet, und stattdessen auf Modelle zurückgegriffen wird, die in den Kapiteln 2 und 3 vorgestellt wurden:

[6] Für eine genaue Aufführung und Herleitung der Schätzgleichungen sei auf Andersen (1985, 1980) verwiesen.
[7] vgl. Verhelst (1995)

- Ein Nachteil ergibt sich aus der Komplexität der Schätzung: Für die Parameterschätzung wird die numerische Approximation mehrfacher Integrale benötigt, wobei keine Informationen über die Genauigkeit der Approximation vorliegt.

- Ein weiterer Nachteil dieser Ansätze ist, dass die Testung von Hypothesen nur beschränkt möglich und zudem deutlich umständlicher ist, als die Hypothesentestung für Modelle, die in Kapitel 3 vorgestellt wurden.
  Soll beispielsweise überprüft werden, ob sich Personen mit einem zum ersten Testzeitpunkt hohen Gesamtpunktwert gleich stark weiter entwickeln wie Personen mit einem niedrigen Gesamtpunktwert, so müssen zwei separate Modelle für die beiden Personengruppen aufgestellt werden. Für beide Gruppen sind Verteilungen der Personenfähigkeiten zu spezifizieren. Nun ist es jedoch fragwürdig, ob die Personenfähigkeiten in einer Gruppe mit niedrigen bzw. hohen Gesamtpunktwerten auch einer Normalverteilung folgen. Eine Fehlspezifikation der Verteilung könnte die Ergebnisse verzerren bzw. zu falschen Ergebnissen führen.
  In den in Kapitel 2 und 3 vorgestellten Modellen müssen keine Verteilungsannahmen getroffen werden. Im LLTM, LLRA und im HLLRA kann nach Zuteilung der Personen in die entsprechenden Gruppen die Testung auf eine gleich starke Entwicklung durch einen Vergleich mit dem restringierten Modell erfolgen. In dem restringierten Modell werden die Veränderungsparameter beider Gruppen gleichgesetzt. Ein Likelihood-Quotiententest gibt Aufschluss darüber, ob die Veränderungen beider Gruppen signifikant voneinander abweichen. Modelle wie das LLTM, LLRA oder auch das HLLRA bieten also eine flexible Hypothesentestung, für die keine Verteilungsannahme notwendig ist.

5 Zusammenfassung und Diskussion

Diese Arbeit bietet einen Einblick in den komplexen Bereich der Veränderungsmessung in der Psychometrie. Im Zentrum stehen dabei Modelle, die auf dem dichotomen Rasch-Modell basieren, welches eines der bekanntesten und am häufigsten eingesetzten Latent-Trait-Modelle ist. Neben Methoden der Veränderungsmessung, die auf dem Rasch-Modell basieren, gibt es andere Ansätze zur Modellierung von Veränderungsprozessen, wie zum Beispiel Latent Growth Curve Modelle, Strukturgleichungsmodelle oder auch Markov-Modelle.

In Kapitel 2 wurden vier Modelle zur Messung von Veränderung vorgestellt, mit denen es möglich ist gruppen- bzw. individuenspezifische Veränderung durch ein oder mehrere Parameter im Modell darzustellen und Hypothesentestung mittels Likelihood-Quotiententest durchzuführen. Die ersten drei Modelle basieren auf den Überlegungen von Fischer (1976, 1983, 1989, 1995) und ermöglichen es, gruppenspezifische Veränderung zu messen.

Das Linear Logistic Test Model (LLTM) bietet neben der Messung gruppenspezifischer Veränderung die Möglichkeit individuenspezifische Eigenschaften zum ersten Testzeitpunkt sowie Schwierigkeiten der Items zu schätzen. Dafür müssen jedoch recht strenge Annahmen an die Testbeschaffenheit in Kauf genommen werden: Das Testmaterial muss Rasch-homogen sein, was sich in der Anwendung häufig als eher schwierig erweist. Zudem kann in manchen Anwendungssituationen Veränderung in mehreren latenten Eigenschaften von Interesse sein. So wirkt sich ein Medikament beispielsweise häufig auf mehrere latente Bereiche aus.
Das Linear Logistic Test Model with Relaxed Assumptions (LLRA) bietet die Möglichkeit der Messung von Veränderung in mehreren latenten Eigenschaften durch Zulässigkeit von mehrdimensional messendem Testmaterial. Doch auch hier muss damit ein Nachteil in Kauf genommen werden: Eine Schätzung von individuenspezifischen und aufgabenspezifischen Eigenschaften ist nicht mehr möglich. Darüber hinaus kann die Zulässigkeit von mehrdimensional messendem Testmaterial zur Folge haben, dass Probleme bei der Parameterschätzung auftreten.
Sowohl im LLTM als auch im LLRA werden die Items zu jedem Zeitpunkt wiederholt. In medizinischen Bereichen, wenn beispielsweise eine neue Therapie oder eine Medikation auf ihre Wirksamkeit hin untersucht werden soll, stellt die Wiederholung derselben Items kein Problem dar. Sind allerdings Erinnerungseffekte möglich, beispielsweise in Bereichen der pädagogischen Psychologie, wenn Lernfähigkeiten erfasst werden soll, so empfiehlt sich die Anwendung von Modellen, in denen verschiedene Items zu den Zeitpunkten vorgelegt werden. Ein solches Modell ist das Hybrid LLRA (HLLRA), welches eine Messung von gruppenspezifischer Veränderung in mehreren latenten Eigenschaften erlaubt. Dieses bietet zudem die Möglichkeit einer effiziente-

ren Schätzung der Parameter als im LLRA. Der größte Nachteil dieses Modells ist, dass der Trendeffekt nicht eindeutig bestimmbar ist, falls die aufgabenspezifischen Parameter nicht aus Voruntersuchungen bekannt sind. Neben der Schätzung von aufgabenspezifischen Parametern ist auch die der personenspezifischen Parameter im HLLRA nicht möglich.

Als letztes Modell von Kapitel 2 wird das Multidimensional Rasch Model for Learning and Change (MRMLC) von Embretson (1991a,b) vorgestellt, das Aussagen über Veränderung auf Basis einzelner Individuen erlaubt. Dieses wurde primär für Bereiche der Messung von Lernfähigkeit konzipiert, kann aber auch auf andere Bereiche angewendet werden. Durch Verwendung eines balancierten Designs bei der Testkonstruktion sollen Erinnerungseffekte durch Wiederholung derselben Aufgaben vermieden werden. Dieses Modell ermöglicht neben der Bestimmung von Aufgabenschwierigkeiten eine individuenspezifische Messung von Personenfähigkeit und Veränderung zu mehreren Zeitpunkten. Der wesentliche Unterschied zu den zuvor vorgestellten Modellen ist, dass das MRMLC verwendet wird, wenn nicht etwa die Veränderung ganzer Personengruppen von Interesse ist, sondern wenn man individuenspezifische Aussagen über Veränderung treffen möchte.

Ein weiteres Modell, welches in mehrerer Hinsicht für die Veränderungsmessung von Bedeutung ist, ist das MRM, das in Kapitel 3 vorgestellt wird. Es bietet die Möglichkeit latente Klassen ausfindig zu machen, innerhalb derer die Veränderung homogen ist, die sich untereinander aber in ihrer Veränderung systematisch unterscheiden. So kann es angewendet werden um Veränderungsmuster in einer Stichprobe zu erkennen.
Zudem besitzt das MRM eine modellprüfende Funktion. Bei Anwendung eines LLTM zur Messung von Veränderung liegen gewöhnlich Hypothesen über die Gruppenzugehörigkeit von Personen in der Stichprobe vor. Die Veränderung innerhalb dieser Gruppen wird als gleich angenommen. Falls es allerdings heterogene Entwicklungen innerhalb einer oder mehrerer Gruppen gibt, könnte es zu Fehlinterpretationen kommen. Das MRM bietet eine Möglichkeit heterogene Entwicklungen innerhalb von Gruppen ausfindig zu machen und damit die Richtigkeit der Homogenität innerhalb von Gruppen abzusichern.

Ein weiterer Ansatz zur Messung von Veränderung wird in Kapitel 4 vorgestellt. In diesem Ansatz wird die mittlere Veränderung einer Personengruppe durch die Differenz der mittleren Personenfähigkeiten zu den Zeitpunkten geschätzt. Dabei wird von einer multivariaten Normalverteilung der Fähigkeitsparameter ausgegangen. Durch Aufstellung der marginalen Likelihood können die Verteilungsparameter, insbesondere die Erwartungswerte geschätzt werden. Dieser Ansatz wird aufgrund der komplexen Schätzung und der begrenzten Hypothesentestung in der Praxis kaum verwendet. Meist wird daher auf die zuerst vorgestellten Modelle zurückgegriffen, die einfacher zu handhaben sind und eine umfangreiche Hypothesentestung erlauben.

Abschließend soll folgende Tabelle einen Überblick über die wichtigsten Anwendungsaspekte der in den Kapiteln 2 und 3 vorgestellten Modelle geben. Weil sie auf ähnlichen Grundlagen basieren, werden die drei Modelle LLTM, LLRA und HLLRA zusammengefasst. Für eine genauere Differenzierung zwischen dem LLTM, LLRA und dem HLLRA wird auf den Überblick auf Seite 33 verwiesen.

	LLTM/LLRA/HLLRA	MRMLC	MRM
Verwendung	Aussagen über Veränderung und Fähigkeiten vordefinierter Gruppen	Aussagen über Veränderung und Fähigkeiten einzelner Individuen	• Identifizierung von Gruppen mit homogener Veränderung • Überprüfung von homogener Veränderung innerhalb vordefinierter Gruppen
Veränderung	Gruppenspezifische Veränderungseffekte mit Trend	Individuenspezifische Veränderung	Erkennung verschiedener Veränderungsmuster innerhalb einer Stichprobe
Voraussetzungen	Rasch-homogenes Testmaterial Rasch-homogene Hybrid-Items Kenntnisse über Gruppenzugehörigkeit der Personen	Rasch-homogenes Testmaterial	Annahme der Rasch-Homogenität innerhalb latenter Gruppen

Tabelle 6: Überblick über die wichtigsten Anwendungsaspekte der Modelle

Ziel dieser Arbeit war es, verschiedene Modelle und Methoden aufzuzeigen, wie Veränderung latenter Eigenschaften erfasst und geschätzt werden kann. Es hat sich gezeigt, dass eine Bewertung, welches Modell 'besser' oder 'schlechter' ist, nicht möglich ist. Die Entscheidung für ein Modell hängt von der konkreten Anwendungssituation ab. Ein Modell kann geeigneter sein als ein anderes, je nachdem welche Ziele man verfolgt, welche Voraussetzungen erfüllt sind und welche Probleme auftreten könnten. Tabelle 6 gibt eine Hilfestellung zur Entscheidungsfindung.

Oftmals kann es von Nutzen sein, mehrere der hier vorgestellten Modelle anzuwenden um zusätzliche Informationen zu gewinnen oder auch um Annahmen zu überprüfen. So ist bei Verwendung eines LLTM die gleichzeitige Anwendung eines MRM zur Überprüfung der homogenen Veränderung innerhalb von Gruppen empfehlenswert. Oftmals können sich die Ergebnisse, die durch Anwendung verschiedener Modelle gewonnen wurden, ergänzen und tiefere Einblicke in die Veränderungsstruktur der gesamten Stichprobe geben.

Literatur

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

Andersen, E. (1980). Comparing latent distributions, *Psychometrika* **45**(1): 121–134.

Andersen, E. (1985). Estimating latent correlations between repeated testings, *Psychometrika* **50**(1): 3–16.

Dempster, A., Laird, N., Rubin, D. et al. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1): 1–38.

Embretson, S. (1991a). A multidimensional latent trait model for measuring learning and change, *Psychometrika* **56**(3): 495–515.

Embretson, S. (1991b). Implications of a multidimensional latent trait model for measuring change, *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions. Washington, DC: American Psychological Association* pp. 184–197.

Fischer, G. (1976). Some probabilistic models for measuring change, *Advances in psychological and educational measurement* pp. 97–110.

Fischer, G. (1977). Linear logistic trait models: Theory and application, *Structural models of thinking and learning* pp. 203–225.

Fischer, G. (1983). Logistic latent trait models with linear constraints, *Psychometrika* **48**(1): 3–26.

Fischer, G. (1989). An IRT-based model for dichotomous longitudinal data, *Psychometrika* **54**(4): 599–624.

Fischer, G. (1995). Linear Logistic Test Models for Change, *Rasch models: Foundations, recent developments, and applications* pp. 157–180.

Fischer, G. and Molenaar, I. (1995). *Rasch models: Foundations, recent developments, and applications*, Springer.

Formann, A. and Spiel, C. (1989). Measuring change by means of a hybrid variant of the linear logistic model with relaxed assumptions, *Applied Psychological Measurement* **13**(1): 91.

Glas, C. (1992). A Rasch model with a multivariate distribution of ability, *Objective measurement: Theory into practice* **1**: 236–258.

Glück, J. and Spiel, C. (1997). Item response models for repeated measures designs: Application and limitations of four different approaches, *Methods of Psychological Research Online* **2**(1): 1–18.

Hoijtink, H. (1995). Linear and Repeated Measures Models for the Person Parameters, *Rasch models: Foundations, recent developments, and applications* pp. 203–214.

Meiser, T. (2007). Rasch Models for Longitudinal Data, *Multivariate and mixture distribution Rasch models: Extensions and applications* .

Meiser, T., Hein-Eggers, M., Rompe, P. and Rudinger, G. (1995). Analyzing homogeneity and heterogeneity of change using Rasch and latent class models: A comparative and integrative approach, *Applied Psychological Measurement* **19**(4): 377.

Meiser, T. and Rudinger, G. (1997). Modeling stability and regularity of change: Latent structure analysis of longitudinal discrete data, *Applications of Latent Trait and Latent Class Models in the Social Sciences, Waxmann, Munster* pp. 389–398.

Meiser, T., Stern, E. and Langeheine, R. (1998). Latent change in discrete data: Unidimensional, multidimensional, and mixture distribution Rasch models for the analysis of repeated observations, *Methods of Psychological Research Online* **3**(2).

Mislevy, R. (1985). Estimation of latent group effects, *Journal of the American Statistical Association* **80**(392): 993–997.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis, *Applied Psychological Measurement* **14**(3): 271.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses, *British Journal of Mathematical and Statistical Psychology* **44**(1): 75–92.

Rost, J. (2004). Lehrbuch Testtheorie, Testkonstruktion.

Rost, J. and Von Davier, M. (1995). Polytomous mixed Rasch models, *Mixture Distribution Rasch Models* pp. 257–268.

Spiel, C. and Glück, J. (1998). Item response models for assessing change in dichotomous items, *International Journal of Behavioral Development* **22**(3): 517.

Strobl, C. (2010). Das Rasch-Modell - Eine verständliche Einführung für Studium und Praxis, *in* M. Spieß (ed.), *Sozialwissenschaftliche Forschungsmethoden*, Rainer Hampp Verlag.

Verhelst, N.D. Glas, C. (1995). Dynamic Generalizations of the Rasch Model, *Rasch models: Foundations, recent developments, and applications* pp. 181–201.

Von Davier, M. and Carstensen, C. (2007). *Multivariate and mixture distribution Rasch models: Extensions and applications*, Springer Verlag.

Von Davier, M. and Rost, J. (1995). Polytomous mixed Rasch models, *Rasch models: Foundations, recent developments, and applications* pp. 371–379.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt, noch nicht einer anderen Prüfungsbehörde vorgelegt und noch nicht veröffentlicht habe.

München, den 14. Juli 2010

Silke Janitza