Ludwig-Maximilians-Universität

Institut für Statistik

Das Partial Credit Modell

Bachelorarbeit

Autor: Anna Theresa Stüber
Betreuung: Prof. Dr. Gerhard Tutz
Dr. Gunther Schauberger

München, 25. Juli 2016

Abstract

Die Item Response Theorie stellt einen Teilbereich der (psychologischen) Testtheorie dar und verfolgt das Ziel mittels wissenschaftlicher Methoden auf die einer Person eigenen Ausprägung eines latenten Merkmales zu schließen. Dieser sind eine Reihe probabilistischer Modelle untergeordnet, mittels derer das eben erwähnte Ziel erreicht werden soll. Eines eben dieser Modelle ist das Partial Credit Modell von Masters, ebenso wie das allgemein gehaltener Generalisierte Partial Credit Modell von Muraki, welche im Rahmen dieser Arbeit vorgestellt werden sollen. Basierend auf dem dichotomen Rasch-Modell begründet sich die Entstehung des ersteren in der Anwendbarkeit auf ordinale Antwortformate.

Prinzipiell wird bei den beiden Modellen die bedingte Wahrscheinlichkeit für das Erreichen eines bestimmten Leistungslevels in Abhängigkeit von zwei Parametern - der Personenfähigkeit und der Aufgabenschwierigkeit - modelliert. Hierbei findet aufgrund des vorliegenden ordinalen Datenformates eine Untergliederung der Aufgabenschwierigkeit in Schwellenparameter statt, sodass für die einzelnen Items eine Gliederung in aufsteigend geordnete Teilkategorien berücksichtigt werden kann. Gleichzeitig lässt das Generalisierte Partial Credit Modell eine flexiblere Modellierung der Wahrscheinlichkeitsfunktion als das Partial Credit Modell zu, indem es neben den Personen- und Schwellenparametern einen Diskriminations-/Steigungsparameter beinhaltet. Damit werden beim Generalisierten Partial Credit Modell prinzipiell zwei Itemparameter betrachtet: die Schwellen-, sowie die eben aufgeführten Steigungsparameter.

Eine zentrale Rolle bei derartigen probabilistsichen Modellen spielt auch das Wissen über diesen eigenen Charakteristiken und daher werden diese Merkmale in einem gesonderten Abschnitt näher erläutert. Für das Partial Credit Modell lassen sich die grundlegenden Annahmen und Eigenschaften durch entsprechende Anpassung vom Rasch-Modell, welches als Konstruktionsgrundlage gilt, ableiten. Dagegen können für das Generalisierte Partial Credit Modell diese Annahmen wegen der allgemein gehalteneren Form nicht mehr im vollen Umfang übertragen werden.

Zur konkreten Schätzung der interessierenden Parameter gibt es verschiedene Möglichkeiten. Hierbei werden drei auf der Maximum-Likelihood-Methode basierende Verfahren zur Itemparameterschätzung für das Partial Credit Modell dargeboten, während sich für das Generalisierte Partial Credit Modell nur eins derer anbietet. Je nach verwendeter Methode können die Personenparameter auf unterschiedliche Weisen bestimmt werden. Diese können in Anbetracht der vorhergehenden Itemparameterschätzung stets auf Grundlage der Maximum-Likelihood-Methode ermittelt werden, teilweise stehen aber auch bayessche Verfahren zur Verfügung.

Die konkrete Durchführung der Parameterschätzung kann mit der vielseitig genutzten Statistik-Software R bewerkstelligt werden. Hierzu werden drei implementierte Pakete vorgestellt, die sich für die Analyse von Daten des Partial Credit Formates eignen. Zusätzlich werden die Pakete für die explizite Analyse eines ausgewählten Datensatzes mittels des Partial Credit, sowie des Generalisierten Partial Credit Modells genutzt.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

IRT	Item Response Theorie
KTT	Klassische Testtheorie
NOM	Normal Ogive Modell
RM	Rasch Modell
PCM	Partial Credit Modell
GPCM	Generalisiertes Partial Credit Modell
RSM	Rating Scale Modell
GRM	Graded Response Modell
ICC	Item Characteristic Curve (Itemspezifische Kurven)
ICCC	Item Category Characteristic Curve
DIF	Differential Item Functioning
IRF	Item-Response-Funktion
ICRF	Item-Category-Response-Funktion
ML	Maximum-Likelihood
JML	Joint/Gemeinsame Maximum-Likelihood
CML	Conditional/Bedingte Maximum-Liklihood
MML	Marginale Maximum-Likelihood
WML	Weighted/Gewichtete Maximum-Likelihood
EAP	Expected A Posteriori
MAP	Maximum A Posteriori
LQ	Likelihood-Quotient
BFGS	Broyden-Fletcher-Goldfarb-Shanno
LLTM	Linear-logistisches Test-Modell
LPCM	Lineares Partial Credit Modell
RCMLM	Random Coefficients Multinomial Logit Modell
FBL	Freiburger Beschwerdenliste
FBL-R	Freiburger Beschwerdenliste, revidiert
FBL-R-ALL	FBL-R-Skala des Allgemeinbefindens
FBL-R-EMO	FBL-R-Skala der Emotionalen Reaktivität

Notation

Buchstabe	Bedeutung
α_i	(Itemspezifischer) Steigungs-/Diskriminationsparameter
β_i	Bei RM: Schwierigkeitsparameter/Lokationsparameter des i-ten Items
β_{ik}, β_{ix}	Bei PCM/GPCM: Schwellen-/Schwierigkeits-/Itemparameter der Aufgabe i zur betrachteten k Kategorien; bei Betrachtung eines bestimmten Scores von x ist Schwierigkeit bezogen auf die entsprechende x-te Kategorie
γ_i	(Itemspezifischer) Rateparameter
δ_{ik}	(Vorläufiger) Schwierigkeitsparameter bezogen auf das Erreichen von Level k bei Aufgabe i
θ_p	Personenparameter der Person p
ν_k	Kategorien-Schwellenparameter der Kategorie k
η	Linearer Prädiktor
Λ, λ	Likelihood- bzw. Log-Likelihood-Funktion
τ_{ik}, τ_{ix}	Summierte Schwellenparameter: Summe über alle Schwellenparameter des Items i die kleiner oder gleich der Kategorie k bzw. dem Score x sind
π_{pik}, π_{pix}	Bedingte (Kategorien-)Wahrscheinlichkeit für Person p Level k bzw. Score x bei Item i zu erreichen
Φ_{pix}	Schwellenwahrscheinlichkeit der Person p bei Item i von Score $x-1$ auf x zu gelangen
m_i	Anzahl der Stufen des Items i
r_p	Gesamt/Testscore bzw. Summe aller von einer Person p bewältigten Stufen bei einem Test mit I Items
s_{ik}, s_{ix}	Anzahl der Personen, die Level k bzw. Score x bei Item i erreichen konnten
x, x_{pi}	Score/absolvierte Stufen; x_{pi} bei Betrachtung des Scores einer bestimmten Person p bei Item i
X_{pi}	Zufallsvariable mit möglichen Realisationen x_{pi}
K_k, K_x	Kategorienkoeffizient der Kategorie k bzw. x
T_k, T_x	Scoring-Funktion

Vektor/Matrix (fett gedruckt)	Bedeutung
$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)$	Vektor mit allen I Diskriminationsparametern
$\boldsymbol{\beta_i} = (\beta_{i1}, \ldots, \beta_{Im_i})$	Vektor mit allen m_i Schwellenparametern des Items i
$\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{1m_1}, \ldots, \beta_{I1}, \ldots, \beta_{Im_I})$	Vektor mit allen m_i Schwellenparametern aller I Items
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)$	Vektor mit Fähigkeitsparametern aller P Personen
$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_L)$	Vektor mit allen L Basisparametern
$\boldsymbol{\tau} = (\tau_{11}, \ldots, \tau_{1m_1}, \ldots, \tau_{I1}, \ldots, \tau_{Im_I})$	Vektor mit jeweils bis zu Kategorie k aufsummierten Schwellenparametern aller I Items
$\boldsymbol{\phi}$	Set von Populationsparametern
$\boldsymbol{r} = (r_1, \ldots, r_P)$	Vektor mit Test-Scores aller P Personen
$\boldsymbol{s_i} = (s_{i1}, \ldots, s_{im_i})$	Vektor mit allen m_i Spaltenrandsummen des Items i
$\boldsymbol{s} = (s_{11}, \ldots, s_{1m_1}, \ldots, s_{I1}, \ldots, s_{Im_1})$	Vektor mit allen m_i Spaltenrandsummen der I Items
$\boldsymbol{x}_p = (x_{p1}, \ldots, x_{pI})$	Vektor mit Scores x_{pi} aller I Items der Person p
$\boldsymbol{x}_i = (x_{1i}, \ldots, x_{Pi})$	Vektor mit Scores x_{pi} aller P Personen zum Item i
$\boldsymbol{X}_{P \times I}$	Antwortmatrix (P Zeilen, I Spalten) mit jeweiligen Einträgen x_{pi}

Indizes	Bedeutung
$b = 1, \ldots, B$	Bootstrap-Stichprobe
$h = 1, \ldots, H$	Subpopulation
$i = 1, \ldots, I$	Items
$j = 0, \ldots, k$	Bewältigte Stufe bei Item i; ggf. x statt k, wenn bestimmter Score betrachtet wird
$k = 0, \ldots, m_i$	Kategorien, Levels
$l = 1, \ldots, L$	Einzelne Kategorie aus allen möglichen Kategorien von I Items
$p = 1, \ldots, P$	Person

1 Einleitung

Seit jeher verfolgt die Menschheit das Interesse die eigenen Kräfte mit denen von Artgenossen zu messen, was sicherlich im hohen Maße auf biologische Einflussfaktoren zurückzuführen ist. Die Macht des Stärkeren war bzw. ist vielerlei entscheidend, etwa bei der Austragung von Eroberungskriegen oder auch auf individueller Ebene, wie z.B. bei der Teilnahme an Wettkämpfen. So lassen sich hierzu beispielhaft die in der Antike beginnenden Olympischen Spiele oder die mittelalterlichen Ritterturniere oder neuzeitliche Autorennen erwähnen. Dabei kommt der Gewinner i.A. als der stärkste oder der schnellste oder kräftigste Teilnehmer zu Tage.

Doch nicht nur auf kriegstechnischer oder sportlicher Ebene kam bzw. kommt es zu Leistungsmessungen. Auch geistige Fähigkeiten sind von großer Bedeutung und stehen teilweise auch in starker Wechselwirkung mit dem physischen Potenzial. Allerdings sind diese psychischen Leistungsparameter nicht einfach mit einer Stoppuhr oder einem Metermaß messbar. Da es sich also um nicht direkt beobachtbare Eigenschaften handelt, spricht man hier auch von latenten Merkmalen. So wurden zahlreiche Möglichkeiten vorgestellt, geistige bzw. generelle psychische Parameter zu quantifizieren. Damit war die Grundidee psychologischer Tests entwickelt: mithilfe wissenschaftlicher Methoden sollen quantitative Aussagen über den relativen Grad der individuellen Merkmalsausprägung möglich gemacht werden (vgl. Lienert, 1998).

Heutzutage haben psychologische Tests ein breites Anwendungsgebiet und werden in allerlei Bereichen, wie etwa innerhalb der Arbeitswelt, in Schulen und Kliniken verwendet. Dabei dienen diese Tests beispielsweise zur Erfassung der Intelligenz oder Leistungsfähigkeit einer Person, zur Ermittlung des Entwicklungsgrades eines Kindes oder zur Feststellung der Eignung eines Job-Bewerbers. Aber wie kann man derartige psychologische Phänomene denn nun überhaupt messen? Mit u.a. dieser Frage beschäftigt sich die psychologische Testtheorie, bei der - je nach zugrundeliegender Datensituation - unterschiedliche Testmodelle zur Verfügung stehen. Im Wesentlichen unterscheidet man zwischen zwei Arten von Testtheorien: der klassischen Testtheorie (KTT) und der Item Response Theorie (IRT). Zu letzterem lässt sich auch das s.g. Partial Credit Modell (PCM) zuordnen, welches den Schwerpunkt dieser Arbeit darstellt.

Um dieses Modell bestmöglich veranschaulichen zu können, wird in Kapitel 2 zunächst die IRT genauer vorgestellt. Im darauffolgenden Kapitel 3 wird kurz auf - das wohl bekannteste Modell der IRT - das Rasch-Modell (RM) näher eingegangen, da sich das PCM weitestgehend von diesem ableiten lässt. Im daran anschließenden Kapitel 4 kommt dann das PCM zur Sprache. Dabei wird sowohl auf die Modellgleichung und grafische Darstellung, als auch auf geltenden Modellannahmen/ -eigenschaften eingegangen. Zudem wird die Verallgemeinerung des PCMs, nämlich das Generalisierte Partial Credit Modell (GPCM), vorgestellt. In Kapitel 5 werden insbesondere drei mögliche Verfahren zur Parameterschätzung im PCM bzw. GPCM dargelegt, welche auf der s.g. Maximum-Likelihood-Methodik basieren. In diesem Zusammenhang wird auch dargelegt, wie die latente Fähigkeit einer Person geschätzt werden kann, wobei diese Art der Schätzungen zumeist auf bayesschen Verfahren beruhen. Anschließend werden in Kapitel 6 drei ausgewählte Pakete, die in der Statistik-Software R implementiert sind, vorgestellt und insbesondere auf die hier zugrundeliegenden Parameterschätzungen eingegangen. Im Kapitel 7 wird schließlich noch das PCM

bzw. GPCM konkret zur Anwendung gebracht, indem die Modelle zur Analyse des Datensatzes der s.g. „Freiburger Beschwerdenliste“ (FBL) genutzt werden. Abschließend werden in Kapitel 8 noch einmal die wichtigsten Punkte dieser Arbeit zusammengefasst und auch die Auswertung der FBL kritisch beurteilt.

Allgemein stützen sich die hier durchgeführten Analysen, sowie die graphischen Darstellungen, auf die Software R [R Core Team, 2013]. Die beigefügte CD-ROM beinhaltet alle zur Veranschaulichung erzeugten Grafiken und vorgenommenen Auswertungen. Eine Übersicht hierzu findet sich am Ende dieser Arbeit.

Gleichzeitig sind bei mathematischen Formeln und Ausdrücken Vektoren klein und fett geschrieben, während Matrizen groß und fett gekennzeichnet sind, um diese von Skalaren, sowie Funktionen unterscheiden zu können. Konkret sind alle in dieser Arbeit verwendeten Buchstaben, Vektoren, Matrizen und Indizes unter dem Oberbegriff „Notation“ aufgelistet, sowie kurz erklärt.

Zunächst werden in der nun anschließenden Sektion der Begriff der IRT, sowie die hierin inbegriffenen statistischen Modelle näher erläutert.

2 Item Response Theorie (IRT)

In diesem Kapitel soll genauer auf die bereits in der Einleitung erwähnte Item Response Theorie (IRT) eingegangen werden. Dieser lassen sich auch die in den darauffolgenden Sektionen detailliert beschriebene Modelle - das Rasch-Modell (RM), das Partial Credit Modell (PCM) und auch das Generalisierte Partial Credit Modell (GPCM) - zuordnen. Die hier nun folgenden Darlegungen zur IRT basieren im Wesentlichen auf Rost (1996, Kapitel 1 und 2) und Becker (2004, Kapitel 3).

Um schließlich auf eben angesprochene IRT genauer eingehen zu können, soll zunächst etwas weiter ausgeholt werden und die Begrifflichkeit einer Testtheorie erfasst werden. Hierbei kennzeichnet der Ausdruck „Test“ nicht etwa das statistische Testen/Prüfen einer Hypothese mittels Stichprobendaten. Vielmehr bezieht sich der Begriff in dieser Arbeit auf psychologische Tests, mithilfe derer psychische und damit nicht direkt messbare Eigenschaften einer Person erfasst werden sollen. Diese Erfassung geschieht mittels s.g. Items. Rost (1996, S. 18) definiert diese wie folgt: „Als Item [...] bezeichnet man die Bestandteile eines Tests, die eine Reaktion oder Antwort hervorrufen sollen, also die Fragen, Antworten, Bilder etc. Wenn auch die Items von Test zu Test sehr unterschiedlich aussehen können, sind sie innerhalb eines Tests sehr ähnlich (homogen), da sie dasselbe Merkmal der Person ansprechen.“

Um nun von einem derartigen psychologischen Test bzw. von den in diesem enthaltenen Items Rückschlüsse auf das zu erfassende Personenmerkmal ziehen zu können, unterscheidet man grundsätzlich zwei Arten von Testtheorien: zum einen gibt es die s.g. „Klassische Testtheorie“ (KTT) und zum anderen eben die „Item Response Theorie“ (IRT). Erstere stellt die ältere der beiden Testtheorien dar und basiert im Wesentlichen auf der Annahme, dass der durch einen Test ermittelte Wert der Merkmalsausprägung einer Person sich aus dem „wahren“ Wert und einem zufälligen Messfehler ergibt. Es wird hierbei davon ausgegangen, dass - bei gleichzeitiger Realisierung eines Sets von Axiomen (siehe dazu Gulliksen, 1950 oder Novick, 1966) - der wahre Wert des Merkmals anhand von Messwiederholungen approximiert werden kann. Trotz der weit verbreiteten Anwendung der KTT sind bei dieser einige Schwächen zu vermerken. Insbesondere sollte hier die Stichprobenabhängigkeit erwähnt werden. Genauer wird in dieser Arbeit allerdings nicht auf die KTT und deren Axiome eingegangen.

Hauptaugenmerk soll nun stattdessen auf die IRT geworfen werden, die nicht wie KTT den Test als Ganzen fokussiert, sondern vielmehr die einzelnen Items eines Tests. Zugleich gilt für die IRT, dass sie im Gegensatz zur KTT - welche als eine direkte Messung angesehen wird - eine indirekte Art der Messung darstellt. „Indirekt“ in dem Sinne, dass die Antwort auf ein Item nicht als direkte Merkmalsausprägung geachtet wird, sondern lediglich als Indikator, um auf eben diese interessierende Ausprägung schließen zu können. Genauer gesagt postuliert die IRT also, dass ein Persönlichkeitsmerkmal das Antwortverhalten einer befragten Person steuert. Da dieses Persönlichkeitsmerkmal/Charakterzug nicht direkt messbar und dementsprechend „versteckt“ ist, wird die vorgestellte Theorie oftmals auch als „Latent Trait“ Theorie bezeichnet. Darüber hinaus wird zudem häufig der Begriff „Probabilistische Testtheorie“ verwendet. Hieran wird der Grundgedanke der IRT ersichtlich: die Modellierung der Wahrscheinlichkeit - im Englischen „probability“ - für eine bestimmte Antwort in Abhängigkeit gewisser zu messender Parameter. Dabei sind stets der s.g. Per-

sonenparameter, welcher als Maß für die Fähigkeit einer Person steht, und der Lokationsparameter, welcher die Schwierigkeit (eines Abschnittes) einer Aufgabe in Bezug zum Personenparameter setzt, inbegriffen. Zusätzlich können im Rahmen der IRT noch bis zu zwei weitere Parameter Berücksichtigung finden. Auf diesen Aspekt wird gegen Ende des Kapitels noch einmal zurückgegriffen. Es sollte noch einmal deutlich gemacht werden, dass die einzelnen Parameter kennzeichnende Größen darstellen, die allerdings nicht direkt messbar sind, sondern geschätzt werden müssen.

Bei der IRT ist es nun prinzipiell von Bedeutung die zugrundeliegenden Antwortformate zu unterschieden. Demzufolge ist es möglich eine Unterteilung anhand der folgenden Eigenschaften vorzunehmen:

1. Gibt es nur zwei Antwortmöglichkeiten (dichotom) oder stehen mehrere Kategorien zur Auswahl (polytom)?

2. Können die jeweiligen Kategorien einer Aufgabe angeordnet werden (ordinal) oder besitzen sie keine natürliche Ordnung (nominal)?

Es ist wichtig die vorliegenden Items einem Antwortformat zuordnen zu können, da in Abhängigkeit der Antwortformate eines Tests unterschiedliche Modelle zur Anwendung kommen können. So ist an dieser Stelle nun anzufügen, dass die IRT nicht ein bestimmtes Modell beschreibt, sondern eine Vielzahl verschiedener Modelle einschließt. Anders ausgedrückt „[...] ist die IRT nicht eine einzelne Theorie, sondern umfasst eine Familie von formalen, mathematischen, probabilistischen Messmodellen, welche postulieren, dass dem beobachtbaren Testverhalten (manifeste Variable) eine Fähigkeit/Eigenschaft bzw. Disposition (latente Variable) zugrunde liegt, die das Testverhalten 'steuert'“ (Rost, 1978, S. 60).

Im Laufe der Geschichte wurden einige Modelle konzipiert, die sich explizit der IRT zuordnen lassen. So begann die Entwicklung der IRT-Modelle bereits in den 40er/50er Jahren mit Vertretern wie Lord, wobei dessen Analysen auf dem s.g. „Normal Ogive“ Modell (NOM) basieren. Die später konstruierten IRT-Modelle bauen allerdings vorwiegend auf der logistische Verteilungsfunktion auf. Diese fand erstmals bei Rasch (1960) mit der Einführung des Rasch-Modells (RM) Verwendung und damit legte er den Grundstein für die Konstruktion weiterer IRT-Modelle auf Basis des Logit-Modells. So konstruierte auch Birnbaum (1968) zwei Modelle für dichotome Antwortformate, die als direkte Erweiterungen des RMs angesehen werden können. Diese sind das 2PL- und 3PL-Modell, welche im nachfolgenden Kapitel 3 noch einmal zur Sprache kommen. Auf der durch das RM vermittelten Basis wurden schließlich in den 80er Jahren auch Modelle entwickelt, die sich auf polytome Antwortformate anwenden lassen. Hier sind etwa das Graded Response Modell (GRM) von Samejima, das Rating Scale Modell (RSM) von Andrich und das Partial Credit Modell (PCM) von Masters zu nennen.

Wie anhand der geschichtlichen Einordnung der IRT-Modelle im vorhergehenden Absatz ersichtlich werde konnte, spielt in den heute vorwiegend verwendeten IRT-Modellen das logistische Regressionsmodell eine zentrale Rolle. Allgemein lässt sich dieses anhand der logistischen Responsefunktion oder der Logit-Linkfunktion folgendermaßen darstellen:

$$h(\eta) = \pi = \frac{e^{\eta}}{1+e^{\eta}} = \frac{exp(\eta)}{1+exp(\eta)} \quad \leftrightarrow \quad g(\pi) = h^{-1}(\eta) = log(\frac{\pi}{1-\pi}) = \eta \tag{2.1}$$

Hieran wird ersichtlich, dass im Gegensatz zur linearen Regression nicht der konkrete Wert einer Responsevariable modelliert wird, sondern die Wahrscheinlichkeit für das konkrete Eintreten einer Ausprägung der Responsevariable. In Anlehnung an Fahrmeir (2009, S. 190f.) können also die nachfolgende Tatsachen berücksichtigt werden. So wird die Wahrscheinlichkeit π durch die Responsefunktion $h(\eta)$ mit dem linearen Prädiktor η verknüpft. Die Responsefunktion - auch Antwortfunktion genannt - stellt dabei auf der ganzen reelen Achse streng monoton wachsende Funktion mit $h(\eta) \in [0, 1]$, $\forall \eta \in \mathbb{R}$. Gleichzeitig wird durch $g = h^{-1}$ die Umkehrfunktion oder Inverse der Responsefunktion repräsentiert und wird spezifisch auch als Linkfunktion bezeichnet. Die Darstellung in Form von $log[\pi/(1 - \pi)]$ ergibt die s.g. logarithmierten Chancen oder kurz Logits. Zudem wird für η das Logit-Modell als Verteilungsfunktion angenommen.

Insgesamt dient Gl. (2.1) als Grundlage zur Konstruktion beinahe aller IRT-Modelle, wobei hier die jeweils interessierende Lösungswahrscheinlichkeit einer Aufgabe in Abhängigkeit von Personen- und Itemparameter(n) modelliert wird. Das sich jeweils im Exponenten befindende η wird entsprechend dem jeweiligen IRT-Modell durch eine spezifische Kombination eben dieser Personen- und Itemparameter ersetzt. Anzumerken ist, dass konkret von allen IRT-Modellen ein Personenparameter pro Proband postuliert wird. Allerdings bedient man sich je nach verwendeten Modell unterschiedlich vieler Itemparameter. Dieser Sachverhalt wird in einem späteren Abschnitt jener Sektion noch einmal aufgegriffen und eingehend erklärt. Das Logit-Modell als Grundlage zur Berechnung von Wahrscheinlichkeiten zu nehmen, ist insbesondere auch deshalb von Vorteil, da der Wertebereich zwischen 0 und 1 liegt und sich zudem auch der Aspekt der strengen Monotonie ausnutzen lässt. An diese Tatsachen wird im Verlauf der Arbeit wiederholt angeknüpft und diese werden auch noch eingehender erklärt.

Explizit sollte hier nun eine Besonderheit aller IRT-Modelle aufgeführt werden: die s.g. „Invarianz Eigenschaft“, entsprechend der die Item- und Personenparameter stichprobenunabhängig sind (vgl. Hambleton, 1991, S. 18) und demzufolge diese Parameter unabhängig voneinander geschätzt werden können. Damit bieten die Modelle der IRT auch einen entscheidenden Vorteil gegenüber der KTT, die stichprobenabhängig ist.

Trotz alledem muss berücksichtigt werden, dass je nach verwendetem IRT-Modell und damit aufgrund unterschiedlicher mathematischer Modellannahmen verschiedene Voraussetzungen erfüllt sein müssen. Dabei gibt es sowohl zentrale Annahmen, die bei allen Modellen gegeben sein müssen, aber auch Bedingungen, die explizit bei der Anwendung eines bestimmten Modells gewahrt sein müssen. Zu Ersterem lassen sich die lokale stochastische Unabhängigkeit und die Homogenität nennen. Beide stellen notwendige Bedingungen für die bereits erwähnte Stichprobenunabhängigkeit dar. Da diese Eigenschaften im nachfolgenden Kapitel 3 und insbesondere 4.4 noch ausführlich erklärt werden, soll hier nun nicht genauer auf diese Annahmen eingegangen werden. Ebenso sei bezüglich der in allen eindimensional konzipierten Modellen geltenden Unidimensionalität auf diese später dargebotenen Sektionen verwiesen.

Zusammenfassend lassen sich gemäß Irtel (1996, S. 46) drei wesentliche Unterscheidungsmerkmale der logistischen IRT-Modelle zu der KTT erkennen:

1. Bei den besagten logistisch konzipierten IRT-Modellen wird zwischen den latenten Personenparametern und den beobachtbaren Variablen im Gegensatz zur KTT kein linearer, sondern

ein logistischer Zusammenhang angenommen.

2. Während die KTT auf direkten Messungen anhand der Testrohwerte aufbaut, geschieht diese Art der Messung bei den IRT-Modellen auf indirektem Wege, basierend auf den einzelnen Items eines Tests.

3. Es wird ein getrennt betrachteter Einfluss der Itemschwierigkeit und der Personenfähigkeit bezogen auf die Lösungswahrscheinlichkeit suggeriert, womit spezifisch objektive Vergleiche möglich werden.

Nun sollen noch kurz einige dieser logistischen IRT-Modelle gennant werden. Wie bereits indirekt erwähnt, kann man die IRT-Modelle nicht nur nach den zugrundeliegenden Antwortformaten unterscheiden. Es gibt eine Reihe von Aspekten, mittels derer sich die Modelle auf verschiedene Weise taxonomisch ordnen lassen. So kann man die Modelle auch hinsichtlich der Zahl an Itemparametern aufgliedern, was eine häufig genutzte Gliederung darbietet. Unter diesem Aspekt fand auch die Einordnung der in Abb. 2.1 dargestellten IRT-Modelle statt. Es sind jeweils die wichtigsten Vertreter genannt und somit gilt es zu beachten, dass in dem Diagramm nicht alle existierenden IRT-Modelle enthalten sind.

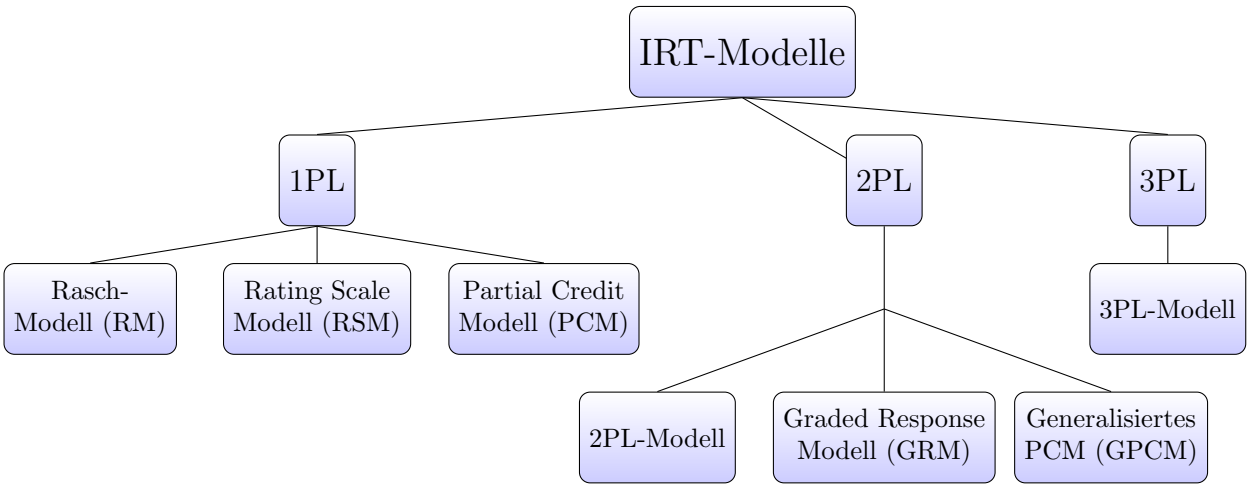

Abb. 2.1: Darstellung der wichtigsten IRT-Modelle, taxonomische Anordnung anhand der Anzahl von Itemparametern

Mit der vorhergehenden Abbildung wird also ersichtlich, dass sich die aufgeführten Modelle nach ihrer Anzahl von einem, zwei oder drei Itemparameter(n) unterscheiden lassen. Hierbei modellieren einparametrische (1PL-)IRT-Modelle das Antwortverhalten mit lediglich einem Itemparameter. Dieser steht in direktem Zusammenhang mit der Positionierung des jeweiligen Items auf der Skala der latenten Variable. Aus diesem Grund wird dieser auch als Lokationsparameter bezeichnet. Grundsätzlich ist dieser in allen aufgezählten Modellen enthalten und wird in direkten Bezug zum Personenparameter gesetzt, welcher als Maß für die Fähigkeit eines betrachteten Probanden steht. Zu dieser Art der IRT-Modelle gehört insbesondere das von Rasch 1960 eingeführte Rasch-Modell und das 1982 von Masters entwickelte Partial Credit Modell, ebenso wie das Rating Scale Modell von Andrich (1978).

Entsprechend der vorhergehenden Erklärung enthalten also zweiparametrische (2PL-)IRT-Modelle zwei Itemparameter. Neben dem bereits genannten Lokationsparameter, wird ein s.g. Steigungs- bzw. Diskriminationsparameter - auch als „Slope Parameter“ bezeichnet - hinzugenommen. Speziell ist dazu zu erwähnen, dass das Generalisierte Partial Credit Modell, dessen Begründer Muraki (1992) ist, den 2PL-IRT-Modellen angehört und letztlich auf dem einparametrischen PCM basiert. Zudem auch die Erweiterung des RMs, das s.g. 2PL-Modell oder auch Birnbaum-Modell, was von Birnbaum (1968) vorgestellt wurde. Außerdem lässt sich noch das Graded Response Modell von Samjima (1969) zu dieser Art der IRT-Modell zuordnen.

Hinzu kommt, dass gerade bei Multiple Choice Tests unterstellt wird, dass eine korrekte Antwort auch erraten werden kann. Diese Tatsache findet bei der Anwendung von 3PL-IRT-Modellen Berücksichtigung, indem das jeweilige Modell nicht nur den Lokations- und Steigungsparameter beinhaltet, sondern zusätzlich einen s.g. Rateparameter, der auch als „Guessing Parameter“ bezeichnet wird. Dieser dritte Itemparameter findet sich z.B. bei der Anwendung von Birnbaum's (1968) eingeführten 3PL-Modell, welches keinen Zweitnamen besitzt. Dieses stellt eine Erweiterung seines 2PL-Modells und damit auch des RMs dar.

Der Einfluss der verwendeten Itemparameter kann insbesondere auch grafisch gut nachvollzogen werden. Zur genaueren Veranschaulichung wird daher in Kapitel 3 auch kurz auf die möglichen Erweiterungen des RMs, das 2PL- und 3PL-Modell nach Birnbaum, eingegangen, und deren Charakteristiken anhand von entsprechenden Grafik verdeutlicht. Zunächst soll dazu aber das RM vorgestellt werden.

3 Das Rasch-Modell (RM)

Im Rahmen der IRT kommt man i.A. nicht um das Modell des Dänen Georg Rasch (1960) herum. Dieses kann explizit zur Auswertung psychologischer Tests, etwa zur Messung der Leistungsfähigkeit oder Intelligenz einer Person, genutzt werden. Um das s.g. unidimensionalen Rasch-Modell (RM) jedoch konkret anwenden zu können, müssen die Antwortkategorien als dichotome Variablen vorliegen. Genauer gesagt bedeutet dies, dass die Antwortmöglichkeiten in einem Test nur als richtig oder falsch eingestuft werden können oder eine Aussage bejaht oder verneint werden kann. Insbesondere liegt diesem Modell damit die Idee zugrunde, dass die Ergebnisse eines Tests nur von zwei wesentlichen Komponenten abhängen. Diese sind die Fähigkeit einer Person, sowie die Schwierigkeit des jeweiligen Items.

Nachfolgend soll eben dieses klassische RM genauer vorgestellt werden, da es - wie bereits zuvor erwähnt - die Basis vieler IRT-Modelle darstellt und insbesondere auch als Grundlage zur Entwicklung des s.g. Partial Credit Modells (PCM) gilt, worauf im Kapitel 4 genauer eingegangen wird. Hierzu werden die Modellgleichung, die im RM geltende Annahmen, sowie die s.g. itemcharakteristischen Kurven (ICCs) und mögliche Parameterschätzungen näher erläutert. Diese Abschnitte lassen sich auch anschaulich mittels Strobl (2012) nachvollziehen. Anschließend werden zusätzlich noch kurz die zwei und drei Itemparameter umfassenden Erweiterungen des RMs vorgestellt, was in Anlehnung an Eid (2014) geschieht.

Mithilfe des RMs lässt sich allgemein die Wahrscheinlichkeit berechnen, dass eine Person p eine Aufgabe i in einem Test lösen kann unter gleichzeitiger Berücksichtigung der latenten Fähigkeit dieser Person θ_p und der Schwierigkeit des betrachteten Items β_i. Dabei bezeichnet p eine der $p = 1, \ldots, P$ Personen und i eines der $i = 1, \ldots, I$ Items. Angemerkt sei hierzu, dass man von dem Personen- oder Fähigkeitsparameter θ_p und vom Aufgaben/Item- oder auch Schwierigkeitsparameter β_i spricht. Diese beiden (latenten) Parametertypen liegen auf einer identischen Skala, die theoretisch von $-\infty$ bis ∞ geht. Zudem kann das dichotome Antwortformat durch Zuweisung einer 0 oder 1 - wobei es sich um die s.g. Dummy-Kodierung handelt - zu der entsprechen dichotomen Zufallsvariable X_{pi} veranschaulicht werden. Folglich wird in Gl. (3.1) also mittels der logistischen Verteilungsfunktion die Wahrscheinlichkeit eine Aufgabe zu lösen unter gleichzeitiger Berücksichtigung der Personenfähigkeit θ_p und der Itemschwierigkeit β_i modelliert.

$$\pi_{pi1} = P(X_{pi} = 1|\theta_p, \beta_i) = \frac{exp(\theta_p - \beta_i)}{1 + exp(\theta_p - \beta_i)} \tag{3.1}$$

Insbesondere gilt dabei stets, dass ein hoher Wert des θ-Parameters eine hohe Personenfähigkeit impliziert, genauso wie ein hoher Wert des β-Parameters auf eine schwierige Aufgabe hindeutet. Umgekehrtes gilt analog. Damit berücksichtigt Gl. (3.1) also, dass die Lösungswahrscheinlichkeit für eine Aufgabe mit entsprechendem Schwierigkeitsgrad β_i zunimmt, je fähiger eine Person ist. Demnach ist in den vorhergehenden Formeln auch inbegriffen, dass für eine Person mit einer Fähigkeit von θ_p bei steigender Aufgabenschwierigkeit die Lösungswahrscheinlichkeit für dieses Item abnimmt. Zudem ist in Gl. (3.1) durch die zugrundeliegende logistische Verteilungsfunktion gewährleistet, dass die Lösungswahrscheinlichkeit π_{ip} im Wertebereich von 0 bis 1 liegt.

Ein besonderes Augenmerk sollte auch auf den Exponenten im Zähler bzw. Nenner geworfen werden. Hierbei ergibt sich, falls eine Person kompetenter ist als eine Aufgabe schwierig ist, eine positive Differenz und damit eine Lösungswahrscheinlichkeit, welche größer als 0.5 ist. Andersherum erhält man eine negative Differenz im Exponenten bzw. eine Lösungswahrscheinlichkeit, die kleiner ist als 0.5, falls ein Item schwieriger ist, als eine Person fähig ist. Sind beide Parameter gleich, so nimmt die Lösungswahrscheinlichkeit einen Wert von 0.5 an. Diese Tatsache findet auch bei den s.g. itemcharakteristischen Kurven Verwendung, welche in einem späteren Abschnitt dieses Kapitels kurz dargestellt und erklärt werden sollen. Zunächst wird aber noch auch auf weitere Darstellunsgmöglichkeiten des RMs eingegangen.

Bisher wurde nur die Berechnung der Wahrscheinlichkeit betrachtet, dass eine Person ein ihr vorgelegtes Item lösen kann. Allerdings lässt sich der umgekehrte Fall - also dass eine Aufgabe von einem Probanden nicht gelöst wird - ebenfalls berechnen. Dies erfolgt über die Ermittlung der Gegenwahrscheinlichkeit, welche der nachstehenden Gl. (3.2) entnommen werden kann. Da die Summe der Wahrscheinlichkeit eines Ereignisses und ihrer Gegenwahrscheinlichkeit bekanntlich 1 bzw. 100% ist, lässt sich π_{pi0} ganz einfach durch Subtraktion der in Gl. (3.1) dargestellten Wahrscheinlichkeit von 1 errechnen. Indem nun die entsprechende Formel eingesetzt und der Minuend auf den gleichen Nenner gebracht wird wie der Subtrahend, ergibt sich schließlich die gesuchte Wahrscheinlichkeit für das Nicht-Lösen einer Aufgabe i durch eine Person p.

$$\pi_{pi0} = P(X_{pi} = 0|\theta_p, \beta_i) = 1 - P(X_{pi} = 1|\theta_p, \beta_i) = \frac{1}{1 + exp(\theta_p - \beta_i)} \tag{3.2}$$

Betrachtet man Gl. (3.1) und Gl. (3.2) der bedingten Wahrscheinlichkeit für eine korrekte und die für eine falsche Antwort noch einmal genauer, so lässt sich erkennen, dass diese sich auch zu einer gemeinsamen Formel zusammenfassen lassen. Damit erhält man die folgende Gl. (3.3) für die jeweilige Wahrscheinlichkeit des Lösens oder Nicht-Lösens einer Aufgabe:

$$\begin{aligned} \pi_{pix} = P(X_{pi} = x_{pi}|\theta_p, \beta_i) &= \frac{exp[x_{pi} \cdot (\theta_p - \beta_i)]}{1 + exp(\theta_p - \beta_i)} \\ &= \begin{cases} \frac{exp[1 \cdot (\theta_p - \beta_i)]}{1 + exp(\theta_i - \beta_j)} = \frac{exp(\theta_p - \beta_i)}{1 + exp(\theta_p - \beta_i)}, & \text{für } x_{pi} = 1 \\ \frac{exp[0 \cdot (\theta_p - \beta_i)]}{1 + exp(\theta_p - \beta_i)} = \frac{1}{1 + exp(\theta_p - \beta_i)}, & \text{für } x_{pi} = 0 \end{cases} \end{aligned} \tag{3.3}$$

Im Zuge der GLMs - Kurzform für generalisierte lineare Modelle - lässt sich die bisher dargestellte logistische Responsefunktion von Gl. (3.1) auch umschreiben zur Logit-Linkfunktion, womit man die logarithmierten Chancen/Odds - für das Eintreten einer korrekten Antwort gegenüber einer falschen - erhält:

$$log(\frac{\pi_{pi1}}{1 - \pi_{pi1}}) = log(\frac{\pi_{pi1}}{\pi_{pi0}}) = \theta_p - \beta_i \tag{3.4}$$

An die möglichen Darstellungen der Modellgleichung soll als Nächstes die grafische Abbildung des RMs näher erläutert werden. Hierzu bedient man sich der bereits angesprochenen itemcharkteristischen Kurven, im Englischen bezeichnet als „Item Charcteristic Curves“ (ICC). Dabei wird die Lösungswahrscheinlichkeit eines Items in Abhängigkeit von der Personenfähigkeit θ_p abgebildet.

Derartige ICCs sind in allen Grafiken von Abb. 3.1 am Ende dieses Kapitels zu sehen. Zunächst sind zur Beschreibung des RMs aber nur die oberen beiden Grafiken von Bedeutung. Erstere entspricht prinzipiell der grafischen Darstellung von Gl. (3.3). Hierbei wird zu einem Item mit Schwierigkeitsgrad $\beta_i = 0$ sowohl die Lösungswahrscheinlichkeit π_{pi1}, als auch die entsprechende Gegenwahrscheinlichkeit π_{pi0} in Abhängigkeit von der latenten Personenfähigkeit dargestellt. Da sich die Wahrscheinlichkeit für das Nicht-Lösen einer Aufgabe direkt aus der Lösungswahrscheinlichkeit dieser Aufgabe ergibt, verzichtet man i.A. auf die Darstellung der Ersteren. Mittels dieser Grafik kann auch erklärt werden, wie sich der Schwierigkeitsparameter einer Aufgabe i ablesen lässt: dieser entspricht genau demjenigen Wert auf der Personenfähigkeits-Achse, bei dem ein Testteilnehmer eine 50%-ige Lösungswahrscheinlichkeit für die betrachtete Aufgabe i hätte. Für diesen Fähigkeitswert ist also die Wahrscheinlichkeit für das Lösen oder Nicht-Lösen eines betrachteten Items gleich groß. Wie in einem der vorhergehenden Abschnitte bereits beschrieben, ist das der Fall, wenn die beiden Parameter θ_p und β_i denselben Wert annehmen. Gleichzeitig befindet sich hier der Wendepunkt einer derartigen ICC.

Da für alle Aufgaben eines Tests dieselbe latente Skala angenommen wird, lassen sich auch die ICCs mehrerer Aufgaben in einer Grafik abbilden. Dies ist in der mittleren Grafik von Abb. 3.1 vorgenommen worden. Anhand des Verlaufs der Kurven sollten einige weitere Eigenschaften des RMs ersichtlich werden: bedingt durch den streng monoton steigenden Verlauf der ICCs - was ein typisches Merkmal logistischer Verteilungsfunktionen ist - nähert sich die Lösungswahrscheinlichkeit mit zunehmender Personenfähigkeit dem Wert 1 an, mit abnehmender Kompetenz geht diese gegen 0. Folglich bringt eine höhere oder niedrigere Personenfähigkeit an den Randbereichen nur eine geringe Änderung der Lösungswahrscheinlichkeit mit sich. Dagegen bedingt eine zunehmende Kompetenz im mittleren Bereich einen vergleichsweise starken Anstieg der Lösungswahrscheinlichkeit. Hierbei dient die Steigung der Kurve also als Maß dafür, wie deutlich das jeweilige Item zwischen den unterschiedlichen Ausprägungen der gleichen Personenfähigkeit differenzieren kann. Demzufolge kann man umso besser zwischen zwei Testindividuen mit unterschiedlich stark ausgeprägter Fähigkeit unterscheiden, je stärker die Steigung im mittleren Bereich der Kurve ist. Aufgrund dessen spricht man allgemein auch von der s.g. Trennschärfe, wobei im RM die Trennschärfe aller Items gleich groß ist.

Auffallend ist in der mittleren Grafik zudem, dass alle Kurven parallel zueinander verlaufen und lediglich entlang der Abszisse verschoben sind. Diese Verschiebung ergibt sich aus dem höheren oder niedrigeren Schwierigkeitsgrad der gestellten Testfrage. Somit bezeichnen entlang der x-Achse nach links verschobene Kurven die Lösungswahrscheinlichkeit für leichtere Items, nach rechts verschobene deuten entsprechend auf schwerere Aufgaben hin. Die Parallelität der Kurven ergibt sich aus der Modellgleichung: für jede Aufgabe gibt es nur einen Parameter, der sich ändern kann, und zwar die Itemschwierigkeit β_i. Inhaltlich gesehen bedingt dies, dass alle Aufgaben die gleiche Trennschärfe besitzen müssen, insofern das RM zur Anwendung kommen soll. Diese Tatsache bezüglich der Parallelität der Kurven kann als strenge Anforderung an einen Test gesehen werden oder bereits als eine Eigenschaft des RMs.

Weitere grundlegende Eigenschaften, die bei Gültigkeit des RMs erfüllt sein müssen, werden nun folgend kurz erläutert. Allerdings gelten beim PCM, welches in Kapitel 4 ausführlich behandelt wird,

diese Annahmen und Eigenschaften des RMs in ähnlicher Weise. Daher sollen die Charakteristiken des RMs lediglich kurz ausgeführt werden. Insbesondere wird nur die inhaltliche Bedeutung, nicht aber die mathematische Darstellung dieser Eigenschaften dargelegt.

Zu einer der Annahmen des RMs gehört die Eindimensionalität, welche die Item- und Personenhomogenität umfasst. Durch die Aufgabenhomogenität wird festgesetzt, dass die Schwierigkeit der Aufgaben für alle Testpersonen identisch sein muss. Umgekehrt muss aber auch gelten, dass die gemessene Personenfähigkeit θ_p unabhängig von der gewählten Aufgabe ist, was als Personenhomogenität bezeichnet wird. Damit impliziert die Eindimensionalität also, dass die Personenfähigkeit θ_p und die Itemschwierigkeit β_i auf einer gemeinsamen latenten Dimension liegen.

Zudem ist auch die lokale stochastische Unabhängigkeit im RM von zentraler Bedeutung. Die stochastische Unabhängigkeit ist gegeben, insofern das Lösen einer Aufgabe nicht von einer vorhergehenden abhängt. Dies ist gewahrt, wenn die Items eines Tests nicht aufeinander aufbauen. Andererseits könnte allerdings auch auf Personenseite diese Annahme verletzt werden, falls beispielsweise die Testteilnehmer von einander abschreiben können. Dies gilt es insbesondere bei der Testplanung zu berücksichtigen und dem ist möglichst entgegenzuwirken. Der Zusatz „lokale “ stochastische Unabhängigkeit bezieht sich auf die Konstanthaltung des Personenparameters. Damit muss die Lösungswahrscheinlichkeit zweier Aufgaben nur unabhängig voneinander sein, solange man eine Person bzw. mehrere Personen mit der gleichen Fähigkeit betrachtet. Genauer gesagt, ist damit also zugelassen, dass eine Person mit einem sehr hohen θ-Wert alle Aufgaben mit hoher Wahrscheinlichkeit löst, im Vergleich zu einem Probanden mit einem niedrigeren Fähigkeitswert.

Des Weiteren spielt die spezifische Objektivität eine zentrale Rolle im RM. Indirekt wurde auf diesen Sachverhalt im vorhergehenden Abschnitt bereits eingegangen. Die spezifische Objektivität garantiert nämlich, dass es beim Vergleich der Fähigkeit zweier oder mehrerer Personen nicht von Bedeutung ist, anhand welchen Items man sie vergleicht. Denn hat ein Testteilnehmer einen höheren Fähigkeitsscore als eine Vergleichsperson, so wird dieser betrachtete Teilnehmer auch mit einer höheren Wahrscheinlichkeit die Aufgabe lösen können als die Vergleichsperson. Umgekehrt ist ein Item mit einem hohem β-Parameter schwieriger zu lösen als eines mit einem niedrigen Aufgabenparameter, unabhängig davon welche Person betrachtet wird.

Außerdem sei erwähnt, dass es für die beiden unbekannten Parameter θ_p und β_i im RM jeweils eine suffiziente Statistik gibt. Allgemein gesprochen wird eine Statistik $T(x)$ für einen nicht bekannten Parameter ν als suffizient bezeichnet, insofern sie genauso viel Information über diesen Parameter ν enthält wie die Stichprobe selbst (vgl. Kauerman, 2014, S. 15). Nun sei nur kurz erwähnt, dass im RM die Gesamtzahl der von einer Person korrekt beantworteten Aufgaben eine suffiziente Statistik für den Fähigkeitsparameter darstellt. In ähnlicher Weise ist die Summe der Personen, welche eine Aufgabe richtig lösen konnten, eine suffiziente Statistik für den Schwierigkeitsparameter. Diese Eigenschaft des RMs wird im nachfolgenden Kapitel 4.4.4 durch Übertragung des Sachverhalts auf das PCM näher erläutert.

Zu alledem lässt sich noch eine Annahme zum Messniveau des RMs treffen. So argumentiert beispielsweise Fischer (1995, Kap. 2), dass das RM nur bis auf Intervallskalenniveau Eindeutigkeit besitzt. Hierbei bezeichnen Intervallskalen metrische Skalen mit nur einem relativen Nullpunkt. Dies lässt auch dadurch begründen, dass sich Absolutskalen - im Gegensatz zu Intervallskalen - mit

absolutem Nullpunkt für die beiden Parameter kaum rechtfertigen lassen, unter dem Aspekt, dass diese Parameter latente, also nicht direkt messbare Eigenschaften darstellen.

Als Nächstes stellt sich noch die Frage, wie die Fähigkeits- und Aufgabenparameter bestimmt werden können. Von besonderem Interesse ist dabei meistens die Schätzung der Aufgabenschwierigkeit. Hierzu sollen die möglichen Methoden zur Parameterschätzung nur kurz namentlich erwähnt werden. Wiederum sei darauf verwiesen, dass in den Kapiteln zum PCM - explizit in Kapitel 5 - hierauf genauer eingegangen wird. Vorwiegend finden im RM drei Schätzmethoden zur Ermittlung der Itemparameter ihre Anwendung, welche auf dem Maximum-Likelihood (ML) Prinzip basieren. Demzufolge werden die Parameter also so geschätzt, dass die Wahrscheinlichkeit die beobachteten Daten zu erhalten, maximiert wird. Die drei besagten Methoden sind die „Gemeinsame ML-Schätzung“ (JML), die „Bedingte ML-Schätzung“ (CML) und die „Marginale ML-Schätzung“ (MML). Diese Schätzansätze unterscheiden sich jeweils im Vorgehen zur Bestimmung der Parameter. Allerdings haben alle gemeinsam, dass hierbei insbesondere von der im RM geltenden lokale stochastische Unabhängigkeit Gebrauch gemacht wird. Im Zuge der JML-Methode werden die Personenparameter gleichzeitig mit denen der Aufgaben ermittelt. Insofern eine der anderen beiden ML-Methoden verwendet wurde, bietet es sich an die Fähigkeitsparameter über die s.g. gewichtete ML-Methode (WML) - Kurzform abgeleitet vom Englischen „weighted“ -, welche von Warm (1985) ausführlich behandelt wird, oder Bayes-Schätzmethoden zu schätzen.

Nun folgend soll noch kurz auf die bereits in Kapitel 2 erwähnten, von Birnbaum (1968) vorgenommenen Erweiterungen des dichotomen RMs eingegangen werden und in diesem Zuge das 2PL- und 3PL-Modell vorgestellt werden. Dies geschieht in Anlehnung an Eid (2014, Kapitel 4.6.1 und 4.6.2). Wie bereits in Kapitel 2 verdeutlicht wurde, wird im Unterschied zum RM im 2PL- oder auch Birnbaum-Modell ein s.g. Diskriminations- oder Steigungsparameter aufgenommen. Dadurch findet eine Gewichtung der Abweichung des Fähigkeitswertes vom Schwierigkeitsparameter mit dem itemspezifischen Faktor α_i statt. Die Modellgleichung lautet dann entsprechend wie folgt:

$$\pi_{pi1} = P(X_{pi} = 1|\theta_p, \alpha_i, \beta_i) = \frac{exp[\alpha_i(\theta_p - \beta_i)]}{1 + exp[\alpha_i(\theta_p - \beta_i)]} \tag{3.5}$$

Gerade bei Multiple Choice Tests setzt man voraus, dass mit einer gewissen Wahrscheinlichkeit auch durch Raten eine Frage korrekt beantwortet werden kann. Berücksichtigt man also zuätzlich einen Rateparameter γ_i, so ergibt sich dann das 3PL-Modell in Form von Gl. (3.6). Nach Rost (1996) kann dieser Modellparameter γ_i entweder geschätzt werden oder präexperimentell vorgegeben werden. Letzteres bedeutet, dass man beispielsweise bei einem vier Kategorien umfassendem Item von $\gamma_i = 1/4 = 0.25$ ausgeht.

$$\pi_{pi1} = P(X_{pi} = 1|\theta_p, \alpha_i, \beta_i, \gamma_i) = \gamma_i + (1 - \gamma_i)\frac{exp[\alpha_i(\theta_p - \beta_i)]}{1 + exp[\alpha_i(\theta_p - \beta_i)]} \tag{3.6}$$

Diese Erweiterungen von Birnbaum lassen sich auch graphisch gut nachvollziehen. Hierzu ist beispielhaft die untere der drei Grafiken in Abb. 3.1 zu betrachten. Zunächst soll auf einige Eigenschaften des 2PL-Modells eingegangen werden. Dazu betrachte man die hellblau eingezeichnete Linie, die zwar denselben Schwierigkeitsparameter $\beta_i = 0$ verglichen mit der grün dargestellten ICC

des RMs hat, allerdings durch Berücksichtigung eines Diskriminationsparameters von $\alpha_i = 2$ eine veränderte Steigung im mittleren Bereich der Kurve aufweist. So ist zu vermerken, dass sich dieselbe Abweichung bzw. die gleiche Distanz von der Itemschwierigkeit $\beta_i = 0$ bei dem RM mit einem Steigungsparameter von $\alpha_i = 1$ anders auswirkt als bei dem 2PL-Modell mit einem Steigungsparameter von $\alpha_i = 2$. Unterhalb des Schwierigkeitsparameters besitzt eine Person eine insgesamt niedrigere Lösungswahrscheinlichkeit bei demjenigen Modell - hier also 2PL-Modell - mit einem höheren Wert des Steigungsparameters. Allerdings gilt oberhalb des Aufgabenwertes β_i Umgekehrtes. Je größer also der Wert des α-Parameters ist, desto besser kann man Personen mit unterschiedlichem Fähigkeitswert unterscheiden bzw. dasjenige Item kann umso besser zwischen den Personen diskriminieren.

Wie dadurch zudem ersichtlich wird, ergibt sich durch die Einführung verschiedener Diskriminationsparameter, dass sich die ICCs der einzelnen Items schneiden und damit unterschiedliche Trennschärfen besitzen. Damit gilt insbesondere, dass die im RM geltende Annahme der spezifischen Objektivität hier nun aufgegeben wird. Denn „testet man nur Personen im oberen Fähigkeitsspektrum, so würde man zu einer anderen Rangordnung der Itemschwierigkeiten gelangen, als wenn man Personen im unteren Fähigkeitsspektrum testet. Die Rangfolge der Itemschwierigkeiten ist somit abhängig von der Auswahl der jeweiligen Personenstichprobe, was zur Konsequenz hat, dass das zweiparametrige logistische Modell keine spezifisch objektiven Messungen ermöglicht“ (vgl. Rost, 1996, S. 134). Des Weiteren ergibt sich, dass weder die Anzahl der gelösten Aufgaben eine suffiziente Statistik für den Fähigkeitsparameter darstellt, noch dass die Anzahl der Probanden, die eine Aufgabe lösen konnten, eine suffiziente Statistik für den Schwierigkeitsparameter ist. Im 2PL-Modell gilt stattdessen, dass die Summe der mit den Steigungsparametern gewichteten Aufgabenparameter eine suffiziente Statistik für den Fähigkeitswert einer Person darbietet. Diese beiden kurz erläuterten Tatsachen lassen sich z.B. mit Rost (1996, Kapitel 3) genauer nachvollziehen, sollen im Folgenden aber nicht weiter ausgeführt werden. Allerdings sei auf eine Problematik der suffizienten Statistik für den Personenparameter im 2PL-Modell hingewiesen: i.A. sind die Diskriminationsparameter nicht bekannt und folglich kann auch die mit den α-Parametern gewichtete Summe der Itemantworten nicht berechnet werden. Dies führt dazu, dass die CML-Schätzung nicht durchgeführt werden kann. Weiterhin sind aber die JML und MML-Methode anwendbar. Auf den mathematischen Hintergrund dieser Schätzmethoden wird in Kapitel 5 schließlich eingegangen.

Nun soll allerdings die Erklärung der grafischen Darstellung weiter ausgeführt werden. So sei noch einmal daran erinnert, dass das 3PL-Modell neben dem Lokations- und Steigungsparameter zusätzlich noch einen Rateparameter γ_i beinhaltet. Dadurch ergibt sich bei grafischer Veranschaulichung dieses Modells - beispielhaft dargestellt anhand der dunkelblauen, gepunkteten Kurve in der untersten Grafik von Abb. 3.1 -, dass bei einem gegen $-\infty$ gehenden Fähigkeitsparameter die Lösungswahrscheinlichkeit nicht wie beim RM gegen 0 strebt, sondern stattdessen gegen γ_i. Nimmt man also - wie bereits zuvor beispielhaft erwähnt - ein vier Kategorien umfassendes Item an, so kann man den Rateparameter auf $\gamma_i = 0.25$ festsetzen.

Auf Basis dieses Wissens soll nun zum PCM von Masters (1982) und dem GPCM von Muraki (1992) übergeleitet werden. Hierzu sei angemerkt, dass der Hauptunterschied zwischen dem RM und dem PCM wohl die Anwendbarkeit auf das jeweilige Datenformat darstellt.

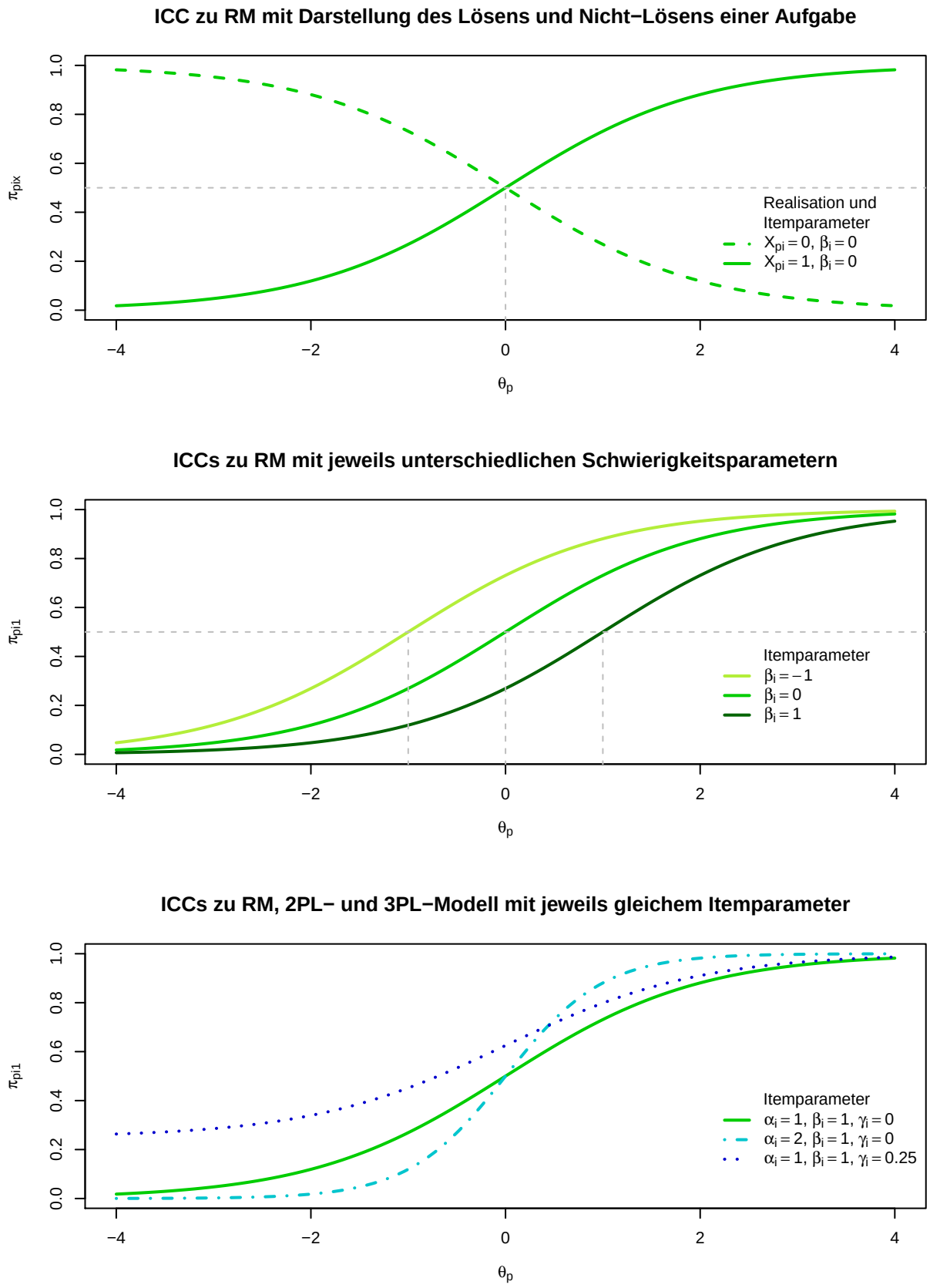

Abb. 3.1: Grafische Veranschaulichungen zu RM von Rasch, sowie zum 2PL- und 3PL-Modell Birnbaum

4 Grundlegendes zum Partial Credit Modell (PCM)

Das in dem vorhergehenden Kapitel beschriebene RM von Rasch (1960) diente Masters (1982) als Grundlage für die Ausarbeitung des s.g. Partial Credit Modells (PCM). Dabei geht das PCM, ebenso wie das RM, davon aus, dass über die gegebene Itemantworten auf eine ihnen gemeinsame, latente Variable geschlossen werden kann. Gerade in Anbetracht der Tatsache, dass es nicht nur dichotome Bewertungsformate gibt - auf die das RM angewendet werden könnte -, hat das PCM Einzug in die IRT genommen. V.a. in der Schule/Universität, im beruflichen Alltag oder insbesondere auch in der Psychologie ist es oftmals notwendig nicht nur zwischen den Ausprägungen falsch/richtig bzw. Ablehnung/Zustimmung zu unterscheiden. So müssen vielmehr auch „Zwischenstufen“ Berücksichtigung finden und damit mehr als zwei Antwortmöglichkeiten bereitgestellt werden. Hat man also Tests, Forschungsfragen, Klausuraufgaben, etc. mit mehrstufigen Antwortkategorien, denen zusätzlich eine gewisse Ordnung zugrunde liegt, so sollte das PCM eine Möglichkeit der Analyse darbieten. Generell ist das PCM bei Variablen mit geordneten Antwortkategorien und damit auf ordinale Daten anwendbar. Dies ist auch der Grund, weshalb das PCM von vielen Autoren als „ordinales Rasch-Modell“ bezeichnet wird (siehe dazu beispielsweise Rost, 1996, Kapitel 3.3.1).

Um die Bedeutung des ordinalen Skalenniveaus und damit einer vergebenen Rangfolge noch einmal zu verdeutlichen, denke man beispielhaft an eine Umfrage nach der Zufriedenheit mit der aktuellen politischen Situation. Hierzu stellt man den Befragten die Antwortkategorien „sehr unzufrieden“, „mittel“ und „sehr zufrieden“ bereit. Damit vergibt folglich eine Person, die „sehr zufrieden“ ankreuzt, der politische Lage eine bessere Bewertung und damit einen höheren Rang, als eine Person die nur mittelmäßig zufrieden mit den aktuellen Umständen ist. Analoges gilt für den Vergleich der abgestuften Antworten „sehr unzufrieden“ und „mittel“. Zugleich sind diese drei Klassifizierungen leicht erweiterbar, etwa indem man das Item um die weniger extremen Antworten „unzufrieden“ und „zufrieden“ ergänzt. Zu beachten ist, dass sich auch diese Antwortmöglichkeiten in die geordnete Struktur der vorhergehenden Kategorien einordnen lassen.

Bei diesen ordinalen Beobachtungsformaten können nach Masters (1982, S. 149ff.) grundsätzlich vier verschiedenen Typen unterschieden werden: „Repeated Trials“, „Counts“, „Rating Scales“ und „Partial Credit“. In den nachfolgenden Kapiteln wird insbesondere auf eben dieses „Partial Credit“-Format eingegangen, die beiden zuerst genannten Typen können beispielsweise mithilfe Thissen (1986) nachvollzogen werden. Ratingskalen können als Spezialfall des Partial Credit Scorings eingeordnet werden und daher lässt sich die Modellgleichung des Rating Scale Modells nach Andrich (1978) mittels einer gewissen Restriktion aus dem PCM herleiten. Auf den Unterschied zwischen RSM und PCM wird gegen Ende von Sektion 4.2 kurz eingegangen.

In diesem Kapitel wird dem Leser nun zunächst aufgezeigt, wie „Partial Credit“-Bewertungen in Matrixnotation repräsentiert werden können und wie sich die Modellgleichung des PCMs schrittweise mithilfe des RMs konstruieren lässt. Zudem wird auf die grafische Darstellung, sowie auf die geltenden Modellannahmen und -eigenschaften näher eingegangen. Außerdem wird noch die Erweiterung des PCMs, das s.g. Generalisierte Partial Credit Modell (GPCM) vorgestellt.

4.1 Darstellung von Partial Credit Scores in Matrixnotation

Diese Sektion baut im Wesentlichen auf Masters (1982, S. 155f.) auf. Das von ihm konstruierte PCM stellt eine Erweiterung des in Kapitel 3 vorgestellten RMs dar und beinhaltet dieses insbesondere auch als Spezialfall. Den Anreiz für Konstruktion des PCMs gab die Annahme durch Partial Credit Scoring präzisere Schätzungen der latenten Personfähigkeit zu erhalten, als durch einfache Einstufung einer Antwort als richtig/falsch bzw. Zustimmung/Ablehnung, was durch das RM postuliert wird. Mithilfe der Partial Credit Bewertung können also auch Fragen mit mehreren zur Verfügung stehenden Antworten berücksichtigt werden oder Items stufenweise beurteilt werden. Beispielsweise kann so eine umfangreiche Mathematikaufgabe nicht nur als falsch oder richtig gelöst angesehen werden, sondern es kann auch zugelassen werden, dass der erste Teil des Items korrekt bearbeitet wurde, im weiteren Verlauf jedoch ein Fehler unterlaufen ist.

Allgemein wird somit als Erstes die Reihenfolge der einzelnen $k = 0, \ldots, m_i$ geordneten Lösungsabschnitte bzw. Kategorien eines Items i bestimmt, wobei wiederum $i = 1, \ldots, I$ gilt. Diese Kategorien werden im Folgenden als Leistungsniveaus/-levels bezeichnet. In aufsteigender Reihenfolge wird dann dem Ausgangsniveau der Wert 0, dem ersten Leistungsniveau der Wert 1, usw. und schließlich dem m_i-ten Level der Wert m_i zugewiesen. Hierbei befindet man sich - je nach betrachtetem Aufgabentyp - in der als Ausgangsniveau bezeichneten Kategorie 0, falls man noch keine Antwort gegeben hat oder man bereits die erste Stufe nicht bewältigen kann oder wenn man die niedrigsten Antwortkategorie angibt. Entsprechend des Indizes von m_i sollte zudem ersichtlich werden, dass es sehr wohl zugelassen ist, dass die einzelnen Items unterschiedlich viele Kategorien aufweisen.

Von diesen Leistungslevels klar abzugrenzen sind die (Lösungs-)Stufen/Schwellen - auch als „Thresholds“ bezeichnet -, die jeweils zwischen zwei benachbarten Kategorien liegen. Ein Item i umfasst entsprechend der Anzahl an Levels also m_i Stufen. Genauer gesagt gelangt man durch Bewältigung der ersten Stufe auf das erste Leistungslevel, durch Meistern der zweiten auf das zweite Level, usw. bis man schließlich durch Bewerkstelligung der m_i-ten Stufe auf das höchste Niveau m_i der Aufgabe kommt. Dieser Sachverhalt kann auch mithilfe der später präsentierten Abb. 4.1 nachvollzogen werden. Allgemein wird im Folgenden die Anzahl erfolgreich absolvierter Stufen einer Person p bei dem jeweils betrachteten Item i als Score x_{pi} bezeichnet, wobei somit gilt $x_{pi} \in k = 0, \ldots, m_i$. Es sollte deutlich werden, dass x_{pi} die aufsummierte Zahl der Schwellen bis hin zu einer betrachteten Kategorie k darstellt. Entsprechend dazu wird X_{pi} als Zufallsvariable mit den möglichen Realisationen x_{pi} definiert, wobei konsequenter Weise auch für $X_{pi} \in k = 0, \ldots, m_i$ gilt.

Da Kenntnisse über das Partial Credit Format unverzichtbar zum Verständnis des PCMs sind, wird zunächst diese Art der Bewertung näher erläutert. Hierzu wird beispielhaft von einem Mathematikitem ausgegangen, welches letztlich drei Rechenschritt - also $m_i = 3$ - bis zur korrekten Lösung verlangt:

$$\sqrt{\frac{21}{0.75} - 12} \; = ?$$

Der erste Schritt, um vom Leistungslevel 0 auf 1 zu gelangen, wäre hier - gemäß mathematischen Rechenregeln - also 21 durch 0.75 zu teilen. Als Nächstes muss man von dem Ergebnis 28 die Zahl

12 subtrahieren, um schließlich im letzten Schritt die Wurzel aus 16 ziehen zu können und damit das höchste Leistungsniveau 3 erlangen zu können. Nur wenn man diese Reihenfolge beachtet und alle Teilschritte hin zum nächst höheren Level meistern kann, erhält man als Endergebnis schließlich die Zahl 4.

Hiermit sollte also deutlich werden, dass jedes einzelne Leistungsniveau bis hin zum Ergebnis separat beurteilt wird und als korrekt oder inkorrekt gelöst eingestuft wird. Wird also ein Teilschritt korrekt absolviert und damit das nächst höhere Level $k+1$ erlangt, so wird diesem nun erreichten Level eine 1 zugewiesen, anderenfalls eine 0. Ist letzteres eingetreten, hat Proband p also eine Stufe nicht meistern können, so kann er auch alle darauf folgenden Teilabschnitte nicht mehr korrekt beantworten und diese werden deshalb ebenfalls mit einer 0 versehen. Sollte man also bereits die erste Lösungsstufe nicht bewerkstelligen können, so verharrt man auf dem untersten Leistungslevel 0. Um das Ganze bildlich darzustellen, kann Abb. 4.1 betrachtet werden.

		Leistungsniveau						
		0		**1**		**2**		**3**
$21/0.75$	$=?$	0	$\xrightarrow{1.Stufe}$	1				
$28-12$	$=?$			1	$\xrightarrow{2.Stufe}$	2		
$\sqrt{16}$	$=?$					2	$\xrightarrow{3.Stufe}$	3

Abb. 4.1: Dreistufiger Lösungsweg am Beispiel eines Mathematikitems

Von dieser beispielhaften Darstellung kann auch direkt auf die Matrixnotation eines derartigen Items übergeleitet werden. Dazu sollte man sich vor Augen führen, dass Niveau 0 einer bereits auf den ersten Lösungsabschnitt gegebenen, falschen Antwort entspricht. Level 1 erreicht man, insofern man $21/0.75$ richtig berechnet, und die anderen zwei Levels erlangt man analog durch korrektes Lösen der jeweiligen Teilschritte. Die Matrixnotation eines solchen Mathematikitems kann dann mithilfe der 0-1-Kodierung bzw. Dummy-Kodierung wie in Tab. 4.1 - beispielhaft anhand von zehn Personen - dargestellt werden und es ergibt sich damit eine $P \times m_i$-Matrix. Hierbei sei darauf hingewiesen, dass der Übersicht halber, Nuller und damit inkorrekte Lösungen, nicht eingetragen sind. Folglich entsprechen diese also theoretisch gesehen den leeren Zellen der Tabelle. Zudem sei angemerkt, dass sich die Anzahl der erfolgreich absolvierten Stufen bzw. der Score x_{pi} aus der Gesamtzahl notierter Einsen - ungeachtet der des Ausgangsniveaus - berechnen lässt. Zugleich stellen die Spaltenrandsummen s_{ik} die Anzahl der Personen dar, welche Level k bei dem betrachteten Item erreichen konnten.

Mithilfe des Scores x_{pi} ist es schließlich auch möglich eine Matrix aufzustellen, die nicht nur die Ergebnisse eines Items, sondern von allen I Aufgaben eines Tests umfasst. Entsprechend ergibt sich eine PxI-Matrix, deren Einträge die jeweilige Scores x_{pi} der Person p bei Item i sind. Auf diese Matrix wird im folgenden Verlauf dieser Arbeit vermehrt zurückgegriffen und wird als Antwort-Matrix $\boldsymbol{X}$ bezeichnet.

Generell kann die vorgestellte Art der stufenweisen Interpretation anhand der Mathematikaufga-

Person ***p***	**Leistungsniveau** **0**	$\overrightarrow{1.Stufe}$	**1**	$\overrightarrow{2.Stufe}$	**2**	$\overrightarrow{3.Stufe}$	**3**	***Score*** x_{pi}
1	1		1		1		1	3
2	1							0
3	1		1		1			2
4	1		1		1		1	3
5	1		1					1
6	1							0
7	1		1		1			2
8	1		1		1		1	3
9	1							0
10	1		1		1			2
Gesamt:	$s_{i0} = 10$		$s_{i1} = 7$		$s_{i2} = 6$		$s_{i3} = 3$	

Tab. 4.1: Matrixnotation eines dreistufigen Lösungsweges am Beispiel eines Mathematikitems

be auf jegliche Items mit ordinalen Antwortmöglichkeiten übertragen werden. So kann auch nochmal auf das Beispiel zu Beginn des Kapitels 4 zurückgegriffen werden. Geht man von der Frage nach der Zufriedenheit mit der aktuellen politischen Situation aus und stellt die fünf Antwortalternativen „sehr unzufrieden“, „unzufrieden“, „mittel“, „zufrieden“ und „sehr zufrieden“ bereit, so wird eine gewisse aufsteigende Rangfolge erkennbar. Hierbei kann nämlich von der eigentlichen Angabe, die ein Befragter gibt, darauf geschlossen werden, welche Stufen dieser genommen hat, um bis zu eben dieser Antwort zu gelangen. Eine Person, die unzufrieden ist mit den politischen Zuständen des Landes, hat so zu sagen eine Stufe mehr bewerkstelligt als eine Person, die sehr unzufrieden ist mit der derzeitigen Situation. Würde man das Ganze analog zu Abb. 4.1 bildlich veranschaulichen, so ergeben sich 5 Levels mit entsprechend $m_i = 4$ Lösungsstufen. Dabei entspricht Level 0 also der starken Unzufriedenheit und geht bis hin zu Level 4, was der einer sehr hohen Zufriedenheit gleich kommt.

Mit diesem Vorwissen zum Partial Credit Format kann nun auch die Konstruktion des PCMs nach Masters (1982) genauer erläutert werden.

4.2 Vom RM zum PCM

Nun folgend soll der Überlegung nachgegangen werden, wie sich das dichotome RM mittels gewisser Abwandlung auf ordinale Daten oder genauer gesagt auf Partial Credit Bewertungen anwenden lässt. Um dies bestmöglich darlegen zu können, wird die Konstruktion des PCMs an Masters (1982) Vorgehen anküpfend erläutert. Des Weiteren wurde auch Masters (1988, S. 282ff.) als Quelle her-

angezogen.

Eine erste Idee bestand darin, jedes Leistungsniveau k einzeln zu betrachten und jeweils zu bestimmen, ob dieses Level k erreicht wurde oder nicht. Wiederum kann man sich also der 0-1-Kodierung bzw. der Dummy-Kodierung bedienen und so für jedes Level k separat festhalten, ob eine Personen p auf dieses gelangen konnte oder nicht. Folglich kann hiermit die bedingte Wahrscheinlichkeit für das Eintreten von $X_{pik} = 1$ berechnet werden, dass also Person p Level k bei der betrachteten Aufgabe i erreicht:

$$P(X_{pik} = 1|\theta_p, \delta_{ik}) = \frac{exp(\theta_p - \delta_{ik})}{1 + exp(\theta_p - \delta_{ik})} \tag{4.1}$$

Wie bereits im RM (vgl. Kapitel 3) bezeichnet i eines der $1, \ldots, I$ Items, p eine der $1, \ldots, P$ Personen und θ_p den zugehörigen Fähigkeitsparameter der betrachteten Person p. Allerdings wird nun zusätzlich das jeweilige Level k der betrachteten Aufgabe i berücksichtigt und folglich muss der Schwierigkeitsparameter um einen Indizes erweitert werden. Da der Schwierigkeitsparameter im Laufe der Konstruktion des PCMs noch abgeändert wird, wird hier zunächst die Bezeichnung δ_{ik} für diesen gewählt. Dieser Parameter bezeichnet also die Schwierigkeit das Level k bei der betrachteten Aufgabe i zu erreichen.

Aufgrund des vorliegenden ordinalen Skalenniveaus gilt stets, dass die Anzahl der Personen, die Level k erreichen, niemals größer sein kann, als die Anzahl der Personen, die Level $k-1$ erreicht haben. Unter Berücksichtigung von Tab.4.1 bedeutet dies also, dass für die Spaltenrandsummen s_{ik} gilt:

$$s_{i0} \geq s_{i1} \geq s_{i2} \geq \cdots \geq s_{im_i}$$

Gleichzeitig gilt, dass s_{ik} eine suffiziente Statistik für δ_{ik} darstellt. Der Begriff einer suffizienten Statistik für einen bestimmten Parameter wurde bereits in Kapitel 3 definiert und bedeutet hier inhaltlich also, dass durch s_{ik} alle relevanten Informationen bezüglich δ_{ik} gegeben sind. Somit kann dann geschlussfolgert werden, dass sich ebenso wie die Spaltenrandsummen auch die (Schätzer der) Schwierigkeitsparameter einer Aufgabe i, aufgeteilt nach dem jeweiligen k-ten Level, anordnen lassen. Dabei ist also das k-te Leistungsniveau einfacher oder zumindest gleichermaßen schwierig zu erlangen wie das $(k+1)$-te Level. Mathematisch ausdrücken lässt sich dies wie folgt darstellen:

$$\delta_{i1} \leq \delta_{i2} \leq \cdots \leq \delta_{im_i}$$

Allerdings ist hier nun auf eine Problematik hinzuweisen: eine zentrale Annahme des RMs stellt die Eindimensionalität dar, gemäß der die dichotome Beobachtung bzw. Zufallsvariable X_{pik} nur von dem jeweiligen Personenparameter θ_p und dem entsprechenden Aufgabenparameter abhängt, unabhängig von jeglichen anderen Einflüssen. Jedoch ist es hier für eine Person - auch ersichtlich anhand von Tab. 4.1 - nicht möglich Level 3 zu erreichen, wenn nicht zuvor Level 1 und daran anschließend Level 2 bewerkstelligt wurden. Offensichtlich unterliegt also z.B. $P(X_{pi3} = 1)$ weiteren Abhängigkeiten als nur der von θ_p und δ_{i3}. Diese hierarchische Abhängigkeit macht also eine Übertragung des RMs auf die Analyse von Partial Credit Scores in dieser Weise fraglich.

Masters (1982) schlägt daher eine andere Vorgehensweise vor. Demzufolge sollte man sich - er-

neut im Bezug auf das drei Kategorien umfassende Beispiel aus Tab. 4.1 - von der Sichtweise der drei aufsteigend angeordneten Levelschwierigkeiten δ_{i1}, δ_{i2} und δ_{i3} abwenden und damit allgemein von dem Parameter δ_{ik}, welcher die Schwierigkeit beschreibt Level k der Aufgabe i zu erreichen. Stattdessen sollten gemäß Masters (1982) die individuellen Schwierigkeiten der einzelnen Lösungsstufen an sich betrachtet werden und damit β_{ij}. Hierbei entspricht j einer der möglichen Stufen bei Item i und der Index läuft entsprechend bis hin zum erreichten Score x_{pi} von Person p bei Item i, womit folglich gilt $j = 0, ..., x_{pi}$ und gleichzeitig $x_{pi} \in k$.

Um diese nun gewählte Vorgehensweise möglichst anschaulich darstellen zu können, soll noch einmal Bezug auf das zuvor erwähnte Mathematikitem aus Abb. 4.1 genommen werden. Hier beschreibt beispielsweise die Schwierigkeit von Stufe 3, wie wahrscheinlich es für eine Person p ist, Leistungsniveau 3 zu erreichen unter gleichzeitiger Berücksichtigung, dass Level 2 bereits erreicht wurde. Anders ausgedrückt wird durch die Schwierigkeit von Stufe 3 also bestimmt, wie wahrscheinlich es ist, dass die betrachtete Person eher auf Niveau 3 gelangt als auf Niveau 2 zu verweilen. Eine mögliche Modellierung dieser bedingten Wahrscheinlichkeit ist in Formel 4.2 dargestellt. Dabei beschreibt $\pi_{pi2} + \pi_{pi3}$ die Wahrscheinlichkeit für Person p entweder Leistungsniveau 2 oder 3 zu erzielen und folglich lässt sich mittels der gesamten Gl. (4.2) die bedingte Wahrscheinlichkeit berechnen, eher die dritte Stufe zu bewältigen und damit eher einen Score von $x_{pi} = 3$ als $x_{pi} = 2$ zu erhalten. Allgemein wird Φ_{pi3} auch als Schwellenwahrscheinlichkeit bezeichnet, da die Wahrscheinlichkeit modelliert wird, die Schwelle von Niveau 2 nach 3 zu überschreiten.

$$\Phi_{pi3} = \frac{\pi_{pi3}}{\pi_{pi2} + \pi_{pi3}} = \frac{exp(\theta_p - \beta_{i3})}{1 + exp(\theta_p - \beta_{i3})} \tag{4.2}$$

Es ist zu beachten, dass zwar explizit durch β_{i3} die Wahrscheinlichkeit von Niveau 2 zu Niveau 3 zu gelangen beinhaltet ist, damit allerdings nichts über die Wahrscheinlichkeit zuvor Level 2 zu erreichen gesagt ist. Insbesondere sei noch auf den hinteren Teil von Gl. (4.2) hingewiesen. Hierbei findet die Tatsache - wie auch im RM - Berücksichtigung, dass die Wahrscheinlichkeit in die höhere Kategorie zu gelangen mit zunehmender Personenfähigkeit θ_p ebenfalls zunimmt. Dies kommt durch die Differenz von $\theta_p - \beta_{i3}$ im jeweiligen Exponenten von Zähler bzw. Nenner zutage. Da insbesondere wieder die logistische Verteilungsfunktion zugrunde liegt und diese streng monoton steigend ist, nimmt mit einem größer werdenden Exponenten schließlich auch die Wahrscheinlichkeit zu das jeweils höhere Leistungslevel zu erreichen. Ist also θ_p größer als β_{i3}, so gilt wie im RM, dass die Wahrscheinlichkeit die Schwelle von Kategorie 2 nach 3 zu überschreiten größer als 0.5 ist. Umgekehrtes gilt entsprechend.

Analog zu Gleichung 4.2 lassen sich auch die Wahrscheinlichkeiten für die Bewältigung der anderen beiden Stufen aufstellen. Damit kann also allgemein formuliert nachfolgende Gl. (4.3) betrachtet werden. Generell wird hierdurch die Wahrscheinlichkeit ausgedrückt, jeweils eher auf ein höheres Niveau und damit einen höheren Score von x zu erlangen, als auf das entsprechende, darunter liegende Niveau zu verweilen bzw. einen niedrigeren Score von $x - 1$ zu erhalten. Letztlich wird also Rasch's Modell aus Gl. (3.1) auf jedes Paar benachbarter Kategorien aus einem Set geordneter

Kategorien angewandt (vgl. Masters, 1988, S. 283).

$$\Phi_{pix} = \frac{\pi_{pix}}{\pi_{pi(x-1)} + \pi_{pix}} = \frac{exp(\theta_p - \beta_{ix})}{1 + exp(\theta_p - \beta_{ix})}, \tag{4.3}$$

$$\text{mit} \quad x = 0, \ldots, m_i \quad \text{und} \quad \pi_{pi(-1)} \equiv 0$$

Zu alldem gilt nun, dass jede Person eine der vier Niveauebenen - einschließlich Level 0 - bei dem beispielhaften Mathematikitem erreichen muss und daraus folgt dann, dass gelten muss:

$$\pi_{pi0} + \pi_{pi1} + \pi_{pi2} + \pi_{pi3} = 1 \tag{4.4}$$

Allgemein formuliert kann somit geschlussfolgert werden, dass bei einer begrenzten Anzahl an m_i möglichen Kategorien für Item i und gleichzeitiger Notwendigkeit der Antwort einer Person p, gilt:

$$\sum_{k=0}^{m_i} \pi_{pik} = 1 \tag{4.5}$$

Anhand von Gl. (4.3) und (4.5) kann nun direkt die Modellgleichung des PCMs hergeleitet werden - was im Detail mithilfe von Rost (1996, S. 201-203) nachvollziehbar ist -, die wie in Gl. (4.6) dargestellt werden kann. Damit wird es also möglich die Wahrscheinlichkeit zu berechnen, dass Person p bei Item i einen Score von x_{pi} erreicht, was der Anzahl an bewerkstelligten Stufen entspricht. Diese Berechnung geschieht in Abhängigkeit vom jeweiligen Personenparameter und Aufgabenparameter, wobei zugleich die Anzahl zur Verfügung stehender Antwortkategorien m_i berücksichtigt wird. Damit gilt es also alle Schwellenparameter des Items i zu berücksichtigen und entsprechend geschieht die Wahrscheinlichkeitsberechnung in Abhängigkeit von $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{im_i})$, welcher als Vektor alle Schwellenparameter zu einem spezifischen Item i enthält. Noch einmal sei darauf hingewiesen, dass sehr wohl zugelassen ist, dass die Aufgaben eines Tests unterschiedlich viele Antwortkategorien - ersichtlich anhand des Indizes von m_i - aufweisen können. Allgemein gesprochen wird die in Gl. (4.6) berechnete Wahrscheinlichkeit als bedingte Kategorienwahrscheinlichkeit bezeichnet und die Funktion als Item-Category-Response-Funktion (ICRF), mit der die Wahrscheinlichkeit einer bestimmten Kategorie bzw. Scores als Funktion der Fähigkeit einer Person und dem jeweiligen Schwellenparameter dargestellt wird.

$$\begin{aligned} \pi_{pix} = P(X_{pi} = x|\theta_p, \boldsymbol{\beta}_{ik}) &= \frac{exp[\sum_{j=0}^{x}(\theta_p - \beta_{ij})]}{\sum_{k=0}^{m_i} exp[\sum_{j=0}^{k}(\theta_p - \beta_{ij})]} \\ &= \frac{exp(x\theta_p - \sum_{j=0}^{x} \beta_{ij})}{\sum_{k=0}^{m_i} exp(k\theta_p - \sum_{j=0}^{k} \beta_{ij})} \\ &= \frac{exp(x\theta_p - \tau_{ix})}{\sum_{k=0}^{m_i} exp(k\theta_p - \tau_{ik})}, \end{aligned} \tag{4.6}$$

$$\text{mit} \quad x \in k = 0, 1, \ldots, m_i \quad \text{und} \quad \sum_{j=0}^{0} (\theta_p - \beta_{ij}) \equiv 0$$

Hierbei bezeichnet τ_{ix} bzw. τ_{ik} die Summe über alle Schwellenparameter die kleiner oder gleich dem Score x sind bzw. der Kategorie k. Auf diesen Parameter wird erst bei der konkreten Analyse des PCMs/GPCMs mittels ausgewählter R-Pakete zurückgegriffen und kann deshalb zunächst ungeachtet bleiben. Zu erwähnen ist an dieser Stelle noch, dass sich anhand von τ_{iK}, also der Summe aller Schwellenparameter des Items i, der s.g. Lokationsparameter errechnen lässt. Dieser Lageparameter des Items auf kontinuierlicher Dimension ergibt sich aus dem arithmetische Mittel von τ_{iK} und es gilt entsprechend $\frac{1}{K}\tau_{iK}$ zu bestimmen. Hieran kann dann Folgendes abgelesen werden: „Ist der Wert negativ, bedeutet dies, dass die Schwellen im Mittel negative Werte annehmen und dass das Item eher im unteren Bereich des latenten Merkmals differenziert. Ist der Wert positiv, diskriminiert das Item eher im oberen Bereich des latenten Merkmals“ (Eid, 2014, S. 237).

Die nachfolgende eingehendere Erklärung der Modellformel - dargestellt in Gl. (4.6) - erfolgt in Anlehnung an Eid (2014, Kapitel 5.4.1.1). Zunächst kann festgehalten werden, dass die Gleichung des PCMs so allgemein formuliert ist, dass sie auch für das niedrigste Level 0 durch Einsetzen eines Scores von $x = 0$ zutreffend ist. Wie aber bereits anhand von Gl. (4.3) ersichtlich wird, gibt es theoretisch gesehen keinen Schwellenparameter für die Kategorie 0. Dies ist insbesondere auch deshalb sinnvoll, da unabhängig von der Fähigkeit einer Person das Ausgangsniveau immer erreicht wird. Deshalb wird der notationellen Einfachkeit halber $\sum_{j=0}^{0}(\theta_p - \beta_{ij}) \equiv 0$ gesetzt. Damit beträgt schließlich die Wahrscheinlichkeit das Ausgangsniveau bzw. das niedrigste Level zu erlangen logischerweise 1. Dazu ist nun auch anzumerken, dass gemäß Muraki (1992, S. 6) β_{i0} kein Lokationsparameter im engeren Sinne darstellt. Dieser Parameter kann theoretisch gesehen jeden Wert annehmen, da sich der Term, welcher diesen Parameter enthält, sowieso aus Gl. (4.6) kürzen lässt:

$$\begin{aligned}\pi_{pix} &= \frac{exp(\theta_p - \beta_{i0}) \cdot exp[\sum_{j=1}^{x_{pi}}(\theta_p - \beta_{ij})]}{exp(\theta_p - \beta_{i0}) + \sum_{k=1}^{m_i} exp[(\theta_p - \beta_{i0}) + \sum_{j=1}^{k}(\theta_p - \beta_{ij})]} \\ &= \frac{exp[\sum_{j=1}^{x_{pi}}(\theta_p - \beta_{ij})]}{1 + \sum_{k=1}^{m_i} exp[\sum_{j=1}^{k}(\theta_p - \beta_{ij})]}\end{aligned} \tag{4.7}$$

Außerdem ist zu vermerken, dass der Nenner der Modellformel in Gl. (4.6) für einen spezifischen Fähigkeitsparameter θ_p konstant über alle Niveaus eines Items i ist. Gleichzeitig sorgt der Nenner somit für eine Normierung und damit dafür, dass die bedingten Kategorienwahrscheinlichkeiten $\pi_{pix} = P(X_{pi} = x|\theta_p, \beta_{ik})$ Werte von 0 bis 1 annehmen können.

Im Gegensatz zum Nenner werden im Zähler nur die Schwellenparameter bis zum jeweils erlangten Score x aufsummiert. Somit hängt der Zähler nur von der Fähigkeit der betrachteten Person θ_p und den bis hin zum Score x aufsummierten Schwellenparametern einer Aufgabe $\sum_{j=0}^{x_{pi}}\beta_{ij}$ ab. Zu beachten ist allerdings, dass der Personenparameter mit dem Score bzw. den bis dahin bewältigten Stufen x gewichtet wird. Diese Gewichtung ergibt sich, da alle Schwellen zwischen den einzelnen Kategorien bis hin zur x-ten mit dem jeweiligen Personenparameter θ_p verglichen werden. In den Worten von Rost (2013) ausgedrückt, „. . . setzt [jede Person] ihren Fähigkeits- oder Attitude-Parameter so oft erfolgreich ein, bis sie ihre Stufe erreicht hat (daher ein x als Koeffizient des Personenparameters)“. Entsprechend wird dann auch der Zähler umso größer, je größer die Differenz zwischen dem Fähigkeitsparameter und den Schwellenparametern ist. Genauer kann dies

auch anhand der später vorgestellten Gl. (4.9) nachvollzogen werden.

Allerdings hängt die bedingte Kategorienwahrscheinlichkeit nicht nur von den aufsummierten Differenzen von Personenparameter und der Schwierigkeit der Stufen bis hin zum Score x ab, sondern es finden durch den Nenner von Gl. (4.6) auch die oberhalb von x liegenden Schwellen Berücksichtigung. Dies bedingt, dass die Wahrscheinlichkeit für eine der mittleren Kategorien bei Zunahme der latenten Variable θ_p nicht ebenfalls immer größer wird, sondern nach Erreichen des jeweiligen Maximums wieder abfällt. Auf diesen Aspekt wird im anschließenden Kapitel 4.3 zurückgegriffen.

Letztlich kann also anhand von Gl. (4.6) auch bestimmt werden, welches Leistungsniveau von einer Person mit einem Fähigkeitswert θ_p am wahrscheinlichsten erreicht wird. Dies hängt aber - wie zuvor erläutert - von der Schwierigkeit aller vorhandenen Schwellen ab.

Kernstück des PCMs stellt gemäß Masters (1997, S. 103) der lokale Vergleich zweier benachbarter Kategorien $x-1$ und x dar, dementsprechend wird hier lokal das binäre RM unter Berücksichtigung des zugehörigen Kategorienparameters β_{ix} postuliert, was in Gl. (4.3) dargestellt wurde. Ähnlich wie in Gl. (3.4) lassen sich somit auch die Adjacent-Category Logits, also quasi die logarithmierte Chancen zweier benachbarter Kategorien k und $k-1$ bzw. bezogen auf den Score $x-1$ und x, bestimmen. Insgesamt stellt also der Quotient $\Phi_{pix}/(1-\Phi_{pix}$ die Chancen (Odds) dar und es wird damit angegeben, in welchem Verhältnis die Wahrscheinlichkeit für das Ereignis Φ_{pix} zu dem Nicht-Eintreten des Ereignisses, also $1-\Phi_{pix}$, steht.

$$\begin{aligned} logit(\Phi_{pix}) &= log(\frac{\Phi_{pix}}{1-\Phi_{pix}}) = log[\frac{\pi_{pix}/(\pi_{pi(x-1)}+\pi_{pix})}{\pi_{pi(x-1)}/(\pi_{pi(x-1)}+\pi_{pix})}] \\ &= log(\frac{\pi_{pix}}{\pi_{pi(x-1)}}) = \theta_p - \beta_{ix} \end{aligned} \tag{4.8}$$

Von besonderer Bedeutung ist beim PCM - wie bereits bei den Eigenschaften der IRT-Modelle in Kapitel 2 erwähnt wurde -, dass die stufenweisen Schwierigkeitsparameter β_{ix} bzw. allgemeiner formuliert β_{ik} separiert betrachtet werden können und damit auch unabhängig von den jeweiligen Personenparametern geschätzt werden. Auf diese Separierbarkeit und die damit einhergehende Suffizienz wird im Kapitel 4.4.4 noch genauer eingegangen. Vermerkt sei, dass die Separierbarkeit der einzelnen Schwierigkeitsparameter β_{ik} den entscheidenden Unterschied zu den Level-Parametern δ_{ik} aus Gl. (4.1) darstellt. Um noch einmal explizit zwischen den beiden Parametern β_{ik} und δ_{ik} zu differenzieren, sei verdeutlicht, dass β_{ik} die Schwierigkeit der k-ten Stufe - wobei $k = 0, \ldots, m_i$ - bei Aufgabe i zum Ausdruck bringt. Folglich kennzeichnet der β-Parameter des PCMs die Lokation der k-ten Schwellen auf der Skala der latenten Variable θ_p. Dagegen steht δ_{ik} für die Schwierigkeit der k-ten Kategorie und ergibt sich damit als die Summe der Schwierigkeiten aller Schwellen, die bis zum Erreichen der k-ten Kategorie gemeistert werden mussten. Insbesondere kann damit auch geschlussfolgert werden, dass die Schwellenparameter eines Items i nicht den Kategorien gemäß aufsteigend sortiert sein müssen, was für die δ-Parameter schon der Fall ist.

Um die Erklärung zu den Schwellenparametern β_{ik} abzurunden, kann auch noch kurz auf den Unterschied zwischen dem PCM und dem RSM von Andrich (1978) eingegangen werden, welches bereits zu Beginn des Kapitels 4 angesprochen wurde. Im PCM werden keine Annahmen bzw. Festsetzungen bezüglich der relativen Schwierigkeit der einzelnen Schwellen der Test-Items zueinander

getroffen. Prinzipiell kann es beispielsweise deutlich schwieriger sein Stufe 1 bei Item i zu bewältigen als Stufe 1 bei Item $\tilde{i}$, wobei gilt $i \neq \tilde{i}$. Dagegen ist das Rating Scale Format gekennzeichnet durch festen Satz geordneter „Bewertungspunkte“, welcher bei allen I Items gegeben ist (vgl. Masters, 1982, S. 162). Demgemäß sollte auch die relative Schwierigkeit der Schwellen sich zwischen den einzelnen Items nicht (bzw. kaum) unterscheiden.

An die theoretischen Aspekte des PCMs anknüpfend, soll im nachfolgenden Kapitel näher auf die grafische Darstellung des PCMs eingegangen werden.

4.3 Grafische Darstellung

Nun folgend sollen für das PCM zwei Möglichkeiten der grafischen Darstellung vorgestellt werden. Dies geschieht in Anlehnung an Rost (1996, Kapitel 3.3.1). Dazu ist es als insbesondere erforderlich die Begriffe „Schwellenwahrscheinlichkeit“ und „Kategorienwahrscheinlichkeit“ noch einmal genau voneinander abzugrenzen. Beide Arten der bedingten Wahrscheinlichkeiten wurden im vorhergehenden Kapitel bereits dargestellt. Die Schwellenwahrscheinlichkeit Φ_{pix} lässt sich in allgemeiner Form Gl. (4.3) entnehmen und die Kategorienwahrscheinlichkeit π_{pix} entspricht der bedingten Wahrscheinlichkeit des PCMs, welche in Gl. (4.6) dargestellt ist. Zunächst soll auf Erstere eingegangen werden.

4.3.1 Anhand der Schwellenwahrscheinlichkeiten

Um die grafische Darstellung des PCMs anhand der s.g. Schwellenwahrscheinlichkeiten verdeutlichen zu können, ist es vorerst erforderlich, den Ausdruck „Schwelle“ und damit von β_{ik} bzw. β_{ix} - bei Betrachtung eines bestimmten Scores x - noch einmal genau zu definieren: „Der Begriff der Schwelle soll suggerieren, dass an diesem Punkt auf dem Kontinuum der Übergang von einer Kategorie zur anderen stattfindet, d.h. die Wahrscheinlichkeit in der folgenden Kategorie zu antworten von diesem Punkt an größer wird als die Wahrscheinlichkeit, in der vorangegangenen Kategorie zu antworten“ (Rost, 1996, S.199). Konsequenter Weise haben auf der jeweiligen Schwelle, d.h. auf der Skala der latenten Variable, die beiden betrachteten Kategorien exakt dieselbe Wahrscheinlichkeit. Damit besteht an diesem Punkt also jeweils eine 50%-ige Wahrscheinlichkeit auf das niedrigere Level $x-1$ oder höhere Level x zweier benachbarter Niveaus zu gelangen. In Anlehnung an Kapitel 4.2, ist β_{ix} also genau derjenige Wert auf Skala der latenten Variable, an dem sich die jeweilige Schwelle befindet.

Bei aufmerksamer Betrachtung dieser Definition kann man direkte Parallelen zum RM (vgl. Kapitel 3) feststellen. Auch hier gibt es eine Schwelle, nämlich diejenige zwischen „falsch“ und „richtig“ bzw. zwischen Ablehnung und Zustimmung und somit zwischen Kategorie 0 und 1. Diese Schwelle entspricht also auch hier dem Schwierigkeitsparameter β_i des RMs und entspricht damit dem Schnittpunkt der beiden logistischen Funktionen in der oberen Grafik von Abb. 3.1 auf der latenten Skala.

An die vorhergehenden Überlegungen anknüpfend kann nun generell die mathematische Formulierung der Schwellenwahrscheinlichkeit hergeleitet werden. Diese kann als relativer Anteil der höheren Kategorienwahrscheinlichkeit π_{pix} an der Gesamtwahrscheinlichkeit zweier benachbarter

Kategorien $\pi_{pi(x-1)}$ und π_{pix} beschrieben werden. Dementsprechend liegt die Schwelle - wie gefordert - bei 0.5, falls die beiden Kategorienwahrscheinlichkeiten gleich sind. Insgesamt entspricht die Schwellenwahrscheinlichkeit Φ_{pix} also genau der in Gl. (4.3) bedingten Wahrscheinlichkeit von zwei benachbarten Kategorien die jeweils niedrigere oder höhere zu erreichen und damit also die sich dazwischen befindende x-te Stufe zu meistern unter gleichzeitig geltender Bedingung, dass Level $x-1$ oder x erlangt wird.

Hier kann noch einmal das RM gesondert betrachtet werden. Bei diesem gilt bei genauerer Überlegung nämlich, dass die Schwellenwahrscheinlichkeit identisch zur Lösungswahrscheinlichkeit ist. Denn es gilt bei detaillierter Betrachtung des mittleren Teils von Gl. (4.3), dass man für $\pi_{pi(x-1)} + \pi_{pix}$ im dichotomen Fall 1 erhält. Damit gilt auch, dass sich die Wahrscheinlichkeiten der beiden Kategorien 0 und 1 an jedem Punkt auf der latenten Skala zu 1 addieren. Dadurch bedingt ergibt sich also, dass die zusätzliche Darstellung der Kategorienfunktion von der 0-Kategorie redundant ist, was bereits anhand der obersten Grafik von Abb. 3.1 dargestellt wurde. Diese Kurve erhält man nämlich durch Spiegelung der 1er-Kategorienfunktion an dem durch β_i gefällten Lot. Wie anhand der nachfolgenden Sektion noch ersichtlich wird, entspricht die Darstellung in der obersten Grafik von Abb. 3.1 damit den Kategorienwahrscheinlichkeiten.

Um nun wiederum auf die Schwellenwahrscheinlichkeiten des PCMs zurückzukommen, nehme man Gl. (4.3) als Grundlage für diese Art der grafischen Darstellung, womit sich schließlich auch ein ähnliches Bild wie zuvor beim RM ergibt. Das Resultat sind demgemäß logistische Wahrscheinlichkeitsfunktionen, wie sie in der oberen Grafik von Abb. 4.2 beispielhaft anhand eines drei Kategorien umfassenden Items auf Skala der Personenfähigkeit dargestellt sind. Entsprechend wird anhand der einzelnen Kurven also die Wahrscheinlichkeit für das Bewältigen der Schwelle zweier benachbarter Kategorien in Abhängigkeit der zugrundeliegenden Personenfähigkeit abgebildet. Hierbei kann auch die jeweilige Schwierigkeit einer Stufe x abgelesen werden, die demjenigen Wert auf der Skala der latenten Variable entspricht, bei dem eine 50%-ige Wahrscheinlichkeit besteht das nächst höhere Niveau $x+1$ zu erreichen oder auf Level x zu verweilen.

Zu beachten ist generell, dass tendenziell nach links verschobene Kurven leichter zu bewältigenden Schwellen entsprechen. Demzufolge kennzeichnen eher rechts auf der Abszisse angeordnete Kurven schwerer zu meisternde Schwellen. Hierdurch kommt die Tatsache zum Ausdruck, dass mit zunehmender Schwierigkeit einer Stufe auch eine höhere Personenfähigkeit zum Erlangen der nächst höheren Niveauebene von Nöten ist. Dabei gilt es zu berücksichtigen, dass die Schwellen eines Items nicht zwingend den Kategorien gemäß aufsteigend arrangiert sein müssen. Zwar ist dies in der oberen Grafik von Abb. 4.2 der Fall und so sind die einzelne β-Parameter des Items den Kategorien entsprechend sortiert. Allerdings kann es auch vorkommen, dass sich beispielsweise die Bewältigung der Schwelle von Kategorie 1 zu 2 schwieriger gestaltet als die Überschreitung der Stufe von Kategorie 0 zu 1 und damit die Anordnung der Kurven auf Skala der latenten Variable andersherum erfolgt. Prinzipiell sollte noch erwähnt werden, dass sich auch die Schwellenwahrscheinlichkeiten mehrerer Items gemeinsam darstellen lassen.

In der nachfolgenden Sektion soll nun auf die bereits mehrfach angesprochene Darstellung des PCMs anhand der Kategorienwahrscheinlichkeiten eingegangen werden.

4.3.2 Anhand der Kategorienwahrscheinlichkeiten

Nun soll eine weitere Möglichkeit der grafischen Darstellung über die einzelnen Kategorienwahrscheinlichkeiten und damit der ICRFs zur Sprache kommen. Insbesondere sollte man sich hierbei im Klaren sein, welche zentrale Bedeutung die zugrundeliegende Ordnung der Antwortkategorien hat. Wiederum auf Skala der latenten Personfähigkeit betrachtet hat dies folgende Auswirkung: bei geringer Kompetenz einer Person dominiert die Wahrscheinlichkeit auf Niveauebene 0 zu verweilen. Mit zunehmender Fähigkeit steigt dann jedoch die Wahrscheinlichkeit an, auf das nächst höhere Level zu gelangen und damit sinkt allmählich die Wahrscheinlichkeit für Level 0 ab. Diese Beobachtung setzt sich für die nachfolgenden Kategorien fort, bis schließlich ab einer gewissen Personenfähigkeit die Wahrscheinlichkeit für die höchste Kategorie bzw. das höchste Leistungsniveau gegen 1 geht und entsprechend die Wahrscheinlichkeit für jegliche andere Kategorien gegen 0 strebt. Diese Veranschaulichung sollte als Grundlage für die im Folgenden vorgestellt grafische Darstellung der Kategorienwahrscheinlichkeiten dienen.

Eine relativ umfassende Erklärung der bedingten Kategorienwahrscheinlichkeiten findet sich bereits in Kapitel 4.2 bei der eingehenden Erklärung von der Modellformel des PCMs in Gl. (4.6). Der Anschaulichkeit halber soll nun diese Gl. (4.6) - wiederum bezogen auf ein dreistufiges und damit vier Kategorien umfassendes Item - aufgeschlüsselt nach den einzelnen vier Kategorien dargestellt werden. Insbesondere sollte hiermit deutlich werden, dass die zugrundeliegenden β-Parameter graphisch gesehen Einfluss auf die relative Höhe der benachbarter Kategorienwahrscheinlichkeiten nehmen.

$$
\begin{aligned}
P(Y_{pi}=0) &= \frac{1}{1+exp(\theta_p-\beta_{i1})+exp(2\theta_p-\beta_{i1}-\beta_{i2})+exp(3\theta_p-\beta_{i1}-\beta_{i2}-\beta_{i3})},\\
P(Y_{pi}=1) &= \frac{exp(\theta_p-\beta_{i1})}{1+exp(\theta_p-\beta_{i1})+exp(2\theta_p-\beta_{i1}-\beta_{i2})+exp(3\theta_p-\beta_{i1}-\beta_{i2}-\beta_{i3})},\\
P(Y_{pi}=2) &= \frac{exp(2\theta_p-\beta_{i1}-\beta_{i2})}{1+exp(\theta_p-\beta_{i1})+exp(2\theta_p-\beta_{i1}-\beta_{i2})+exp(3\theta_p-\beta_{i1}-\beta_{i2}-\beta_{i3})},\\
P(Y_{pi}=3) &= \frac{exp(3\theta_p-\beta_{i1}-\beta_{i2}-\beta_{i3})}{1+exp(\theta_p-\beta_{i1})+exp(2\theta_p-\beta_{i1}-\beta_{i2})+exp(3\theta_p-\beta_{i1}-\beta_{i2}-\beta_{i3})}.
\end{aligned}
\tag{4.9}
$$

Stellt man diese Kategorienwahrscheinlichkeiten als Kurven - s.g. „Item Category Characteristic Curves“ (ICCCs) - abermals auf Skala der latenten Variable θ_p dar, so ergeben sich derartige Grafiken, wie sie mittig und unten in Abb. 4.2 dargeboten sind. Die Aufgrund der Abhängigkeit der bedingten Kategorienwahrscheinlichkeiten von der latenten Personenfähigkeit spricht man hier allgemein auch von der s.g. Kategoriencharakteristik (-funktion) (vgl. Eid, 2014, S.237). Es sei darauf hingewiesen, dass die obere, sowie die mittlere Grafik aus Abb. 4.2 in direkten Zusammenhang stehen, während die unterste Grafik der Erklärung eines später folgenden Sachverhaltes dient. Anhand der oberen beiden Grafiken lassen sich nun beispielhaft die nachfolgenden Eigenschaften der Kategoriencharakteristiken nachvollziehen (vgl. Eid, 2014, S.237):

1. Mit zunehmender Personenfähigkeit nimmt die Wahrscheinlichkeit für die erste Kategorie

bzw. das Ausgangsniveau ab.

2. Die Wahrscheinlichkeiten für die mittleren Kategorien nehmen in aufsteigender Reihenfolge bis zu einem Maximum hin zu und anschließend wieder ab.

3. Dagegen nimmt die Wahrscheinlichkeit für das Erreichen der höchsten Leistungsebene mit Zunahme der latenten Variable ebenfalls zu.

4. Die Schnittpunkte der Kategoriencharakteristiken kennzeichnen die jeweiligen Schwellenparameter β_{ix}. Entsprechend ist hier die Wahrscheinlichkeit für die eine oder andere zweier benachbarter Kategorien gleich groß und beträgt entsprechend 50%. Somit gilt:

$$\theta_p = \beta_{ix} \Leftrightarrow \pi_{pix} = \pi_{pi(x-1)}$$

5. Durch den Bezug zu den den Schwellenparameter lassen sich die jeweiligen Schnittpunkte der Kategoriencharakteristiken auch als Wendepunkte der Schwellenwahrscheinlichkeiten deuten. Daraus kann abgeleitet werden, dass bei einem höher gelegenen Wert der Personfähigkeit verglichen mit dem Wert des Wendepunktes und damit von der Schwelle β_{ix}, die Präferenz hin zur jeweils höheren Kategorie steigt. Dies lässt sich wiederum auf zweierlei Weisen mathematisch ausdrücken:

$$\theta_p < \beta_{ix} \Leftrightarrow \pi_{pix} < \pi_{pi(x-1)}$$

$$\theta_p > \beta_{ix} \Leftrightarrow \pi_{pix} > \pi_{pi(x-1)}$$

Insbesondere sollte auch deutlich werden, dass zwar die Kategorienwahrscheinlichkeiten an sich in aufsteigender Reihenfolge geordnet sind, keinesfalls aber die einzelnen β_{ix}-Parameter. Es kann durchaus vorkommen, dass $\beta_{i(x+1)}$ einen kleiner Wert aufweist als der Schwellenparameter einer darunter liegenden Kategorie β_{ix} und somit die Bewältigung einer betrachteten Schwelle schwieriger ist als die einer darüber liegenden. Diese Tatsache lässt sich insbesondere anhand der untersten Grafik von Abb. 4.2 nachvollziehen. Hier ist nun beispielhaft ein drei Kategorien umfassendes Item mit ungeordneten Schwellenparametern dargestellt, wobei also gilt $\beta_{i2} < \beta_{i1} < \beta_{i3}$. Demzufolge besitzt die Kategoriencharakteristik der Kategorie 1 an keinem Punkt auf Skala der latenten Variable einen höheren Wert als die Kategoriencharakteristik der übrigen Kategorien des Items. Folglich wird also Kategorie 1 gemieden und mit zunehmender Personenfähigkeit ergibt sich dann, dass die Wahrscheinlichkeit für Kategorie 0 in Kategorie 2 übergeht. Bei derartigen Items bietet es sich zumeist an, gemiedene Kategorien zu entfernen und zu Prüfen, „ob die Antwortskala hierdurch bessere Eigenschaften aufweisen würde“ (Eid, 2014, S. 239).

Durch die vorhergehende Vorstellung der Modellformel, sowie durch die eben vorgestellten grafischen Veranschaulichungen, lassen sich bereits einige Eigenschaften des PCMs erkennen. Auf eben diese Eigenschaften und dem PCM zugrundeliegende Annahmen soll im anschließenden Kapitel genauer eingegangen werden.

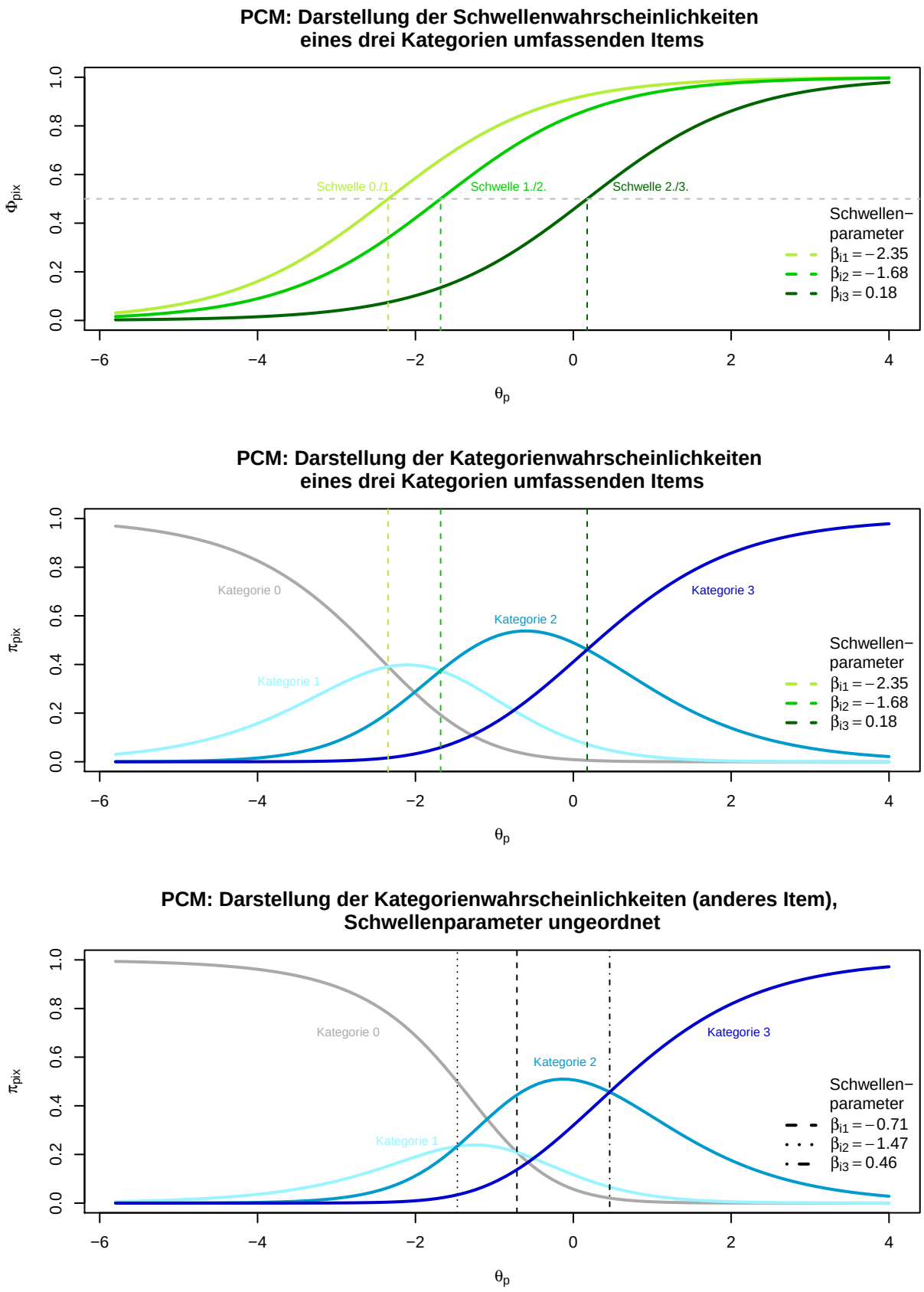

Abb. 4.2: Grafische Veranschaulichung des PCMs anhand der Schwellenwahrscheinlichkeiten, sowie der Kategorienwahrscheinlichkeiten von (zwei) beispielhaften Items

4.4 Annahmen und Eigenschaften

Da es sich bei dem PCM um eine mathematisches, probabilistisches Messmodell handelt, ist es wohl auch naheliegend, dass einige Annahmen erfüllt sein müssen, um das PCM letztlich zur Anwendung bringen zu können. Sind diese Annahmen gegeben, so können dann auch aussagekräftige Schlüsse aus den Parameterschätzungen gezogen werden. Zugleich bringt das PCM auch einige nützliche Eigenschaften mit sich, die insbesondere bei der später dargelegten Parameterschätzung ausgenutzt werden. Neben der Tatsache, dass zur Anwendung des PCMs ordinale Daten vorliegen müssen, können im Wesentlichen fünf weitere Annahmen und Eigenschaften des PCMs ausgemacht werden. Dazu wird in den folgenden Abschnitten auf die Eindimensionalität, lokale stochastische Unabhängigkeit, Suffizienz, spezifische Objektivität und das Messniveau näher eingegangen.

4.4.1 Eindimensionalität

Von grundlegender Bedeutung für die Anwendung des PCMs ist u.a. die Eindimensionalität, mit der die zentrale Annahme zum Ausdruck kommt, dass das Antwortverhalten einer Person nur von einer einzigen latenten Variable gesteuert wird. Dies bedeutet beispielhaft dargestellt also, dass von einem Mathematikitem verlangt wird, nur die mathematische Fähigkeit einer Person anzusprechen und somit nicht zusätzlich sprachliche Fertigkeiten zum Lösen einer Aufgabe verlangt werden. Die anschließende kurze Erläuterung zur Eindimensionalität erfolgt durch Übertragung dieses in Strobl (2012, S. 23) und Koller (2012, S. 15f.) dargelegten Sachverhalts für das RM auf das PCM.

Die Annahme der Eindimensionalität umfasst die s.g. Item- und Personenhomogenität, welche - wie in Kapitel 2 erwähnt - in allen IRT-Modellen vorausgesetzt werden. Hierbei wird durch die Aufgabenhomogenität festgesetzt, dass die Schwierigkeit der Aufgaben für alle Testpersonen identisch sein muss. Umgekehrt muss aber auch gelten, dass die gemessene Personenfähigkeit θ_p unabhängig von der gewählten Aufgabe ist, was als Personenhomogenität bezeichnet wird.

Damit impliziert die Eindimensionalität also, dass die Personenfähigkeit und die Aufgabenschwierigkeit bzw. genauer ausgedrückt die Schwellenparameter der einzelnen Aufgabe auf einer gemeinsamen latenten Dimension liegen. In der Modellgleichung wird dies durch die Subtraktion der beiden Parameter(-typen) voneinander ersichtlich, in der grafischen Darstellung sowohl von den Schwellen-, als auch von den Kategorienwahrscheinlichkeiten anhand der gemeinsamen Abtragung der Parameter auf der Abszisse. Damit wird auch klar, dass sich die Anordnung der Fähigkeitsparameter der jeweils betrachteten Personen auf der x-Achse nicht ändern darf, egal für welches Item. Diese Reihenfolge hängt also allein von der Lösungskompetenz der entsprechenden Probanden ab, die anhand der bewältigten Aufgabenschwellen gemessen werden soll.

4.4.2 Spezifische Objetivität

Eng mit der zuvor beschriebenen Annahme der Eindimensionalität bzw. deren Voraussetzungen der Item- und Personenhomogenität verknüpft, ist die im RM wie PCM geltende Eigenschaft der spezifischen Objektivität. Wiederum lässt sich diese Charakteristik, die anschaulich für das RM in Strobl (2012, S.20-23), Koller (2012, S. 19-22) und Irtel (1996) beschrieben wird, auf das PCM übertragen.

Garantiert wird mit dieser Annahme - wie dem Begriff bereits abzuleiten ist -, dass spezifisch objektive Vergleiche mittels des PCMs möglich sind. Genauer gesagt, wird dadurch einerseits gewährleistet, dass der Vergleich zweier Personen bzw. deren Fähigkeit nicht von den betrachten Items abhängt, und andererseits, dass der Vergleich zweier Items anhand ihrer Schwellenparameter unabhängig von den Personen ist, die diese Aufgaben bearbeitet haben.

Um das Ganze mathematisch zu veranschaulichen, betrachte man noch einmal die in Gl. (4.8) dargebotenen Adjacent-Categorie Logits. Mithilfe derer kann besonders gut die Annahme der spezifischen Objektivität verdeutlicht werden. Vergleicht man nämlich zwei verschiedene Personen p und $\tilde{p}$, so gilt dann Folgendes beim Vergleich anhand zweier benachbarter Kategorien:

$$\begin{aligned} &log[\frac{P(X_{pi}=x)}{P(X_{pi}=x-1)}] - log[\frac{P(X_{\tilde{p}i}=x)}{P(X_{\tilde{p}i}=x-1)}] = \\ &log[\frac{P(X_{pi}=x)/P(X_{pi}=x-1)}{P(X_{\tilde{p}i}=x)/P(X_{\tilde{p}i}=x-1)}] = -(\theta_p - \theta_{\tilde{p}}) \end{aligned} \tag{4.10}$$

Wie anhand von Gl. (4.10) ersichtlich wird, hängt der Vergleich zweier Personen einzig von deren θ-Parametern, nicht aber von der Kategorien der jeweiligen Aufgabe bzw. den Schwellenparametern ab, anhand deren man die Fähigkeit der beiden Probanden untersucht.

In analoger Weise kann auch gezeigt werden, dass beim Vergleich zweier Items i und $\tilde{i}$ bzw. genauer gesagt deren Schwellenparameter es keinen Einfluss hat, welche Personen die jeweiligen Items bearbeitet haben:

$$\begin{aligned} &log[\frac{P(X_{pi}=x)}{P(X_{pi}=x-1)}] - log[\frac{P(X_{p\tilde{i}}=x)}{P(X_{p\tilde{i}}=x-1)}] = \\ &log[\frac{P(X_{pi}=x)/P(X_{pi}=x-1)}{P(X_{p\tilde{i}}=x)/P(X_{p\tilde{i}}=x-1)}] = -(\beta_{ix} - \beta_{\tilde{i}x}) \end{aligned} \tag{4.11}$$

Somit gilt stets, dass leichtere Aufgaben insgesamt eine höhere Lösungswahrscheinlichkeit aufweisen als schwierigere. Dies führt dann zu der bereits erwähnten Verschiebung der Schwellenparameter entlang der Abszisse. In der oberen Grafik von Abb. 4.2 ist dieser Sachverhalt gut erkennbar. Damit kann schließlich geschlussfolgert werden, dass unter der gegebenen Annahme spezifischer Objektivität die logistischen Kurven - dargestellt anhand der Schwellenwahrscheinlichkeit - entweder identisch sind oder aber parallel zueinander verlaufen, was abhängig von den jeweiligen Schwellenparametern ist. Jedoch schneiden sie sich keinesfalls.

Gleichzeitig wäre die Annahme spezifischer Objektivität auch verletzt, insofern eine Aufgabe für einzelne Personengruppen leichter zu lösen ist, als für andere. Beispielsweise könnten Aufgaben für Probanden mit unterschiedlicher Ethnizität bzw. Muttersprache unterschiedlich schwierig sein. Damit könnte sich also bei einer Aufteilung der Testteilnehmer nach deutschsprachig und nichtdeutschsprachig für die zweite Gruppe erhebliche Verständnisfragen herauskristallisieren, welche nicht im Zusammenhang mit der eigentlichen Testfrage stehen. Dies bezeichnet man als „Differential Item Functioning“ oder kurz DIF. Derartige Aufgaben müssen zur Anwendung des PCMs aus dem Test entfernt werden. Wie dieses DIF erkannt werden kann, wird kurz in Kapitel 6 bei der

Vorstellung ausgewählter R-Pakete erklärt.

Zunächst sollen nun aber weitere Annahmen des PCMs präsentiert werden. Hierzu wird im Folgenden die lokale stochastische Unabhängigkeit genauer behandelt.

4.4.3 (Lokale) Stochastische Unabhängigkeit

Wie bereits in Kapitel 2 beschrieben wurde, zählt zu einer der grundlegenden Bedingungen der IRT-Modelle die Annahme der lokalen stochastischen Unabhängigkeit. Diese ist unverzichtbar für die Gewährleistung der Stichprobenunabhängigkeit der IRT-Modelle, durch welche ein klarer Vorteil gegenüber der KTT gegeben ist. Erneut lässt sich diese Eigenschaft, welche ausführlich in Strobl (2012, S. 16-20) und Koller (2012, S. 16ff.) behandelt wird, vom RM auf das PCM übertragen.

Ganz allgemein gesprochen ist die Unabhängigkeit zweier Ereignisse A und B wie folgt definiert: „Sind zwei Ereignisse [...] unabhängig [...], so ist es für die Wahrscheinlichkeit von A ohne Bedeutung, ob B eintritt, d.h. $P(A|B) = P(A)$“ (Fahrmeir, 2010, S. 206). Im PCM ist die stochastische Unabhängigkeit nun gegeben, insofern das Lösen einer Aufgabe nicht von einer vorhergehenden abhängt bzw. falls das Bewältigen einer Aufgabenschwelle unabhängig von der eines anderen Items möglich ist. Hätte man beispielsweise zwei Mathematikitems, bei denen man das Ergebnis der ersten Aufgabe zum Lösen der zweiten Aufgabe benötigt, so wäre das Prinzip der stochastischen Unabhängigkeit verletzt.

Andersherum kann die Annahme der stochastischen Unabhängigkeit aber auch auf der Personenseite verletzt sein. Dies wäre der Fall, falls eine Person bei einem anderen Testteilnehmer abschreiben kann. Diese Tatsache muss insbesondere bei der Planung der Testdurchführung Berücksichtigung finden.

Insofern die stochastische Unabhängigkeit gegeben ist, kann schließlich auch die Wahrscheinlichkeitberechnung aus Gl. (4.6) nicht nur bezüglich eines Items und einer Person betrachtet werden, sondern durch Multiplikation der Einzelwahrscheinlichkeiten auf beliebig viele P Personen und/oder I Items erweitert werden. Hiervon wird v.a. im nachfolgenden Unterkapitel 4.4.4 zur Suffizienz Gebrauch gemacht.

Der Zusatz „lokale“ stochastische Unabhängigkeit bezieht sich auf die Konstanthaltung des Personenparameters. Demzufolge müssen zwei Aufgaben nur voneinander unabhängig sein, solange man Personen mit der gleichen Fähigkeit betrachtet. Genauer gesagt, ist damit also zugelassen, dass eine Person mit einem sehr hohen θ-Wert alle Aufgaben mit hoher Wahrscheinlichkeit lösen kann, bzw. jeweils eine recht hohes Leistungslevel erreichen kann im Vergleich zu einem Probanden mit einem niedrigeren Fähigkeitswert. Insgesamt lässt sich schlussfolgern, dass bei Gültigkeit der Annahme lokaler stochastischer Unabhängigkeit die Inter-Itemkorrelation (nahezu) gänzlich durch die Personenfähigkeit erklärt werden kann.

4.4.4 Suffizienz

Eine der wohl wichtigsten Eigenschaften des PCMs und allgemein von den IRT-Modellen ist die Separierbarkeit der Parameter und die damit einhergehende Suffizienz. Wie im RM gibt es auch im PCM für die beiden unbekannten Parameter θ_p und β_{ix} jeweils eine suffiziente Statistik. Auch

hier lässt sich die Suffizienz analog zu Kapitel 3 definieren: eine Statistik $T(x)$ für einen nicht bekannten Parameter ν wird als suffizient bezeichnet, insofern sie genauso viel Information über diesen Parameter ν enthält wie die Stichprobe selbst (vgl. Kauerman, 2014, S. 15).

In dieser Sektion soll nun entsprechend gezeigt werden, dass die einzelnen Zeilenrandsummen von Tab. 4.1 bzw. der Score x eine suffiziente Statistik für den Fähigkeitsparameter θ_p repräsentieren und die entsprechenden Spaltenrandsummen s_{ix} als suffiziente Statistik für den jeweiligen Schwellenparameter β_{ix} angesehen werden können. Der nachfolgende Beweis zur im PCM geltenden Suffizienz geschieht in Anlehnung an Masters (1982, S. 159ff.).

Zunächst wird die Wahrscheinlichkeit modelliert, dass sich für eine Person p ein beliebiger Antwortvektor $\boldsymbol{x}_p = (x_{p1}, \ldots, x_{pI})$ bei einem I Aufgaben umfassenden Test ergibt. Um diese Wahrscheinlichkeit zu berechnen werden die Einzelwahrscheinlichkeiten der $i = 1, \ldots, I$ Items - wie sie in Gl. (4.6) dargestellt sind - miteinander multipliziert, was aufgrund der zuvor erläuterten lokalen stochastischen Unabhängigkeit möglich ist. Dementsprechend wird zur Berechnung der gesuchten Wahrscheinlichkeit dann auf den Vektor $\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{1m_i}, \ldots, \beta_{I1}, \ldots, \beta_{Im_i})$ mit den einzelnen Schwellenparametern aller I Aufgaben bedingt und es ergibt sich:

$$\begin{aligned} P(\boldsymbol{x}_p|\theta_p, \boldsymbol{\beta}) &= \prod_{i=1}^{I}\Big[\frac{exp[\sum_{j=0}^{x_{pi}}(\theta_p - \beta_{ij})]}{\sum_{k=0}^{m_i} exp[\sum_{j=0}^{k}(\theta_p - \beta_{ij})]}\Big] \\ &= \frac{exp[\sum_{i=1}^{I}\sum_{j=0}^{x_{pi}}(\theta_p - \beta_{ij})]}{\Psi_p}, \\ \text{mit} \quad \Psi_p &= \prod_{i=1}^{I}\Big[\sum_{k=0}^{m_i} exp[\sum_{j=0}^{k}(\theta_p - \beta_{ij})]\Big] \end{aligned} \tag{4.12}$$

Da im Folgenden verschiedene Personen und Items betrachtete werden, erstreckt sich die Summe im Zähler jetzt bis zum jeweiligen Score der p-ten Person beim i-ten Item und daher findet eine entsprechende Kennzeichnung mit x_{pi} statt. Zudem wird der Übersicht halber der Nenner in der letzten Zeile von Gl. (4.12) als Ψ_p zusammengefasst. Im Nachfolgenden Beweis der Suffizienz ist dieser dann sowieso zweitrangig, da Ψ_p gekürzt werden kann.

Geht man davon aus, dass der Testscore r_p eines betrachteten Teilnehmers bekannt ist, so lässt sich Gl. (4.12) faktorisieren. Vorher definiere man nun allerdings noch den Gesamt- bzw. Testscore r_p einer Person p bei einem I Items umfassenden Test. Dieser ergibt sich aus der Summe aller insgesamt von dem betrachteten Probanden bewältigten Stufen eines Tests, also als Summe der einzelnen Scores $\boldsymbol{x}_p$ und somit gilt $r_p = \sum_{i=1}^{I} x_{pi}$. Nimmt man nun die bereits angesprochene Faktorisierung von Gl. (4.12) vor, so ergibt sich Nachfolgendes:

$$P(\boldsymbol{x}_p|\theta_p, \boldsymbol{\beta}) = P(\boldsymbol{x}_p|r_p, \boldsymbol{\beta}) \cdot P(r_p|\theta_p, \boldsymbol{\beta}) \tag{4.13}$$

Schließlich kann noch die in vorhergehender Formel enthaltene bedingte Wahrscheinlichkeit für eine Person p einen Score von r in einem Test mit I Aufgaben zu erreichen, folgendermaßen errechnet werden:

$$
\begin{aligned}
P(r_p|\theta_p, \boldsymbol{\beta}) &= \frac{\sum_{\boldsymbol{x}_p}^{r_p} exp[\sum_i^I \sum_{j=0}^{x_{pi}} (\theta_p - \beta_{ij})]}{\Psi_p} \\
&= \frac{exp(r_p \cdot \theta_p)}{\Psi_p} \sum_{\boldsymbol{x}_p}^{r_p} exp(-\sum_{i=1}^{I} \sum_{j=0}^{x_{pi}} \beta_{ij})
\end{aligned}
\tag{4.14}
$$

Hierbei bezeichnet $\sum_{\boldsymbol{x}_p}^{r_p}$ entsprechend die Summe über alle diejenigen Antwortvektoren $\boldsymbol{x}_p$, sodass sich insgesamt ein Score von r_p ergibt. Demnach kann schließlich auch die bedingte Wahrscheinlichkeit berechnet werden, eine bestimmten Antwortvektor $\boldsymbol{x}_p$ zu erlangen, gegeben eines Gesamtscores von r_p. Hierzu ist es lediglich erforderlich Gl. (4.13) umzustellen und entsprechend dividiere man die bedingte Wahrscheinlichkeit aus Gl. (4.12) durch die von Gl. (4.14). Somit lässt sich dann Folgendes errechnen:

$$
\begin{aligned}
P(\boldsymbol{x}_p|r_p, \boldsymbol{\beta}) &= \frac{P(\boldsymbol{x}_p|\theta_p, \boldsymbol{\beta})}{P(r_p|\theta_p, \boldsymbol{\beta})} \\
&= \frac{exp(r_p\theta_p) \cdot exp(-\sum_{i=1}^{I} \sum_{j=0}^{x_{pi}} \beta_{ij})}{exp(r_p\theta_p) \cdot \sum_{\boldsymbol{x}_p}^{r_p} exp(-\sum_{i=1}^{I} \sum_{j=0}^{x_{pi}} \beta_{ij})} \\
&= \frac{exp(-\sum_{i=1}^{I} \sum_{j=0}^{x_{pi}} \beta_{ij})}{\sum_{\boldsymbol{x}_p}^{r_p} exp(-\sum_{i=1}^{I} \sum_{j=0}^{x_{pi}} \beta_{ij})}
\end{aligned}
\tag{4.15}
$$

Wie der vorhergehenden Gl. (4.15) zu entnehmen ist, ist zur Berechnung der bedingten Wahrscheinlichkeit einen bestimmten Antwortvektor $\boldsymbol{x}_p$ in einem Test zu erhalten, gegeben dem Gesamtscore r_p, nicht von Bedeutung, welchen Fähigkeitswert die jeweils betrachtete Person p besitzt. Es wird lediglich der Testscore r_p benötigt. Der Antwortvektor $\boldsymbol{x}_p$ beinhaltet also keine zusätzlichen Informationen über die Fähigkeit einer Person p, die nicht bereits mittels des Testscores r_p vermittelt werden konnten und demgemäß stellt r_p also eine suffiziente Statistik für θ_p dar. Anders ausgedrückt bedeutet dies, dass für die Anwendung des PCMs nicht von Bedeutung ist, welchen Fähigkeitsparameter eine Person besitzt, die einen Score von r_p bei einem Test erreicht, sondern ist nur abhängig von der relativen Schwierigkeit der einzelnen Stufen der I Items.

Führt man diese Überlegungen weiter, so kann auch die bedingte Wahrscheinlichkeit für die ganze Matrix $\boldsymbol{X}_{P\times I}$ gegebener Antworten - also eine Matrix bestehend aus den Antwortvektoren $\boldsymbol{x}_p$ aller P betrachteten Personen - gegeben dem Vektor mit den jeweiligen Gesamtscore der einzelnen Teilnehmer $\boldsymbol{r} = (r_1, \ldots, r_P)$ formuliert werden:

$$
P(\boldsymbol{X}|\boldsymbol{r}, \boldsymbol{\beta}) = \prod_{p=1}^{P} \Big[\frac{exp(-\sum_{i=1}^{I} \sum_{j=0}^{x_{pi}} \beta_{ij})}{\sum_{\boldsymbol{x}_p}^{r_p} exp(-\sum_{i=1}^{I} \sum_{j=0}^{x_{pi}} \beta_{ij})} \Big] \tag{4.16}
$$

In analoger Weise kann ebenso gezeigt werden, dass die jeweilige Anzahl der Personen, die die einzelnen Stufen eines Items i bewerkstelligen können, eine suffiziente Statistik für die Schwellenschwierigkeiten darstellen. Wiederum soll dies schrittweise gezeigt werden. Dazu überlege man sich

zunächst, wie sich die bedingte Wahrscheinlichkeit für das Beobachten eines bestimmten Antwortvektors $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{Pi})$ zu einem Item i von P Personen ermitteln lässt, was in Gl. (4.17) dargestellt ist. Dabei müssen nun auch alle P Fähigkeitsparameter berücksichtigt werden. Diese sind in dem Vektor $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)$ zusammengefasst. Außerdem gilt zu beachten, dass nun mehr nur die m_i Schwellenparameter des betrachteten Items i von Interesse sind, welche in dem Vektor $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{im_i})$ inbegriffen sind.

$$\begin{aligned} P(\boldsymbol{x}_i|\boldsymbol{\theta}, \boldsymbol{\beta}_i) &= \prod_{p=1}^{P} \Big[\frac{exp[\sum_{j=0}^{x_{pi}} (\theta_p - \beta_{ij})]}{\sum_{k=0}^{m_i} exp[\sum_{j=0}^{k} (\theta_p - \beta_{ij})]} \Big] \\ &= \frac{[exp(\sum_{p=1}^{P} x_{pi}\theta_p)]\,[exp(-\sum_{p=1}^{P}\sum_{j=0}^{x_{pi}} \beta_{ij})]}{\Psi_i}, \\ \text{mit} \quad \Psi_i &= \prod_{p=1}^{P} \Big[\sum_{k=0}^{m_i} exp[\sum_{j=0}^{k} (\theta_p - \beta_{ij})] \Big] \end{aligned} \tag{4.17}$$

Zugleich kann auch die Wahrscheinlichkeit ermittelt werden einen spezifischen Vektor $\boldsymbol{s}_i = (s_{i1}, \ldots, s_{im_i})$ bei dem betrachteten Item i zu erhalten. Dabei beinhaltet dieser Vektor $\boldsymbol{s}_i$ die einzelnen Summen der Personen, die ein Level k bzw. Score x erreicht haben und somit die j-te Stufe bis hin zu Kategorie k bzw. Score x bewerkstelligen konnten. Es ergibt sich also Folgendes:

$$P(\boldsymbol{s}_i|\boldsymbol{\theta}, \boldsymbol{\beta}_i) = \frac{[\sum_{\boldsymbol{x}_i}^{\boldsymbol{s}_i} exp(\sum_{p=1}^{P} x_{pi}\theta_p)]\,[exp(-\sum_{p=1}^{P}\sum_{j=0}^{x_{pi}} \beta_{ij})]}{\Psi_i} \tag{4.18}$$

Entsprechend stellt in Gl. (4.18) die Summe $\sum_{\boldsymbol{x}_i}^{\boldsymbol{s}_i}$ die Summe aller Antwortvektoren dar, die den Vektor $\boldsymbol{s}_i$ bilden. Wiederum kann nun die bedingte Wahrscheinlichkeit berechnet werden, den Antwortvektor $\boldsymbol{x}_i$ zu erhalten, gegeben den Vektor $\boldsymbol{s}_i$. Hierzu dividiere man die modellierte Wahrscheinlichkeit aus Gl. (4.17) durch die von Gl. (4.18), womit sich schließlich das Folgende ergibt:

$$\begin{aligned} P(\boldsymbol{x}_i|\boldsymbol{\theta}, \boldsymbol{s}_i) &= \frac{P(\boldsymbol{x}_i|\boldsymbol{\theta}, \boldsymbol{\beta}_i)}{P(\boldsymbol{s}_i|\boldsymbol{\theta}, \boldsymbol{\beta}_i)} \\ &= \frac{exp(\sum_{p=1}^{P} x_{pi}\theta_p)}{\sum_{\boldsymbol{x}_i}^{\boldsymbol{s}_i} exp(\sum_{p=1}^{P} x_{pi}\theta_p)} \end{aligned} \tag{4.19}$$

Mit Gl. (4.19) wird nun ersichtlich, dass die Schwierigkeit der einzelnen Aufgabenstufen β_{ix} durch Bedingen auf den Vektor $\boldsymbol{s}_i$ gänzlich unberücksichtigt bleibt. Damit ist es bei Anwendung des PCMs also nicht nötig all die Information, die einer Datenmatrix - in Form von Tab. 4.1 - bezüglich der Schwierigkeit der einzelnen Schwellen eines Items entnommen werden können, genau zu kennen. Stattdessen reicht es aus die Anzahl der Personen zu kennen, die die jeweilige Stufe eines Items bewerkstelligen konnten und damit stellt die einzelnen Spaltenrandsummen - zusammengefasst als Vektor $\boldsymbol{s}_i$ - eine suffiziente Statistik für den jeweiligen Schwellenparameter β_{ix} dar. Insgesamt umfasst also der Vektor $\boldsymbol{s} = (s_{11}, \ldots, s_{1m_1}, \ldots, s_{I1}, \ldots, s_{Im_I})$ all die nötigen Informationen, die einer Datenmatrix $\boldsymbol{X}$ über die Schwellenparameter $\boldsymbol{\beta}$ entnommen werden können.

Insbesondere wird hier auch eine für die Praxis relevante Tatsache deutlich: insofern Daten mittels

des PCMs ausgewertet werden können, ist die Möglichkeit gegeben, die Kompetenz einer Person unabhängig von den genutzten Items zu beurteilen. Genauer gesagt können dann statistisch gesehen äquivalente Fähigkeitsmessungen mittels Tests durchgeführt werden, die sich entweder in der Anzahl an Items und/oder der Stufenstruktur der einzelnen Aufgaben und/oder der Schwierigkeit der Bewältigung einzelner Schwellen unterscheiden.

Um dieses Kapitel zu den grundlegenden Eigenschaften bzw. Annahmen des PCMs abzurunden, gilt es noch auf eine weiter bedeutende Charakteristik des PCMs einzugehen, welche in der nachfolgenden Sektion dargelegt wird.

4.4.5 Messniveau

Abschließend sei noch das zugrundeliegende Messniveau zu beschreiben. Bezüglich dessen lässt sich die in Kapitel 3 getroffene Annahme direkt auf das PCM übertragen: das PCM besitzt nur bis auf Intervallskalenniveau, die metrische Skalen mit nur einem relativen Nullpunkt kennzeichnen, Eindeutigkeit. Gemäß Fischer (1995, Kap. 2) lässt sich dies dadurch begründen, dass sich Absolutskalen - im Gegensatz zu Intervallskalen - mit absolutem Nullpunkt für die Fähigkeits- und Schwellenparameter kaum rechtfertigen lassen, unter dem Aspekt, dass diese Parameter latente, also nicht direkt messbare Eigenschaften darstellen, ihnen insbesondere also kein natürlicher Nullpunkt zugeordnet werden kann.

Im Rahmen der Beschreibung des PCMs sollte noch auf eine generell gehalteneres, weniger restriktives Modell, das s.g. Generalisierte Partial Credit Modell (GPCM), eingegangen werden. Dieses lässt im Gegensatz zum PCM unterschiedliche Steigungsparameter zu, wodurch sich teilweise auch veränderte Annahmen/Eigenschaften ergeben. Eben dieses GPCM wird im nachfolgenden Kapitel noch genauer eingegangen.

4.5 Generalisiertes Partial Credit Modell (GPCM)

In ähnlicher Weise wie Birnbaum Erweiterungen zu Rasch's RM konstruierte, erarbeitete auch Muraki (1992) ein verallgemeinertes Modell zum PCM, das s.g. Generalisierte Partial Credit Modell (GPCM). Grundlegender Unterschied zwischen dem 2PL-Modell und dem GPCM ist das zugrundeliegende Datenformat: Birnbaum's Modell lässt sich bei Items mit dichotomer Antwortstruktur anwenden, während das von Muraki bei polytomen Items genutzt werden kann. Nun folgend soll auf eben dieses GPCM kurz eingegangen werden und die Unterschiede bzw. die Gemeinsamkeiten mit dem PCM herausgearbeitet werden. Dazu wird im Wesentlichen Bezug auf Muraki (1992) und Muraki (1997) genommen.

Wie bereits in Kapitel 2 anhand von Abb. 2.1 ersichtlich wird, handelt es sich dabei um ein 2PL-IRT-Modell. Entsprechend beinhaltet das GPCM sowohl einen Lokations-, als auch einen Steigungsparameter. Daher konstruierte Muraki - wie Masters (1982) das PCM - auch das GPCM schrittweise, wobei eben ein Steigungs-/Diskriminationsparameter α_i ergänzt wurde. Das GPCM entspricht letztlich also dem PCM unter Hinzunahme des Diskriminationsparameters α_i, der entsprechend des Indizes zwischen den einzelnen Items eines Tests variieren kann.

$$\pi_{pix} = P(X_{pi} = x|\theta_p, \alpha_i, \beta_{ik}) = \frac{exp[\sum_{j=0}^{x} \alpha_i(\theta_p - \beta_{ij})]}{\sum_{k=0}^{m_i} exp[\sum_{j=0}^{k} \alpha_i(\theta_p - \beta_{ij})]} \tag{4.20}$$

Theoretisch kann der Steigungsparameter α_i jeglichen Wert zwischen $-\infty$ und ∞ annehmen. Nach De Ayala (2009) sollte dieser aber in einem Bereich zwischen 0.8 und 2.5 liegen. Letztlich beeinflusst dieser Steigungsparameter, den Grad bzw. das Ausmaß, mit welchem die kategoriellen Antworten zwischen den Items variieren, wenn sich der Fähigkeitsparameter ändert. Der α-Parameter im GPCM bringt also wie beim 2PL-Modell zum Ausdruck, wie stark die einzelnen Items, mittels der gegebenen Antworten der Probanden, zwischen den jeweiligen Personenfähigkeiten unterscheiden bzw. diskriminieren können.

Zudem soll noch kurz eine weitere Darstellung, welche auch beim PCM dargeboten wurde, berücksichtigt werden. Auch für das GPCM kann wieder ein lokaler Vergleich zweier benachbarter Kategorien $x-1$ und x vorgenommen werden. Infolgedessen ergeben sich die Adjacent-Category Logits. Diese nehmen für das GPCM die folgende Form an:

$$logit(\Phi_{pix}) = log(\frac{\pi_{pix}}{\pi_{pi(x-1)}}) = \alpha_i(\theta_p - \beta_{ix}) \tag{4.21}$$

Etwas weiter ausholend soll eine weitere Darstellungsmöglichkeit für das GPCM bzw. durch entsprechende Anpassung auch für das PCM vorgestellt werden. Häufig geht man - gerade beim GPCM - in ähnlicher Weise vor wie Andrich (1978) bei dem von ihm konstruierten RSM und unterteilt den Schwierigkeitsparameter folgendermaßen:

$$\beta_{ix} = \beta_i - \nu_k \tag{4.22}$$

Die Erläuterungen hierzu beziehen sich auf Muraki (1997, S. 154ff.). Der Kategorien-Schwellenparameter ν_k lässt sich dann als relative Schwierigkeit der Kategorie k verglichen mit den anderen Kategorien eines Items i interpretieren oder als Abweichung von dem Item-Lokationsparameter β_i. Demgemäß müssen also die einzelnen ν_k's nicht unbedingt der Reihenfolge der Kategorien folgend sortiert sein. Um diesen Parameter letztlich auch eindeutig bestimmen zu können wird die folgende Bedingung gestellt:

$$\sum_{k=2}^{m_i} \nu_k = 0 \tag{4.23}$$

Unter Berücksichtigung dieser Aufteilung des β-Parameters kann dann der Exponent - hier bezogen auf den Zähler - von Gl. (4.20) umgeschrieben werden zu:

$$\sum_{j=0}^{x} \alpha_i[T_x(\theta_p - \beta_i) + K_x], \tag{4.24}$$

$$\text{mit} \quad K_x = \sum_{j=0}^{x} \nu_x$$

Andrich (1978) bezeichnet T_x und K_x als Scoring-Funktion und Kategorienkoeffizient. Für das GPCM - sowie auch das PCM - ist die Scoring-Funktion T_k bzw. T_x eine lineare, ganzzahlige

Funktion der Form $T_k = k$ bzw. $T_x = x$. Hierbei gilt allgemein, dass das GPCM nur dann als Modell für geordnete, kategoriale Antworten dient, wenn die Scoring-Funktion eine linear steigende Funkion darstellt, also wenn $T_k > T_k - 1$ für alle Kategorien innerhalb eines Items i gilt und insofern $\alpha_i > 0$ ist. Entsprechend kann der erwartete Wert der Scoring-Funktion - wie nachfolgend dargestellt - berechnet werden. Hierbei bezeichnet $\tilde{T}(\theta)$ die s.g. Item-Response-Funktion (IRF) für ein Item mit polytomen Bewertungsformat und zugleich gegebenen Scoring-Funktionen T_k, wobei $k = 1, \ldots, m_i$ gilt. Damit kann die IRF als Regression des Item-Scores bezüglich der Fähigkeitsskala angesehen werden (vgl. Muraki, 1997, S. 156). Im GPCM ergibt sich gemäß Muraki (1993) die IRF als bedingtes, arithmetisches Mittel von Item-Scores zu einem gegeben θ_p.

$$\tilde{T}(\theta_p) = \sum_{k=1}^{m_i} T_k \pi_{pik} \tag{4.25}$$

Unter Berücksichtigung der bisher vorgestellten Informationen zum GPCM lässt sich das PCM also als Spezialfall aus diesem ableiten, indem man für alle Items $i = 1, \ldots, I$ den gleichen Steigungsparameter von $\alpha_i = 1$ annimmt. Hier sei erwähnt, dass bei der Vorstellung der R-Pakete in Kapitel 6 zudem eine weniger restriktive Form des PCMs vorgestellt wird, die zunächst allerdings unberücksichtigt bleibt.

In jedem Fall sollte dementsprechend also ersichtlich werden, dass sich bei der grafischen Darstellung der Kategorienwahrscheinlichkeiten des GPCMs im Vergleich zu der des PCMs lediglich der Anstieg der Kurven ändert, nicht aber die Lokation der Schwellenparameter auf Skala der latenten Variable. Weiterhin entspricht der β-Parameter also demjenigen Wert auf der Personenfähigkeitsskala, bei dem eine 50%-ige Chance für das höhere oder niedrigere Level zweier benachbarter Kategorien besteht. Damit ist der Lokationsparameter bei der grafischen Darstellung der Kategorienwahrscheinlichkeiten anhand des Schnittpunkts zweier benachbarter Kurven abzulesen. Demzufolge werden die Kurven des GPCMs mit $\alpha_i < 1$ flacher im Vergleich zu denen des PCMs und damit die Prädiktion des Fähigkeitsparameters ungenauer. Umgekehrt lässt sich festhalten, dass bei $\alpha_i > 1$ der Anstieg der Kurven verglichen mit dem des PCMs zunimmt und damit Aussagen über die Kompetenz einer Person umso präziser werden. Es lässt sich also wiederum Bezug zu der in Kapitel 3 erwähnten Trennschärfe nehmen, derzufolge man umso besser zwischen zwei Testindividuen mit unterschiedlich stark ausgeprägter Fähigkeit unterscheiden kann, je stärker die Steigung im mittleren Bereich der Kurve ist. In der Praxis verfolgt man also das Ziel, möglichst Items mit hohem α_i zu ermitteln, um somit mehr bzw. genauere Informationen zu einer Person zu erfassen.

Nun sollte noch geklärt werden, inwiefern die Annahmen und Eigenschaften, die im PCM gelten - siehe dazu Kapitel 4.4 -, auf das GPCM übertragbar sind. Da das GPCM, ebenso wie das PCM, ein eindimensionales IRT-Modell darstellt, lässt sich die Annahme der Eindimensionalität, welche in Sektion 4.4.1 genauer behandelt wurde, direkt auf das GPCM anwenden. Wie bereits in Kapitel 3 bei der Vorstellung des 2PL-Modells von Birnbaum (1968) ersichtlich wurde, gilt schließlich auch beim GPCM, dass aufgrund der Einführung des Diskriminationsparameters α_i keine objektiven Vergleiche mehr möglich sind. Wegen des Steigungsparameters werden nämlich unterschiedliche Trennschärfen der einzelnen Items zugelassen und damit ist die Annahme der spezifischen Objekti-

vität hier nicht gegeben. Dementgegen ändert die Einführung des Diskriminationsparameters aber nichts an der Gültigkeit der lokalen stochastischen Unabhängigkeit. Anders ausgedrückt ist die lokale stochastische Unabhängigkeit eine notwendige Bedingung für die Anwendbarkeit des GPCMs. Dieser Sachverhalt wurde bereits in Sektion 2 dargelegt, wobei darauf hingewiesen wurde, dass die angesprochene Eigenschaft für die in IRT-Modellen geltende Stichprobenunabhängigkeit unbedingt erfüllt sein muss. Bezüglich der Suffizienz lässt sich wohl wie beim 2PL-Modll von Birnbaum (1968) annehmen, dass jeweils die Summe der mit den Steigungsparametern gewichteten Schwellenparameter eine suffiziente Statistik für den Fähigkeitswert einer Person darstellt. Aufgrund fehlender Quellen kann dieser Aspekt jedoch nur durch Übertragung der Annahme, welche in dargestellt in Rost (1996) ausführlich beschrieben wird, vom 2PL-Modell auf das GPCM bewerkstelligt werden. Allerdings lässt sich dies nicht aussagekräftig belegen. Die Annahme bezüglich des Messniveaus ändert sich bei der Einführung des Steigungsparameters nicht und damit ist diese Annahme des PCMs, welche in Kapitel 4.4.5 dargelegt wurde, direkt übertragbar auf das GPCM.

Mit diesen grundlegendem Wissen zum PCM, sowie GPCM kann im nachfolgenden Kapitel näher auf die möglichen Parameterschätzungen der beiden Modelle eingegangen werden. Hierzu sind einige der vorgestellten Annahmen unverzichtbar, worauf aber explizit bei der Vorstellung der einzelnen Methoden eingegangen wird.

5 Schätzmethoden für das PCM/GPCM

Bisher wurde eingehend das PCM, sowie hieran anknüpfend das GPCM vorgestellt und insbesondere die Eignung für die Analyse von Tests mit ordinalen Items. Doch wie kann man denn nun die Schwierigkeit einer Aufgabenschwelle ermitteln oder wie geht die Fähigkeit einer Person aus den jeweiligen Antwortverhalten hervor? Um dies zu bestimmen, bedient man sich der Parameterschätzung. Hierbei gibt es verschiedene Ansätze, wobei diese sich zumeist an der Maximum-Likelihood-Methodik oder an bayesschen Verfahren orientieren. Die meisten der in diesem Kapitel ausführlicher dargestellten Methoden basieren auf der Maximum-Likelihood (ML) Schätzung, womit insbesondere die Itemparameter abgeschätzt werden. Für die Schätzung der Personenparameter wird zudem kurz auf zwei bayessche Verfahren eingegangen.

Zunächst soll kurz der Grundgedanke der ML-Methode umrissen werden. Prinzipiell ergibt sich für jede Person eine Stichprobe von $x_{p1}, x_{p2}, \ldots, x_{pI}$ Antworten - zusammengefasst in dem Vektor $\boldsymbol{x}_p$ - zu den insgesamt I Aufgaben und für jedes Item erhält man eine Stichprobe $s_{i1}, s_{i2}, \ldots, s_{im_i}$, welche die Anzahl der Personen, die die jeweilige Kategorie k des Items i erreichen konnten, widerspiegelt und im Vektor $\boldsymbol{s}_i$ inbegriffen sind. Nun kann man die Wahrscheinlichkeit für das Auftreten einer derartigen Stichprobe berechnen, wobei gleichzeitig davon ausgegangen wird, dass die einzelnen x_{pi}'s bzw. die s_{ik}'s einer jeweiligen Wahrscheinlichkeitsdichte von $f(x_{pi}|\theta_p)$ bzw. $g(S_{ik}|\beta_{ik})$ folgen. Nachdem im PCM die lokale stochastische Unabhängigkeit gilt, lässt sich dann allgemein das Auftreten der entsprechenden Stichprobe als „Produkt der Wahrscheinlichkeiten für das Auftreten jedes einzelnen Elements der Stichprobe“ (Kolanoski, 2008) ermitteln. Die so errechnete Wahrscheinlichkeit wird als Likelihood-Funktion bezeichnet, welche im Folgenden stets durch ein Λ gekennzeichnet wird.

Gemäß Eid (2014, S. 158) soll hierbei der Begriff „Likelihood“ verdeutlichen, dass nicht die Wahrscheinlichkeit für das Eintreten eines bestimmten Ereignisses ermittelt wird, sondern lediglich der Parameterwert geschätzt wird, der am „naheliegensten“ bzw. „plausibelsten“ erscheint - im Englischen „likely“ - unter Berücksichtigung der bereits eingetretenen Itemantworten. Entsprechend bedient man sich zwar in beiden Fällen derselben Gleichung, doch besteht ein konzeptueller Unterschied.

Um nun eben denjenigen bzw. diejenigen Modellparameter zu ermitteln, die am „plausibelsten“ für die vorliegende Stichprobe sind, gilt es das Maximum der Likelihood-Funktion zu ermitteln. Hierin begründet sich der Begriff der Maximum-Likelihood (ML) Schätzung. Es sei darauf hingewiesen, dass man aus numerischen Gründen zumeist den Logarithmus der Likelihood-Funktion zur Bestimmung des Maximums verwendet, welcher in dieser Arbeit durch den Parameter $\lambda = log(\Lambda)$ gekennzeichnet ist. Um dieses Maximum zu berechnen, bedarf es der ersten Ableitung der Likelihood-Funktion Λ bzw. der Log-Likelihood-Funktion λ. Diese wird gleich 0 gesetzt und nach dem/den gesuchten Parameter(n) aufgelöst. Generell handelt es sich hierbei um eine Reihe an partiellen Ableitungen der Log-Likelihood bezüglich $\boldsymbol{\beta}$. Dies wird auch als s.g. Score-Funktion $s(\boldsymbol{\beta})$ bezeichnet,

woraus dann also der ML-Schätzer folgendermaßen hervor geht:

$$s(\boldsymbol{\beta}) := \frac{\partial}{\partial \boldsymbol{\beta}} \lambda = 0 \tag{5.1}$$

Demgemäß ist $s(\boldsymbol{\beta})$ also ein Vektor, bestehend aus den ersten Ableitungen zu jedem Element aus $\boldsymbol{\beta}$. Durch Null setzen der Score-Funktion ergibt sich ein nichtlineares Gleichungssystem, welches mit numerischen Verfahren - etwa der Newton-Raphson-Methode - gelöst werden kann.

Da mittels der ersten Ableitung allgemein Extremwerte bestimmt werden, wozu auch das/die Minima einer Funktion gehören, muss zusätzlich bestimmt werden, um welche Art von Extremwert es sich konkret handelt. Dies geschieht in diesem Fall über die s.g. Hesse-Matrix $H(\boldsymbol{\beta})$, die existiert, insofern die untersuchte Funktion zweimal stetig differenzierbar ist. Demgemäß sind in dieser Matrix die zweiten partiellen Ableitungen zusammengefasst. Die Hesse-Marix lässt sich also folgendermaßen definieren:

$$H(\boldsymbol{\beta}) = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \lambda = \frac{\partial}{\partial \boldsymbol{\beta}} s(\boldsymbol{\beta}) \tag{5.2}$$

Damit sich schließlich zeigt, dass es sich bei den ermittelten Extremwerten um Maxima handelt, muss gelten, dass die Hesse-Matrix immer negativ definit ist. Die Definitheit wird dabei mithilfe der Determinante bestimmt.

Auf Basis dieses Wissens können nachfolgend drei mögliche Methoden der Itemparameterschätzung vorgestellt werden. Diese sind die gemeinsame, die bedingte und die marginale ML-Schätzung, abgekürzt mit JML, CML und MML. Bei ersterer Methode werden die Fähigkeits- und Aufgabenparameter gleichzeitig geschätzt. Bei den anderen beiden Methoden werden zunächst nur die Itemparameter geschätzt, wobei hier jeweils ein unterschiedliches Vorgehen gewählt wird, um die Personenparameter anfangs unberücksichtigt zu lassen. Detailliert werden diese Schätzungen in den nachfolgenden Sektionen beschrieben. Es sei vermerkt, dass alle drei Methoden auf das PCM anwendbar sind, sich für das GPCM nur die MML-Methode eignet. Aus diesem Grund werden die drei Methoden im Wesentlichen anhand des PCMs dargestellt, wobei bei der MML-Methode kurz das Vorgehen bei der Schätzung von Itemparametern des GPCMs umrissen wird.

Gleichzeitig sei noch einmal darauf hingewiesen, dass im PCM insbesondere die Annahme der spezifischen Objektivität zu gelten hat. Es wird also vorausgesetzt, dass kein DIF vorliegt und sich somit die Lösungswahrscheinlichkeit eines Items bzw. die Bewältigung einer Schwelle bei Personen mit gleich großer Fähigkeit nicht zwischen einzelnen Subgruppen unterscheidet. Zudem gilt zu beachten, dass bei der nachfolgenden Vorstellung der Schätzmethoden von vollständigen Daten ausgegangen wird. Zum Umgang mit diesen fehlenden Werten gibt es unterschiedliche Vorgehensweisen, auf einige dieser wird bei der Vorstellung der R-Pakete in Kapitel 6 noch genauer eingegangen.

Zusätzlich ist noch anzumerken, dass zur eindeutigen Identifizierbarkeit der Parameter eine Normierung vorgenommen werden muss. Hierbei gibt es prinzipiell zwei Möglichkeiten, bei der entweder die s.g. Summennormierung oder eine Fixierung gewählt werden können. Bei Ersterem wird die Summe der Schwellenparameter innerhalb jedes Items auf 0 festgesetzt. Dagegen wird bei Letzterem ein Schwierigkeitsparameter als Referenz gewählt - i.A. der des Ausgangslevels - und dessen Wert auf 0 fixiert.

5.1 Gemeinsame ML-Schätzung (JML)

Eine Möglichkeit die β-Parameter des PCMs zu schätzen bietet die s.g. unbedingte bzw. gemeinsame - im Englischen „joint“ - Maximum-Likelihood (JML) Methode. Die nachfolgende Erklärung zur JML orientiert sich an Masters (1982, S.163-166) und Masters (1997, S. 109f.) und wird hier nur für das PCM dargestellt.

Bei der JML-Methode wird nun die Likelihood der gesamten Datenmatrix $\boldsymbol{X}$ als Produkt der Einzelwahrscheinlichkeiten π_{pix} über alle $p = 1, \ldots, P$ Personen und alle $i = 1, \ldots, I$ Items aufgestellt. Es wird also insbesondere wieder die Eigenschaft der lokalen stochastischen Unabhängigkeit ausgenutzt. In diesem Fall ergibt sich somit die Likelihood Λ_{JML} wie in nachstehender Formel dargestellt:

$$\Lambda_{JML} = \prod_{p=1}^{P} \prod_{i=1}^{I} \pi_{pix} = \frac{exp[\sum_{p=1}^{P} \sum_{i=1}^{I} \sum_{j=0}^{x_{pi}} (\theta_p - \beta_{ij})]}{\prod_{p=1}^{P} \prod_{i=1}^{I} \{\sum_{k=0}^{m_i} exp[\sum_{j=0}^{k} (\theta_p - \beta_{ij})]\}} \tag{5.3}$$

Entsprechend lässt sich nun die Log-Likelihood λ_{JML} durch logarithmieren des Ausdrucks aus Gl. (5.3) berechnen:

$$\lambda_{JML} = log(\Lambda_{JML}) = \sum_{p=1}^{P} \sum_{i=1}^{I} x_{pi}\theta_p - \sum_{p=1}^{P} \sum_{i=1}^{I} \sum_{j=1}^{x_{pi}} \beta_{ij} - \sum_{p=1}^{P} \sum_{i=1}^{I} log\{\sum_{k=0}^{m_i} exp[\sum_{j=0}^{k} (\theta_p - \beta_{ij})]\}, \tag{5.4}$$

$$\text{wobei } \sum_{j=0}^{x_{pi}} \beta_{ij} = \sum_{j=1}^{x_{pi}} \beta_{ij}, \text{ da } \beta_{i0} \equiv 0$$

Hierbei kann Gl. (5.4) noch weiter vereinfacht werden, v.a. mittels der im PCM vorhandenen suffizienten Statistiken für die beiden Parameter θ_p und β_{ix}. So gilt $r_p = \sum_{i=1}^{I} x_{pi}$, wobei sich also der Test-Score r_p einer Person p aus der Gesamtzahl der vom Probanden bewältigten Stufen ergibt. Außerdem stellt $\sum_{j=1}^{x_{pi}} \beta_{ij}$ die aufsummierten Schwierigkeiten der Schwellen von Item i dar, welche von Teilnehmer p absolviert werden konnten. Dies kann auch für alle P Personen ermittelt werden und folglich erhält man $\sum_{p=1}^{P} \sum_{j=1}^{x_{pi}} \beta_{ij}$. Da zugleich gilt, dass s_{ix} die Zahl der Personen umfasst, welche die k-te Kategorie erreichen konnten und folglich die j Schwellen bis hin zu Kategorie k bei Item i meistern konnten, kann der vorhergehende Ausdruck auch umgeschrieben werden zu $\sum_{p=1}^{P} \sum_{j=1}^{x_{pi}} \beta_{ij} = \sum_{j=1}^{m_i} s_{ij}\beta_{ij}$. Unter Berücksichtigung der beiden vorgestellten Tatsachen lässt sich Gl. (5.4) folgendermaßen umformen:

$$\lambda_{JML} = \sum_{p=1}^{P} r_p\theta_p - \sum_{i=1}^{I} \sum_{j=1}^{m_i} s_{ij}\beta_{ij} - \sum_{p=1}^{P} \sum_{i=1}^{I} log\{\sum_{k=0}^{m_i} exp[\sum_{j=0}^{k} (\theta_p - \beta_{ij})]\} \tag{5.5}$$

Hierzu sei nun Folgendes angemerkt: da in Gl. (5.5) die Beobachtung r_p nur einmal in Kombination mit dem Fähigkeitsparameter θ_p und die Beobachtung s_{ik} nur einmal in Kombination mit dem Schwellenparameter β_{ik} vorkommt, wird der objektiven Vergleich von Personen und Items ermöglicht. Die Log-Likelihood nimmt diese Form nämlich einzig dann an, wenn die Parameter des Modells linear im Argument der Exponentialfunktion sind (vgl. Masters, 1982).

Schließlich lassen sich nun auch die erste und zweite Ableitung von λ_{JML} nach θ_p und jeweiligem β_{ij} ermitteln. Um dies anschaulich darzustellen, wird zunächst jeweils nur die erste Ableitung von

demjenigen Teil des Terms aus Gl. (5.4), welcher sowohl den Fähigkeitsparameter, als auch den Schwellenparameter enthält. Entsprechend wird also zuerst nur die Ableitung vom logarithmierten Teil der Log-Likelihood gebildet, um so stückweise die Ableitung nach den jeweiligen Parametern zu konstruieren. Demgemäß ergibt sich dann beim Ableiten nach θ_p zunächst Nachstehendes:

$$\begin{aligned}\frac{\partial}{\partial\theta_p} log\{\sum_{k=0}^{m_i} exp[\sum_{j=0}^{k}(\theta_p-\beta_{ij})]\} &= \frac{\sum_{k=0}^{m_i} k\cdot exp[\sum_{j=0}^{k}(\theta_p-\beta_{ij})]}{\sum_{k=0}^{m_i}[exp\sum_{j=0}^{k}(\theta_p-\beta_{ij})]} \\ &= \sum_{k=0}^{m_i} k\pi_{pik} = \sum_{k=1}^{m_i} k\pi_{pik}\end{aligned} \tag{5.6}$$

Analog kann auch die erste Ableitung des logarithmierten Ausdrucks von λ_{JML} nach dem jeweiligen Schwellenparameter β_{ij} ermittelt werden:

$$\begin{aligned}\frac{\partial}{\partial\beta_{ij}} log\{\sum_{k=0}^{m_i} \exp[\sum_{h=0}^{k}(\theta_p-\beta_{ih})]\} &= \frac{-\sum_{k=j}^{m_i} exp[\sum_{h=0}^{k}(\theta_p-\beta_{ih})]}{\sum_{k=0}^{m_i} exp[\sum_{h=0}^{k}(\theta_p-\beta_{ih})]} \\ &= -\sum_{k=j}^{m_i}\pi_{pik}\end{aligned} \tag{5.7}$$

Bei letzterer Formel ist zu beachten, dass die Schwierigkeit β_{ij} von Schwelle j nur dann vorkommt, falls $k \geq j$ gilt. Daraus kann dann geschlussfolgert werden, dass sich bei der Ableitung von $\sum_{k=0}^{m_i}\beta_{ij}$ nach β_{ij} die Summe $\sum_{k=0}^{m_i}$ abkürzen lässt zu $\sum_{k=j}^{m_i}$.

Anschließend kann nun auch die vollständige erste Ableitung von λ_{JML} nach den latenten Parametern unter Berücksichtigung von Gl. (5.6) und Gl. (5.7) vereinfacht berechnet werden:

$$\frac{\partial}{\partial\theta_p}\lambda_{JML} = r_p - \sum_{i=1}^{I}\sum_{k=1}^{m_i} k\pi_{pik}, \qquad p = 1,\ldots,P \tag{5.8}$$

$$\frac{\partial}{\partial\beta_{ij}}\lambda_{JML} = -S_{ij} + \sum_{p=1}^{P}\sum_{k=j}^{m_i}\pi_{pik}, \qquad i = 1,\ldots,I;\ j \in k = 1,\ldots,m_i \tag{5.9}$$

Bildet man nun auch die zweite Ableitung von λ_{JML}, leitet man also Gl. (5.8) und Gl. (5.9) nochmals nach dem jeweiligen Parameter ab, so ergeben sich die nachfolgenden beiden Gleichungen. Dieser Schritt ist für den Nachweis des Maximums erforderlich, soll jedoch bei den in den nachfolgenden Kapitel vorgestellten ML-Methoden nicht explizit dargeboten werden. Hier wird nun allerdings ersichtlich, dass durch das Nullsetzen von Gl. (5.8) und Gl. (5.9) mit anschließendem Auflösen nach den gesuchten Parametern jeweils das Maximum bestimmt wird, da die zweite Ableitung von λ_{JML} in beiden Fällen echt kleiner 0 ist:

$$\frac{\partial^2}{\partial^2\theta_p}\lambda_{JML} = -\sum_{i=1}^{I}\Big[\sum_{k=1}^{m_i} k^2\pi_{pik} - \Big(\sum_{k=1}^{m_i} k\pi_{pik}\Big)^2\Big] \tag{5.10}$$

$$\frac{\partial^2}{\partial^2 \beta_{ij}} \lambda_{JML} = -\sum_{p=1}^{P} \left[\sum_{k=j}^{m_i} \pi_{pik} - \left(\sum_{k=j}^{m_i} \pi_{pik} \right)^2 \right] \tag{5.11}$$

Anschließend an den Nachweis der Maxima, lassen sich nun die einzelnen θ- bzw. β-Parameter bestimmen, indem man Gl. (5.8) und Gl. (5.9) jeweils gleich 0 setzt und nach dem Fähigkeitsparameter einer Person p oder einem Schwellenparameter des Items i auflöst. Dies geschieht mittels iterativer Verfahren, wie zum Beispiel mithilfe der Newton-Raphson-Methode. Hierbei werden die Personen- und Itemparameter folgendermaßen bestimmt:

$$\hat{\theta}_p^{\{n+1\}} = \hat{\theta}_p^{\{n\}} - \frac{r_p - \sum_{i=1}^{I} \sum_{k=1}^{m_i} k \pi_{pik}^{\{n\}}}{-\sum_{i=1}^{I} \left[\sum_{k=1}^{m_i} k^2 \pi_{pik}^{\{n\}} - \left(\sum_{k=1}^{m_i} k \pi_{pik}^{\{n\}} \right)^2 \right]} \tag{5.12}$$

$$\hat{\beta}_{ij}^{\{n+1\}} = \hat{\beta}_{ij}^{\{n\}} - \frac{-S_{ij} + \sum_{p=1}^{P} \sum_{k=j}^{m_i} \pi_{pik}^{\{n\}}}{-\sum_{p=1}^{P} \left[\sum_{k=j}^{m_i} \pi_{pik}^{\{n\}} - \left(\sum_{k=j}^{m_i} \pi_{pik}^{\{n\}} \right)^2 \right]} \tag{5.13}$$

In der vorhergehenden Formel bezeichnen $\hat{\theta}_p$ und $\hat{\beta}_{ij}$ die jeweiligen geschätzten Parameter. Zugleich wird durch ein geschweifte Klammern gekennzeichnete n die n-te Iteration des Algorithmus angezeigt. Entsprechend kennzeichnet $n+1$ dann die darauffolgende Iteration. Bei jeder Iteration wird hier die Differenz zwischen dem neuen und dem vorhergehenden Schätzer bestimmt, also $(|\hat{\theta}_p^{\{n+1\}}| - |\hat{\theta}_p^{\{n\}}|)$ bzw. $(|\hat{\beta}_{ij}^{\{n+1\}}| - |\hat{\beta}_{ij}^{\{n\}}|)$. Die Iterationen wiederholen sich schließlich solange, bis ein im Vorhinein festgesetzter, minimaler Wert bei diesen Differenzen erreicht wird und damit die Schätzungen konvergieren. Der in der letzten Iteration errechnete Wert entspricht dann somit dem Maximum der Likelihood.

Die Nutzung der JML-Schätzung birgt jedoch einige Probleme. So müssen zunächst Aufgaben, die von keiner bzw. von allen Personen beantwortet werden konnten, aus der zu analysierenden Datenmatrix entfernt werden. Hier würden sich nämlich für die Itemparameter Schätzwerte von jeweils $-\infty$ bzw. ∞ ergeben. Inhaltlich lässt sich dies folgendermaßen begründen: wird eine Aufgabe von allen Personen gelöst, dann kann nicht abgeschätzt werden, wie leicht dieses Item im Vergleich zu anderen ist. Dieses könnte eine extrem leichte Frage darstellen, aber andererseits auch nur ein klein wenig einfacher als ein anderes Item sein, welches zumindest von einer Person nicht beantwortet werden konnte. Andersherum kann man nicht abschätzen, wie schwierig eine Frage wirklich ist, wenn sie von keinem der Testteilnehmer beantwortet werden konnte. Andersherum müssen auch Personen die keines der Items oder alle Items korrekt beantworten konnten von den Analysen ausgeschlossen werden. Hier gilt analog wie vorher, dass die Fähigkeit dieser Personen im Vergleich zu den anderen Probanden nicht abgeschätzt werden kann.

Prinzipiell ist die praktische Anwendung der JML-Methode insbesondere deswegen wenig verbreitet, da sich ein Problem bezüglich der Konsistenz der Parameterschätzer ergibt. Diese Problematik lässt sich auch mithilfe von Fischer (1995) oder Sijtsma (2006) nachvollziehen. Generell wird der mit der JML einhergehende Nachteil als „Incidental Parameter Problem“ bezeichnet (vgl. hierzu Koller, 2012, S. 36): steigt die Anzahl der Personen, welche den Test durchführen, steigt also der Stichprobenumfang, so führt dies theoretisch zu einer immer genauer werdenden Schätzung der Itemparameter. Genauso gilt, dass bei einer größer werdenden Anzahl der Items, die Schätzung der

Personenparameter genauer wird. Problem hierbei ist jedoch, dass man die Stichprobenzahl verhältnismäßig leicht erhöhen kann, indem man zusätzliche Probanden den Test durchführen lässt. Allerdings weist ein Test im Normalfall eine fixe Zahl an Fragen auf, die nicht nach Belieben erweitert werden können. Hieraus kann abgeleitet werden, dass durch Erhöhung des Stichprobenumfangs jede zusätzliche Person einen weiteren θ-Parameter einbringt, ohne die Möglichkeit die Schätzgenauigkeit des ergänzten Parameters durch mehr Items zu erhöhen. Folglich führt die gleichzeitige Maximierung von Personen- und Aufgabenparametern zur Inkonsistenz der Schwierigkeitsparameter für eine endliche Anzahl an Items. Dabei gilt, dass „Ein Schätzer [...] konsistent [ist], wenn er für immer größere Stichproben immer genauer wird. Mit anderen Worten kann man die Schätzung beliebig genau machen, indem man die Stichprobe weit genug erhöht“ (Pernerstorfer, 2005, S. 2). Zusammenfassend lässt sich also festhalten, dass die Itemparameterschätzer inkonsistent für $P \to \infty$ und gleichzeitig fixierter Anzahl an Items I, obwohl die Konsistenz für $P \to \infty$, $I \to \infty$, $P/I \to \infty$ gegeben wäre (vgl. Fischer, 1995, S. 43). Zudem gilt es zu beachten, dass die mittels der JML-Methode ermittelten Schätzer einen Bias aufweisen. Anhand von Simulationsstudien konnte gezeigt werden, dass diese Verzerrung, welche insbesondere bei einer kleinen Anzahl an Items zutage kommt, signifikant reduziert werden kann durch Multiplikation des errechneten Schätzwertes mit dem Korrekturfaktors $(I-1)/I$ (vgl. Masters, 1997, S. 110). Diese korrigierten Schätzer sind zumeist äquivalent zu denjenigen, welche sich durch die bedingte ML-Methode ermitteln lassen. Diese CML-Methode wird im nun nachfolgenden Kapitel dargestellt.

5.2 Bedingte ML-Schätzung (CML)

In diesem Kapitel wird nun die bedingte - im Englischen „conditional“ - Maximum-Likelihood (CML) Schätzung vorgestellt. Diese Methode bedient sich insbesondere der suffizienten Statistik r_p für den Personenparamter θ_p und beruht im Wesentlichen auf der in Kapitel 4.4.2 beschriebenen spezifischen Objektivität. Hieran wird eventuell schon deutlich, dass sich die CML-Schätzung effektiv auf das PCM anwenden lässt, allerdings keine Möglichkeit der Parameterschätzung im GPCM bietet.

Im Gegensatz zur gemeinsamen ML-Schätzung handelt es sich bei der in dieser Sektion dargebotenen bedingten und der im nachfolgenden Kapitel vorgestellten marginalen ML-Schätzung um zweistufige Verfahren. Während bei der JML die beiden latenten Parameter gleichzeitig geschätzt werden, werden bei der CML- und MML-Methode im ersten Schritt nur die Itemparameter geschätzt, wobei die Fähigkeitsparameter auf verschiedene Weisen zunächst unberücksichtigt bleiben. Diese θ-Parameter werden dann erst im zweiten Schritt ermittelt. Damit werden durch die nachfolgend beschriebenen Methoden konsistente Itemparameterschätzer dargeboten. In Anlehnung an Masters (1997, S. 108f.) soll zunächst das Vorgehen bei der CML-Methode näher erläutert werden.

Bei dieser ML-Methode bedient man sich Gl. (4.15), mittels derer die bedingte Wahrscheinlichkeit für einen Antwortvektor $\boldsymbol{x}_p$ einer Person p ermittelt werden kann. Diese hängt nicht von θ_p ab, sondern es bedarf einzig dem Testscore r_p der betrachteten Person. Von dieser Gleichung kann nun auch die Wahrscheinlichkeit abgeleitet werden, in einer bestimmten Kategorie k des Items i zu

antworten, gegeben dem Gesamtscore r_p:

$$\begin{aligned}\pi_{rik} &= \frac{exp(-\sum_{j=0}^{k}\beta_{ij})\cdot\sum_{\boldsymbol{x}_{p,q\neq i}}^{r-k} exp(-\sum_{q\neq i}^{I}\sum_{j=0}^{x_{pq}}\beta_{qj})}{\sum_{h=0}^{m_i}\left[exp(-\sum_{j=0}^{h}\beta_{ij})\cdot\sum_{\boldsymbol{x}_{p,q\neq i}}^{r-h} exp(-\sum_{q\neq i}^{I}\sum_{j=0}^{x_{pq}}\beta_{qj})\right]} \\ &= \frac{exp(-\sum_{j=0}^{k}\beta_{ij})\cdot\kappa_{r-k,i}}{\kappa_r}\end{aligned} \tag{5.14}$$

Es sei darauf hingewiesen, dass $\sum_{\boldsymbol{x}_{p,q\neq i}}^{r-k}$ die Summe über alle Antwortvektoren $\boldsymbol{x}_{p,q\neq i}$ kennzeichnet, bei denen das i-te Item ausgeschlossen ist und sich der Score $r-k$ ergibt. Schließlich lässt sich nun mithilfe von Gl. (4.15) und Gl. (5.14) die bedingte Likelihood über P Personen und variierenden Testscores konstruieren, was in Gl. (5.15) dargestellt ist. Bezogen auf die zweite Zeile dieser bedingten Likelihood stellt $M = \sum_{i=1}^{I} m_i$ den maximal möglichen Score eines Tests dar. Unter Berücksichtigung dessen kann man dann schlussfolgern, dass sich das Produkt $\prod_{p=1}^{P}\kappa_r$ im Nenner der ersten Zeile von Gl.(5.15) zu $\prod_{r}^{M-1}\kappa_r^{P_r}$ umschreiben lässt. Hierbei bezeichnet P_r die Anzahl der Personen, die den spezifischen Score r erreicht haben.

$$\begin{aligned}\Lambda_{CML} &= \prod_{p=1}^{P}\frac{exp(-\sum_{i=1}^{I}\sum_{j=0}^{x_{pi}}\beta_{ij})}{\kappa_r} \\ &= \frac{exp(-\sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=0}^{x_{pi}}\beta_{ij})}{\prod_{r}^{M-1}\kappa_r^{P_r}}\end{aligned} \tag{5.15}$$

Anschließend kann wiederum durch logarithmieren des Ausdrucks Λ_{CML} die Log-Likelihood λ_{CML} ermittelt werden, wobei sich Nachfolgendes ergibt:

$$\lambda_{CML} = log(\Lambda_{CML}) = -\sum_{i=1}^{I}\sum_{j=1}^{m_i} s_{ij}\beta_{ij} - \sum_{r}^{M-1} P_r log(\kappa_r) \tag{5.16}$$

Hierbei bezeichnet s_{ij} wiederum die Gesamtzahl der Personen, die Schwelle j oder eine höhere bewältigen konnten. Da zudem das Ausgangsniveau - also Kategorie 0 - von allen Personen erreicht wird, beginnt der Indizes j in diesem Fall erst bei 1. Erneut wurde hier ausgenutzt, dass $\sum_{p=1}^{P}\sum_{j=1}^{x_{pi}}$ äquivalent ist zu $\sum_{j=1}^{m_i]s_{ij}\beta_{ij}}$, was bereits in der vorhergehenden Sektion 5.1 begründet wurde.

Um nun die gesuchten β-Parameter bestimmen zu können, gilt es die erste Ableitung von λ_{CML} bezüglich der einzelnen β_{ik}'s zu ermitteln und anschließend zu prüfen, ob die zweite Ableitung kleiner 0 ist, um so garantieren zu können, dass es sich bei den mittels der ersten Ableitung geschätzten Itemparametern tatsächlich um ein Maximum handelt. Beide Ableitungen sind nachfolgend angegeben, wobei nicht im Detail auf diese eingegangen werden soll. Angemerkt sei, dass hier $\sum_{j=k}^{m_i}\pi_{rij}$ die bedingte Wahrscheinlichkeit einer Person p mit einem Testscore von r_p bezeichnet, Level k oder ein höheres Leistungsniveau bei einem Item i zu erreichen.

$$\frac{\partial}{\partial\beta_{ik}}\lambda_{CML} = -S_{ik} - \sum_{r}^{M-1}\frac{P_r}{\kappa_r}\left(\frac{\partial}{\partial\beta_{ik}}\kappa_r\right) = -S_{ik} + \sum_{r}^{M-1} P_r\sum_{j=k}^{m_i}\pi_{rij}, \tag{5.17}$$

$$\frac{\partial^2}{\partial^2 \beta_{ij}} \lambda_{CML} = -\sum_{r}^{M-1} P_r \Big(\sum_{j=k}^{m_i} \pi_{rij} \Big) \Big(1 - \sum_{j=k}^{m_i} \pi_{rij} \Big) \tag{5.18}$$

Um schließlich die Schwellenparameter schätzen zu können muss Gl. (5.17) maximiert werden. Wie auch schon bei der JML-Schätzung werden auch hier die β-Parameter mittels eines interativen Verfahrens ermittelt. Dafür eignet sich wiederum die numerische Optimierung mittels des Newton-Raphson-Verfahrens. Hier sei noch darauf hingewiesen, dass diese Schätzmethode der β-Parameter im Gegensatz zur JML konsistente Schätzer liefert, was allgemein von Andersen (1970, S. 285-288) für die CML-Methode gezeigt wird. Zu beachten ist, dass auch bei dieser ML-Methode Parameter von Aufgaben, die von keinem oder allen Personen gelöst werden konnten, nicht ermittelbar sind. Hier gilt dieselbe Begründung, die bereits in Sektion 5.1 zur JML dargelegt wurde.

In einem zweiten Schritt können anschließend die Personenparameter bestimmt werden. Hierzu verwendet man i.A. die unbedingte ML-Methode und nutzt dabei die im ersten Schritt geschätzten Itemparameter. Allerdings muss man sich vor Augen führen, dass die Verwendung der geschätzten Itemparameter zu Ermittlung der Personenparameter eine zusätzlich Unsicherheit bei der Schätzung dieser θ-Parameter bewirkt. Alternativ bietet sich auch an die Personenparameter mittels der s.g. gewichteten - im Englischen „weighted“ - Maximum-Likelihood (WML) Methode zu schätzen. Die WML-Methode soll in dieser Arbeit nicht genauer ausgeführt werden. Es sei lediglich gesagt, dass hierbei die Likelihood mit der Quadratwurzel der Diagonalelemente von der beobachteten Fisher-Informationsmatrix gewichtet wird (vgl. Welchowski, 2014, S. 7). Dies baut auf der von Warm (1985) vorgeschlagenen WML-Methodik zur Reduzierung der Verzerrung, welche mit der ML-Schätzung einhergeht, auf.

Allgemein gilt gemäß Andersen (1970) für die CML-Methode, dass - bei Gültigkeit der hier erfüllten Annahmen (siehe dazu Andersen, 1970, S. 286) - die Itemparameter-Schätzungen für $P \to \infty$ und bei gleichzeitig fester Anzahl von I Items konsistent sind und zudem auch unverzerrt, asymptotisch effizient, sowie asymptotisch normalverteilt. Eine weitere Methode konsistente Schätzer zu erhalten, bietet die im nachfolgenden Abschnitt vorgestellte marginale ML-Schätzung.

5.3 Marginale ML-Schätzung (MML)

Eine andere Möglichkeit die Parameter von IRT-Modellen zu schätzen, stellt die marginale Maximum-Likelihood (MML) Methode dar. Diese ist besonders nützlich in Hinsicht auf Modelle, welche keine „einfache“ suffiziente Statistik für den Fähigkeitsparameter θ_p aufweisen und somit nicht mittels der CML-Methode geschätzt werden können. Hierzu sei insbesondere das GPCM zu nennen. Zudem ist wiederum auf den bereits angesprochenen Vorteil gegenüber der JML-Schätzung hinzuweisen: die MML-Methode erlaubt konsistente Schätzungen der Itemparameter, selbst dann wenn ein Test eine feste Zahl an Items hat und nur die Zahl der befragten Personen steigt (vgl. Sijtsma, 2006, S. 85). Einen deutlichen Vorteil stellt auch die Tatsache dar, dass mittels der MML auch Personen, die keine bzw. alle Aufgaben korrekt lösen konnten, bzw. Aufgaben, die von keinem oder allen Probanden richtig beantwortete werden konnten, Berücksichtigung finden. Das genaue Vorgehen bei dieser Art der Parameter-Schätzung soll mittels Masters (1997, S. 110ff.) und Johnson (2007, S. 7f.) verdeutlicht werden.

Bei der MML-Schätzung handelt es sich ebenso wie bei der CML-Schätzung um ein zweistufiges Verfahren. Im ersten Schritt werden also wiederum nur die Itemparameter geschätzt. Dabei betrachtet man die Personenparameter als „nebensächlich“ bzw. als „Stör-Parameter“ und lässt diese in der Likelihood-Funktion unberücksichtigt, indem man davon ausgeht, dass es sich bei den betrachteten Personen um eine zufällige Stichprobe aus einer Population handelt, in welcher die Fähigkeiten gemäß einer bestimmten Dichtefunktion $f(\theta)$ verteilt sind. Im Allgemeinen erachtet man latenten Variablen als stetig verteilte Zufallsvariablen, was auch bei den hier gesuchten, latenten Fähigkeits- und Schwellenparametern durchaus Sinn macht. Entsprechend kann man annehmen, dass die θ-Parameter beispielsweise normalverteilt sind und somit gilt $\theta \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, wobei μ den Erwartungswert und σ^2 die Varianz repräsentieren. Auch wenn bei dieser Art der Itemparameterschätzung die Personenparameter unberücksichtigt bleiben, ist eine korrekte Verteilungsannahme der Fähigkeiten unerlässlich. Anderenfalls können sich starke Verzerrungen der Schätzung ergeben. Insgesamt lässt sich also festhalten, dass die Itemparameter feste Effekt darstellen, während die Personenparameter als zufällig angenommen werden.

Zunächst überlege man nun konkret, dass sich mithilfe der Wahrscheinlichkeit π_{pix} die Wahrscheinlichkeit für den beobachteten Antwortvektor $\boldsymbol{x}_p$ der Person p gegeben der Fähigkeit dieser Person θ_p und einem Set von Itemparametern $\boldsymbol{\beta}$ folgendermaßen ermitteln lässt, was ausgeschrieben schließlich Gl. (4.12) entspricht:

$$P(\boldsymbol{x}_p|\theta_p, \boldsymbol{\beta}) = \prod_{i=1}^{I} \pi_{pix} \tag{5.19}$$

Nimmt man nun an, dass eine Person zufällig aus einer Population mit stetig verteilter Fähigkeitsdichte $f(\theta)$ gezogen wurde, so ergibt sich demgemäß die marginale Wahrscheinlichkeit für den Antwortvektor $\boldsymbol{x}_p$ einer Person p als:

$$P(\boldsymbol{x}_p|\boldsymbol{\phi}, \boldsymbol{\beta}) = \int_{\theta} P(\boldsymbol{x}_p|\theta_p, \boldsymbol{\beta}) f(\theta, \boldsymbol{\phi}) d\theta \tag{5.20}$$

Hierbei bezeichnet $\boldsymbol{\phi}$ ein Set von Populationsparametern, beispielsweise mit $\boldsymbol{\phi} = (\mu, \sigma^2)$,wobei μ den Fähigkeits-Mittelwert eben dieser Population und σ^2 die Varianz der Fähigkeitswerte widerspiegelt. Zieht man nun zufällig P Personen aus der betrachteten Population, so kann die marginale Likelihood Λ_{MML} wie in Gl. (5.21) dargestellt werden. Hierbei sind nun die einzelnen „Stör-“ Parameter θ_p gänzlich ausgeschlossen.

$$\Lambda_{MML} = \prod_{p=1}^{P} P(\boldsymbol{x}_p|\boldsymbol{\phi}, \boldsymbol{\beta}) \tag{5.21}$$

Schließlich erhält man die Item- oder auch die Populationsparameterschätzer durch Ableiten der Log-Likelihood $\lambda_{MML} = log(\Lambda_{MML})$ nach dem jeweiligen Parameter und darauffolgender Bestimmung des Maximums. Entsprechend gilt es also folgende Likelihood-Gleichungen zu lösen:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \lambda_{MML} \quad \text{und} \quad \frac{\partial}{\partial \boldsymbol{\phi}} \lambda_{MML} \tag{5.22}$$

In der Praxis zeigt sich jedoch ein deutlicher Nachteil dieser Methode: bei der Bestimmung der

einzelnen Likelihoods müssen bei P betrachteten Personen jeweils P Integrale bestimmt werden. Auch wenn dies mittels geeigneter Software-Pakete bewerkstelligt werden kann, ist die Prozedur rechenintensiv und damit zeitaufwendig. Eine Alternative bietet der s.g. EM-Algorithmus, welcher durch Anwendung auf das GPCM von Muraki (1992) eingehend erläutert wird.

Für das PCM lässt sich der EM-Algorithmus - was die Kurzform des eigentlichen Begriffs Expectation-Maximization-Algorithmus darstellt - in Anlehnung an Masters (1997, S. 111f.) zusammenfassen. Anstatt also die Gl. (5.22) nach dem jeweils gesuchten Parameter aufzulösen, bedient man sich des von Bock (1981) dargelegten EM-Algorithmus. Ginge man davon aus, dass neben den jeweiligen Antworten x_{pi} auch die Personenparameter $\boldsymbol{\theta}$ bekannt wären, so ließe sich die gemeinsame Likelihood schreiben als $\Lambda_{MML}(\boldsymbol{\beta}, \boldsymbol{\phi}|\boldsymbol{X}, \boldsymbol{\theta})$ und damit die Log-Likelihood als $\lambda_{MML} = log\Lambda_{MML}(\boldsymbol{\beta}, \boldsymbol{\phi}|\boldsymbol{X}, \boldsymbol{\theta})$. Tatsächlich kann dieser Ausdruck aber nicht direkt maximiert werden, da dies von dem Wissen über die Fähigkeitsparameter $\boldsymbol{\theta}$ abhängt. Alternativ kann allerdings der Erwartungswert $E(\lambda_{MML})$ gegeben aktuellen Schätzungen der einzelnen Personenparameter (θ_p) maximiert werden. Hierzu wird zunächst die marginale posteriori-Dichte von θ_p gegeben dem Test-Score $\boldsymbol{x}_p$ der Person p bestimmt, wobei „provisorische“ Schätzer für die Item- und Populationsparameter verwendet werden. Dies bezeichnet man als den s.g. Expectation- oder kurz E-Schritt. Anschließend maximiert man im Maximization- oder kurz M-Schritt die zuvor aufgestellte gemeinsame Log-Likelihood, um so besser angepasste Schätzer für die Item- und Populationsparameter zu erlangen:

$$\frac{\partial}{\partial \boldsymbol{\beta}} E(\lambda_{MML}) = 0 \quad \text{und} \quad \frac{\partial}{\partial \boldsymbol{\phi}} E(\lambda_{MML}) = 0 \tag{5.23}$$

Diese zwei Schritte werden solange wiederholt, bis Konvergenz erreicht ist und sich damit der Schätzer gemäß eines Konvergenzkriteriums nahe genug an den tatsächlichen Wert des Parameters angenähert ist.

Im engsten Sinne lässt sich auch für das GPCM die MML-Methode in der beschrieben Form zum PCM bewerkstelligen. Dabei gilt es jedoch auch den jeweiligen Steigungsparameter der einzelnen Items zu berücksichtigen und demgemäß auch bei der Aufstellung der Likelihood-Funktion die jeweiligen α-Parameter zu beachten, sowie zusätzlich die partiellen Ableitungen von der Log-Likelihood nach dem itemspezifischen Diskriminationsparameter, wobei dem Vektor $\boldsymbol{\alpha}$ alle I Steigungsparameter angehören:

$$\frac{\partial}{\partial \boldsymbol{\alpha}} \lambda_{MML} \tag{5.24}$$

Schließlich gilt es auch hier wieder bei Verwendung des EM-Algorithmus die nachfolgende partiellen Ableitungen zu bestimmen:

$$\frac{\partial}{\partial \boldsymbol{\alpha}} E(\lambda_{MML}) \tag{5.25}$$

Analoges Vorgehen wie beim PCM liefert so schließlich ein Set von Itemparameter-Schätzern - wobei hier die Steigungsparameter inbegriffen sind - sowie die posteriori-Verteilung der Fähigkeitsparameter. Zur eingehenden Erläuterung der MML-Methode angewandt auf das GPCM sei auf Muraki (1992, S. 9-14) oder Muraki (1997, S. 156-159) verwiesen.

Mittels der MML-Methode ergeben sich also insbesondere konsistente Itemparameterschätzer. Allerdings hängt die Unverzerrtheit der MML-Schätzer davon ab, ob die Verteilung der Personen-

fähigkeit $f(\theta)$ korrekt spezifiziert wurde. Im Falle einer falschen Annahme bezüglich der Dichtefunktion ergeben sich konsequenter Weise verzerrte Schätzungen (siehe hierzu Mair, 2009, S. 9). Insofern die angesprochenen Verteilung korrekt spezifiziert ist, sind die über die CML- und die MML-Methode ermittelten Schätzungen der Itemparameter asymptotisch äquivalent. Hierauf wird in Kapitel 7 bei der konkreten Anwendung der Schätzmethoden noch einmal eingegangen.

Um nun im zweiten Schritt auch die Personenparameter ermitteln zu können bieten sich verschiedene Verfahren an, auf welche in Anlehnung an De Ayala (2009, 75-78) kurz eingegangen wird: im Rahmen der ML-Schätzung kann u.a. auch auf die bereits in der vorhergehenden Sektion vorgestellte WML-Methode zurückgegriffen werden. Generell ist die Schätzung über ML-Methoden kritisch zu betrachten, da für Personen mit einem Score von 0 oder perfektem Score keine endlichen Fähigkeitsparameter berechnet werden können.

Allerdings gibt es auch einige bayessche Verfahren, um die latenten Parameter zu ermitteln. Die zugrundeliegende Strategien bauen hauptsächlich auf der Information bezüglich der im Vorhinein bzw. vor der Beobachtung der Stichprobe gegebenen (a priori) Wahrscheinlichkeitsverteilung des θ-Parameter auf. Diese wurde insbesondere auch bei der vorhergehenden Schätzung der Itemparameter über die MML-Methode ausgenutzt. Das Ergebnis von der Integration der priori-Verteilung mit den beobachteten Daten ist die posteriori- Verteilung. Damit lässt sich also die posteriori-Verteilung durch Einarbeitung der priori-Verteilung in die beobachteten Daten - z.B. Likelihoodfunktion - beschreiben. Nimmt man beispielsweise die Normalverteilung als priori-Verteilung an, und integriert diese Information in die Likelihood, so kann die Unsicherheit bezüglich der Fähigkeitsparameter in der sich ergebenden posteriori-Verteilung reduziert werden. Um nun die Lokation des Fähigkeitsparameters $\hat{\theta}_p$ einer Person p abzuschätzen, kann der Modus oder der Median der posteriori-Verteilung herangezogen werden. Im Falle einer symmetrischen Verteilung weisen der diese denselben Wert auf.

In diesem Zusammenhang sind hier nun die bayesschen Verfahren „Maximum A Posteriori„ (MAP) und „Expected A Posteriori“ (EAP) aufzuführen, die generell zur Ermittlung der Personenparameter im PCM oder GPCM genutzt werden können. Bei Ersterem wird jeweils der Modus der posteriori-Verteilung für die Schätzung der Fähigkeitsparameter $\hat{\theta}_p$ genutzt, während bei Letzterem der Median verwendet wird. Insbesondere beim Vergleich der ML-Methode (mit Newton-Verfahren) und der EAP konnte in verschiedenen Kontexten des Testens gezeigt werden, dass EAP gegenüber der ML-Schätzung zu favorisieren ist (vgl. Chen, 1998, S. 571). Zudem bieten EAP und MAP den klaren Vorteil auch die Fähigkeitsparameter bei allen Antwortmustern, explizit als auch bei einem Score von 0 oder perfektem Ergebnis, abschätzen zu können, d.h. für alle Personen kann über die bayesschen Verfahre ein finiter Parameter ermittelt werden.

Gleichzeitig gibt es einige Gesichtspunkte, in den sich MAP und EAP unterscheiden. Im Wesentlichen sollte hier aufgeführt werden, dass MAP eine iterative Methode der Parameterschätzung darstellt, während EAP ein nicht-iteratives Verfahren ist, welches auf numerischer Quadratur/Integration beruht. Zudem erweist sich die mittlere quadratische Abweichung der Schätzungen resultieren aus dem EAP-Verfahren als geringer gegenüber der MAP-Technik. Weitere Unterschiede, sowie die mathematische Methodik der Schätzungen können mittels De Ayala (2009) nachvollzogen werden. Zudem gilt es zu beachten, dass sich durch die Verwendung der geschätzten Itemparameter zur Ermittlung der Personenparameter zusätzliche Unsicherheiten bei letzteren auftreten.

6 Schätzung von PCM/GPCM mittels ausgewählter R-Pakete

An die Erklärung möglicher Schätzmethoden der Parameter, welche im PCM bzw. GPCM vorhanden sind, anschließend, sollen in diesem Kapitel nun R-Pakete vorgestellt werden, die für die Analyse von ordinalen Antwortformaten und damit explizit für das Partial Credit Format geeignet sind. Ganz allgemein handelt es sich bei R um eine Statistik-Software, mittels derer insbesondere statistische Analysen, sowie auch die Erzeugung hochwertiger Grafiken möglich ist. Die Programmiersprache R ist eng verwandt mit der 1980 konstruierten Programmiersprache S und die Syntax weist Ähnlichkeiten mit C auf (vgl. R Core Team, 2016, S. 1).

In R selber werden Einheiten reproduzierbaren Codes in Form von Paketen, im Englischen bezeichnet als „Packages“, zur Verfügung gestellt. Hier werden neben Funktionen auch Dokumentationen zur Handhabung der jeweiligen Funktionen, sowie Beispieldaten bereitgestellt (vgl. Wickham, 2015). Auch für die Anwendung des PCMs bzw. GPCMs gibt es in R zur Verfügung stehende Pakete. Im Rahmen dieser Arbeit sollen nun drei derer vorgestellt werden und gesondert auf die implementierte Funktion zur Konstruktion des PCMs bzw. GPCMs, sowie auf einige gebräuliche, weiterführende Funktionen aus dem jeweiligen Paket eingegangen werden. Im anschließenden Kapitel 7 werden diese auch noch konkret zur Anwendung gebracht.

Hierzu wird nun auf das eRm-, das ltm- und das TAM -Paket näher eingegangen. Angemerkt sei vorab, dass das eRm - Kurzform für „Extended Rasch Modeling“ - nur Analysen mittels des PCMs, nicht aber des GPCMs ermöglicht. Die Namen der anderen beiden Pakete ergeben sich aus „Latent Trait Models (under IRT)“ und „Test Analysis Modules“. Mit diesen Paketen kann sowohl das PCM, als auch das GPCM zur Anwendung gebracht werden. Nun folgend soll zunächst auf das eRm genauer eingegangen werden.

Zusätzlich sollte vorab noch darauf hingewiesen werden, dass in Hinsicht auf die Modellierung des PCMs der α-Parameter im eRm-Paket über alle Items hinweg unabänderlich auf 1 festgesetzt ist. Dagegen ist bei den anderen beiden R-Paketen eine weniger restriktive Form des PCMs zugelassen, demgemäß auch ein Steigungsparameter geschätzt wird, der zwar über alle Items hinweg gleich ist, allerdings nicht unbedingt einen Wert von 1 annehmen muss. Wie diese weniger restriktive Form des PCMs zustande kommt, wird explizit in Sektion 6.2 aufgezeigt-

6.1 Paket 'eRm'

Das R-Paket eRm bietet - wie bereits erwähnt - nur die Möglichkeit das PCM auf einen Datensatz mit ordinalen Antwortformat anzuwenden, nicht aber das GPCM. Hierbei werden die Itemparameter mittels der CML-Methode ermittelt, welche in Kapitel 5.2 bereits dargestellt wurde. Die Fähigkeits- bzw. Personenparameter werden auf Basis der geschätzten Aufgabenparameter über die unbedingte ML ermittelt. Die nun folgende ausführlichere Darstellung des eRm-Paketes geschieht in Anlehnung an Mair (2007b) und Kiefer (2016b). Anzumerken sei, dass in der zugrundeliegenden Datenmatrix fehlende Werte zugelassen sind, welche mit 'NA' kodiert werden. Zudem können mit-

tels des eRm-Paketes auch Fähigkeitswerte von Personen ermittelt werden, die einen Testscore von 0 hatten oder einen perfekten Score erreichten.

Allgemein wird in diesem Paket die bisher bekannte Modellformel des PCMs, wie sie Gl. (4.6) dargestellt ist, quasi in abgewandelter Form genutzt. Generell basiert die Parameterisierung der im eRm-Paket verfügbaren Modelle auf dem s.g. „linear-logistischen Test-Modell“ (LLTM). Die Besonderheit dieses Modells ist die Annahme/lineare Bedingung, dass sich die Itemparameter aus einer Linearkombination von s.g. Basisparametern und einer gegebenen Gewichtung pro Item zusammensetzen. Nimmt man diese Veränderung angewandt auf das PCM vor, so ergibt sich das „Lineare Partial Credit Modell“ (LPCM) von Fischer (1994). Hierbei dient dieses Modell hauptsächlich der Erfassung von Veränderungen - beispielsweise von Behandlungseffekten - durch die Verwendung des Konzeptes s.g. virtueller Items. Durch geeignete Festlegung des Designs ist das LPCM wiederum in das PCM überführbar. Um dies zu veranschaulichen sei nachfolgend die Modellgleichung des LPCMs mit gewohnter Benennung gegeben, wobei durch die lineare Restriktion jedem τ_{ik} alle möglichen $i \times m_i$ Kombination zugeordnet werden. Die nun folgende Erläuterung dazu basiert auf Mair (2007a, S. 37) und Fischer (1994).

$$P(X_{pi} = x|\theta_p, i) = \frac{exp(x\theta_p + \tau_{ix})}{\sum_{k=0}^{m_i}(k\theta_p + \tau_{ik})}, \tag{6.1}$$

$$\text{mit} \quad i = 1, \ldots, I, \quad k = 1, \ldots, m_i \quad \text{und} \quad \tau_{ix} = \sum_{l=1}^{L} w_{ixl}\eta_l$$

Hierbei bezeichnet man η_l zumeist als „Basisparameter“ (vgl. Mair, 2007a, S.28) und es gilt $L = \sum_{i=1}^{I} m_i$. Derweilen stellen die w_{ixl}'s festgelegte Gewichte der Basisparameter dar. Insgesamt findet also eine linearen Zerlegung der einzelnen τ_{ik}-Parameter in gewichtete Teilkompetenzen statt. Gleichzeitig gilt hier nun - im Vergleich zur bisher bekannten Modellformel des PCMs -, dass die einzelnen τ-Parameter nun zur jeweiligen Personenfähigkeit addiert werden und demzufolge die Leichtigkeit anstelle der Schwierigkeit betrachtet wird. Um das Modell identifizierbar zu machen, werden die Parameter wiederum folgendermaßen normiert: $\tau_{i0} = 0$ für $i = 1, \ldots, I$ und $\sum_{i=1}^{I}\sum_{k=0}^{m_i}\tau_{ik} = 0$.

Für die konkrete Schätzung der $\boldsymbol{\tau}$-Parameter - mit $\boldsymbol{\tau} = (\tau_{11}, \ldots, \tau_{1m_1}, \ldots, \tau_{I1}, \ldots, \tau_{Im_I})'$ - wird nun also die Linearkombination $\hat{\boldsymbol{\tau}} = \boldsymbol{W}\hat{\boldsymbol{\eta}}$ verwendet, wobei $\boldsymbol{W}$ der Designmatrix entspricht und $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_L)'$ einem Vektor mit allen Basisparametern. Insbesondere kann also abgeleitet werden, dass für das PCM die Basisparameter den einzelnen Schwellenparametern entsprechen. Insgesamt sollte also auch deutlich werden, dass $\boldsymbol{W}$ einer $L \times (L-1)$-Matrix entspricht, also zeilenweise auf die einzelnen Kategorien zu den I Items bezieht und spaltenweise auf die Basisparameter.

$$\boldsymbol{W} = \begin{pmatrix} -1 & -1 & \ldots & -1 \\ 1 & & & \\ & 1 & \multirow{2}{*}{0} & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix}$$

Damit ist die Designmatrix $\boldsymbol{W}$ des PCMs - bei dem keine konkrete Gewichtung der einzelnen Teilkomponenten stattfindet - unter Berücksichtigung der aufgezeigten Summennormierung - wie vorhergehend abgebildet darstellbar.

Zur eingehenderen Vertiefung des LPCMs, sowie die dazu entsprechend angepassten CML- und MML-Methoden können anschaulich anhand von Fischer (1994) nachvollzogen werden. Im eRm-Paket wird zur konkreten Schätzung die CML-Methode verwendet, da diese gemäß Mair (2007b, S. 6) abgesehen von den wünschenswerten Eigenschaften einer Schätzung, auch nahe an dem Konzept der spezifischen Objektivität orientiert ist und sich auf Basis der Schätzung direkt der Likelihood-Quotienten-Test durchführen lässt, welcher im weiteren Verlauf dieses Kapitels noch eingehend beschrieben wird.

Der Befehl zur konkreten Anwendung des PCMs lautet in diesem Paket nun folgendermaßen, wobei basierend auf der CML-Methode - in der Art wie in 5.2 präsentiert ist - der Newton-Raphson-Algorithmus angewandt wird, um die Schwellenparameter zu bestimmen:

```
PCM(X, W, se = TRUE, sum0 = TRUE, etaStart)
```

Hierbei bezeichnet 'PCM' den eigentlichen Funktionsaufruf, während in der Klammer dargestellte Variablen Parameter sind, die der Funktion übergeben werden müssen bzw. können. Wie anhand des vorhergehenden Satzes bereits ersichtlich wird, sind einige der Parameter optional. Bei den nachfolgenden Darlegungen zu gebräuchlichen Funktionen der drei R-Pakete wird zumeist nicht auf alle Parameter eingegangen, die jeweils übergeben werden können, sondern es erfolgt eine Beschränkung auf die wesentlichsten. Weitere mögliche Einstellungen lassen sich aber anhand der Handbücher zu den entsprechenden R-Paketen eruieren.

Bei dem vorliegenden Befehl ist nur die Übergabe der Datenmatrix 'X' erforderlich, die restlichen Parameter können zur Konkretisierung zusätzlich angegeben werden. Diese Datenmatrix beinhaltet als Einträge die jeweiligen Scores - beginnend bei Kategorie 0 - aller Person, welche zeilenweise abgetragen werden, und der entsprechenden Items, welche spaltenweise angegeben sind. Folglich entspricht diese Matrix einer Matrix in der Form von $\boldsymbol{X}$, wie sie in Kapitel 4 vorgestellt wurde. Des Weiteren bezeichnet 'W' die Designmatrix $\boldsymbol{W}$ des PCMs wie sie in einem der vorhergehenden Absätze dieser Sektion dargestellt wurde und wird von R automatisch erzeugt, insofern sie nicht als Parameter übergeben wird. Insgesamt entspricht diese also einer optionalen Designmatrix zur Normierung auf ein bestimmtes Item, was bei konkreter Umsetzung in der Analyse mittels des LPCMs übergehen würde. Zudem wird durch 'se' vom Benutzer angegeben, ob die Standardfehler ausgegeben werden sollen oder nicht. Der Default-Wert ist hierbei 'TRUE', was beutet, dass die Standardfehler bei der Ausführung des PCM-Befehls aufgeführt werden, falls der logische Wert nicht vorab als 'FALSE' festgesetzt wird. Mittels 'sum0' wird die Normierung bestimmt, wobei durch den Standardwert 'TRUE' - mit angepasster Designmatrix 'W' - die Itemparameter so normiert werden, dass sie in der Summe 0 ergeben. Anderenfalls wird der erste Itemparameter auf 0 gesetzt und der Schätzwert aller anderen in Abhängigkeit von diesem bestimmt. Durch 'etaStart' kann ein Vektor von Itemparametern, welche als Startwerte benutzt werden, übergeben werden. Insofern 'etaStart' nicht angegeben wird, wird der Nullvektor verwendet.

Nach der Anwendung der Funktion 'PCM' auf einen konkreten Datensatz erhält man ein Objekt

der Klasse 'eRm', wozu u.a. folgende Werte ausgegeben werden: der Wert der Log-Likelihood 'loglik', die Anzahl der Iteration 'iter', bis die CML konvergiert, sowie die Gesamtzahl der Parameter 'npar', die geschätzt wurden. Außerdem werden natürlich die mittels des Newton-Raphson-Verfahrens - angewandt auf die CML-Methode - geschätzten τ-Parameter und - wenn nicht anders festgesetzt - die Standardfehler dieser ausgegeben. Das Konvergenzkriterium wird dabei durch die s.g. nichtlineare Minimierung/Optimierung bestimmt, was hier nun aber nicht weiter ausgeführt wird. Generell biete es sich an den 'summary()'-Befehl zu nutzen und so eine übersichtliche Darstellung der Schätzwerte $\hat{\tau}_{ik}$, sowie der zugehörigen Standardfehler und das daraus ermittelbare 0.95%-Konfidenzintervall zu erhalten. Es sei erwähnt, dass hierbei sowohl die Schwierigkeits-, als auch Leichtigkeitsparameter angegeben werden, welche simple Änderung des Vorzeichens, also Multiplikation mit -1 ineinander überführbar sind.

Anstatt sich nun die einzelnen Schwellenparameter $\hat{\beta}_{ik}$ aus den jeweiligen Kategorien-Schwierigkeitsparametern $\hat{\tau}_{ik}$ zu berechnen, kann man sich der folgenden, im eRm-Paket implementierten Funktion bedienen:

```
thresholds(object)
```

Hierbei muss 'object' einem Objekt der Klasse 'eRm' entsprechen. Nach Aufruf dieser Funktion wird eine Matrix ausgegeben, deren Einträge den jeweiligen Schwellenparametern der betrachteten Aufgaben entsprechen. Zusätzlich wird eine Spalte 'location' ausgegeben, welche zeilenweise das arithmetische Mittel der Schwellenparameter einer Aufgabe enthält und damit dem in Kapitel4.2 vorgestellten Lokalisationsparameter entspricht.

Die so ermittelten Schwellenparameter stellen auch diejenigen Parameter dar, welche bei der grafischen Veranschaulichung des PCMs in Form der Schwellenwahrscheinlichkeit - wie sie in Kapitel 4.3.1 dargestellt wurden - genutzt werden. Hier kann nun direkt übergeleitet werden zur grafischen Darstellung mittels des eRm-Paketes. Durch Aufruf der Funktion 'plotICC()' kann man sich die Kategorienwahrscheinlichkeiten der einzelnen Items als Grafik(en) ausgeben lassen. Die Funktion sieht i.A. folgendermaßen aus:

```
plotICC(object, item.subset,...)
```

Bei dem Parameter 'object' handelt es sich um ein Objekt der Klasse 'eRm', welcher notwendigerweise übergeben werden muss. Mithilfe des Arguments 'item.subset' kann man diejenigen Items bestimmen, welche in der erzeugten Grafik angezeigt werden sollen. Wenn nicht anders angegeben werden die Grafiken zu allen I Items ausgegeben. Die übrigen Parameter, welche der 'plotICC'-Funktion theoretisch übergeben werden können entsprechen weitestgehend den gebräulichen Grafik-Argumenten.

Auf Basis der bereits geschätzten Itemparamter können im zweiten Schritt schließlich die Personenparameter über die unbedingte Log-Likelihood ermittelt werden und so die einzelnen $\hat{\theta}_p$'s ermittelt werden. Da durch Simulationsstudien gezeigt werden konnte, dass bayessche Methoden Schätzer mit höherer Genauigkeit hervorbringen, ist es geplant derartige Methoden in zukünftigen Versionen des Paketes zu implementieren (vgl. Mair, 2007a, S. 37). Gesondert zu erwähnen ist, dass es mittels des eRm-Paketes möglich ist auch die Fähigkeitswerte für Personen zu ermitteln, die bei

einer (oder mehreren) Aufgaben einen nicht-beobachteten Score (NA-Wert) eingetragen haben und auch für Probanden, die eine Testscore von 0 oder einen perfekten Score aufweisen. Dies geschieht über s.g. Spline-Interpolation. Dabei bezeichnet ein Spline im mathematischen Sinne eine Funktion, welche stückweise - also jeweils zwischen zwei Knotenpunkten - aus Polynomen von höchstens n-ten Grades zusammengesetzt ist. An diesen Knotenpunkten werden bestimmte Differenzierbarkeitseigenschaften gefordert, sodass ein glatter Verlauf der Kurve gewährleistet ist. Damit können folglich die fehlenden Fähigkeitswerte bzw. diejenigen von Personen mit einem Testscore von 0 oder voller Punktzahl angenähert werden.

Die Schätzung der Personenparameter erfolgt im eRm-Paket letztlich über nachfolgenden Befehl, wobei es sich bei dem übergebenen Parameter 'object' wiederum um ein Objekt der Klasse 'eRm' handelt:

```
person.parameter(object)
```

Die Funktion 'person.parameter()' erzeugt ein Objekt der Klasse 'ppar'. Angemerkt sei nun, dass bei der Aufstellung der Likelihood Personen mit dem gleichen Testscore entfernt werden, damit die Schätzung der Fähigkeitsparameter schneller bewerkstelligt werden kann. Demzufolge ist die ausgegebene Log-Likelihood 'loglik' auf die verringerten Daten bezogen. Zudem bezieht sich auch die Anzahl der geschätzten Parameter 'npar' auf die reduzierten Daten. Weitere Rückgabe-Werte sind die Anzahl an Iterationen bis zur Konvergenz 'niter', die über die Likelihood ermittelten Fähigkeitsparameter 'thetapar', sowie die zugehörigen Standardfehler 'se.theta'. Zusätzlich können auch die entsprechende Hesse-Matrix 'hessian', 'theta.table', 'pers.ex' und 'X.ex' abgerufen werden. Dabei beinhaltet 'theta.table' alle Fähigkeitsparameter der Personen aus dem Orginaldatensatz, sowie die zugehörigen Standardfehler und die Grenzen des 2.5%- und 97.5%-Konfidenzintervalles. Zudem können auch diejenigen Personen bzw. der Indizes dieser Probanden 'pers.ex' angezeigt werden, die aufgrund eines Testscores von 0 oder vollem Score ausgeschlossen wurden. Die entsprechende Matrix zu diesen ausgeschlossenen Teilnehmern ist in 'X.ex' inbegriffen.

Bisher noch recht wenig zur Sprache kamen mögliche Tests zur Überprüfung der Modellgültigkeit bzw. der Modellannahmen. Im Rahmen dessen gilt es vorwiegend die Annahme der spezifischen Objektivität zu Prüfen und somit genauer gesagt zu testen, ob „Differential Item Functioning“ (DIF) - siehe dazu 4.4.2 - vorliegt und damit die Items für verschiedene Gruppen/Subpopulationen unterschiedlich schwierig zu lösen sind. Allgemeiner gesprochen dürfen sich bei der Gültigkeit des PCMs die Aufgabenparameter der einzelnen Teilgruppen nicht systematisch unterscheiden. Insofern DIF vorliegt, „kann kein fairer Vergleich der Merkmalsausprägungen mittels der entsprechenden Skala vorgenommen werden. Ein Vergleich einer Merkmalsausprägung zw. verschiedenen Stufen einer Moderatorvariable wäre verzerrt, weil in den Gruppen das entsprechende Einzelitem in unterschiedlicher Weise die zu messende latente Merkmalsausprägung anzeigt“ (Wirtz, 2014, S. 404). Damit kann konkret für das PCM also auch getestet werden, ob die Steigung aller kurven - dargestellt anhand der Schwellenwahrscheinlichkeiten - gleich ist bzw. ob alle Items die gleiche Trennschärfe aufweisen. Zum konkreten Prüfen der Annahme der spezifischen Objektivität stehen im eRm-Paket nun verschiedene Tests zur Verfügung: der Likelihood-Quotienten-Test (LQ-Test), Martin-Löf-Test, Wald-Test (global und lokal), sowie der χ^2-Anpassungstest. Im Rahmen dieser Arbeit soll nur auf

den zuerst genannten eingegangen werden, wobei die nachfolgenden Schilderungen zum LQ-Test in Anlehnung an Strobl (2012) und Koller (2012) erfolgen.

Die Basis des LQ-Tests bildet die aus der CML-Schätzung resultierende Likelihood (siehe dazu Kapitel 5.2). Die zugrunde liegende Idee ist dabei folgende: man berechnet einerseits die Likelihood für den gesamten Datensatz und dann andererseits auch noch die Likelihoods getrennt für die gebildeten zwei oder mehr Gruppen. Wenn nun die Parameterschätzung für die Gesamt- und die Teilgruppen gleich sind, dürften sich auch die jeweiligen Likelihoods nicht unterscheiden. Damit können also zwei Fälle eintreten:

1. Die Itemschwierigkeiten in den Gruppen sind gleich: Insofern man die einzelnen Likelihoods der Gruppen multipliziert, ergibt sich die gleiche Gesamtlikelihood, die man bei einer Parameterschätzung ohne Gruppenbildung erhalten würde.

2. Die Itemschwierigkeiten in den Gruppen sind ungleich: Man erhält unterschiedliche Parameterschätzwerte in den Gruppen, d.h. die geschätzte Itemschwierigkeit zwischen den einzelnen Gruppen unterscheidet sich deutlich. Durch Multiplikation der Gruppenlikelihoods steigt so schließlich auch die Gesamtlikelihood.

Im eRm-Paket ist der LQ-Test nun folgendermaßen implementiert, wobei sich der Name der Funktion von der englischen Bezeichnung Likelihood-Ratio-Test (LR-Test) ableiten lässt:

```
LRtest(object, splitcr = "median", se = TRUE)
```

Als 'object' wird wiederum ein Objekt der Klasse 'eRm' - welches mit der Funktion 'PCM()' erstellt wurde - übergeben. Zugleich wird durch 'splitcr' das Kriterium festgelegt, nach welchem die einzelnen Subpopulationen gebildet werden sollen. Die zugrundeliegende Standardeinstellung ist hierbei der s.g. Mediansplit. Die Zuordnung zu den jeweiligen Gruppen erfolgt dann anhand des Medians der von den betrachteten Probanden erzielten Testscores r_p, wobei $p = 1, \ldots, P$ gilt. Dementsprechend werden zwei Gruppen gebildet, wobei eine Gruppe all diejenigen Personen umfasst, die einen Testscore kleiner oder gleich dem Wert des Medians erreicht haben, und in der zweiten Gruppe sind somit all diejenigen Testteilnehmer inbegriffen, die einen größeren Testscore als der Median aufweisen. Auch andere Gruppierungen können vorgenommen werden, beispielsweise nach dem arithmetischen Mittel durch Angabe von 'splitcr=„mean“' oder indem man alle Personen mit demselben Testscore zu Subpopulation mittels 'splitcr=„all.r“' zugeteilt werden. Auch kann die Einteilung anhand eines Vektors festgesetzt werden, der jede Personen explizit einer Gruppe zuweist, wobei dieser Vektor vom Typ 'numeric', 'character' oder 'factor' sein kann.

Im Rahmen des LQ-Tests wird also die Gesamtpopulation in h Subpopulationen - wobei gilt $h = 1, \ldots, H$ - gemäß eines bestimmten Kriteriums unterteilt. Unter Gültigkeit des PCMs erwartet man nun, dass die geschätzten Itemparameter der einzelnen Gruppen in etwa gleich sind und insbesondere dem β-Parameter - geschätzt auf Basis der Gesamtpopulation - entsprechen. Damit lassen sich die Null- und Alternativhypothese folgendermaßen aufstellen:

$$\mathrm{H}_0: \ \beta_{ij} = \beta_{ij}^{(1)} = \cdots = \beta_{ij}^{(H)} \qquad \textit{vs.} \qquad \mathrm{H}_1: \ \exists u, v: \ \beta_{ij}^{(u)} \neq \beta_{ij}^{(v)}; \ \ u, v \in h$$

Beim LQ-Test werden die geschätzten Aufgabenparameter der Gruppen $\hat{\beta}_{ij}^{(h)}$ nun mit dem der Gesamtstichprobe $\hat{\beta}_{ij}$ verglichen. Als Schätzer ergeben sich dann diejenigen Werte, welche die bedingte Likelihood maximieren. Dabei gehen in die Schätzung nur noch die in der entsprechenden Gruppe enthaltenen Personen ein, nicht alle P Testteilnehmer. Der Likelihood-Quotient ergibt sich dann, indem man die Gesamtlikelihood der geschätzten Itemparameter $\hat{\beta}_{ij}$ durch das Produkt der Gruppenlikelihoods mit den geschätzten Aufgabenparameter der einzelnen Gruppen $\hat{\beta}_{ij}^{(h)}$ teilt:

$$\mathrm{LQ} = \frac{\Lambda_{CML}}{\prod_{h=1}^{H} \Lambda_{CML}^{(h)}} \tag{6.2}$$

Würde nun das Rasch-Modell Gültigkeit besitzen, so ergibt sich für den LQ in etwa der Wert 1, da es - abgesehen von zufälligen Abweichungen - keinen Unterschied machen würde, ob man eine nach Gruppen getrennte oder eine gemeinsame Schätzung durchführen würde. Dieses Szenario wird als Nullhypothese des Tests angenommen. Erhält man ein signifikantes Ergebnis, so deutet dies dementsprechend darauf hin, dass das PCM nicht gültig ist. Dementsprechend gilt für die Alternativhypothese, dass in den gebildeten Subpopulationen jeweils unterschiedliche Itemparameter $\hat{\beta}^{(h)}$ zu den gegebenen Daten innerhalb der Gruppe passen. Damit ergibt sich dann, dass das Produkt im Nenner von Gl. (6.2) größer wird und man insgesamt schließlich einen Likelihood-Quotienten erhält, der kleiner als 1 und damit als die Nullhypothese ist.

Allerdings kann der Likelihood-Quotient LQ nicht direkt zum Testen der Hypothese angewendet werden. Jedoch kann angenommen werden, dass die Teststatistik T einer χ^2-Verteilung mit entsprechender Anzahl an Freiheitsgraden folgt. Dabei errechnen sich die Freiheitsgrade aus der Differenz zwischen der aufsummierten Anzahl von Parametern in den einzelnen Subpopulationen abzüglich der Anzahl an Parametern der Gesamtpopulation.

$$T = -2ln(LQ) = 2(\sum_{h=1}^{H} \lambda_{CML}^{(h)} - \lambda_{CML}) \tag{6.3}$$

Mithilfe der Teststatistik T kann nun der statistische Test folgendermaßen konstruiert werden: besitzt das PCM Gültigkeit, beträgt also $LQ = 1$, so ergibt sich $T = 0$. Andrerseits erhält man bei Verletzung des Modells $LQ < 1$ und damit für die Teststatistik $T > 0$. Ergibt sich also ein großer Wert für T - beispielsweise größer als das 95%-Quantil der χ^2-Verteilung mit entsprechender Anzahl an Freiheitsgraden - so liegt eine eben beschriebene, signifikante Modellverletzung vor.

Nachteil an diesem R-Paket ist - wie bereits erwähnt - die Tatsache, dass ordinale Daten basierend auf dem PCM analysiert werden können, allerdings nicht mittels des GPCMs. Hierzu können allerdings die nachfolgend vorgestellten Pakete, das ltm- und das TAM-Paket genutzt werden. In dem nun anschließend Kapitel wird zunächst auf Ersteres eingegangen.

6.2 Paket 'ltm'

Das ltm-Paket wurde zur Analyse von multivariaten dichotomen, sowie polytomen Daten entwickelt und baut auf Anwendung von „Latente-Variablen-Modellen". Hierbei werden je nach Skalenniveau der beobachteten und der latenten Variable unterschiedliche Modellierungsannahmen getroffen. Um dies zu veranschaulichen ist nachfolgend Tab. 6.1 dargeboten, welche in Anlehnung an Cai (2012) erstellt wurde. Allgemein gesprochen stellen Latente-Variablen-Modelle multivariate Regressionsmodelle dar, die die Möglichkeit bieten von stetigen oder kategorialen Antworten/-mustern auf unbeobachtete Kovariablen zu schließen (vgl. Rizopoulos, 2006, S.1). Konkret wird im ltm-Paket nun das Latente-Variablen-Modell unter Betrachtung der IRT umgesetzt.

	Latente Variable	
Beobachtete Variable	Diskret	Stetig
Diskret	Latente Klassenanalyse	Item Response Theorie/ Latent Trait Analyse
Stetig	Latente Profileanalyse	Faktorenanalyse

Tab. 6.1: Unterschiedliche Anwendung von Latenten-Variablen-Modellen je nach Ausprägung (stetig oder diskret) der beobachteten und der latenten Variable

Die nun anschließende Beschreibung geschieht in Anlehnung an Rizopoulos (2006). Die Grundidee der latenten Variablenanalyse ist es für ein gegebenes Set an Antwortvariablen $x_1, \ldots, x_K$ ein Set latenter Variablen $\boldsymbol{\theta} = \theta_1, \ldots, \theta_P$ zu finden - wobei $P << K$ -, sodass dieses Set im Wesentlichen dieselbe Information wie die latente Variablen selbst beinhaltet. Während im ltm-Paket vorwiegend das GRM zur Analyse polytomer, ordinaler Daten bereitgestellt wird, kann auch das PCM bzw. GPCM, sowie das RSM herangezogen werden.

Die Parameterisierung orientiert sich bei der Anwendung des PCMs bzw. GPCMs an der Modellformel des GPCMs, wie sie in Gl. (4.20) dargestellt ist. Hierbei kann der Nutzer festlegen, ob das PCM in seiner restriktivsten Form, also mit einem α-Parameter, der für alle Items bei 1 liegt, anzuwenden ist oder ob die Analyse mittels des PCMs, mit einem für alle Items konstanten, aber angepassten α-Parameter durchgeführt werden soll. Des Weiteren kann aber auch vom PCM in seiner allgemeinsten Form, also dem GPCM, mit α_i-Parameter, der für jedes Item separat geschätzt wird, Gebrauch gemacht werden.

Hierzu soll nun kurz auf den wesentliche Unterschied in der Verwendung des PCMs in seiner restriktivsten Form oder der ebenso möglichen, weniger restriktiven Form verdeutlicht werden. Das PCM in seiner restriktivsten Form weist einen Steigungsparameter von $\alpha_i = 1$ auf und setzt voraus, dass die Fähigkeiten der Personen der Standardnormalverteilung $\theta_p \sim N(0, 1)$ folgen. Diese Form des PCMs wird im ltm-Paket mit 'rasch' bezeichnet. Im Vergleich hierzu geht man nun bei der weniger restriktive Form - im ltm-Paket bezeichnet mit '1PL' - davon aus, dass die Personenparameter normalverteilt sind mit Erwartungswert 0, setzt die Varianz σ^2 jedoch als nicht bekannt voraus, also gilt $\theta_p \sim N(0, \sigma^2)$. Diese Annahme ist äquivalent zu derjenigen, die Standardnormalverteilung der Personenparameter anzunehmen und demzufolge aber einen Steigungsparameter $\alpha_i = \alpha = \sigma$

zuzulassen. Die Äquivalenz dieser beiden Ansätze ist in der nachfolgenden Gl. (6.4) dargestellt. Hierbei gilt zunächst $\theta_p \sim N(0, \sigma^2)$ und $\alpha_i = 1$.

$$\begin{aligned} \alpha_i(\theta_p - \beta_{ij}) &= \alpha_i(\theta_p - \beta_{ij})\frac{\sigma}{\sigma} \\ &= \alpha_i \cdot \sigma\Big(\frac{\theta_p}{\sigma} - \frac{\beta_{ij}}{\sigma}\Big) \\ &= \sigma(\tilde{\theta_p} - \tilde{\beta_{ij}}) \end{aligned} \tag{6.4}$$

$$\rightarrow \tilde{\alpha_i} = \alpha_i \cdot \sigma = 1 \cdot \sigma = \sigma = \tilde{\alpha}, \quad \tilde{\theta_p} \sim N(0, 1)$$

Nachdem die zugrundeliegende Parameterisierung im ltm-Paket über das GPCM erfolgt, ist es naheliegend, dass die Berechnung der gesuchten Itemparameter auf Basis der MML-Methode geschieht, in der Form wie sie in Kapitel 5.3 bereits ausführlich dargestellt wurde. Hierbei werden die einzelnen Fähigkeitsparameter als standardnormalverteilt angenommen, wobei je nach Modell schließlich die α-Parameter festgelegt werden. Gleichzeitig erfolgt die numerische Integration über die s.g. „Gauß-Hermite-Quadratur“, wobei die zu integrierende Funktion in eine Gewichtsfunktion und ein spezielles Polynom jeweils an bestimmten Auswertungspunkten zerlegt wird. Zu beachten ist hierbei, dass die vom Benutzer festgelegt Anzahl eben dieser Auswertungspunkte, Einfluss auf die Parameterschätzungen, die Standardfehler, sowie den Wert der Log-Likelihood haben kann (vgl. Rizopoulos, 2006, S. 4). Für eine genauere Darstellung dieser Methode kann beispielsweise Liu (1994) genutzt werden. Zusätzlich sei noch gesagt, dass die Maximierung der integrierten Log-Likelihood in Hinsicht auf die Itemparameter - also die logarithmierte Form von Gl. (5.21 - über den s.g. „Broyden-Fletcher-Goldfarb-Shanno“ (BFGS) Algorithmus erfolgt. Dieser ist der Gruppe der „Quasi-Newton-Verfahren“ zuzuordnen und bietet die Möglichkeit nicht-lineare Optimierungsprobleme zu lösen. Hierbei bezeichnet das Quasi-Newton Verfahren eine Variante des Newton-Verfahrens, wobei die Hessematrix eines Funktionals nicht direkt erforderlich ist, sondern bei dem stattdessen die Hesse-Matrix iterativ approximiert wird.

Die nun nachfolgenden Erklärungen zum ltm-Paket bzw. zu den implementierten Funktionen, die bei der Analyse mittels des PCMs und GPCMs hauptsschlich von Bedeutung sind, stützen sich auf Rizopoulos (2006) und Rizopoulos (2015). Um zunächst das PCM bzw. GPCM auf die ordinalen Daten anzuwenden, bedient man sich der nachstehenden Funktion:

```
gpcm(data, constraint = c("gpcm", "1PL", "rasch"), IRT.param = TRUE,
     start.val = NULL, na.action = NULL, control = list())
```

Hierbei muss es sich bei dem übergebenen Parameter 'data' entweder um einen s.g. 'data.frame' oder eine numerische Matrix handeln und damit wird letztlich also wieder eine Matrix der Form $\boldsymbol{X}$ verlangt. Mittels 'constraint' kann der Benutzer über einen 'character'-Wert angeben, welches Modell auf die gegebenen Daten angewandt werden soll. Wie bereits erwähnt kann man den Steigungsparameter des PCMs konsequent auf 1 festlegen, wobei hierzu der Parameter 'constraint = "rasch“ übergeben wird, oder aber man lässt einen α-Parameter zu, der für alle Items gleich ist, indem man 'constraint = "1PL“ wählt. Die Analyse der Daten mittels des GPCMs und damit α_i-

Parametern, die für jedes Item gesondert geschätzt werden, ist durch Angabe von 'constraint = "gpcm" möglich. Die übrigen Parameter sind optional, so kann etwa mittels 'control' eine Liste von Einstellungen übergeben werden, beispielsweise wie viele Iterationen der Quasi-Newton Algorithmus, der zur Schätzung der Itemparameter genutzt wird, maximal durchlaufen soll. Als Optimierungsmethode wird - wie bereits erwähnt - zumeist das BFGS-Verfahren verwendet, welches auch in dem eRm-Paket als Standard festgelegt ist. Gesondert betrachtet werden sollte nur noch kurz der Parameter 'na.action', anhand dessen sich festlegen lässt, wie mit den fehlenden Daten bzw. 'NA'-Werten umgegangen werden soll. Durch den Default-Wert 'na.action=NULL' werden fehlende Werte zugelassen.

Das mit dem 'gpcm'-Befehl erzeugt Objekt beinhaltet u.a folgende Komponenten: eine Liste 'coefficents', welche aufgeschlüsselt nach den einzelnen Aufgaben die geschätzten Schwellenparameter $\tau_i k$ und den jeweils zugehörigen Diskriminationsparameter umfasst, die Log-Likelihood 'loglik', sowie die approximierte Hesse-Matrix 'hessian' bei erreichter Konvergenz. Will man sich statt der aufsummierten Schwellenparameter τ_{ik} die einzelnen Schwierigkeitsparameter β_{ik} ausgeben lassen, ist es empfehlenswert die Funktion 'coefficients()' auf das erzeugte Objekt der Klasse 'gpcm' anzuwenden.

Des Weiteren sind noch einige weitere Informationen bei dem Aufruf der 'gpcm()'-Funktion inbegriffen und so kann man sich beispielsweise auch die Liste 'GH' ausgeben lassen, welche die zwei Bestandteile beinhaltet, die bei der Gauß-Quadratur verwendet werden: der erste Eintrag 'Z' umfasst die numerische Matrix, welche die jeweiligen Abszissen-Werte bzw. den Auswertungspunkte enthält, und der zweite Eintrag 'GHw' birgt einen nummerischen Vektor mit den zugehörigen Gewichtungen. Daneben sind noch weitere Werte des Objekts 'gpcm' abrufbar, die hier aber nun nicht weiter von Bedeutung sind.

Zusätzlich kann auch die Varianz-Kovarianz-Matrix ausgegeben werden durch Aufrufen der nachfolgenden Funktion. Hierbei sollte es sich bei 'object' um ein Objekt der Klasse 'gpcm' handeln. Insofern der logische Parameter 'robust' auf 'TRUE' gesetzt wird, wird der Sandwich-Schätzer verwendet. Nach Ausführung der Funktion erhält man eine numerische Matrix, welche die Varianz-Kovarianz-Matrix der aus der MML-Schätzungen hervorgehenden Parameter repräsentiert.

```
vcov(object, robust = FALSE, ...)
```

Das mittels der 'gpcm()'-Funktion angepasste PCM bzw. GPCM kann im ltm-Paket auch grafisch dargestellt werden. Hierzu bedient man sich der nachstehenden Funktion:

```
plot(x, type = c("ICC", "IIC", "OCCu", "OCCl"), items = NULL,
     categorie = NULL, ...)
```

Wiederum erfordert 'x' also die Angabe eines Objekts der Klasse 'gpcm'. Mittels 'type' lässt sich festlegen, in welcher Form das angepasste Modell dargestellt werden soll. Hierbei kann das Modell mittels der Einstellung 'ICC' über die Kategorienwahrscheinlichkeiten - wie in Kapitel 4.3.2 beschrieben - dargestellt werden. Auf die übrigen Werte, die als Grafik-Typen gewählt werden können, sollen hier nicht weiter eingegangen werden, jedoch können anhand von Rizopoulos (2015) nähere Informationen dazu gewonnen werden. Anhand des Parameters 'items' kann man diejenigen

Aufgaben auswählen, die grafisch dargestellt werden sollen. Mittels 'categorie' lässt sich in Form eines Skalars festsetzen, welche Antwort-Kategorie eines Items jeweils abgebildet wird. Insofern die Standard-Einstellung 'NULL' vorliegt, werden alle Kategorien dargeboten. Alle weiteren Parameter, die der Funktion 'plot()' übergeben werden können, sind Grafik-Einstellungen und werden hier nicht weiter ausgeführt.

Zudem gilt es nach der Itemparameterschätzung auch diejenigen Schätzer für die Fähigkeit der Probanden zu ermitteln. Hierzu kann die im ltm-Paket bereitgestellte Funktion 'factor.scores()' genutzt werden. Dabei sind im Folgenden nur einige wesentliche Parameter dargestellt, die der Funktion übergeben werden können.

```
factor.scores(object, method = c("EB", "EAP", "MI"), prior = TRUE, ...)
```

Wie üblich wird der Funktion zunächst der Name des Objekts übermittelt, welches von der Klasse 'gpcm' sein muss. Mittels des Parameters 'method' kann die gewünschte Methode zur Schätzung der Personenparameter festgelegt werden. Dabei stehen die drei folgenden Verfahren zur Auswahl: Empirischer Bayes (EB), Exected A Posteriori (EAP) und Multiple Imputation (MI). Zu Ersterem lässt sich beispielsweise die im vorhergehenden Kapitel 5.3 beschriebene MAP-Methode zuordnen. Ebenfalls in dem angesprochenen Kapitel wurde das EAP-Verfahren dargestellt. Zusätzlich sei gesagt, dass sich gerade beim Auftreten fehlender Werte die Anwendung der MI-Methode eignet, deren Schätzwertbestimmung zumeist auf der Simulation eines plausiblen multivariaten Verteilungsmodells beruht. Insofern die Methode 'EB' gewählt wurde, kann mittels 'prior' festgelegt werden, ob als a priori Verteilung die Normalverteilung für die Fähigkeitsparameter angenommen werden soll, um schließlich den jeweiligen Modus der posteriori Verteilung zu ermitteln. Hierbei wird durch die Standard-Einstellung 'prior = TRUE' eben dies festgesetzt.

Nach Aufruf der Funktion 'factor.scores' erfolgt die Ausgabe einer Liste der Klasse 'fscores', mit insbesondere der folgenden Komponenten: 'score.dat' umfasst als 'data.frame' alle beobachteten Antwortmuster, mit sowohl den beobachteten, als auch erwarteten Häufigkeiten. Hierzu werden die entsprechenden Fähigkeitsparameter, sowie der Standardfehler angegeben.

Die Dichte der Fähigkeitsparameter lässt sich grafisch mittels Kerndichteschätzer darstellen, wobei dies auf dem entsprechende, gewählten Verfahren zur Schätzung der Wahrscheinlichkeitsverteilung beruht. Gemäß eines s.g. Kerns, welcher selbst eine Dichte darstellt, und einer festgesetzten Bandbreite ergibt sich der Kerndichteschätzer quasi aus der Überlagerung von entsprechend skalierten Kernen,die gemäß einer Stichprobenrealisation positioniert werden. Zur genaueren Erklärung der Kerndichteschätzung sei beispielsweise auf Schomaker (2008, S. 42ff.) verwiesen. Hierbei stehen auch im ltm-Paket verschiedene Kerne bei Dichteschätzung zur Verfügung. So kann u.a. der Gauß-Kern, welcher dem Default-Wert der Funktion entspricht, oder etwa der Epanechnikov-Kern gewählt werden. Die Bandbreite kann mit dem Parameter 'bw' angepasst werden. Durch Aufruf der Funktion 'plot()' mit entsprechender Angabe eines Objekts 'x' der Klasse 'fscores' erfolgt die Ausgabe einer Grafik mit der Kerndichteschätzung der Fähigkeitsparameter.

```
plot(x, bw = "nrd0", kernel = "gaussian", ...)
```

Eine weitere nützliche Funktion im ltm-Paket stellt 'anova()' dar. Hiermit können explizit zwei

Modelle verglichen werden bzw. genauer gesagt, lässt sich bestimmen, ob ggf. eines der Modelle besser zu den vorliegenden Daten passt. Insgesamt bedeutet dies hier also, dass man das PCM, welches das Modell unter der Nullhypothese darstellt, und GPCM zur Anwendung bringt und anschließend bestimmt, ob die Berücksichtigung des Diskriminationsparameters einen signifikanten Unterschied birgt. Um diese Signifikanz der hierarchisch geschachtelten Modelle - s.g. „nested models" - zu testen, bedient man sich des Likelihood-Quotienten-Tests. Konkret wird beim Likelihood-Quotienten-Test die Differenz der χ^2-Werte von den beiden betrachteten Modellen gebildet, sowie die Differenz der entsprechenden Freiheitsgrade, welche sich konsequenter Weise nur um einen Freiheitsgrad unterscheiden (vgl. Gonzalez, 2001, S. 263). Bei normalverteilten Daten und einer ausreichend großen Stichprobe ist der Betrag der errechneten Differenz selbst χ^2-verteilt mit einem Freiheitsgrad. In R kann dieser Test schließlich mittels der implementierten Funktion 'anova()' durchgeführt werden:

```
anova.gpcm(object1, object2, simulate.p.value = FALSE, B = 200,
           seed = NULL, ...)
```

Des Weiteren bietet diese Funktion die Möglichkeit die p-Werte des Likelihood-Quotienten-Tests zu schätzen, indem man sich des parametrischen Bootstrappings bedient, welches insbesondere aufgrund der nicht bekannten Verteilung der Prüfgröße von Nutzen ist. „Bootstrapping" bezeichnet hierbei eine computergestützte Methode des Resamplings, wobei „man unter einem Resampling-Verfahren eine statistische Methode, die auf Basis einer Ausgangsstichprobe [...] durch wiederholtes Erzeugen einer neuen Stichprobe nach einem vorgeschriebenen Muster Informationen über die Zusammensetzung der zugrundeliegenden Verteilung liefert" (Pauls, 2003, S. 1). Die Anzahl der Stichproben unter der Nullhypothese lässt sich durch Festsetzen von 'B' auf den gewünschten Wert bewerkstelligen, die Standardeinstellung beträgt hierbei 'B=200'. Der p-Wert wird schließlich über die nachfolgende Gl. (6.5) approximiert, wobei T_{obs} den Wert der Likelihood-Quotienten-Statistik des Orginal-Datensatzes bezeichnet und T_b denjenigen Wert der Statistik, der sich für die i-te Bootstrap-Stichprobe ergibt.

$$[1 + \sum_{b=1}^{B} I(T_b > T_{obs})]/(B + 1) \tag{6.5}$$

Insofern 'simulate.p.value = TRUE' gesetzt wurde, kann mittels der 'plot()'-Funktion ein QQ-Plot ausgegeben werden, bei dem die Likelihood-Quotienten-Statistik der Bootstrap-Stichproben mit der asymptotischen χ^2-Verteilung verglichen wird. Anzumerken sei noch, dass es sich beim Bootstrapping zwar jeweils um eine (Pseudo-)Zufallsstichproben handelt, diese jedoch auf einem gewissen Algorithmus beruhen und so durch Festlegung eines Startwertes reproduzierbar werden. Dazu kann der Parameter 'seed' verwendet werden, wobei - wenn nicht anders angegeben - ein zufälliger Startwert gewählt wird.

Auch der nachfolgend vorgestellte Test zum Prüfen der Anpassungsgüte - im Englischen bezeichnet mit „Goodness of Fit" - eines Modells, stützt sich auf das Bootstrapping. Will man mittels des ltm-Paketes testen, wie gut das angepasste PCM oder GPCM die Beobachtungen erklären kann, so verwendet man die nachstehende Funktion. Demgemäß besagt die Nullhypothese, dass die Zufallsvariable die angegebene Verteilung besitzt und damit das angepasste Modell ausreichend gut zu den Daten passt. Diesbezüglich lässt sich dann auch die entsprechende Alternativhypothese formulieren.

```
GoF.gpcm(object, simulate.p.value = TRUE, B = 99, seed = NULL)
```

Diese Funktion führt für das mittels 'object' übergebene Modell - also hier PCM oder GPCM - einen parametrischen Bootstrap-Test basierend auf Pearson's χ^2-Statistik durch, welcher folgendermaßen definiert ist:

$$\sum_{q=1}^{2^I} \frac{[O(q) - E(q)]^2}{E(q)} \tag{6.6}$$

Hierbei steht q für ein bestimmtes Antwortmuster, I stellt wiederum die Gesamtzahl an Items dar und $O(q)$, sowie $E(q)$ repräsentieren jeweils die beobachteten bzw. erwarteten Häufigkeiten. Die übrigen Parameter, welche der Funktion 'GoF.gpcm()' übergeben werden können, entsprechen in analoger Weise den bei der 'anova()'-Funktion verwendbaren Parametern. Das genau schrittweise Vorgehen beim Bootstrap-Test lässt sich mittels Rizopoulos (2015) nachvollziehen. Die wichtigsten Rückgabewerte des erzeugten Objektes der Klasse 'GoF.gpcm' sind der Wert der χ^2-Statistik für die beobachteten Daten, die Anzahl der Bootstrap-Stichproben 'B' und insbesondere der p-Wert des Tests 'p.value'. Anhand Letzterem kann entsprechend abgelesen werden, ob sich die beobachteten Häufigkeiten signifikant von den erwarteten unterscheiden.

Anschließend an die Erläuterungen zum ltm-Paket soll nachfolgend noch ein weiteres R-Paket vorgestellt werden, mithilfe dessen die Analyse von Daten durch das PCM bzw. GPCM ermöglicht wird.

6.3 Paket 'TAM'

Um das Kapitel der R-Pakete abzuschließen, soll im Folgenden noch das recht neue Paket TAM vorgestellt werden, dass sowohl die Analyse ordinaler Daten mittels des PCMs, als auch des GPCMs zulässt. Für das PCM ist es möglich die Parameter über die JML-Methode, wie sie in Kapitel 5.1 dargeboten wurde, zu schätzen. Dagegen bietet sich die MML-Methode, welche in Kapitel 5.3 vorgestellt wurde, sowohl für das PCM, als auch das GPCM an. Bei der anschließenden Schätzung der Personenparameter werden verschiedene Methoden bereitgestellt: insofern man die MML-Methode zur Bestimmung der Itemparameter genutzt hat, kann man sich der unbedingten ML-Methode bedienen oder der gewichteten ML-Methode, wie sie in Kapitel 5.2 beschrieben wurde, oder der Exected A Posteriori (EAP) Methode, wobei dieses bayessche Verfahren in Sektion 5.3 erläutert wurde. Dagegen erfolgt die Personenparameterschätzung bei Anwendung der JML-Methode gleichzeitig mit der Itemparameterschätzung, was bereits in Kapitel 5.1 erklärt wurde.

Generell stellt das TAM-Paket wohl dasjenige der drei vorgestellten Pakete dar, in welchem das „Basismodell“ so allgemein formuliert ist, dass sich durch jeweilige Spezifikation eine Reihe unterschiedlicher IRT-Modelle anwenden lassen und sich insbesondere auch mehrdimensionale Modelle konstruieren lassen. Dieses „Basismodell“ ist das s.g. „Random Coefficients Multinomial Logit“ Modell (RCMLM). Konkret lässt sich das unidimensionale RCMLM nachfolgend in Anlehnung an Adams (1997, S. 49-52) und Volodin (2002, S. 2-5) beschreiben. Hierbei soll konkret geklärt werden, wie das PCM aus diesem hervorgeht.

Wiederum geht man von $p = 1, \ldots, P$ Personen, sowie von $i = 1, \ldots, I$ Items aus, wobei jedes dieser $k = 1, \ldots, m_i$ Kategorien aufweist. Hierbei bringt jedes Individuum ein Satz an Item-Antworten $\boldsymbol{I}_p$ mit sich, was sich für alle P Personen als Vektor $\boldsymbol{I}$ zusammenfassen lässt. Gleichzeitig wird ein Vektor von Zufallsvariablen $\boldsymbol{X_{pi}} = (X_{pi1}, \ldots, X_{pim_i})'$ eingeführt, dessen Einträge X_{pij} den Wert 1 annehmen, falls die Antwort von Person p zu Item i in Kategorie j - wobei wiederum gilt $j \in k$ - fällt und anderenfalls wird dieser der Wert 0 zugewiesen. Eine Antwort in der Kategorie $j = 0$ bzw. im Ausgangsniveau wird prinzipiell mit einer 0 versehen und bewirkt damit, dass diese zur Referenzkategorie wird, was zur Modellidentifikation erforderlich ist. Damit kann das Antwortmuster einer Person schließlich in dem Zufallsvariablen-Vektor $\boldsymbol{X}_p = (\boldsymbol{X}'_{p1}, \ldots, \boldsymbol{X}'_{pI})'$ zusammengefasst werden.

Die Personen selber werden anhand eines D-dimensionalen Fähigkeitsparameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)$ beschrieben, welcher als zufällig verteilt mit einer Populationsdichte, die durch die multivariate Normalveteilungs-Dichtefunktion beschrieben wird, angesehen wird. Hierbei bezeichnet $\boldsymbol{\mu}$ den Mittelwertsvektor und $\boldsymbol{\Sigma}$ die Kovarianzmatrix des Zufallsvektors $\boldsymbol{\theta}$. Entsprechend gilt also Folgendes:

$$g(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\boldsymbol{\pi})^{D/2}|\boldsymbol{\Sigma}|^{1/2}} exp[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})] \tag{6.7}$$

Im RCMLM sind nun die Itemparameter im Vektor $\xi = (\xi_1, \ldots, \xi_L)$' zusammengefasst. Bei der anschließend präsentierten Modellierung der Antwortwahrscheinlichkeiten werden Linearkombinationen der ξ's genutzt, um die empirische Charakteristik der Antwortkategorien eines jeden Items zu beschreiben. Die jeweiligen linearen Kombinationen werden mittels der Design-Vektoren $\boldsymbol{a}'_{ij}$ definiert, wobei die einzelnen Vektoren zur Design-Matrix $\boldsymbol{A} = (\boldsymbol{a}'_{11}, \ldots, \boldsymbol{a}'_{1m_1}, \ldots, \boldsymbol{a}'_{I1}, \ldots, \boldsymbol{a}'_{Im_i})$ zusammengefasst werden können. Definiert man die Anzahl der Kategorien eines Tests als $K = \sum_{i=1}^{I} m_i$, so entspricht die Designmatrix also einer $K \times L$-Matrix. Aufgrund dieses Ansatzes, den Itemparameteren ein lineare Modell zugrundezulegen, ermöglicht das RCMLM die Anwendung vieler existierender IRT-Modelle.

Des Weiteren beinhaltet das RCMLM eine Scoring-Funktion, die es erlaubt den jeweiligen Antwortscore bzw. das Leistungsniveau zu beschreiben, welches sich zu jeder Antwortkategorie der D-Dimensionen zuordnen lässt. Damit wird es möglich die Notation eines Antwortscores einzuführen, durch den das Scorelevel für die Dimension d der beobachteten Antwort in Kategorie k des Items i angegeben wird. Falls die Antwortkategorie eines bestimmten Items einer betrachteten latenten Dimension d nicht angehört, so wird der Score dieser Dimension entsprechend auf 0 gesetzt. Die einzelnen Score c_{pik} können in dem Vektor $\boldsymbol{c}_{ik} = (c_{1ik}, \ldots, c_{Dik})'$ zusammengefasst werden und diese wiederum lassen sich gemeinsam in Form einer Matrix $\boldsymbol{C} = (\boldsymbol{c}'_{11}, \ldots, \boldsymbol{c}'_{1m_1}, \ldots, \boldsymbol{c}'_{I1}, \ldots, \boldsymbol{c}_{Im_I})'$ betrachten. Generell wird durch die Einführung von $\boldsymbol{c}$ als Scoring-Funktion eine flexiblere Beziehung zwischen der Qualität einer Antwort und dem Leistungslevel, welches von dieser reflektiert wird, zugelassen. Allgemein wird mit der Einführung der Scoring-Matrix $\boldsymbol{C}$, sowie der Design-Matrix $\boldsymbol{A}$ die Möglichkeit eröffnet ein generelles gemischtes multinomiales Logit-Regressionsmodell zu konstruieren, welches eine Vielzahl existierender IRT-Modelle bzw. insbesondere der Rasch-Familie angehörende Modelle beinhaltet.

Aufbauend auf den eben beschriebenen Bezeichnungen kann nun die Wahrscheinlichkeit einer

bestimmten Item-Antwort mittels des RCMLM folgendermaßen dargestellt werden:

$$P(X_{pik} = x_{pik}; \boldsymbol{A}, \boldsymbol{C}, \xi|\theta_p) = \frac{exp[\boldsymbol{x}_{pik}(c'_{ij}\theta_p + \boldsymbol{a}'_{ij}\xi)]}{\sum_{u=1}^{m_i} exp[\boldsymbol{x}_{piu}(c'_{iu}\theta_p + \boldsymbol{a}'_{iu}\xi)]} \tag{6.8}$$

Damit lässt sich auch die Wahrscheinlichkeit für einen bestimmten Antwortvektor $\boldsymbol{x}$ der p-ten Person modellieren, wobei $\boldsymbol{\Omega}$ das Set aller möglichen Antwortvektoren bezeichnet:

$$P(\boldsymbol{X}_p = \boldsymbol{x}_p; \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi}|\boldsymbol{\theta}) = \frac{exp[\boldsymbol{x}'_p(\boldsymbol{C}\boldsymbol{\theta} + \boldsymbol{A}\boldsymbol{\xi})]}{\boldsymbol{\Psi}(\boldsymbol{\theta}, \boldsymbol{\xi})}, \tag{6.9}$$

$$\text{mit} \quad \boldsymbol{\Psi}(\boldsymbol{\theta}, \boldsymbol{\xi}) = \prod_{i=1}^{I} \sum_{j=1}^{m_i} exp(\boldsymbol{c}'_{ij}\boldsymbol{\theta} + \boldsymbol{a}'_{ij}\boldsymbol{\xi}) = [\sum_{\boldsymbol{z}\in\Omega} exp[\boldsymbol{z}'(\boldsymbol{C}\theta_p + \boldsymbol{A}\boldsymbol{\xi})]]^{-1}$$

Folglich kann man schließlich auch die marginale Formulierung der Wahrscheinlichkeit eines Antwortvektors $\boldsymbol{x}_p$ bilden, wobei man davon ausgeht, dass die Fähigkeitsparameter $\boldsymbol{\theta}$ einen Mittelwert von $\boldsymbol{\mu} = \mathbf{0}$ und Varianz $\boldsymbol{\Sigma}$ aufweisen:

$$P(\boldsymbol{X}_p = \boldsymbol{x}_p; \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathbb{R}^D} P(\boldsymbol{X}_p = \boldsymbol{x}_p; \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\xi}|\boldsymbol{\theta}) f(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{\theta} \tag{6.10}$$

Generell sollte anhand der vorhergehenden Gleichungen ersichtlich werden, wie sich das PCM basierend auf dem RCMLM konstruieren lässt. In Anlehnung an Volodin (2002, S. 5) gilt Folgendes: die $\boldsymbol{A}$-Matrix besteht aus einer Abfolge von Blöcken, die gemäß den einzelnen Items ergeben. Die Größe dieser Blöcke entspricht der jeweiligen Anzahl der Kategorien m_i innerhalb der Items. Damit ergibt sich schließlich für das PCM, dass Elemente oberhalb der Hauptdiagonalen 0 sind, während die Elemente, welche auf eben dieser Hauptdiagonalen liegen gleich -1 sind. Das negative Vorzeichen ergibt sich, da anderenfalls die Leichtigkeitsparameter anstatt der Schwierigkeitsparameter ermittelt werden. Beispielhaft ist eine derartige Matrix $\boldsymbol{A}$ für zwei Items mit je drei und vier Kategorien nachstehend abgebildet, ebenso wie der sich für das PCM ergebende $\boldsymbol{C}$-Vektor, dessen Einträge sich prinzipiell aus den Sequenzen aufeinanderfolgender ganzzahliger Werte für jedes der Items zusammensetzt.

$$\boldsymbol{A}_{PCM} = \begin{pmatrix} -1 & & & & & & & \\ -1 & -1 & & & & & 0 & \\ -1 & -1 & -1 & & & & & \\ & & & -1 & & & & \\ & & & -1 & -1 & & & \\ 0 & & & -1 & -1 & -1 & & \\ & & & -1 & -1 & -1 & -1 & \end{pmatrix}, \quad \boldsymbol{C}_{PCM} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

Zu beachten gilt, dass $\boldsymbol{A}_{PCM}$ und $\boldsymbol{C}_{PCM}$ nicht direkt in dieser Form in Gl. (6.10) einsetzbar sind, da hiermit keine eindeutige Identifizierbarkeit des Models gegeben ist. Eine genauere Ausführung

hierzu findet sich in Volodin (2002, S. 6-11). Das GPCM lässt sich durch entsprechende Anpassung von $\boldsymbol{C}$ generieren.

An die Beschreibung der Parameterisierung des PCMs im TAM-Paket anschließend soll nun näher auf das TAM-Paket an sich eingegangen werden, wobei die nachfolgende Beschreibung in Anlehnung an Kiefer (2013) und Kiefer (2016a) geschieht. Zunächst soll kurz auf die im TAM-Paket implementierte Funktion 'tam.jml2()' zur Parameterschätzung des PCMs über die JML-Methode eingegangen werden. Nachfolgend ist wiederum die Funktion mit den (optionalen) Parametern dargestellt. Dieser Methode können prinzipiell relativ viele weitere Werte übergeben werden, die aber hier nicht alle von grundlegender Bedeutung sind. Zur genaueren Vertiefung bietet sich dazu allerdings Kiefer (2016a, S. 88-93) an.

```
tam.jml2(resp, adj = .3, bias = TRUE, pweights = NULL, ...)
```

Mit 'resp' wird eine Matrix mit den Itemantworten und damit wieder eine Matrix mit der Gestalt von $\boldsymbol{X}$ übergeben. Bei 'adj' handelt es sich um eine Anpassungs-Konstante, die zu extremen Scores - also bei perfektem Score oder einem von 0 - subtrahiert bzw. addiert wird. Standardgemäß ist hier der Wert 0.3 festgelegt. Durch Übergabe des logischen Wertes 'bias' kann der Benutzer angeben, ob der JML-Bias durch Multiplikation mit $(I-1)/I$ reduziert werden soll oder nicht. Zugleich bezeichnet 'pweights' einen optionalen Vektor, mittels dessen einzelnen Personen eine gewisse Gewichtung zugeteilt werden kann. Nach korrektem Ausführen der Funktion erhält man eine Liste, die u.a. folgende Einträge beinhaltet: 'item' gibt in Form einer Tabelle die geschätzten τ-Parameter 'xsi', sowie deren Namen 'xsi.label' und den zugehörigen Standardfehler 'xsi.se' wider, während 'theta' die gleichzeitig geschätzten Personenparameter enthält und 'WLE' die unter Berücksichtigung der Anpassungskonstante 'adj' geschätzten Personenparameter.

Nun soll die mittels des TAM-Pakets mögliche MML-Schätzung, die sowohl für das PCM, als auch das GPCM möglich ist, beschrieben werden. Wie auch beim ltm-Paket wird hierbei zur numerischen Integration die Gauß-Quadratur genutzt. Erneut sollen bei der dargestellten Funkion nur auf die wichtigsten zu übergebenden Parameter eingegangen werden, wobei angemerkt sei, dass zahlreiche zusätzliche Einstellungen vorgenommen werden können, die sich im Handbuch zum TAM-Paket nachlesen lassen.

```
tam.mml(resp, irtmodel = "1PL", constraint = "cases", est.variance = FALSE,
        pweights = NULL, item.elim = TRUE, control = list(), ...)
tam.mml2(resp, irt.model = "2PL", ...)
```

Die 'tam.mml()'-Funktion erkennt automatisch, ob die in Form eines 'data.frames' übergebenen Daten vom Partial Credit Format sind oder dichotom und entsprechend die Analyse mittels des RMs bewerkstelligt werden soll. Die Analyse der Daten mittels des PCMs mit einem fixen Diskriminationsparameter erfolgt über die Festsetzung des Parameters 'irtmodel = "1PL". Will man hingegen zulassen, dass für jedes Item ein eigener Steigungsparameter geschätzt wird und infolge dessen das GPCM zur Anwendung bringen, so wird die 'tam.mml.2pl()'-Funktion aufgerufen und 'irtmodel = "2PL" angegeben.Über den Parameter 'constraint' lässt sich wiederum die Normierung festlegen. Mit 'est.variance' kann bestimmt werden, ob die zugehörige Kovarianzmatrix geschätzt

werden soll oder nicht. Mittels der Einstellung 'person.weights' kann optional ein Vektor mit Gewichtungen der einzelnen Personen übergeben werden. Über den logischen Parameter 'item.elim' lässt sich angeben, ob ein Item, welches nur 0-Einträge hat, von der Analyse ausgeschlossen werden soll. Standargemäß ist hier 'TRUE' eingestellt. Zudem kann mittels 'control()' eine Reihe von Argumenten übergeben werden, mit denen sich Einstellungen bezüglich des Algorithmus festlegen lassen. Beispielsweise lassen sich Konvergenzkriterien oder die maximale Anzahl an Iterationen bestimmen.

Nach Anwendung der Funktion 'tam.mml()' bzw. 'tam.mml2()' können eine Vielzahl an Rückgabewerten ausgegeben werden, wobei hier wiederum eine Beschränkung auf die wesentlichsten erfolgt. Die gewichteten Schwellenparameter β_{ik} werden in Form eines Vektors 'xsi' ausgegeben, die wie vorhergehend erläutert den ξ's entsprechen. Unter 'Item Parameters -A*Xsi' finden sich die aufsummierten, gewichteten Schwellenparameter. Gleichzeit kann man auch die ermittelte Kovarianzmatrix kann über 'variance' abrufen. Die Personenparameterschätzungen lassen sich mit 'person' anzeigen. Hier erfolgt die Ausgabe in Form einer Matrix, bei welcher der jeweilige Score des Probanden und der zugehörige Mittelwert der posteriori-Verteilung 'EAP', sowie die Standardabweichung 'SD.EAP' aufgelistet werden. Zudem kann auch die automatisch generierte Design-Matrix 'A' abgerufen werden, ebenso wie die zuvor mit $\boldsymbol{C}$ beschriebene Scoring-Matrix 'B'.

Wiederum kann man sich auch des 'summary()'-Befehls bedienen und so eine übersichtliche Darstellung der Ergebnisse nach Anwendung der 'tam.mml()'- bzw. 'tam.mml2()'-Funktion erhalten. Hierbei wird u.a die Anzahl der Iterationen, die Gesamtzahl der Auswertungspunkte der numerischen Integration, die Log-Likelihood, die Zahl der Personen, sowie Items und hierzu auch die Anzahl der Schwellen- und Steigungsparameter ausgegeben. Ebenso nützlich sind die Angabe des AICs, BICs, sowie des Bias-korrigierten AICs und adjustierten BICs. Des Weiteren können hier auch wieder die einzelnen gewichteten Schwellenparameter β_{ik} oder aber auch das aufsummierte Analogon, also die τ-Parameter, betrachtete werden.

Die angezeigten Standardfehler werden über die nachstehende Funktion berechnet. Hierbei werden die Kovarianzen zwischen den Parameterschätzungen bei der Berechnung der Standardfehler nicht berücksichtigt. Zugleich erfolgt die Ermittlung der Standardfehler über numerische Differenzierung. Bei der Funktion 'tam.se()' wird mit 'tam.obj' dasjenige Objekt übergeben, welches zuvor mittels der Funktion 'tam.mml()' erzeugt wurde. Durch 'numdiff.parm' lässt sich festlegen, mit welcher Schrittweite der Parameter die numerische Differenzierung stattfinden soll, falls dies Standardeinstellung nicht abgeändert wird liegt dies bei 0.001.

```
tam.se(tamobj, numdiff.parm = 0.001, ...)
```

Wiederum gilt es die Personenparameter zu ermitteln. Insofern dies nicht über die EAP-Methode geschehen soll, wobei sich die zugehörigen Parameter dann bereits nach Aufruf der 'tam.mml()'-Funktion abrufen lassen, kann die nachstehende 'IRT.factor.scores()'-Funktion genutzt werden. Hierbei wird mittels 'object' das sich aus nach der konkreten Durchführung der MML-Schätzung ergebende Objekt der Klasse 'tam.mml' übergeben. Mit dem Parameter 'type' lässt sich die gewünschte Methode zur Personenparameterschätzung festlegen. Dabei kann durch entsprechende Angabe die EAP, die MLE- oder WLE-Methode genutzt werden.

```
IRT.factor.scores(object, type= "EAP", ...)
```

Zudem stellt auch dieses Paket die Möglichkeit bereit zwei - hier also über die MML-Methode geschätzte - Modelle auf Basis des Likelihood-Quotienten-Tests zu vergleichen. Somit kann also wiederum getestet werden, ob ein unter der Nullhypothese angenommenes Untermodell, hier entsprechend das PCM, ausreichend gut an die Daten angepasst ist oder ob ggf. das GPCM, durch Berücksichtigung von pro Item variierenden Steigungsparametern, einen signifikanten Unterschied aufweist. Dieser Test wurde bereits beim ltm-Paket in Sektion 6.2 eingehende beschrieben und soll daher nicht weiter ausgeführt werden. Konkret wird dieser Test im TAM-Paket mittels der nachstehenden Funktion ausgeführt. Dabei werden durch den Parameter 'object' zwei Modelle verlangt, die es zu vergleichen gilt, entsprechend wäre es in diesem Fall über die MML-Methode erzeugte Objekte, angepasst für das PCM und GPCM.

```
anova(object, ...)
```

Auf Basis der vermittelten Kenntnisse zu den drei Paketen, mittels derer die Analyse von Daten des Partial Credit Formates möglich ist, sollen im nachfolgenden Kapitel diese bzw. die inbegriffenen Funktionen genutzt werden, um die Daten der s.g. „Freiburger Beschwerdenliste“ zu analysieren.

7 Datenbeispiel

Im letzten Kapitel dieser Arbeit sollen die in der Theorie dargestellten Methoden zum PCM und GPCM noch auf einen Datensatz praktisch angewendet werden bzw. insbesondere die in der vorhergehenden Sektion vorgestellten R-Pakete zur Anwendung gebracht werden. Dazu wird zunächst ein kurzer Überblick zu den vorliegenden Daten der s.g. Freiburger Beschwerdenlist (FBL) gegeben und in diesem Rahmen das Ganze deskriptiv veranschaulicht. Anschließend werden die vorgestellten R-Pakete genutzt, um einen Teil der Daten konkret auszuwerten. Im Wesentlichen sollen hierbei die sich ergebenden Schätzer - ermittelt über die CML- und die MML-Methode - verglichen werden, sowie die konkreten Ergebnisse, die durch Anwendung des PCMs bzw. des GPCMs zutage kommen.

Konkret werden hierzu zwei verschiedene Teildatensätze der FBL genutzt. Einer dieser wird dabei relativ ausführlich analysiert, während die dem anderen Teildatensatz zugehörigen Ergebnisse vorwiegend der vergleichenden Veranschaulichung dienen sollen. Die jeweiligen Fragen, die den beiden Teildatensätzen zugeteilt sind, werden ebenfalls in der nun anschließenden Sektion dargeboten.

7.1 Beschreibung des FBL-Datensatzes und deskriptive Analyse

Im Folgenden wird zunächst der zugrundeliegende Datensatz anhand von Fahrenberg (1994) beschrieben. Hierbei handelt es sich bei dem besagten Datensatz um die Angaben von insgesamt 2070 Personen aus den alten und neuen Bundesländern auf Fragen der s.g. Freiburger Beschwerdenliste, kurz FBL. Dieser Fragebogen wird seit 1975 vielfach genutzt, um insbesondere körperliche Beschwerden bei Jugendlichen - ab 16 Jahren -, sowie Erwachsenen zu erfassen. Hierbei geht es v.a. darum aktuelle, situativ-bedingte und chronisch-habituelle Beschwerden abzufragen. Bis heute wurde die FBL jedoch weiterentwickelt und normiert. Mittlerweile umfasst der Fragebogen 80 Items, der nach zwei Schlüsseln eingeteilt werden kann. Einer davon ist die revidierte Form, welche als FBL-R bezeichnet wird, und besitzt neun Skalen. Diese ergeben sich durch Einteilung in funktionelle Syndrome bzw. Organsysteme und sind nachfolgend aufgelistet: Allgemeinbefinden, Müdigkeit, Herz-Kreislauf, Magen-Darm, Kopf-Hals-Reizsyndrom, Anspannung, Emotionale Reaktivität, Schmerz und Sensorik. Auf Basis dessen lässt sich auch der Summenwert der Items bilden, welcher den Index der körperlichen Beschwerdenhaftigkeit darstellt.

Hierbei liegen bei jedem Item 5 Antwortkategorien, die aufsteigend nummeriert sind mit Zahlenwerten von 1 bis 5 vor. Allgemein bezeichnet ein hoher Zahlenwert eine niedrige Ausprägung, beispielsweise kann Kategorie 5 für die Antwort „selten“ stehen, wohingegen ein niedriger Zahlenwert eine starke oder häufig auftretende Kategorie umfasst.

Der im Rahmen dieser Arbeit analysierte Datensatz mit den Antworten von 2070 Personen stützt sich auf die Einteilung und damit auf die 80 Items des vorgestellten FBL-R-Schlüssels. Hier wird nun das PCM bzw. GPCM im Wesentlichen auf die Skala des Allgemeinbefindens - nachfolgend bezeichnet mit FBL-R-ALL - angewandt. Diese umfasst insgesamt acht Items, welchen die in Tab. 7.1 dargestellten Fragen angehören. Zudem werden auch noch kurz die Ergebnisse der Analyse zur Skala der Emotionalen Reaktivität präsentiert. Die entsprechenden Fragen dieser Skala lassen sich Tab. 7.2 entnehmen.

Item-Nr.	Benennung	Fragestellung
FBLI_8	Kopfschmerzen	„Haben Sie Kopfschmerzen?“
FBLI_9	Appetitmangel	„Haben Sie Appetitmangel?“
FBLI_56	Wetterfühligkeit	„Sind Sie wetterfühlig?“
FBLI_74	Kalte Hände	„Haben Sie selbst bei warmer Witterung kalte Hände?“
FBLI_75	Empfindliche Haut	„Haben Sie empfindliche Haut?“
FBLI_78	Schmerzempfindlichkeit	„Sind Sie schmerzempfindlich?“
FBLI_79	Stress	„Haben Sie das Gefühl im Stress zu sein?“
FBLI_80	Gesundheit	„Haben Sie sich in letzter Zeit Sorgen um Ihre Gesundheit gemacht?“

Tab. 7.1: FBL-R-ALL: Revidierte Form der Freiburger Beschwerdenliste mit den acht Items der Skala Allgemeinbefinden

Item-Nr.	Benennung	Fragestellung
FBLI_57	Aufregung	„Spüren Sie es am ganzen Körper, wenn Sie sich über etwas aufregen?“
FBLI_58	Tränen	„Kommen Ihnen in bestimmten Situationen die Tränen?“
FBLI_59	Stottern	„Kommt es in bestimmten Situationen vor, dass Sie zu stottern beginnen?“
FBLI_60	Erröten	„Erröten Sie?“
FBLI_61	Luft wegbleiben	„Bleibt Ihnen in aufregenden Situationen die Luft weg?“
FBLI_62	Herzklopfen	„Spüren Sie bei Aufregung Herzklopfen?“
FBLI_63	Stuhldrang	„Pflegt sich bei Ihnen in aufregenden Situationen Stuhldrang einzustellen?“
FBLI_64	Weiche Knie	„Beginnen Sie bei Aufregung zu zittern oder bekommen Sie ’weiche Knie’?“

Tab. 7.2: FBL-R-EMO: Revidierte Form der Freiburger Beschwerdenliste mit den acht Items der Skala Emotionale Reaktivität

Wie bereits erwähnt, ist die individuelle Einstufung in fünf Kategorien möglich. Bei Betrachtung der Skala des Allgemeinbefindens liegt den beiden Items 8 und 9 eine andere Einteilung der Kategorien zugrunde als den übrigen. Die unterschiedlichen Ausprägungen, die sich anhand der verschiedenen Kategorien widerspiegeln, sind in Tab. 7.3 aufgelistet, sowie zusätzlich auch diejenigen der FBL-R-EMO-Fragen.

Mit diesem Wissen zu den vorliegenden Daten kann nun das konkrete Vorgehen bei der Analyse der Teildatensätze geschildert werden. So wurden im ersten Schritt zunächst einige Anpassungen vorgenommen. Zum einen wurden die anfänglich gegebenen Kategorien, welche Werte von 1 bis 5 aufwiesen, um jeweils 1 erniedrigt und erstrecken sich damit von 0 bis 4. Diese Veränderung wurde durch die Anforderungen der R-Pakete impliziert, welche eine Nullkategorie verlangen, soll aber

FBL-R-ALL-Items	FBLI_8, FBLI_9	FBLI_56, FBLI_74, FBLI_75, FBLI_78, FBLI_79, FBLI_80
FBL-R-ALL-Items		FBLI_57, FBLI_58, FBLI_59, FBLI_60, FBLI_61, FBLI_62, FBLI_63, FBLI_64
Kategorie 1	„fast täglich“	„sehr stark“
Kategorie 2	„etwa 3-mal pro Woche“	„stark“
Kategorie 3	„etwa 2-mal pro Monat“	„mittel“
Kategorie 4	„etwa 2-mal pro Jahr“	„kaum“
Kategorie 5	„praktisch nie“	„praktisch nie“

Tab. 7.3: Bedeutung der Kategorien von FBL-R-ALL und FBL-R-EMO zugehörigen Items

hier nun nicht weiter beachtet werden. Zum anderen wurden alle diejenigen Personen entfernt, für die bei einem der FBL-R-Items keine Angabe vermerkt ist. Dies begründet sich darin, dass die verschiedenen R-Pakete - wie in Kapitel 6 dargelegt - unterschiedliches Vorgehen beim Umgang mit diesen fehlenden Werten aufweisen. Damit reduziert sich der zu analysierende Datensatz von 2070 Personen auf 2035 für die FBL-R-ALL-Fragen und auf 2032 Personen für die FBL-R-EMO-Fragen. Nun folgend soll zunächst kurz deskriptiv auf die zugrundeliegende Datensituation eingegangen werden.

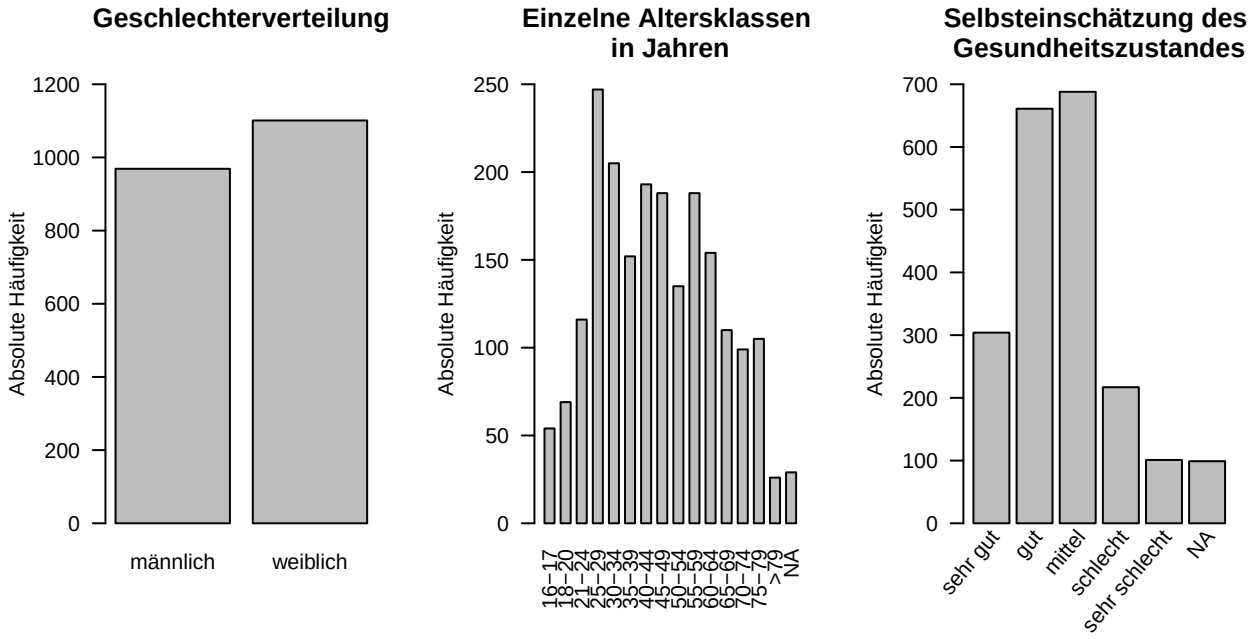

Abb. 7.1: Darstellung der absoluten Häufigkeiten (bearbeiteter Datensatz): Geschlechterverteilung, Altersklassen und Selbsteinschätzung des eigenen Gesundheitszustandes

Neben den Ergebnissen bei der Abfragung der Beschwerden umfasst der FBL-R-Datensatz noch eine Reihe weiterer Informationen. So sind u.a. auch das Geschlecht, die Altersgruppe, der Familienstand, Hausarzt-, Facharzt- und Zahnarztbesuche, die Medikamenteneinnahme von Schlaf-/Be-

ruhigungsmitteln, anderen Medikamenten, sowie homöopathischen Präparaten und die Zufriedenheit etwa mit der finanziellen oder gesundheitlichen Situation inbegriffen. Auch wurde beispielsweise abgefragt, ob man sich in psychotherapeutischer Behandlung befindet, an einer chronischen Krankheit leidet oder etwa wie man den eigenen gesundheitlichen Zustand bewertet.

Um einen groben Überblick über die vorliegenden Daten zu vermitteln, sind in Abbildung 7.1 Säulendiagramme dargestellt. So wird mittels der ersten Grafik links ersichtlich, dass beim bearbeitete Datensatz der Frauenanteil überwiegt, wobei den Frauen ein absoluter Wert von 1101 und den Männern einer von 969 zukommt. In der nächsten Grafik bezüglich der Aufteilung befragter Personen auf die einzelnen Altersgruppen ist kein klares Muster zu erkennen. Die Verteilung auf die einzelnen Gruppen ist eher linkssteil und multimodal, wobei ein relativ hoher Anteil der Befragten in die Altersklasse von 25 bis 29 Jahren fällt. Gleichzeitig sollte beachtet werden, dass die Klassenbreite der ersten beiden Gruppen sich von den übrigen unterscheidet. Beispielhaft sind in der Grafik rechts noch die absoluten Häufigkeiten bezüglich Selbsteinschätzung des eigenen Gesundheitszustandes dargestellt. Demzufolge sehen die meisten der Befragten ihren Gesundheitszustand als „mittel“ bis „gut“ an.

Um die deskriptive Analyse abzurunden sind nachfolgend die absoluten Häufigkeiten der angegebenen Kategorien pro Item, welche sich auf das Allgemeinbefinden und die Emotionale Reaktivität beziehen, in Abb. 7.2 und 7.3 abgetragen. Besonders auffallend bei ersterer sind hierbei FBLI_9 und FBLI_74 mit einem relativ hohen Anteil an Personen, die sich Kategorie 1 zugeordnet haben. Demzufolge verzeichnen 1548 Personen „fast täglich“ Appetitmangel zu haben und so vermerken zudem 1050 Befragte selbst bei warmer Witterung kalte Hände zu haben. Für die anderen Items ist im weitestgehenden Sinne keine derartige Präferenz einer bestimmten Kategorie zu erkennen.

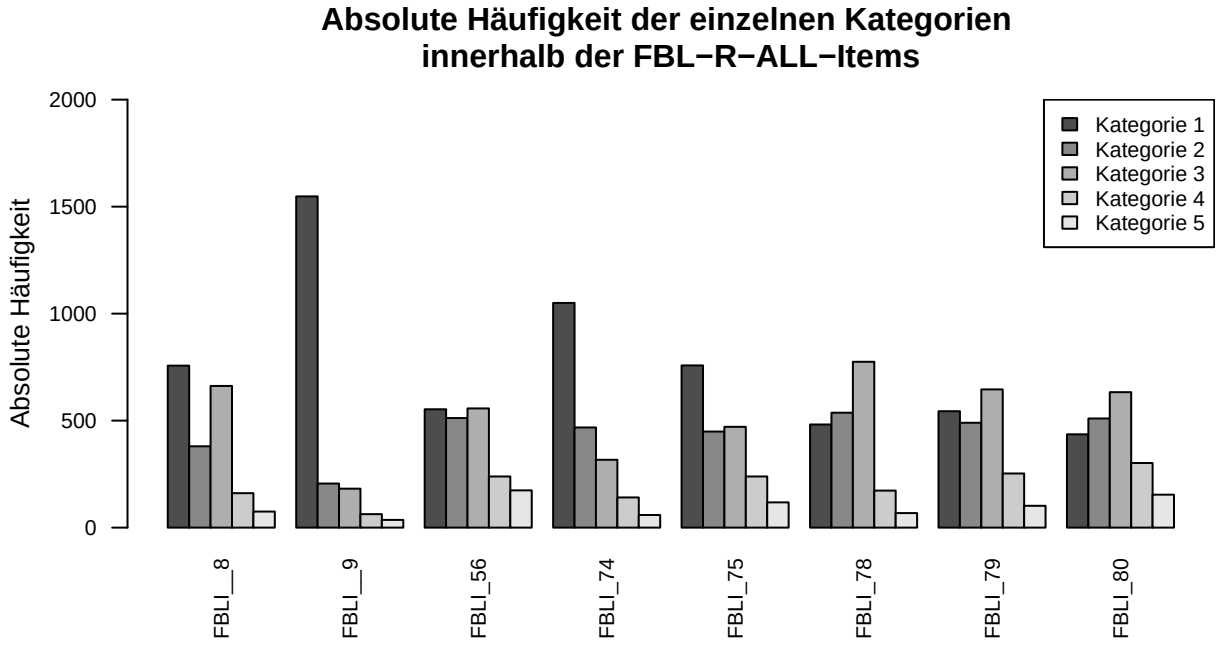

Abb. 7.2: Darstellung der absoluten Häufigkeiten (bearbeiteter Datensatz): ausgewählte Kategorien innerhalb der FBL-R-ALL-Items

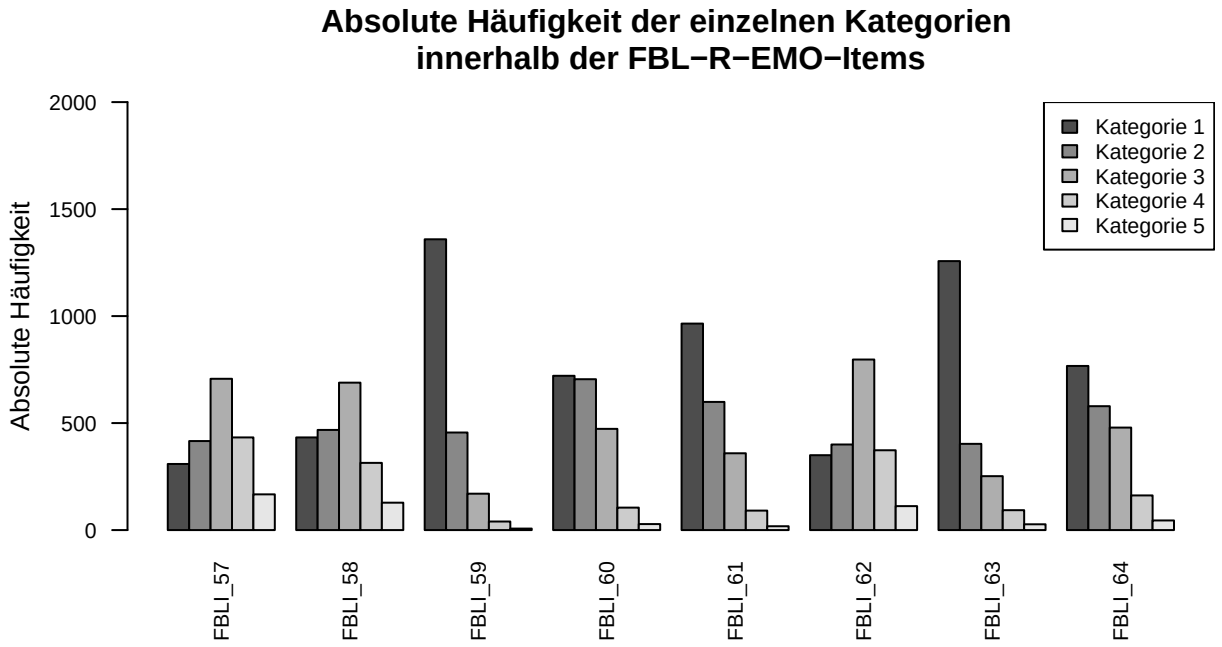

Abb. 7.3: Darstellung der absoluten Häufigkeiten (bearbeiteter Datensatz): ausgewählte Kategorien innerhalb der FBL-R-EMO-Items

Dagegen stechen in Abb. (7.3) die ersten Kategorien von Item FBLI_59 und FBLI_63 besonders hervor. Demgemäß gaben 1359 Personen an, dass sie in bestimmten Situationen sehr stark zu stottern beginnen, und 1257 Personen verzeichnen demnach sehr starken Stuhldrang, sobald sie in aufregende Situationen kommen. Gerade bei FBLI_57, FBLI_58 und FBLI_62 zeigt sich eine hohe Präferenz der mittleren Kategorie. Insbesondere die Items FBLI_59, FBLI_60, FBLI1, FBLI_63 und FBLI_64 weisen eine sehr geringe Zahl an Befragten auf, die sich der letzten Kategorie zuordnen und damit „praktisch nie“ bei den jeweiligen Fragen angeben würden.

7.2 Auswertung des FBL-R-Datensatzes

An diesen Überblick anschließend sollen im Folgenden nun die im vorherigen Kapitel 6 präsentierten R-Pakete konkret zur Analyse der FBL-R-Daten verwendet werden. Hierzu wird sowohl das PCM, als auch das GPCM angewandt. Insbesondere sollen zur Parameterschätzung die CML- und die MML-Methode genutzt werden, dagegen wird die JML-Methode aufgrund der in Sektion 5.1 dargelegten Nachteile ungeachtet bleiben. Generell sei noch einmal darauf hingewiesen, dass die CML-Schätzung einzig mittels des eRm-Paketes möglich ist und dabei nur die Modellierung des PCMs mit einem konstanten Steigungsparameter von $\alpha = 1$ für alle Items bereitgestellt wird. Allerdings bieten das ltm- und TAM-Paket eine Bewerkstelligung der Analyse mittels des PCMs, bei welchem auch ein angepasster α-Parameter zulässig ist, und zusätzlich des GPCMs, für welches jeweils mittels des beschriebenen EM-Algorithmus auch die α_i-Parameter geschätzt werden. Diese weniger restriktive Form des PCMs kommt durch die Annahme bezüglich der Verteilung der

Personenparameter zutage, welche ausführlich in Sektion 6.2 dargestellt wurde.

Entsprechend dem jeweils zugrundeliegenden Modell liefert die MML-Schätzung für das ltm- und das TAM-Paket - bis auf die zweite Nachkommastelle genau - dieselben Schätzwerte. Dies lässt sich dadurch begründen, dass in den beiden Paketen zwar jeweils unterschiedliche Parameterisierungen der Modelle vorliegen, diese aber beide auf die allgemeine Modellformel zurückführbar sind - wie es v.a. für das PCM in dem vorhergehenden Kapitel gezeigt wurde -, und insbesondere auch dieselbe Annahme bzgl. der Verteilung der Fähigkeitsparameter getroffen wird, sowie derselben Algorithmus zur Bestimmung der Schwellenparameter genutzt wird. So kann in beiden Paketen ausgewählt werden, ob die restriktivste Form des PCMs verwendet werden soll, indem man die Standardnormalverteilung der Personenparameter voraussetzt, also $\theta_p \sim N(0,1)$. Gleichzeitig besteht aber auch die Möglichkeit die weniger restriktive Form des PCMs zur Anwendung zu bringen, wobei die Personenparameter als normalverteilt mit Mittelwert $\mu = 0$ und Varianz σ^2 angenommen werden. Wie in Kapitel 6.2 gezeigt wurde, ist letzteres äquivalent zur Annahme standardnormalverteilter Personenparameter und eines geschätzten, aber für alle Items konstanten α-Parameters. Diese weniger restriktive Form des PCMs wird im Folgenden stets bei der Anwendung der MML-Methode genutzt.

Auch wenn dies im vorherigen Kapitel nicht derartig ausführlich behandelt wurde, lässt sich anfügen, dass die aus der MML-Methode resultierenden Schätzungen des GPCMs für das ltm- und TAM-Paket ebenfalls nahezu identische Schätzwerte liefert. Wobei dies in analoger Weise wie beim PCM begründbar ist.

In der ersten nachfolgenden Sektion 7.2.1 sollen die Äquivalenz der CML- und MML-Schätzung anhand der FBL-R-ALL-Daten aufgezeigt werden, ebenso werden hierzu die konkret errechneten Schätzwerte dargeboten und im vorherigen Kapitel 6 vorgestellte Tests durchgeführt. Zum Vergleich wird in der nachfolgenden Sektion 7.2.2 eine deutlich kürzer gehaltene Auswertung der Skala FBL-R-ALL-EMO präsentiert.

Auf die Analyse der Personenparameter soll hier weitestgehend verzichtet werden. Allerdings sollte festgehalten werden, dass in diesem Sinne nun nicht mehr von Fähigkeitsparametern gesprochen werden kann, sondern hier durch die θ-Parameter einzig das Allgemeinbefinden bzw. die emotionale Reaktivität der im Datensatz inbegriffenen Personen beschrieben werden sollte.

7.2.1 Analyse des Allgemeinbefindens (FBL-R-ALL)

Zunächst soll nun anhand dieser Auswertung der in Sektion 5.3 erwähnte Sachverhalt der asymptotischen Äquivalenz von CML- und MML-Schätzungen der Itemparameter veranschaulicht werden. Um dies darzustellen, wird auf die aufsummierten Schwellenparameter τ_{ik} Bezug genommen, allerdings würde man mittels der β-Parameter zu demselben Ergebnis gelangen. Zugleich sei darauf hingewiesen, dass die asymptotische Äquivalenz unter der angenommenen Nebenbedingung einer korrekt spezifizierten Verteilung der Personenparameter bei der MML-Schätzung zu zeigen ist. Insgesamt muss für diese Äquivalenz also gelten, dass für eine gegen ∞ strebende Zahl an Beobachtungen der Quotient der jeweiligen Parameterschätzer der CML- und MML-Methode gegen 1 geht. Konkret wird nun zur Analyse das PCM auf die FBL-R-ALL-Daten angewandt.

Grundsätzlich gilt Folgendes: will man von der weniger restriktiven Form mit $\alpha_i = \alpha = \sigma$, welche

bei den jeweiligen MML-Schätzungen angenommen wird, auf die restriktive Form, welche bei der CML-Methode vorausgesetzt wird, rückschließen, so muss man gemäß Gl. (6.4) die über die MML-Methode geschätzten Schwellenparameter mit dem entsprechenden α_{MML}-Parameter gewichten. Damit lässen sich die Schätzwerte der restriktiven und der weniger restriktiven Form des PCMs ineinander überführen.

Unter Berücksichtigung dessen gilt es im vorliegenden Fall die geschätzten τ-Parameter mit dem Steigungsparameter von $\hat{\alpha}_{MML} = 0.614$ zu gewichten. Damit ergeben sich die in nachfolgender Tabelle 7.4 dargestellten Schätzwerte, die jeweils aus der CML-, welche mit dem eRm-Paket bewerkstelligt wurde, und MML-Methode, welche mit dem ltm-Paket durchgeführt wurde, hervorgehen. Die angegebenen Werte sind auf die dritte Nachkommastelle genau gerundet und die summierten Schwellenparameter der MML-Methode sind um den entsprechenden α-Parameter angepasst. Die nachfolgend dargestellten $\hat{\tau}_{ij,MML}$ bezeichnen also die gewichteten, aufsummierten Schwellenparameter. Gleichzeitig kennzeichnet die erste Schwelle den Übergang von Kategorie 0 zu 1, usw.

	PCM							
	Schwelle 1		Schwelle 2		Schwelle 3		Schwelle 4	
Item i	$\hat{\tau}_{i1,CML}$	$\hat{\tau}_{i1,MML}$	$\hat{\tau}_{i2,CML}$	$\hat{\tau}_{i2,MML}$	$\hat{\tau}_{i3,CML}$	$\hat{\tau}_{i3,MML}$	$\hat{\tau}_{i4,CML}$	$\hat{\tau}_{i4,MML}$
FBLI_8	-0.017	0.499	-0.957	0.036	0.298	1.800	1.094	3.151
FBLI_9	1.551	2.051	1.488	2.493	2.567	4.120	3.320	5.467
FBLI_56	-0.717	-0.183	-1.253	-0.247	-0.629	0.874	-0.340	1.699
FBLI_74	0.219	0.721	0.314	1.304	1.042	2.555	2.012	4.094
FBLI_75	-0.185	0.329	-0.626	0.363	-0.121	1.370	0.597	2.634
FBLI_78	-0.914	-0.369	-1.721	-0.699	-0.417	1.111	0.526	2.607
FBLI_79	-0.689	-0.153	-1.410	-0.400	-0.681	0.830	0.217	2.271
FBLI_80	-1.015	-0.461	-1.721	-0.694	-1.229	0.290	-0.602	1.452

Tab. 7.4: Summierte Schwellenparameter τ_{ik} der FBL-R-ALL-Items (gewichtet: α-Parameter eingerechnet) geschätzt über CML- und MML-Methode

Anhand dieser Tabelle kann bereits eine erste Feststellung gemacht werden: so scheinen sich für die Abstände der summierten Schwellenparameter - geschätzt über die CML- und der MML-Methode - jeweils sehr ähnliche Werte zu ergeben. Insgesamt ist es also naheliegend zu vermuten, dass die Parameterschätzungen der CML-Methode durch lineare Transformation aus der MML-Methode hervorgehen bzw. umgekehrt. Dies lässt sich grafisch bestätigen, indem man die jeweiligen Schätzer - resultierend aus der CML- und MML-Methodik - gegeneinander abträgt. Derartige Scatterplots sind - aufgeteilt nach den vier Schwellen - in nachstehender Abb. 7.4 dargestellt.

Die gegeneinander abgetragene Parameterschätzer liegen auf einer Geraden, was bedeutet, dass diese asymptotisch äquivalent sind bis auf eine im PCM zulässige, lineare Transformation. Diese Transformation lässt sich durch die Verschiebung des Nullpunktes, welcher durch die unterschiedliche Normierung der Parameter in den zwei Paketen zutage kommt, begründen. Vergleicht man

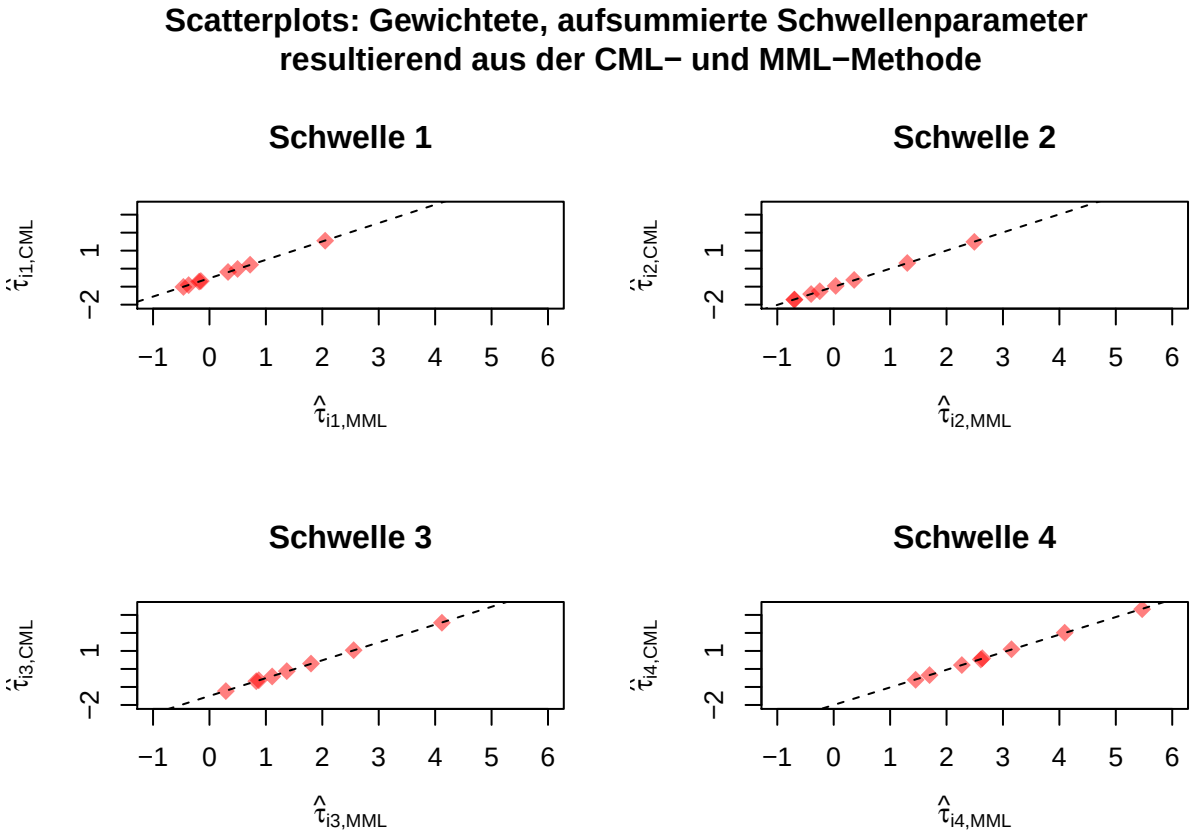

Abb. 7.4: Scatterplots: Abtragung der aus CML- und MML-Methode resultierenden (gewichteten) τ-Parameter der FBL-R-ALL-Items gegeneinander, aufgeteilt nach den einzelnen vier Schwellen

nun also explizit das eRm-Paket, welches die CML-Methode zur Parameterschätzung nutzt, mit dem ltm- oder TAM-Paket, in denen jeweils die MML-Schätzung zur Anwendung kommt, so muss noch eine weitere Tatsache berücksichtigt werden: es liegen jeweils unterschiedliche Konventionen bezüglich der Restriktion zur Identifizierbarkeit des Modells vor. Hierbei gilt im eRm-Paket, dass die Summe aller betrachteten τ-Parameter 0 ergeben muss, womit also gilt $\sum_i \sum_j \hat{\tau}_{ij,CML} = 0$. Beim ltm- und TAM-Paket geschieht die Restriktion hingegen über die Annahme bezüglich der Personenparameter, für die hier nun $\theta_p \sim N(0, \sigma^2)$ gilt. Letztlich ergibt sich bei diesen beiden Paketen dann also beim Aufsummieren aller τ-Parameter ein unbekannter Wert $x_{0,MML}$, genauer gesagt gilt dementsprechend $\sum_i \sum_j \hat{\tau}_{ij,MML} = x_{0,MML}$.

Unter Berücksichtigung aller zuvor genannten Aspekte sollte bei der Anwendung der jeweiligen Pakete gelten, dass die CML- und die MML-Methode - von Schätzungenauigkeiten abgesehen - äquivalente Schätzungen liefern, die durch folgende Umrechnung ineinander überführbar sind:

$$\hat{\tau}_{ij,CML} \approx \hat{\tau}_{ij,MML} \cdot \hat{\alpha}_{MML} - x_{0,MML} \tag{7.1}$$

Letztlich kann hiermit auf die asymptotische Äquvalenz der Schätzer geschlossen werden, womit schließlich - wie zuvor gefordert - der Quotient aus den jeweiligen Schätzern gegen 1 strebt.

Nun folgend sollen auch die Schwellenparameter der einzelnen Kategorien und damit explizit die geschätzten β-Parameter dargestellt werden, um einen Eindruck von dem relativen „Schwierigkeitsgrad“ zu bekommen. Zur konkreten Veranschaulichung der geschätzten Schwellenparameter kann Abb. 7.5 betrachtet werden. Hier sind die aus der MML-Schätzung resultierenden Schwellenparameter des FBL-R-ALL-Items mitsamt der zugehörigen Standardfehler dargestellt. Demgemäß würden sich also die CML-Schätzer durch Berücksichtigung des Abstandes $x_{0,MML}$ bzw. durch Angleichen des Nullpunktes ergeben, womit sich also lediglich die Position der Schätzwerte auf der y-Achse ändert. In Abb. 7.5 sind nun die Schwellenparameterwerte an sich farblich gekennzeichnet, wobei ein hellerer Blauton eine verhältnismäßig „schwierigere“ Stufe kennzeichnet, dunklere Blautöne repräsentieren entsprechend „leichtere“ Schwellen. Zudem zeigen sich für einige Kategorienübergänge größere Unsicherheiten - ersichtlich anhand des Konfindenzintervalles - als für andere. Dies lässt sich durch die jeweils gegebene Anzahl an Beobachtungen erklären, wozu man die untere Grafik von Abb. 7.2 heranziehen kann. Demgemäß ergeben sich für Kategorien mit relativ gesehen wenig Beobachtungen größere Unsicherheiten in der Schätzung der Schwierigkeitsparameter.

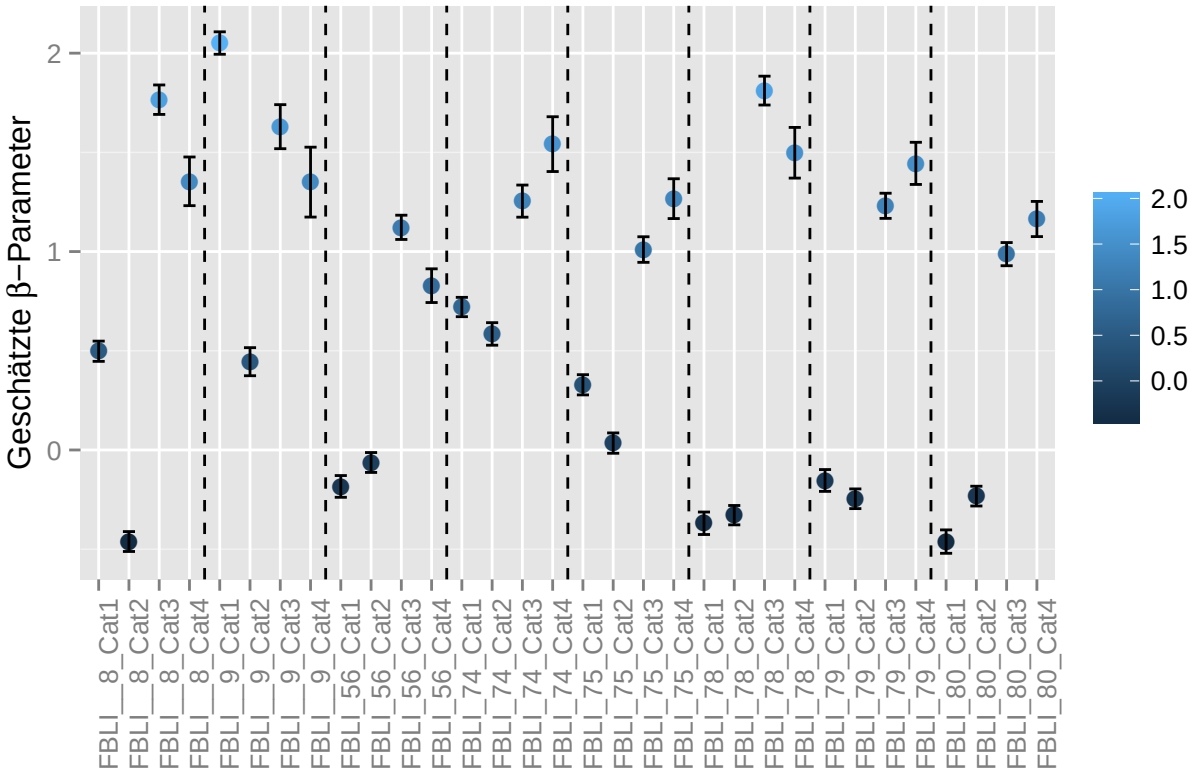

Abb. 7.5: Konkrete Darstellung der aus der MML-Methode resultierenden (gewichteten) Schwellparameterwerte der FBL-R-ALL-Items mit zugehöriger Standardabweichung

Hier sollte nun zunächst Folgendes festgehalten werden: die aus der MML-Methode resultierenden Schätzungen für das PCM, sowie auch die des GPCMs, ergeben durch Anwendung des ltm- und des TAM-Paketes - abgesehen von Schätzungenauigkeiten - die gleichen Werte. Gleichzeitig sind die Schätzer der MML-Methode und die der CML-Methode für das PCM durch entsprechende Anpas-

sung ineinander überführbar und damit unter Berücksichtigung der jeweiligen Zentrierung nahezu identisch. Damit sollten die nachfolgenden durchgeführten Tests, welche in den verschiedenen Paketen implementiert sind, auch zu ähnlichen Ergebnissen kommen und auch die jeweiligen grafischen Darstellungen entsprechend aufeinander anpassbar sein. Zunächst sollen nun anschließend auch die aus der Anwendung des GPCMs resultierenden Schätzungen präsentiert werden.

Basierend auf der MML-Methode werden die Schätzwerte, welche aus der Analyse der Daten mittels des GPCMs hervorgehen, berechnet. Um diese Werte nun einmal konkret vorzustellen, ist die nachfolgende Tab. 7.5 dargeboten, bei der die Schwellenparameter β_{ik} präsentiert werden. Zugleich sind für das PCM die Schätzwerte resultierend aus der weniger restriktiven Form und folglich mit einem für alle Items konstanten Steigungsparameter angegeben. Konkret sind die Schwellenparameter in ihrer ungewichteten Form dargestellt, d.h. der jeweilige Diskriminationsparameter ist nicht eingerechnet, sondern der entsprechenden mit „α“ bzw. „α_i“ gekennzeichneten Spalte zu entnehmen.

	PCM					**GPCM**				
Item	β_{i1}	β_{i2}	β_{i3}	β_{i4}	α	β_{i1}	β_{i2}	β_{i3}	β_{i4}	α_i
FBLI_8	0.812	-0.754	2.873	2.201	0.614	0.796	-0.744	2.849	2.196	0.621
FBLI_9	3.340	0.720	2.650	2.194	0.614	3.398	0.718	2.675	2.208	0.603
FBLI_56	-0.298	-0.104	1.825	1.344	0.614	-0.316	-0.097	1.766	1.342	0.646
FBLI_74	1.174	0.950	2.038	2.506	0.614	1.032	0.907	1.948	2.426	0.683
FBLI_75	0.536	0.055	1.641	2.058	0.614	0.719	0.022	1.795	2.194	0.518
FBLI_78	-0.601	-0.537	2.948	2.436	0.614	-0.624	-0.402	2.506	2.245	0.776
FBLI_79	-0.249	-0.402	2.002	2.348	0.614	-0.039	-0.688	2.778	3.004	0.386
FBLI_80	-0.751	-0.380	1.604	1.892	0.614	-0.763	-0.295	1.370	1.741	0.812

Tab. 7.5: Schwellenparameter β_{ik} der FBL-R-ALL-Items zu PCM und GPCM (ungewichtet: α- bzw. α_i-Parameter nicht eingerechnet) geschätzt über MML-Methode

Zu beachten ist, dass die Anwendung des GPCMs auf die FBL-R-ALL-Daten für alle inbegriffenen Items einen Diskriminationsparameter $\alpha_i < 1$ hervorbringt. Bezogen auf die Trennschärfe bedeutet dies, dass bei der grafischen Darstellung anhand der Kategorienwahrscheinlichkeit die Kurven relativ flach verlaufen, die Steigung im mittleren Bereich der Kurve also eher gering ausfällt und sich somit die Prognose der latenten Variable anhand der Items eher schwierig gestaltet. Während die Diskriminationsparameter der ersten vier Items - also von FBLI_8, FBLI_9, FBLI_56 und FBLI_74 - relativ ähnlich zu dem konstanten Steigungsparamter $\alpha_{PCM} = 0.614$ sind, unterscheiden die der anderen vier Items sich deutlich stärker von diesem. Gerade die Steigungsparameter der Items FBLI_78 und FBLI_80 weisen einen wesentlich höheren Wert auf, während der des Items FBLI_79 äußerst niedrig angesiedelt ist. Da sich die Unterschiede in den Schätzungen - hervorgehend aus dem PCM und dem GPCM - besonders gut grafisch darstellen lassen, sind die nachfolgenden Abb. 7.6 und 7.7 dargeboten, wobei eben gerade für die ersten vier Items kaum Unterschiede im Anstieg der Kurven zu erkennen sind.

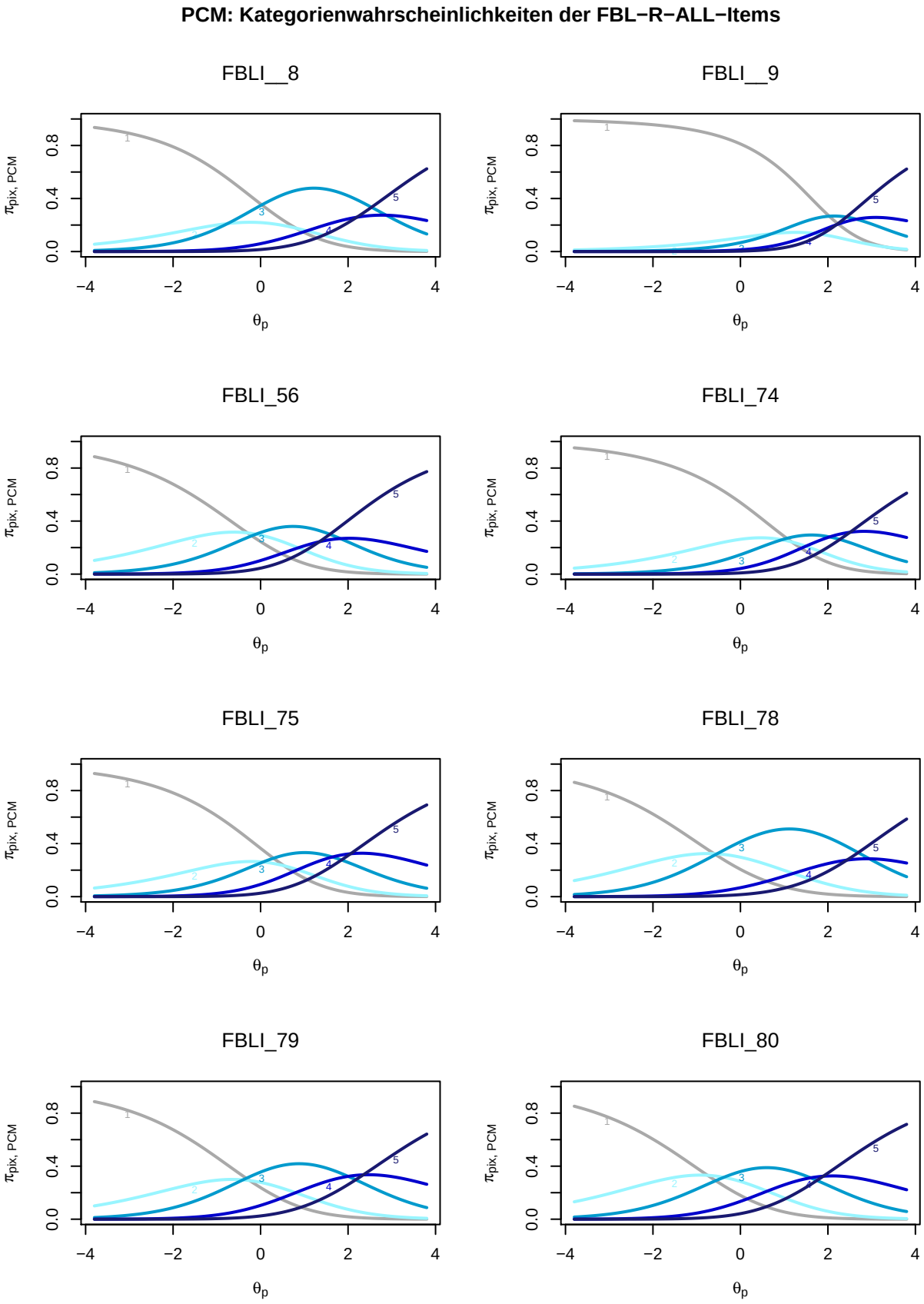

Abb. 7.6: Grafische Veranschaulichung des PCMs (konstanter α-Parameter) angewandt auf die FBL-R-ALL-Daten anhand der Kategorienwahrscheinlichkeiten

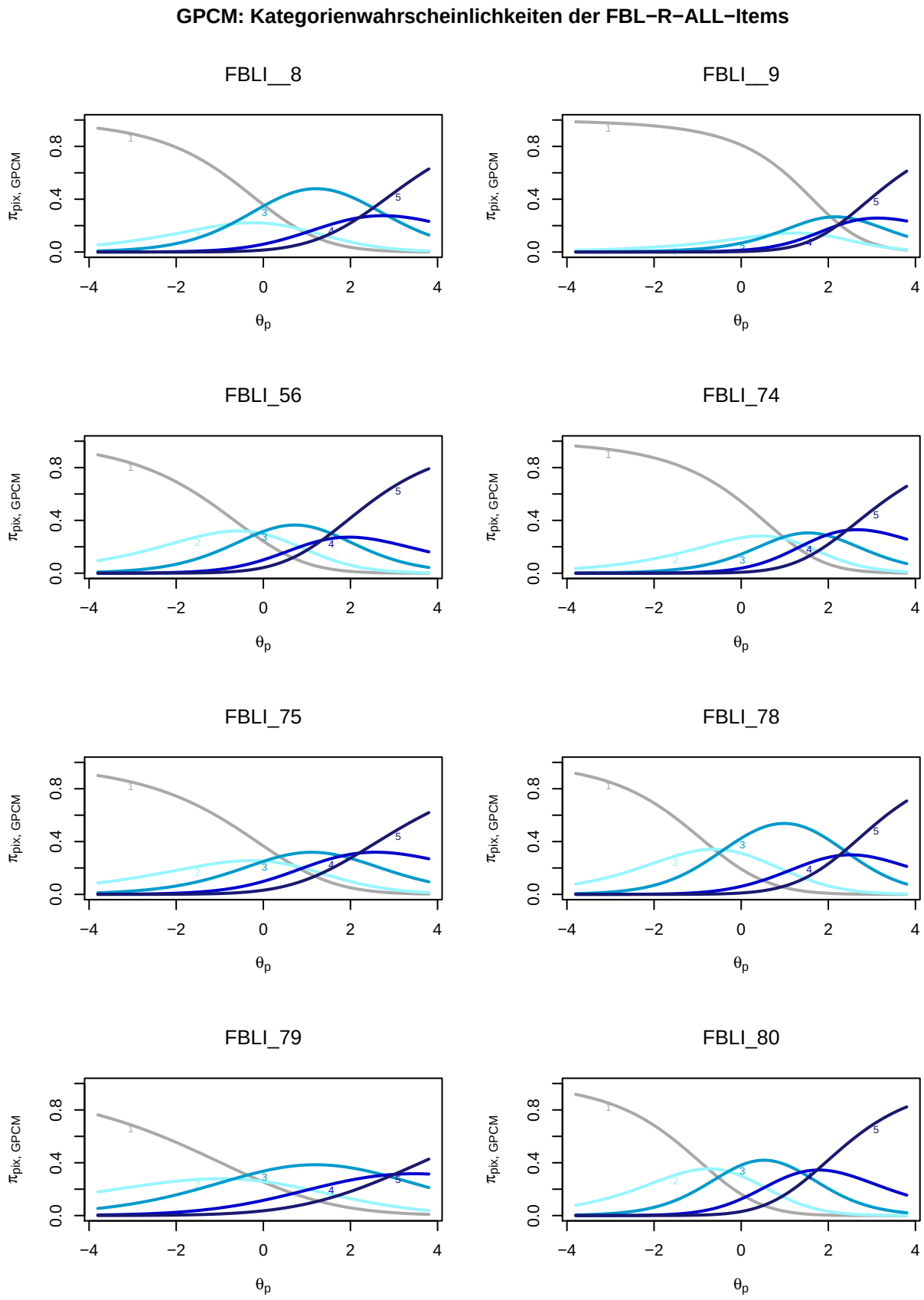

Abb. 7.7: Grafische Veranschaulichung des GPCMs (α_i-Parameter) angewandt auf die FBL-R-ALL-Daten anhand der Kategorienwahrscheinlichkeiten

Anhand von Tab. 7.5, sowie den vorhergehenden Abb. 7.6 und 7.7 sollten einige weitere Dinge ersichtlich werden, die bereits in den vorherigen Kapiteln angesprochen wurden. Einerseits sollte somit gelten, dass - wie in 4.5 erläutert - durch Anwendung des GPCMs lediglich der Diskriminationsparameter α_i hinzukommen sollte, sich aber die Lokalisation der Schwellenparameter auf Skala der latenten Variable und infolgedessen die β_{ik}'s nicht ändern sollten. Dies ist weitestgehend auch erfüllt, allerdings zeigt sich gerade für das Item FBLI_79 ein größerer Unterschied zwischen den $\beta_{ik,PCM}$ und $\beta_{ik,GPCM}$, was wahrscheinlich im hohen Maße auf die niedrige Trennschärfe des Items zurückzuführen ist.

Zudem ist zu vermerken, dass relativ viele ungeordnete Schwellenparameter auftreten. Somit weist der β-Parameter zweier benachbarter Kategorien für die entsprechend höhere Kategorie β_{ix} einen niedrigeren Wert auf als der Schwellenparameter der darunter liegenden Kategorie $\beta_{i(x-1)}$. Grafisch drückt sich dies - wie in Kapitel 4.3.2 beschrieben - dann darin aus, dass eine derartige Kategorie x „gemieden“ wird und dementsprechend die Kategoriencharakteristik von $x-1$ in $x+1$ übergeht, d.h. ein Score von x ist nie wahrscheinlicher als das Erreichen der übrigen Scores. Unter Beachtung dieses Aspekts lässt sich die Frage aufstellen, ob es bei einigen der Items nicht erforderlich wäre die „gemiedenen“ Kategorie zu entfernen bzw. ob es nicht sinnvoll wäre einige der Kategorien zusammenzufassen.

Die bisherigen Analysen werfen zudem die Fragen auf, welches Modell - also das PCM oder GPCM - besser zur Analyse der vorliegenden FBL-R-ALL-Daten geeignet ist und zudem, ob die beiden Modelle überhaupt ausreichend gut an die Daten angepasst sind. Um dies zu klären, werden nachfolgend die in Kapitel 6 vorgestellten Funktionen der jeweiligen R-Pakete genutzt. Zunächst soll der Frage nachgegangen werden, welches der beiden Modelle - also PCM oder GPCM - die vorliegenden Daten besser beschreibt bzw. ob es einen signifikanten Unterschied macht, eine für jedes Item angepassten α_i-Parameter zuzulassen. Um dies zu testen kann man sich der im ltm- und TAM-Paket implementierten 'anova()'-Funktion bedienen. Hierzu sei zunächst auf das jeweilige AIC des PCMs und GPCMs hingewiesen: für das PCM ergibt sich - basierend auf der Analyse mittels des ltm-Paketes - 'AIC = 41997.79', während man für das GPCM 'AIC = 41921.84' erhält. Bereits an dem niedriger angesiedelten Wert des AIC's für das GPCM könnte die Vermutung aufkommen, dass ggf. ein signifikanter Unterschied zwischen der Anwendung des PCMs und des GPCMs bei der Anwendung auf die gegebenen Daten vorliegt. Führt man also die 'anova()'-Funktion - wobei hier der ausgegebene p-Wert auf Basis von 'B = 200' Boostrapstichproben angenähert wird, mit „seed = 123“ - konkret durch, so ergibt sich als Wert der LQ-Teststatistik 89.95 und schließlich zum Signifikanzniveau von 0.01, dass die Nullhypothese verworfen werden muss. Demgemäß sind die Modellanpassungen von PCM und GPCM nicht äquivalent und damit die Verwendung des GPCM bzw. die Berücksichtigung des α_i-Parameters zur Beschreibung der Daten signifikant besser geeignet.

Konkret lässt sich aber auch testen, wie gut die Modelle einzeln betrachtet zu den Daten passen. Diese Beurteilung geschieht über das Testen der Anpassungsgüte, beispielsweise mithilfe der im ltm-Paket zur Verfügung gestellten 'GoF.gpcm()'-Funktion. Hier wird also basierend auf der parametrischen Bootstrap-Approximation Pearson's χ^2-Test verwendet - wie er in Sektion 6.2 beschrieben ist -, um die Anpassungsgüte des jeweils betrachteten Modells zu beurteilen. Beide Male wurde

hier bei der konkreten Durchführung mittels der im ltm-Paket integrierten 'GoF.gpcm()'-Funktion die Standardeinstellung von 'B = 99' Bootstrap-Stichproben beibehalten und der Reproduzierbarkeit halber 'seed = 123' gewählt. Bei der Anwendung dieser Funktion auf das PCM-Objekt ergab sich 'Tobs = 823326' als Wert der Teststatistik. Bei Betrachtung des p-Wertes von 0.01 muss dabei die Nullhypothese zu einem Signifikanzniveau von 0.05 verworfen werden und damit die Annahme, dass das PCM ausreichend gut an die vorliegenden Daten angepasst ist. Obwohl das GPCM - wie im vorhergehenden Absatz dargestellt - dem PCM vorgezogen wird, muss auch hier die Feststellung gemacht werden, dass die Nullhypothese aufgrund eines p-Wertes von 0.01 verworfen wird, wobei die Teststatistik einen Wert von 'Tobs = 920313' aufweist. Diese Ergebnisse machen die Anwendung des PCMs, als auch des GPCMs zur Analyse der FBL-R-ALL-Daten fraglich.

Damit sich die Analyse der FBL-Daten nicht nur auf einen Teildatensatz beschränkt, werden in der nachfolgenden Sektion noch kurz die Ergebnisse der Analyse mittels des PCMs und GPCMs bezüglich der Emotionalen Reaktivität vorgestellt.

7.2.2 Analyse der Emotionalen Reaktivität (FBL-R-EMO)

Abschließend sollen zusammengefasst die Auswertungen der FBL-R-EMO-Daten in ähnlicher Weise wie in der vorhergehenden Sektion dargestellt werden. Um die konkrete Fragestellung, welche hinter den inbegriffenen Items steckt, nachvollziehen zu können, sei auf Tab. 7.2 zu Beginn des Kapitels verwiesen. Zunächst sollen die sich für die Items der Emotionalen Reaktivitätsskala ergebenden Schwellenparameter β_{ik} dargeboten werden. Diese sind wiederum in ungewichteter Form für das PCM und das GPCM in Tab. 7.6 dargestellt.

	PCM					**GPCM**				
Item	β_{i1}	β_{i2}	β_{i3}	β_{i4}	α	β_{i1}	β_{i2}	β_{i3}	β_{i4}	α_i
FBLI_57	-1.087	-0.760	0.912	1.911	0.886	-1.079	-0.746	0.888	1.894	0.921
FBLI_58	-0.689	-0.458	1.367	1.952	0.886	-0.665	-0.526	1.454	2.053	0.782
FBLI_59	1.293	1.771	2.782	3.568	0.886	1.908	2.198	3.417	4.334	0.582
FBLI_60	-0.259	0.766	2.518	2.773	0.886	-0.186	0.937	3.311	3.383	0.565
FBLI_61	0.385	1.010	2.469	3.197	0.886	0.258	0.936	2.254	3.004	1.121
FBLI_62	-0.850	-0.882	1.280	2.265	0.886	-0.907	-0.682	1.037	1.967	1.379
FBLI_63	1.253	1.062	2.112	2.795	0.886	1.873	1.203	2.498	3.286	0.594
FBLI_64	0.020	0.491	1.982	2.645	0.886	-0.171	0.465	1.602	2.297	1.600

Tab. 7.6: Schwellenparameter β_{ik} der FBL-R-EMO-Items zu PCM und GPCM (ungewichtet: α- bzw. α_i-Parameter nicht eingerechnet) geschätzt über MML-Methode

Auffallend ist in Tab. 7.6 bereits, dass - verglichen mit den Ergebnissen zu dem FBL-R-ALL-Teildatensatz - der geschätzten Steigungsparameter des PCMs einen höheren Wert mit $\alpha = 0.886$ aufweist und es damit bei dieser Skala leichter ist, mittels der jeweils angegebenen Antwortkategorie zwischen den einzelnen Personenparametern zu differenzieren.

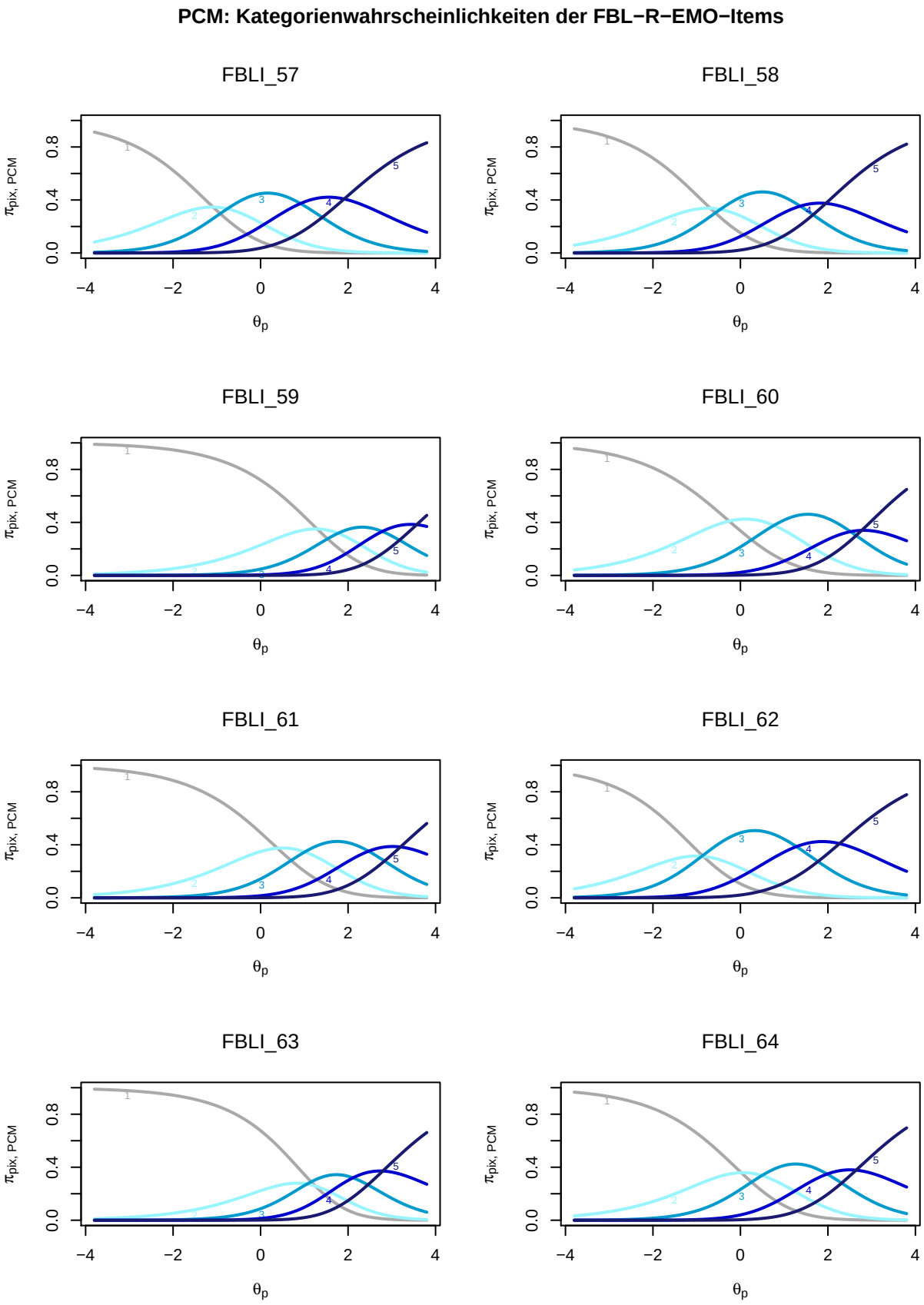

Abb. 7.8: Grafische Veranschaulichung des PCMs (konstanter α-Parameter) angewandt auf die FBL-R-EMO-Daten anhand der Kategorienwahrscheinlichkeiten

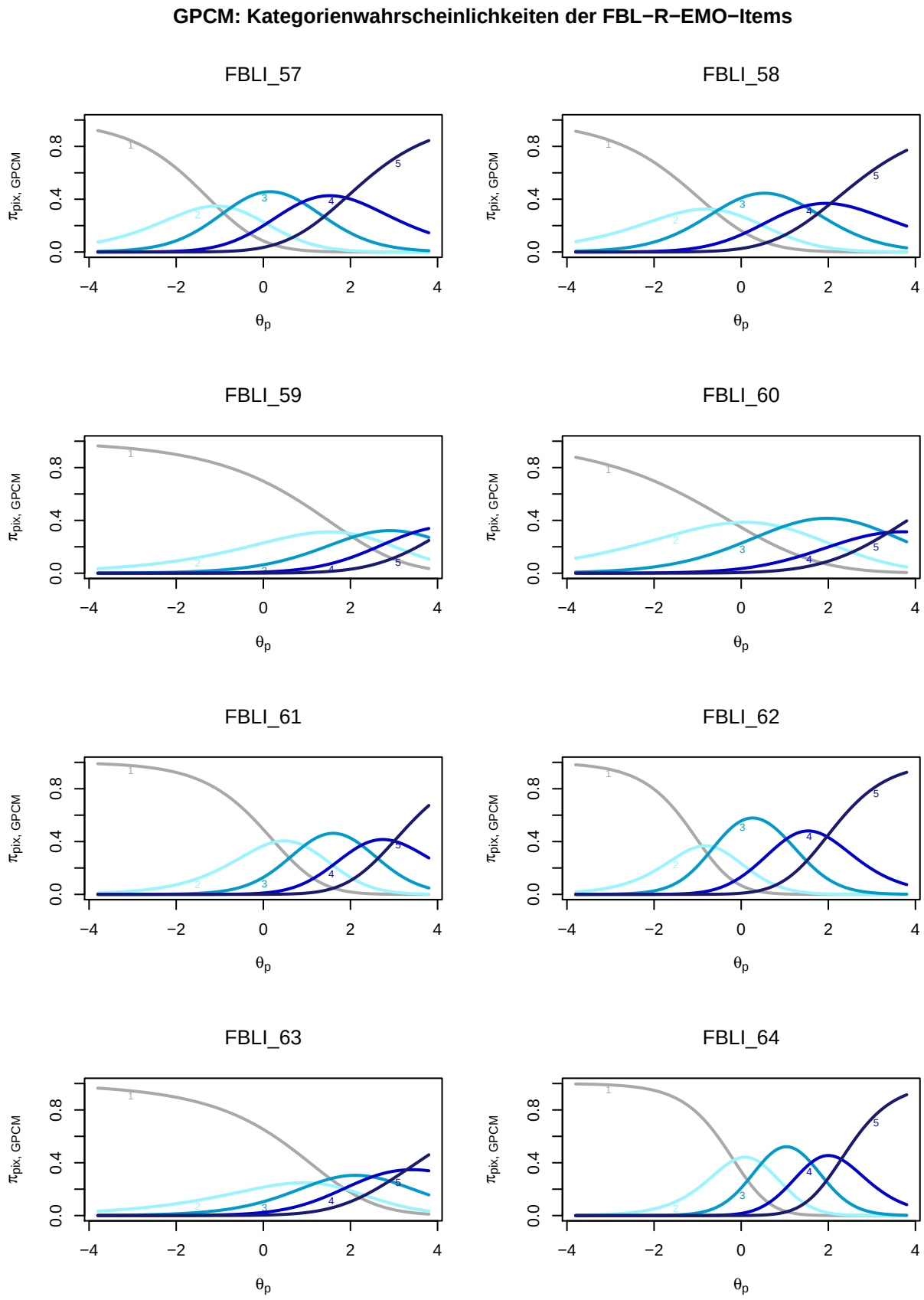

Abb. 7.9: Grafische Veranschaulichung des GPCMs (α_i-Parameter) angewandt auf die FBL-R-EMO-Daten anhand der Kategorienwahrscheinlichkeiten

Zudem lässt sich auch bei der Analyse mittels des GPCMs festhalten, dass insbesondere mithilfe der Items FBLI_61, FBLI_62 und FBLI_64 aufgrund eines Diskriminationsparameters, der größer ist als 1, zwischen den verschiedenen Ausprägungen der latenten Variable differenziert werden kann. Außerdem sind hier die Schwellenparameter beinahe aller Items - jeweils ausgenommen FBLI_63 - den Kategorien gemäß aufsteigend angeordnet und damit ist die Problematik „gemiedener" Kategorien weitestgehend nicht gegeben. Diese Feststellungen lassen sich auch gut anhand der Abb. 7.8 und 7.9 nachvollziehen.

Führt man nun auch wieder den LQ-Test zwischen zwei geschachtelten Modellen - hier mittels der 'anova.gpcm()'-Funktion des ltm-Paketes - durch, so ergibt sich ein p-Wert, der kleiner als 0.01 ist, und damit ein signifikanter Unterschied in der Verwendung der itemspezifischen Diskriminationsparameter α_i. Infolgedessen scheint es auch bei der Skala der Emotionalen Reaktivität ratsam zu sein, das GPCM zur Analyse der Daten zu verwenden. Bereits anhand der AIC's hätte eine derartige Vermutung aufkommen können. So ergibt sich für das PCM eine Wert von 'AIC = 38684.42', während man für das GPCM 'AIC = 38404.73' erhält.

Untersucht man nun wiederum, ob die Modelle überhaupt ausreichend gut an die Daten angepasst sind und verwendet dazu die im ltm-Paket integrierte 'GoF.gpcm()'-Funktion mit einer Standardeinstellung von 'B = 99' Bootstrapstichproben und gleichzeitiger Wahl von 'seed = 123', so ergibt sich bezüglich der Anpassungsgüte Folgendes: gemäß der parametrischen Bootstrap-Approximation an Pearson's χ^2-Test wird in beiden Fällen die Nullhypothese zu einem Signifikanzniveau von 0.05 verworfen. Konkret ergibt sich bei der Anwendung der 'GoF.gpcm()'-Funktion auf das PCM-Objekt eine Teststatistik von 'Tobs = 954642' und führt damit zu einem p-Wert von 0.03. Dagegen erhält man für das GPCM als Wert der Teststatistik 'Tobs = 1479216' mit einem zugehörigem p-Wert von 0.03. Auch für die Skala der Emotionalen Reaktivität ist somit die Anwendung des PCMs bzw. des GPCMs zur Analyse der inbegriffenen Daten fraglich.

Damit verdeutlichen die Tests zur Anpassungsgüte der Modelle insgesamt, dass mittels der errechneten Schätzer keine genauen Prognosen anhand gegebener Scores gemacht werden können. Insofern sollte geprüft werden, ob ein anderes Modell eine bessere Anpassung an die vorliegenden Daten garantiert oder aber, ob im Vorhinein bei der Konstruktion der Fragen bzw. den möglichen Antwortkategorien einige Tatsache berücksichtigt werden müssten. Diese Aspekte werden im nachfolgenden Kapitel noch kurz diskutiert, wobei hier ebenso die wichtigsten Punkte dieser Arbeit noch einmal zusammengefasst werden.

8 Résumé

Abschließend sollen noch einmal die wichtigsten Punkte dieser Arbeit zusammengefasst werden. Zunächst sei dazu als Oberbegriff die (psychologische) Testtheorie zu nennen, die das Ziel verfolgt über wissenschaftlich - bzw. hier statistische - Methoden von einer erfassbaren Aussage auf die individuelle, latente Ausprägung des Merkmals einer Person, beispielsweise die politische Einstellung, zu schließen. Im Rahmen der Testtheorie unterscheidet man prinzipiell zwei Arten: zum einen die Klassische Testtheorie (KTT) und zum anderen die Item Response Theorie (IRT). Trotz der weiten Verbreitung sollte bei der Anwendung der KTT die Stichprobenabhängigkeit als klarer Nachteil nicht außer Acht gelassen werden, welcher für die IRT nicht gegeben ist. Die IRT als solche fokussiert im Gegensatz zur KTT nicht den Test als Ganzen, sondern auf die einzelnen, dem Test inbegriffenen Items. Dem Oberbegriff der IRT lassen sich einige probabilitische Modelle zuordnen, wobei die meisten derer einen Funktionstyp logistischer Art aufweisen. In ihrer Anwendbarkeit sind sie prinzipiell entsprechend des vorliegenden Datenformates zu unterscheiden. So lässt sich etwa vom Rasch-Modell (RM) bei dichotomen Antwortmustern Gebrauch machen, während sich das im Rahmen dieser Arbeit vorgestellte Partial Credit Modell, ebenso wie das Generalisierte Partial Credit Modell zur Analyse polytomer, ordinal skalierter Items eignet.

Um diesem Datenformat gerecht zu werden, wird pro Item nicht nur ein einziger Schwierigkeitsparameter - wie im RM - betrachtet, sondern es werden die einzelnen vorliegenden Kategorien berücksichtigt, indem die Schwierigkeit der jeweiligen Stufen zwischen diesen Kategorien geachtet wird. Diese Parameter werden als Schwellenparameter bezeichnet, welche wiederum in wahrscheinlichkeitsfunktionalen Zusammenhang mit den einzelnen Personenparametern gesetzt werden. Hiermit kann letztlich durch die Einbettung in das logistische Regressionsmodell die Wahrscheinlichkeit für das Erreichen eines bestimmten Testscores - der sich additiv aus der Anzahl der insgesamt bewältigten Stufen zusammensetzt - in Abhängigkeit der benannten Parameter modelliert werden. Während das PCM nur die einzelnen Schwellenparameter im Bezug auf die Itemparameter berücksichtigt, wird im GPCM durch die Hinzunahme eines s.g. Diskriminations- bzw. Steigungsparameters eine flexiblere Modellierung zugelassen.

Da die Konstruktion des PCMs im Wesentlichen auf dem RM aufbaut, ist es naheliegend, dass einige der geltenden Annahmen des RMs auf das PCM übertragen werden können. Tatsächlich ist dies - unter Berücksichtigung einiger Abwandlungen - sogar gänzlich möglich und damit werden dem PCM folgende Eigenschaften zuteil: die Eindimensionalität, gemäß der das Antwortverhalten einer Person nur von einem einzigen latenten Parameter gesteuert wird, die spezifische Objektivtät, welche objektive Vergleiche von Personen bzw. von Aufgaben garantiert, ebenso wie die aus der Stichprobenunabhängigkeit resultierende Eigenschaft der lokalen stochastischen Unabhängigkeit, die u.a. bei der Parameterschätzung ein nützliches Werkzeug darstellt. Des Weiteren lassen sich in diesem Zusammenhang noch die Suffizienz, worunter allgemein zusammengefasst ist, dass es für die beiden im PCM inbegriffenen Parameter jeweils eine suffiziente Statistik gibt, erwähnen. Zusätzlich gilt für das Messniveau, dass dieses bis auf Intervallskalenniveau Eindeutigkeit besitzt.

Für das GPCM können dagegen nicht alle Eigenschaften in dieser Form, wie sie beim PCM vorliegen, übernommen werden. Aufgrund der Berücksichtigung des Steigungsparameters ist es

v.a. nicht mehr möglich spezifisch objektive Vergleiche von Personen und/oder Items anzustellen. Zudem kann man sich nicht mehr der suffizienten Statistiken in dem Ausmaße bedienen, wie es für das PCM der Fall ist.

Vorrangiges Interesse besteht nun an der konkreten Parameterschätzung, welche im PCM, sowie GPCM möglich sind. Für das PCM bieten sich zur Schätzung der Itemparameter im Wesentlichen drei Verfahren an, welche alle auf der Maximum-Likelihood (ML) basieren: die gemeinsame ML-Methode (JML), die bedingte ML-Methode (CML) und die marginal ML-Methode (MML). Dagegen stellt für das GPCM nur letztere ein sinnvolle Möglichkeit der Itemparameterschätzung dar. Prinzipiell unterscheiden sich die drei genannten Verfahren in ihrem konkreten Vorgehen. Bei der JML-Methode werden dabei alle Parameter - also Personenparameter eingeschlossen - gleichzeitig maximiert und auf dieser Basis geschätzt. Bei dem allgemein vorliegenden Szenario, demgemäß ein Test eine fixe Anzahl an Items aufweist und lediglich die Zahl der Testteilnehmer erhöht wird/ werden kann, zeigt sich der deutliche Nachteil der JML-Methode in resultierenden, inkonsistenten Itemparameterschätzungen. Daher wird zumeist auf eine der anderen beiden Methoden zurückgegriffen, die jeweils zweistufige Verfahren darstellen und die Personenparameter zunächst auf unterschiedliche Weise unberücksichtigt lassen. Die CML-Methode geht hierbei so vor, dass auf die suffiziente Statistik für die Personenparameter bedingt wird und damit diese Parameter zunächst unberücksichtigt bleiben können. Bei der MML-Methode wird dagegen davon ausgegangen, dass die Personenparameter einer bestimmten Verteilung folgen - zumeist geht man von der Normalverteilung mit Erwartungswert 0 und Varianz σ^2 oder gar der Standardnormalverteilung aus - und werden damit aus der ML herausintegriert. Sowohl das CML-, als auch das MML-Verfahren liefert konsistent Schätzungen und die beiden Methoden sind bei korrekter Spezifizierung der Personendichte für das MML-Vorgehen asymptotisch äquivalent.

Zur an die CML- bzw. MML-Methode anschließenden Ermittlung der Personenparameter können wiederum verschiedene Ansätze gewählt werden. Hierbei stehen insbesondere die unbedingte, sowie die gewichtete ML-Methode (WML) zur Verfügung. Bei der MML-Methode können aufbauend auf der priori-Verteilung der Personenparameter hauptsächlich zwei bayessche Verfahren ausgenutzt werden: die „Expected A Posteriori“ (EAP) und „Maximum A Posteriori“ (MAP) Methode, wobei die jeweiligen EAP-Werte die Median und die MAP-Werte den jeweiligen Modus der a-posteriori-Verteilung der Personenparameter darstellen. Hierbei bieten die bayesschen Verfahren den klaren Vorteil, dass auch die Fähigkeiten von Personen mit vollem Gesamtscore oder einem von 0 abgeschätzt werden können.

Die Analyse von Daten mittels des PCMs bzw. GPCMs kann über die Statistik-Software R durch Nutzung implementierter Pakete durchgeführt werden. Hierzu bieten sich z.B. das eRm-, das ltm- oder das TAM-Paket an. Grundsätzlich kann das eRm-Paket nur das PCM für gegebene Daten anpassen, während die beiden anderen Pakete sowohl das PCM, als auch das GPCM bereitstellen. Innerhalb dieser Pakete sind die Modelle jeweils über verschiedene Ausgangsmodelle integriert. So lässt sich im eRm-Paket das linear-logitische Testmodell als Basis ausmachen, von welchem sich das Lineare Partial Credit Modell und durch gewisse Restriktion dessen das Partial Credit Modell an sich ableiten lässt. Im ltm-Paket stellt das Generalisiert Partial Credit Modell die Grundlage der Analyse von Daten mit Partial Credit Format dar, von dem sich durch entsprechende Anpassung

des Steigungsparameters das Partial Credit Modell konstruieren lässt. Im TAM-Paket ist das Basismodell durch das s.g. Random Coefficients Multinomiale Logit-Modell gegeben. Durch die jeweilige Wahl der inbegriffenen Matrizen kann hiervon sowohl das Partial Credit, als auch das Generalisierte Partial Credit Modell abgeleitet werden.

Diese Pakete wurden abschließend zur Analyse der s.g. Freiburger Beschwerdenliste mittels des PCMs und GPCMs genutzt. Generell beinhaltet dieser Datensatz die in Form von Kategorien angegebenen Antworten von insgesamt 2070 Personen zu aktuellen, situativ-bedingten und chronisch-habituellen Beschwerden. Die 80 Items des Datensatzes in seiner revidierten Form können dabei in neun Skalen untergliedert werden. In dieser Arbeit wurden dabei die Skala des Allgemeinbefindens, sowie auch die der Emotionalen Reaktivität näher untersucht. Hierbei konnte auf Basis des ersten Teildatensatzes die asymptotische Äquivalenz der CML- und der MML-Methode verdeutlicht werden. Für beide genannten Skalen wurde konkret sowohl das PCM als auch das GPCM zur Analyse der Daten herangezogen.

In diesem Zusammenhang wurde jeweils auch mittels der parametrischen Bootstrap-Approximation an Pearson's χ^2-Test geprüft, welches der beiden (geschachtelten) Modelle jeweils besser zur Analyse der Daten geeignet ist. Genauer gesagt wurde hiermit getestet, ob die Berücksichtigung eines itemspezifischen Steigungsparameters und damit das GPCM einen signifikanten Unterschied im Vergleich zum PCM birgt. Für beide Teildatensätze konnte so gezeigt werden, dass zu einem Signifikanzniveau von 0.05 jeweils das GPCM präferiert werden sollte.

Daran anschließend wurde auch die Anpassungsgüte der jeweiligen Modelle geprüft. Hierbei ergab sich für beide Skalen, also die des Allgemeinbefindens und die der Emotionalen Reaktivität, dass weder das Partial Credit Modell, noch das Generalisierte Partial Credit Modell ausreichend gut an die Daten angepasst sind. Damit stellt sich allgemein die Frage, ob die beiden genannten Modell zur Analyse der Freiburger Beschwerdenliste geeignet sind oder ggf. ein anderes Modell bevorzugt werden sollte. In diesem Rahmen könnte eventuell das Rating Scale Modell, bei welchem vorausgesetzt wird, dass die Distanz zwischen den einzelnen Antwortkategorien gleich groß ist, einen Ausweg bieten.

Zudem kann auch eine im Vorhinein genauere Definition der einzelnen Kategorien erforderlich sein. So könnten beispielsweise die einzelnen Abstufungen von „sehr stark“, „stark“, „mittel“, „kaum“ und „praktisch nie“ von den einzelnen Befragten unterschiedlich aufgefasst werden. Generell ist es bei der konkreten Auswertung mittels des PCMs oder GPCMs nämlich ratsam „gemiedene“ Kategorien zu entfernen. Dies wäre dann ggf. auch bei der Konstruktion von Items nachfolgender Befragungen zu berücksichtigen und demzufolge sollten vielleicht bestimmte Kategorien zusammengefasst oder klarer von den übrigen abgegrenzt werden. Genauso problematisch könnte in die andere Richtung gehend aber auch die recht grobe Einteilung in die Kategorien „fast täglich“, „etwa 3-mal pro Woche“, „etwa 2-mal pro Monat“, „etwa 2-mal pro Jahr“ und „praktisch nie“ sein. Hierbei wäre es eventuell erforderlich, mehr Kategorien und damit mehr Abstufungen der verfügbaren Antworten bereitzustellen. In diesem Hinblick sollte wohl als kritischer Punkt auch aufgeführt werden, dass die Zuteilung zu den einzelnen Kategorien anhand der Selbsteinschätzung der Befragten geschieht und damit sehr subjektiv ausfällt. So könnte es beispielsweise passieren, dass Frauen ihre Beschwerden insgesamt als schwerwiegender bewerten, als dies Männer tun würden. Damit ist

generell die Frage nach einer ojektiveren Möglichkeit der Beschwerden-Abfrage gegeben.

Ein Problem könnte allerdings auch darin liegen, dass die Skalen nicht korrekt zugeteilt sind. In diesem Sinne könnte eine bestimmte Frage ggf. nicht auf beispielsweise das Allgemeinbefinden oder die Emotionale Reaktivität abzielen, sondern eher auf eine andere oder zusätzliche latente Variable. Wie anhand letzterem zu erkennen ist, kann es vorkommen, dass mittels eines bestimmten Items auf mehrdimensionaler Ebene mehrere latente Personenmerkmale abgefragt werden. Damit wäre die Annahme der Eindimensionalität nicht gegeben und demzufolge das PCM bzw. das GPCM nicht zur Analyse der Daten geeignet. Stattdessen müsste man hier auf mehrdimensionale IRT-Modell zurückgreifen.

Obwohl die Anwendung des PCMs, sowie des GPCMs zur Analyse der Freiburger Beschwerdenliste fraglich ist bzw. insbesondere Prognosen auf Basis der geschätzten Parameter kritisch bewertet werden sollten, stellen die beiden Modelle dennoch eine gute Möglichkeit bereit, latente Eigenschaften auf eine quantitativ, messbare Ebene zu überführen. Um noch einmal Bezug zur Einleitung zu nehmen, sei daran erinnert, dass gerade Persönlichkeitsmerkmale wie die politische Einstellung, die Intelligenz, das Allgemeinbefinden, etc. nicht einfach mittels einer Stoppuhr oder einem Metermaß messbar sind und damit komplexere Lösungsansätze verlangen. Diese finden sich z.B. im Rahmen der psychologischen Testtheorie und hier insbesondere in Form der IRT-Modelle, denen gleichzeitig nützliche mathematische Eigenschaften zu eigen sind. Je nach zugrundeliegendem Datenformat können so v.a. das PCM und das GPCM aussagekräftige Ergebnisse liefern. Auch bietet hierzu die Statistik-Software R einfache Modellierungs- und umfassende Analysemöglichkeiten mittels der beiden Modelle, welche in Form von verschiedenen Paketen bereitgestellt sind. Zusammenfassend lässt sich das PCM und in generell gehaltenerer Form auch das GPCM als geeignetes Werkzeug zur Analyse von ordinalen Daten, genauer gesagt von Daten des Partial Credit Formates, ansehen und empfiehlt sich insbesondere mithilfe von R-Paketen anzuwenden.

Literaturverzeichnis

Adams, R.J. und Wilson, M. u. W. M. (1997). Multilevel Item Response Models: An Approach to Errors in Variables Regression. *Journal of Educational and Behavioral Statistics 22*, 47–76.

Andersen, E. (1970). Asymptotic Properties of Conditional Maximum-Likelihood Estimators. *Journal of the Royal Statistical Society 32*(2), 283–301.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika 43*, 561–573.

Becker, J. (2004). *Computergestütztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT).* Dissertation, Freie Universität, Berlin.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In M. Lord, F.M. und Novick (Ed.), *Statistical theories of mental test scores (S. 395-479). Reading: Addison-Wesley.*

Bock, R.D. und Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika 46*, 443–459.

Cai, L. (2012). Latent Variable Modeling. *Shanghai Arch Psychiatry 24*, 118–120.

Chen, S.K. und Hou, L. u. D. B. (1998). A comparison of Maximum Likelihood Estimation and Expected A Posteriori Estimation in CAT using the Partial Credit Model. *Eductional and Psychological Measurement 58*(4), 569–595.

De Ayala, R. (2009). *The Theory and Practice of Item Response Theory.* New York: The Guilford Press.

Eid, M. und Schmidt, K. (2014). *Testtheorie und Testkonstruktion.* Bachelorstudium Psychologie - Band 20. Göttingen: Hogrefe Verlag.

Fahrenberg, J. (1994). *Die Freiburger Beschwerdenliste (FBL). Form FBL-G und revidierte Form FBL-R.* Göttingen.

Fahrmeir, L. und Künstler, R. u. P. I. u. T. G. (2010). *Statistik - Der Weg zur Datenanalyse (7.Auflage).* Berlin Heidelberg: Springer-Verlag.

Fahrmeir, L. und Kneib, T. u. L. S. (2009). *Regression - Modelle, Methoden und Anwendung.* Berln Heidelberg: Springer-Verlag.

Fischer, G.H. und Molenaar, I. (1995). *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer New York.

Fischer, G.H. und Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika 59*(2), 177–192.

Gonzalez, R. und Griffin, D. (2001). *Testing Parameters in Structural Equation Modeling: Every 'One' Matters*, Volume 6.

Gulliksen, H. (1950). *Theory of Mental Tests*. Wiley Publications in Psychology. New York: Wiley.

Hambleton, R.K. und Swaminathan, H. u. R. H. (1991). *Fundamentals of Item Response Theory*. Measurement Methods for the Social Science. Newbury Park, CA: SAGE Publications.

Irtel, H. (1996). *Entscheidungs- und testtheoretische Grundlagen der psychologischen Diagnostik*. Mannheim: P. Lang.

Johnson, M. (2007). Marginal Maximum Likelihood Estimation of Item Response Models in R. *Journal of Statistical Software 20*.

Kauerman, G. und Hothorn, T. (2014). *Vorlesungsskript: Statistik IV*.

Kiefer, T und Robitzsch, A. u. W. M. (2013). TAM (Test Analysis Modules) - Tutorial 5: Partial Credit Model. Website. online verfügbar unter: `http://www.edmeasurementsurveys.com/TAM/Tutorials/5PartialCredit.htm`; abgerufen am: 15.06.2016.

Kiefer, T. und Robitzsch, A. u. W. M. (2016a). R Core Team.

Kiefer, T und Robitzsch, A. u. W. M. (2016b). *Package 'eRm' (Version 0.15-6)*. R Core Team.

Kolanoski, H. (2008). Vorlesungsskript: Die Maximum-Likelihood-Methode. online verfügbar unter: `https://www-zeuthen.desy.de/~kolanosk/smd_ss02/skripte/ml.pdf`; abgerufen am: 25.05.2016.

Koller, I. und Alexandrowicz, R. u. H. R. (2012). *Das Rasch-Modell in der Praxis - Eine Einführung mit eRm*. Wien: Facultas Verlags- und Buchhandels AG.

Lienert, G.A. und Raatz, U. (1998). *Testaufbau und Testanalyse*. Grundlagen Psychologie. Weinheim: Beltz, PsychologieVerlagsUnion.

Liu, Q. und Pierce, D. (1994). A note on Gauss-Hermite quadrature. *Biometrika 81*, 624–629.

Mair, P. und Hatzinger, R. (2007a). CML based estimation of extended Rasch models with the eRm package in *R*. *Psychology Science 49*, 26–43.

Mair, P. und Hatzinger, R. (2007b). Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software 20*.

Mair, P. und Hatzinger, R. u. M. M. (2009). Extended Rasch Modeling: The R Package eRm. PDF-Dateianhang zum Programmpaket eRm; online verfügbar unter: `https://cran.r-project.org/web/packages/eRm/vignettes/eRm.pdf`; abgerufen am: 20.06.2016.

Masters, G.N. und Wright, B. (1997). The Partial Credit Model. In R. Van der Linden, W.J. und Hambleton (Ed.), *Handbook of Modern Item Response Theory (S. 101-121). New York: Springer-Verlag*.

Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika 47*(2), 149–174.

Masters, G. (1988). The analysis of partial credit scoring. *Applied Measurement in Eduction 1*, 279–297.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement 16*, 159–176.

Muraki, E. (1993). Information functions of the generealized partial credit model. *Applied Psychological Measurement 17*, 351–363.

Muraki, E. (1997). A Generalized Partial Credit Model. In R. Van der Linden, W.J. und Hambleton (Ed.), *Handbook of Modern Item Response Theory (S. 153-164). New York: Springer-Verlag.*

Novick, M. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology 3*, 1–18.

Pauls, T. (2003). *Resampling-Verfahren und ihr Anwendung in der nicht-parametrischen Testtheorie.* Dissertation, Heinrich Heine Universität, Düsseldorf.

Pernerstorfer, G. (2005). Erwartungstreue und Konsistenz. Website. online verfügbar unter: `http://www.mathe-online.at/materialien/georg.pernerstorfer/files/Kap1/erwtreue_konsistenz.pdf`; abgerufen am: 24.05.2016.

R Core Team (2016). *The R language definition (Version 3.3.0).*

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests.* Studies in mathematical psychology. Copenhagen: Danmarks Paedagogiske Institut.

Rizopoulos, D. (2006). ltm: An R for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software 17.*

Rizopoulos, D. (2015). *Package 'ltm' (Version 1.0-0).* R Core Team.

Rost, D.H. und Spada, H. (1978). Probabilistische Testtheorie. In K. J. Klauer (Ed.), *Handbuch der pädagogischen Diagnostik 1 (Bd. 1, S. 59-97). Düsseldorf: Schwann.*

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion.* Aus dem Programm Huber: Psychologie-Lehrbuch. Bern: Hans Huber Verlag.

Rost, J. (2013). Rasch-Modell, ordinales. In M. Wirtz (Ed.), *Dorsch Lexikon der Psychologie (16.Auflage, S.364). Bern: Hans Huber Verlag.*

Schomaker, M. und Toutenburg, H. u. W. M. u. H. C. (2008). *Deskriptive Statistik: Eine Einführung in Methoden und Anwendungen mit R und SPSS.* Springer-Lehrbuch. Springer Berlin Heidelberg.

Sijtsma, K. und Junker, B. (2006). Item Response Theory: Past Peformance, Present Developments, and Future Expectations. *Behaviormetrika 33*(1), 75–102.

Strobl, C. (2012). *Das Rasch-Modell : Eine verständliche Einführung für Studium und Praxis (2.Auflage)*. Sozialwissenschaftliche Forschungsmethoden. Mering: Rainer Hampp Verlag.

Thissen, D. und Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika 51*(4), 567–577.

Volodin, N. und Adams, R. (2002). The Estimation of Polytomous Item Response Models with many Dimensions. Website. online verfügbar unter: `http://research.acer.edu.au/ar_misc/14`; abgerufen am: 01.07.2016.

Warm, T. (1985). *Weighted likelihood estimation of ability in item response theory with tests of finite length*. Dissertation, The University of Oklahoma - Graduated College, Norman, OK.

Welchowski, T. (2014). *Person parameter estimation in the polytomous Rasch model*. Master-arbeit, Ludwig-Maximilians Universität, München.

Wickham, H. (2015). *R Packages*. Sebastopol, CA: O'Reilly Media.

Wirtz, M. und Böcker, M. (2014). Differential Item Functioning (DIF). In M. Wirtz (Ed.), *Dorsch Lexikon der Psychologie (17.Auflage, S.404). Bern: Hans Huber Verlag.*

Inhalt der CD-ROM

Ordner	**Unterordner**	**Dateien**
Bachelor-Arbeit		BA_mit_Matr.Nr
		BA_ohne_Matr.Nr
Daten		fgjn93fr19_aa
		fgjn93fr19_ad
		fgjn93fr19_kb
		fgjn93fr19_pd
		fgjn93fr19_readme
Grafiken	Grafiken_Theorie	1_RM
		2_PCM
	Grafiken_Auswertung	1_Deskriptiv_allgemein
		2_Deskriptiv_FBL-R-ALL
		3_Deskriptiv_FBL-R-EMO
		4_Scatterplot_CML-MML_beta
		5_Scatterplot_CML-MML_tau
		6_beta-Parameter
		7_PCM_FBL-R-ALL
		8_GPCM_FBL-R-ALL
		9_PCM_FBL-R-EMO
		10_GPCM_FBL-R-EMO
R-Code		BA_Theorie
		BA_Auswertung

Eidesstattliche Erklärung

Hiermit versichere ich, Anna Theresa Stüber, die vorliegende Arbeit selbständig und ohne fremde Hilfe angefertigt zu haben. Die verwendete Literatur und sonstige Hilfsmittel sind vollständig angegeben.

München, 25. Juli 2016

.....
Unterschrift