Nonparametrische Tests im Rasch-Modell

- Teststärken unter verschiedenen Modellverletzungen -

Diplomarbeit am Institut für Statistik

der Ludwig-Maximilians-Universität München

von

Pascal Jordan

2. Februar 2010

Erster Gutachter : **Prof. Dr. Helmut Küchenhoff**

Zweiter Gutachter : **Prof. Dr. Markus Bühner**

Zusammenfassung

Drei nonparametrische Testklassen für das Rasch-Modell werden in dieser Arbeit vorgestellt. Alle drei Klassen ermöglichen Hypothesentests, die bereits für geringen Stichprobenumfang Gültigkeit besitzen und nicht auf Parameterschätzungen angewiesen sind.

Der Hauptteil der Arbeit befasst sich mit der Ergründung der Teststärke von vier ausgewählten Statistiken $(Y, T_{11}, \phi, \sigma_r^2)$ der kombinatorischen Testklasse anhand der Simulation verschiedener Alternativmodelle. Der Andersen-Likelihood-Quotienten-Test (λ) dient dabei als parametrische Vergleichsbasis.

Die Ergebnisse zeigen, dass Y - die einzige Prüfgröße mit (geringfügigen) Diskrepanzen bezüglich des α-Fehlers - unter jeder Bedingung höhere Teststärken als λ aufweist. Beide Größen reagieren stark auf variierende Itemdiskrimination und in deutlich abgeschwächter Form auf lokal abhängige Daten.

Die T_{11}-Größe reagiert auf jede realisierte Modellverletzung sensitiv. Außer in einer speziellen Bedingung ist sie ferner der ähnlich reagierenden ϕ-Statistik überlegen.

Für die Prüfgröße σ_r^2 zeigen sich hohe Teststärken im Fall von zwei Testhälften, die unterschiedliche Konstrukte messen, bzw. im Fall von zwei Testhälften, die über stark unterschiedliches Diskriminationspotential verfügen.

Eine Anwendung der Prüfgrößen auf die Subskalen des Intelligenz-Struktur-Tests (IST 2000 R) spiegelt die Resultate der Simulation wider. Nur die beiden Prüfgrößen Y und T_{11} lehnen für alle Subskalen, in Einklang mit nonparametrischen deskriptiven Methoden basierend auf der dritten Testklasse, die Rasch-Konformität ab. Der anschließende Versuch, anhand einer Itemselektionsprozedur Rasch-Skalierbarkeit zu erlangen, scheitert für die meisten Subskalen.

Vorwort

Diese interdisziplinäre Diplomarbeit wurde im Zeitraum zwischen dem 6.August 2009 und dem 2.Februar 2010 an der Ludwig-Maximilians-Universität München geschrieben. Ihr ist eine CD beigefügt, die den Programmiercode der Simulationen und den Programmiercode der Datensatzauswertung beinhaltet.

Ich möchte mich herzlich bei Professor Dr. Helmut Küchenhoff und Professor Dr. Markus Bühner für Ihre Unterstützung bedanken.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Hinter vielen psychologischen Testverfahren steht implizit die Annahme, dass eine latente - d.h. nicht beobachtbare - Größe das Antwortverhalten „steuert“. Beispielsweise soll in einem Test, der aus Rechenaufgaben unterschiedlicher Komplexität besteht, die „numerische Intelligenz“ gemessen werden. Man erhofft sich, anhand des gezeigten Antwortmusters einer Person Rückschlüsse auf deren numerische Fähigkeit ziehen zu können. Hierfür stehen Wahrscheinlichkeitsmodelle zur Verfügung, die es erlauben, mit Hilfe des beobachteten Antwortmusters Schätzwerte für die latente Größe anzugeben. Damit diese Werte adäquate Approximationseigenschaften besitzen ist es allerdings notwendig, dass die Daten nicht in Widerspruch zu dem postulierten Modell stehen.

Ein probabilistisches Modell spezifiziert - für eine zufällig der Population entnommene Person - die gemeinsame Wahrscheinlichkeitsfunktion der manifesten Testdaten $\mathbf{x}$ sowie der latenten Größe $\boldsymbol{\theta}$ und setzt somit an die Wahrscheinlichkeitsfunktion der beobachtbaren Daten $f(\mathbf{x}) = \int f(\mathbf{x}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$ gewisse testbare Restriktionen.

Die probabilistischen Modelle lassen sich zunächst grob anhand der Dimensionalität der latenten Variable klassifizieren. Modelle mit eindimensionaler Struktur erweisen sich dabei für die Anwendung als besonders nützlich. Eindimensionalität ermöglicht die eindeutige Interpretation eines Testergebnisses sowie den adäquaten Vergleich der Testwerte zweier Personen (Stout, 1987; Bühner, 2006, Kapitel 7).

Das Rasch-Modell besitzt unter den eindimensionalen probabilistischen Modellen eine Sonderstellung. Die latente Variable wird auf Intervallskalenniveau gemessen. Die „übliche Praxis“, für jeden Probanden den Summenwert als Grundlage für sein

Testergebnis zu verwenden, erfährt im Rasch-Modell durch die Suffizienz des Summenscores ihre Rechtfertigung. Ferner sind Vergleiche zwischen Items invariant gegenüber der konkreten Personenauswahl. Charakteristika der Items lassen sich z.B. auch in anhand des Summenscores selektierten Stichproben konsistent schätzen[1].
Diese Eigenschaften werden durch starke Annahmen erkauft. Neben dem Ausschluss von Ratetendenzen der Personen wird eine strenge parametrische Form für die Item-Response-Funktionen (diese geben für jeden Wert der unbeobachtbaren Größe die bedingte Wahrscheinlichkeit an, das Item zu lösen) angenommen. An einen konkreten Datensatz sind somit starke Einschränkungen durch das Rasch-Modell gesetzt. Ein Modelltest ermöglicht die Überprüfung dieser Restriktionen[2]. Bei signifikantem Wert wird die Nullhypothese der Gültigkeit des Rasch-Modells verworfen. Die vorliegende Arbeit befasst sich nun mit einer speziellen Testklasse, den nonparametrischen Tests.

Kapitel 2 beschäftigt sich zunächst, nach einer kurzen Einführung in die Item-Response-Theorie, mit den Annahmen, die das Rasch-Modell kennzeichnen. Ein Teil dieser Annahmen findet sich auch in anderen probabilistischen Modellen wieder. Um die für das Rasch-Modell spezifischen Eigenschaften, insbesondere die Suffizienz des Summenscores, näher herauszustellen, wird dieses anschließend in zwei Obermodelle eingebettet. Neben der expliziten Kennzeichnung des Rasch-Modells ermöglicht dieser Vorgang auch eine Übertragung der Aussagen und Tests der Obermodelle. Ferner lassen sich anhand der Wertigkeit der Obermodelle Modellverletzungen bezüglich des Schweregrades einordnen. Abschließend für dieses zweite Kapitel erfolgt die Darstellung der Parameterschätzung im Rasch-Modell mittels Maximum-Likelihood. Sie bildet die Grundlage für die parametrischen Modelltests.

Nonparametrische Modelltests werden in **Kapitel 3** erörtert. Es handelt sich um drei Testklassen. Die erste Testklasse benutzt die suffizienten Statistiken des Rasch-Modells, um eine Aussage über die bedingte Verteilung der Datenmatrix zu formulieren. Ein Monte-Carlo-Algorithmus zur Approximation des p-Werts einer bezüglich dieser bedingten Verteilung verwendeten Prüfgröße rundet die Darstellung dieser sogenannten kombinatorischen Rasch-Modell-Tests ab. Anschließend erfolgt die Darstellung der zweiten Testklasse. Die Mantel-Haenszel-Tests werden zunächst als allgemeine Tests für dreidimensionale Kontingenztabellen beschrieben und darauffolgend für das Rasch-Modell konkretisiert. Während die ersten beiden Testklassen spezifisch

[1] Siehe Kapitel 2.4 bzw. die Darstellung des Likelihood-Quotienten-Tests in Kapitel 4.2.
[2] Genauer ermöglicht ein Modelltest „nur“ die Falsifikation des Rasch-Modells.

für das Rasch-Modell sind, erlauben die Tests der dritten Gruppe die Überprüfung recht allgemeiner Eigenschaften von Item-Response-Modellen. Allen drei Klassen gemeinsam hingegen ist der Fokus auf Teststatistiken, die bereits für geringen Stichprobenumfang anwendbar sind und keine Asymptotik erfordern.

Befassen sich die ersten drei Kapitel mit theoretischen Darlegungen, so ist das zentrale Element der Kapitel 4 und 5 die Beurteilung der Güte der kombinatorischen Testklasse. **Kapitel 4** beschreibt vier häufig verwendete Alternativmodelle, die unterschiedliche Formen der Abweichung vom Rasch-Modell repräsentieren. Sie bilden den Ausgangspunkt für die Simulation der Teststärke der kombinatorischen Modelltests in **Kapitel 5**. Im Gegensatz zu den übrigen nonparametrischen Testklassen, welche auf weitgehend bekannte Teststatistiken für kategoriale Daten hinauslaufen, sind die Eigenschaften der kombinatorischen Prüfgrößen relativ unbekannt. Die erforderliche Rechenleistung gestattet es ferner erst seit kurzem, ihr Verhalten für größere Datensätze zu betrachten.

Beschränkt sich Kapitel 5 im Wesentlichen auf die kombinatorischen Modelltests, so sollen in **Kapitel 6** mehrere nonparametrische Methoden gemeinsam zur Analyse eines Datensatzes eingesetzt werden. Insbesondere wird der These der Rasch-Skalierbarkeit des IST 2000 R (Intelligenz-Struktur-Test), eines nach der klassischen Testtheorie konstruierten Intelligenztests, nachgegangen. Zur weitergehenden Untersuchung dieser Frage eignen sich - neben den in Kapitel 3 dargelegten Prüfgrößen - nonparametrische deskriptive Methoden, die in diesem Kapitel vorgestellt werden sollen.

Kapitel 7 diskutiert abschließend - im Kontext der wesentlichen Erkenntnisse aus der praktischen Analyse des Datensatzes und der Simulationsstudie - die Rolle nonparametrischer Methoden.

Notation

Im Folgenden wird das Antwortverhalten einer Gruppe von n Personen auf eine k-elementige Menge dichotomer Items mittels Item-Response-Modellen betrachtet.

Beobachtbare Größen sind hierbei:

X_{vi}	Indikatorfunktion des Ereignisses: „v-te Person löst i-tes Item“
$\mathbf{X_v}$	Vektor aus Indikatorfunktionen der Person v mit j-tem Eintrag X_{vj}
$\mathbf{X}$	Gesamte Datenmatrix, $(\mathbf{X})_{vi} = X_{vi}$
$\boldsymbol{x_{(i)}}$	Die i-te Spalte der Datenmatrix $\boldsymbol{X}$

Direkte Funktionen dieser beobachtbaren Größen sind:

R_v	Summenscore der v-ten Person: $R_v = \sum_i X_{vi}$
C_i	Anzahl Personen, die Item i lösen: $C_i = \sum_v X_{vi}$
n_{ijh}	Anzahl Personen aus Gruppe h, die Item i lösen und Item j nicht lösen

Unbeobachtete Größen, die durch das spezifische Modell genauer festgelegt werden, sind:

$\boldsymbol{\theta_v}$	Vektor aus „Fähigkeitsparametern“ der v-ten Person
$f_i(\boldsymbol{\theta})$	Bedingte Wahrscheinlichkeit Item i zu lösen (gegeben $\boldsymbol{\theta}$)
$\omega_{ij}(\boldsymbol{\theta})$	Bedingtes Odds-Ratio: $\omega_{ij}(\boldsymbol{\theta}) := \frac{f_i(\boldsymbol{\theta})(1-f_j(\boldsymbol{\theta}))}{f_j(\boldsymbol{\theta})(1-f_i(\boldsymbol{\theta}))}$
$G(\boldsymbol{\theta})$	Verteilungsfunktion der latenten Variable $\boldsymbol{\theta}$

Des Weiteren gelten die folgenden Konventionen:

$\mathbf{1}$	Vektor, dessen Komponenten den Wert 1 besitzen
I_A	Indikatorfunktion des Ereignisses A
$\lvert A\rvert$	Anzahl der Elemente der Menge A
$\boldsymbol{A}^-$	Generalisierte Inverse der Matrix $\boldsymbol{A}$
$L^1(\mu)$	Menge aller Funktionen f mit der Eigenschaft $\int \lvert f\rvert \mathrm{d}\mu < \infty$

Mehrdimensionale Größen (Matrizen, Vektoren) werden stets in Fettdruck gesetzt. Die Darstellung von Matrizen erfolgt in Großbuchstaben.

Neben den in dieser Notation benannten Größen können in den einzelnen Kapiteln zusätzliche Variablen definiert werden. Die Darstellungen dieses Abschnitts beziehen sich lediglich auf die Basisgrößen.
Gelegentlich erfordert der Kontext zudem keine Indizierung der Personen. X_i bezeichnet dann den Response einer zufällig gezogenen Person auf Item i.

Kapitel 2

Grundlagen des Rasch-Modells

2.1 Item-Response-Theorie

Die Item-Response-Theorie (IRT) beschäftigt sich mit der Modellierung des Antwortverhaltens einer Gruppe von Personen auf eine Menge an Items.
Die kategorialen Antwortoptionen des i-ten Items, mit den Werten $0, 1, \ldots, m_i$ gekennzeichnet, bilden die potentiell möglichen Ausprägungen der beobachtbaren Zufallsvariable X_{vi}, welche die gewählte Kategorie der v-ten Person auf dem i-ten Item repräsentiert. Über alle Items zusammengeführt bezeichnet $\boldsymbol{X_v}$ das gesamte Antwortverhalten der v-ten Person.
Neben diesen beobachtbaren Größen postuliert die IRT - als zentrales Element - die Existenz einer potentiell mehrdimensionalen, latenten Variable $\boldsymbol{\theta_v}$, die die Wahrscheinlichkeitsfunktion des Vektors $\boldsymbol{X_v}$ über eine Mischverteilung bestimmt:

$$P(\boldsymbol{X_v} = \boldsymbol{x_v}) = \int P(\boldsymbol{X_v} = \boldsymbol{x_v} | \, \boldsymbol{\theta_v}) \mathrm{d}G(\boldsymbol{\theta_v})$$

Auf dieser Grundlage basierend können die einzelnen IRT-Modelle[1] anhand ihrer Annahmen bezüglich der latenten Variable und der bedingten Wahrscheinlichkeitsfunktion unterschieden werden. Häufig findet sich die Annahme einer eindimensionalen latenten Größe sowie die Annahme der bedingten Unabhängigkeit der Items gegeben $\boldsymbol{\theta}$. Es soll jedoch an dieser Stelle deutlich werden, dass diese Annahmen nicht zum

[1] Auch die klassische Testtheorie lässt sich in diesen dargestellten Rahmen einbetten (Holland und Hoskens, 2003).

Grundgerüst der IRT gehören, sondern bereits spezielle Ausgestaltungen beinhalten. Die konkrete Form dieser Ausgestaltung für das Rasch-Modell wird im folgenden Abschnitt behandelt. Es erfolgt dabei eine Beschränkung auf dichotome Daten ($m_i = 1$). Für diesen Spezialfall ist die Item-Response-Funktion $f_i(\boldsymbol{\theta})$ definiert als bedingte Wahrscheinlichkeit, das Item zu lösen:

$$f_i(\boldsymbol{\theta_v}) := P(X_{vi} = 1 | \, \boldsymbol{\theta_v})$$

Sie wird in der folgenden Darstellung des Rasch-Modells eine zentrale Rolle einnehmen.

2.2 Annahmen und Implikationen des Rasch-Modells

Das Rasch-Modell, als Verteter der Item-Response-Modelle, ist durch folgende Annahmen gekennzeichnet (Fischer, 1995):

(1) θ_v ist eindimensional;

(2) streng monoton wachsende, stetige Item-Response-Funktionen $f_i(\theta_v)$;

(3) lokale stochastische Unabhängigkeit, d.h.

$$P(\mathbf{X_v} = \mathbf{x_v} \, | \, \theta_v) = \prod_i P(X_{vi} = x_{vi} \, | \, \theta_v);$$

(4) $\forall i : \lim_{\theta_v \to +\infty} f_i(\theta_v) = 1, \lim_{\theta_v \to -\infty} f_i(\theta_v) = 0;$

(5) Suffizienz des Summenscores, d.h. $P(\mathbf{X_v} = \mathbf{x_v} \, | \, R_v = r_v \, , \theta_v)$ ist unabhängig von θ_v.

Die ersten beiden Annahmen besagen, dass der Test genau *eine* Fähigkeit misst und dass die Lösungswahrscheinlichkeit mit steigender Fähigkeit wächst. Ferner sind keine Sprünge in der Lösungswahrscheinlichkeit zu verzeichnen. In Kombination mit Annahme (3) ergeben sich damit Aussagen über die Assoziationsstruktur der Items.

Lokale stochastische Unabhängigkeit bedeutet, dass innerhalb einer Gruppe von Personen mit gleicher Fähigkeit die Items unabhängig sind. Marginal betrachtet, d.h. über die Fähigkeitsverteilung in der Population gemittelt, ergibt sich jedoch eine nichtnegative Kovarianz:

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \mathrm{Cov}(E(X_i|\,\Theta), E(X_j|\,\Theta)) + E(\mathrm{Cov}(X_i, X_j|\,\Theta)) && (2.1) \\ &\overset{(3)}{=} \mathrm{Cov}(E(X_i|\,\Theta), E(X_j|\,\Theta)) = \mathrm{Cov}(f_i(\Theta), f_j(\Theta)) && (2.2) \end{aligned}$$

Der letzte Term stellt eine Kovarianz zwischen zwei Funktionen von einer skalaren Zufallsvariable dar. Da $f_i(.)$ und $f_j(.)$ nach Annahme (2) monoton wachsende Funktionen sind, folgt eine nichtnegative Kovarianz (siehe auch Kapitel 3.3).
Während die ersten drei Annahmen die Assoziation der Items „erklären“, impliziert Annahme (4) den Ausschluss von Ratetendenzen. Dies ist eine Annahme, die theoretisch sehr strittig erscheint (Kubinger und Draxler, 2007). Die Existenz von Rateverhalten würde eine untere Schranke $c_i > 0$ mit der Eigenschaft $f_i(\theta) \geq c_i$ zur Folge haben. Das steht jedoch im Widerspruch zur Forderung (4).
Die fünfte Annahme stellt gewissermaßen den Kern des Rasch-Modells dar. Für die Inferenz auf den Fähigkeitsparameter ist allein der Summenscore R_v ausreichend. Zwei Personen mit gleichem Summenscore erhalten, unabhängig davon welche spezifischen Items sie gelöst haben, den gleichen geschätzten Fähigkeitswert. Auch wenn es übliche Testpraxis ist, für den Rückschluss den Summenscore zu verwenden, so ist dieses Vorgehen nur mit einem Rasch-Modell theoretisch zu rechtfertigen[2].

Bei Gültigkeit dieser fünf Annahmen lässt sich stets der Parameter θ so transformieren (siehe Fischer, 1995, S.16f), dass die Item-Response-Funktionen die logistische Form annehmen und jedes Item durch einen Schwierigkeitsparameter β_i gekennzeichnet ist. Für eine Person mit Fähigkeit θ_v folgt für die Wahrscheinlichkeit, das i-te Item zu lösen:

$$f_i(\theta_v) = P(X_{vi} = 1|\,\theta_v, \beta_i) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)} \tag{2.3}$$

Unter der Annahme der Unabhängigkeit des Antwortverhaltens verschiedener Personen sowie Annahme (3) resultiert für die Wahrscheinlichkeit der gesamten Datenmatrix $\mathbf{X}$:

$$P(\mathbf{X}\,|\,\theta_1, \ldots, \theta_n\,, \boldsymbol{\beta}) = \prod_{v=1}^{n} \prod_{i=1}^{k} \frac{\exp(x_{vi}(\theta_v - \beta_i))}{1 + \exp(\theta_v - \beta_i)} \tag{2.4}$$

[2]Für die *Ordnung* der Personen ist die Verwendung des Summenwerts jedoch bereits ab einem „Monotone Homogeneity Model“ (siehe Kapitel 2.3) gerechtfertigt.

Durch Umformung ergibt sich aus (2.4):

$$P(\mathbf{X} \mid \theta_1, \ldots, \theta_n, \beta_1, \ldots, \beta_k) = v(\boldsymbol{\theta}, \boldsymbol{\beta}) \exp\left(\sum_v r_v \theta_v - \sum_i c_i \beta_i\right) \tag{2.5}$$

Es liegt eine Exponentialfamilie vor, in der die Zeilen- sowie Spaltensummen suffiziente Statistiken bilden. Für die Schätzung der Item- und Personenparameter sind allein die Randsummen der Datenmatrix $\boldsymbol{X}$ ausreichend. Weiterhin ist die suffiziente Statistik der Fähigkeitsparameter nicht von den Itemparametern abhängig (und umgekehrt). Auf diese Eigenschaften lassen sich sowohl die zahlreichen Testmöglichkeiten des Rasch-Modells als auch die konsistente Schätzung der Itemparameter zurückführen. Gegenüber anderen, weniger restriktiven Modellen (siehe Abschnitt 2.3) bietet das Rasch-Modell zudem den Vorteil, dass die Messung der Fähigkeit auf Intervallskalenniveau erfolgt. Eine Eigenschaft, die insbesonders im Bereich des adaptiven Testens sowie bei der Verbindung von Testwerten, die aus unterschiedlichen Testformen entstammen (Dorans u.a., 2007), eine große Rolle spielt.
Weiterhin ergeben sich aus (2.3) die folgenden Eigenschaften bezüglich der Response-Funktionen zweier beliebig gewählter Items i und j:

- *Latente proportionale Odds*, d.h. die bedingten Odds, Item i zu lösen sind ein von θ unabhängiges Vielfaches der bedingten Odds, Item j zu lösen.

$$\omega_{ij}(\theta) = \exp(\beta_j - \beta_i) \tag{2.6}$$

- *Gleichmäßige Itemordnung*, d.h. (o.B.d.A. sei $\beta_i \geq \beta_j$)

$$\omega_{ij}(\theta) \leq 1 \quad \forall\theta \tag{2.7}$$

Die Eigenschaft der gleichmäßigen Itemordnung wird im folgenden Kapitel zur Einbettung des Rasch-Modells in Obermodelle wieder aufgegriffen. Sie stellt, wie auch (2.6), empirisch prüfbare Restriktionen an die Daten. Ein auf der Eigenschaft (2.6) basierender Test wird in Kapitel 3.3 diskutiert.

Aus den starken Modellannahmen folgen somit zahlreiche, sowohl aus statistischer Sicht (Exponentialfamilie) als auch aus anwendungsorientierter Perspektive (Summenscore beschreibt die Person vollständig) relevante Eigenschaften, die dem Rasch-Modell einen herausragenden Stellenwert unter den Item-Response-Modellen zuweisen. Molenaar (1995, S.5) bemerkt hierzu:

„*Whenever possible, it is thus recommended to find a set of items that satisfies the RM, rather than find an IRT model that fits an existing item set.*“

An dieser Stelle sei jedoch erwähnt, dass es bezüglich der Wertigkeit des Rasch-Modells durchaus kontroverse Ansichten gibt. Zum Einen wird die Notwendigkeit des Rasch-Modells aufgrund seiner Kompatibilität mit strengen Forderungen an den Prozess des Messens[3] betont. Zum Anderen ergibt eine Analyse realer Datensätze selten ein Rasch-Modell, so dass der Ausschluss von Items zur Erzielung eines Rasch-Modells die Reliabilität der Messung verringert. Mitunter wird die Möglichkeit, ein alternatives Modell zu finden, welches die Daten adäquat beschreibt und somit insbesonders keine Items ausschließt, als ein besseres Vorgehen betrachtet.

„*The perspective taken here is that the goal of the use of IRT models is to describe the interaction between each examinee and test item as accurately as possible [...] This perspective is counter to the one that proposes that strict mathematical criteria are needed to define measurement and only models that meet the criteria are acceptable.*“ (Reckase, 2009, S.21)

Im folgenden Abschnitt werden - unabhängig von dieser Kontroverse - zwei größere Klassen von IRT-Modellen dargestellt. Das Rasch-Modell lässt sich als spezielles Submodell dieser beiden Klassen auffassen. Es wird hierbei erkennbar, dass eine zentrale Eigenschaft des Rasch-Modells, über die konkurrierende Modelle dieser Klassen nicht verfügen, durch die speziellen suffizienten Statistiken gegeben ist. Ferner soll verdeutlicht werden, dass diese beiden Klassen - trotz im Vergleich zum Rasch-Modell abgeschwächter Annahmen - noch bezüglich einer ordinalen Messintention „brauchbare“ Modelle generieren. Gerade im Hinblick auf die erwünschte Teststärke eines Tests auf Rasch-Modell-Konformität erweist sich dies als relevant. Verletzt ein Alternativmodell auch die Annahmen des (größeren) Obermodells, so ist es wichtiger, dieses nicht als Rasch-Modell zu klassifizieren (d.h. einen β-Fehler zu vermeiden), als bei einem aus der Klasse der Obermodelle stammenden Alternativmodell. Im letzteren Fall besitzen die aus dem Rasch-Modell resultierenden ML-Schätzwerte $\hat{\theta}_v$, die aus einer streng monotonen Funktion des Summenwerts resultieren, trotz Modellverletzung noch eingeschränkte Gültigkeit im Sinne einer ordinalen Messung.

[3]Eine Herleitung des Rasch-Modells anhand meßtheoretischer Forderungen bzw. Annahmen findet sich z.B. bei Fischer (1995).

2.3 Obermodelle

Das Rasch Modell ist ein „Monotone Homogeneity Model“(MHM) (Mokken, 1971; Sijtsma und Molenaar, 2002), d.h. es folgt den Annahmen:

(U) Eindimensionalität: θ ist eine skalare Größe;

(LI) Bedingte Unabhängigkeit:

$$P(\mathbf{X_v} = \mathbf{x_v} \,|\, \theta) = \prod_i P(X_{vi} = x_{vi} \,|\, \theta)\,;$$

(M) Monotonie der Item-Response-Funktionen, d.h.

$$\theta_1 < \theta_2 \Rightarrow f_i(\theta_1) \leq f_i(\theta_2) \quad \forall i.$$

Die Erfüllung dieser Forderungen erlaubt die Ordnung der Personen mittels des Summenscores (Grayson, 1988). Θ ist stochastisch anhand des Summenwertes R geordnet, d.h. für $a \leq b$ gilt:

$$P(\Theta > c \,|\, R = a) \leq P(\Theta > c \,|\, R = b) \quad \forall c \tag{2.8}$$

Dieses Resultat liefert die Grundlage für eine ordinale Messung der Fähigkeit. In Gruppen mit höherem Summenscore fällt der Anteil jener Personen, die ein bestimmtes Fähigkeitsniveau überschreiten, größer aus als in Gruppen mit niedrigem Score.
Postuliert man in einem MHM zusätzlich

(NI) Nicht überschneidende Item-Response-Funktionen, d.h. $\forall i$, $\forall j$,

$$(f_i(\theta) \leq f_j(\theta) \quad \forall \theta) \vee (f_i(\theta) \geq f_j(\theta) \quad \forall \theta)\,,$$

so ergibt sich das „Double Monotonicity Model“(DMM). In einem DMM besitzen die Items für jedes Fähigkeitsniveau die gleiche Reihenfolge bezüglich ihrer Schwierigkeit. Das DMM ermöglicht die Ordnung der Items nach Itemscore. Unabhängig von der Populationsverteilung spiegelt die Ordnung der Items gemäß der empirisch beobachteten relativen Häufigkeit die Ordnung der Items für beliebiges θ wider:

$$f_i(\theta) \leq f_j(\theta) \Leftrightarrow P(X_i = 1) \leq P(X_j = 1)$$

Eine invariante Itemordnung kann u.a. zur Identifikation „unplausibler“ Antwortmuster hilfreich sein (Sijtsma und Molenaar, 2002).
Aus (2.7) folgt unmittelbar, dass das Rasch-Modell Annahme (NI) erfüllt. Die Annahmen des MHM finden sich außerdem explizit in den entsprechenden Annahmen des Rasch-Modells wieder. Somit stellt das Rasch-Modell ein spezielles DMM dar. Diese Beziehung ermöglicht es Methoden, die zur Überprüfung eines DMM dienen, auch auf das Rasch-Modell anzuwenden[4].

Die Forderungen (U), (LI), (M) und (NI) setzen keine konkrete parametrische Gestalt der Item-Response-Funktionen voraus. Das Rasch-Modell hingegen spezifiziert nach (2.3) die Item-Response-Funktionen parametrisch. Durch die Schwierigkeitsparameter β_i ist das Rasch-Modell eindeutig bestimmt. Der folgende Abschnitt behandelt zunächst die möglichen Schätzmethoden dieser Parameter, um anschließend auf die Klasse der parametrischen Modelltests einzugehen.

2.4 Schätzung der Itemparameter

Zur Schätzung der Itemparameter stehen drei Maximum-Likelihood-basierte Methoden zur Verfügung[5].

Joint-Maximum-Likelihood (JML)

In der JML-Methode werden die Itemparameter simultan mit den Personenparametern durch Maximierung von (2.4) geschätzt. Da für $n \to \infty$ (k fest) die Anzahl zu schätzender Parameter unbeschränkt wächst, führt diese Maximierung nicht zu konsistenten Schätzern der Itemparameter. Aus diesem Grund weicht man meist auf eine der folgenden Methoden aus.

[4]Siehe hierzu Kapitel 3.3 sowie Kapitel 6.

[5]Eine bayesianische Schätzung findet sich bei Swaminathan und Gifford (1982).

Conditional-Maximum-Likelihood (CML)

Die CML-Methode nutzt die Suffizienz des Summenscores. Die Personenparameter werden durch Bedingen auf den Summenscore eliminiert[6]:

$$P(\mathbf{X_v} = \mathbf{x_v} \,|\, R_v = r_v \,, \theta_v) = \frac{P(\mathbf{X_v} = \mathbf{x_v}, R_v = r_v \,|\, \theta_v)}{P(R_v = r_v \,|\, \theta_v)} \tag{2.9}$$

Der Nenner von (2.9) ist:

$$P(R_v = r_v \,|\, \theta_v) = \sum_{\mathbf{x'1}=r_v} P(\mathbf{X_v} = \mathbf{x} \,|\, \theta_v) = \sum_{\mathbf{x'1}=r_v} \prod_{i=1}^{k} \frac{\exp(x_i(\theta_v - \beta_i))}{1 + \exp(\theta_v - \beta_i)}$$

Da jeder Responsevektor $\boldsymbol{x}$ der Summe den gleichen Wert der suffizienten Statistik $r_v = \sum_i x_{vi}$ besitzt, folgt:

$$P(R_v = r_v \,|\, \theta_v) = v(\theta_v, \boldsymbol{\beta}) \sum_{\mathbf{x'1}=r_v} \exp\left(\theta_v r_v - \textstyle\sum_i x_i \beta_i\right) \tag{2.10}$$

Der Vorfaktor sowie der erste Term des Exponenten sind für alle zu betrachtenden Responsevektoren gleich, somit ergibt sich durch Einsetzen von (2.10) in (2.9):

$$P(\mathbf{X_v} = \mathbf{x_v} | R_v = r_v \,, \theta_v) = \frac{\exp(-\sum_i x_{vi}\beta_i)}{\sum_{\mathbf{x'1}=r_v} \exp(-\sum_i x_i \beta_i)} \tag{2.11}$$

Bei Unabhängigkeit der Responsevektoren verschiedener Personen resultiert:

$$P(\boldsymbol{X} \,|\, \boldsymbol{r} \,, \boldsymbol{\beta}) = \prod_v \frac{\exp(-\sum_i x_{vi}\beta_i)}{\sum_{\mathbf{x'1}=r_v} \exp(-\sum_i x_i \beta_i)} = \frac{\exp(-\sum_i c_i \beta_i)}{\prod_v (\sum_{\mathbf{x'1}=r_v} \exp(-\sum_i x_i \beta_i))} \tag{2.12}$$

Maximierung von (2.12) liefert den konsistenten und asymptotisch normalverteilten CML-Schätzer (Pflanzagl, 1994). Die Existenz des CML-Schätzers ist unter Regularitätsbedingungen asymptotisch gesichert.
Im Gegensatz zur MML-Schätzung (siehe nächster Abschnitt) bildet (2.12) eine Exponentialfamilie. Dies ist Grundlage für die Vielfalt an Teststatistiken im Rasch-Modell. Item-Response-Modelle, in denen die Separierbarkeit der Item- und Personenparameter nicht gegeben ist, können auf diese Schätzmethode nicht zurückgreifen. Um auch in diesem Fall die theoretischen Nachteile der JML-Schätzung zu umgehen, wird der Personenparameter als Zufallsvariable betrachtet und herausintegriert.

[6]Nach Definition ist die bedingte Wahrscheinlichkeit des Responsevektors $\boldsymbol{X_v}$ gegeben die suffiziente Statistik R_v unabhängig von θ_v.

Hiermit bleibt die Anzahl zu schätzender Parameter - bei steigendem Stichprobenumfang n - konstant. Zugleich bedeutet dies jedoch, dass der Modellfit nicht seperat von der Verteilungsannahme für den Parameter θ beurteilt werden kann. Auch für das Rasch-Modell ist diese Vorgehensweise möglich. Sie wird im Folgenden als dritte ML-basierte Methode beschrieben.

Marginal-Maximum-Likelihood (MML)

In der MML-Schätzung wird zusätzlich zu (2.4) eine Verteilung[7] für den latenten Parameter θ postuliert, $\Theta_v \overset{i.i.d}{\sim} G$. Die zu maximierende Likelihood lautet dann:

$$\prod_v P(\mathbf{X_v} = \mathbf{x_v}) = \prod_v \int_{-\infty}^{\infty} \prod_i \frac{\exp(x_{vi}(\theta_v - \beta_i))}{1 + \exp(\theta_v - \beta_i)} \mathrm{d}G(\theta_v) \tag{2.13}$$

Bei Gültigkeit dieses erweiterten Modells ist der MML-Schätzer konsistent sowie asymptotisch normalverteilt. Ferner sind MML- und CML-Schätzer asymptotisch äquivalent (Pflanzagl, 1994).
Ein Nachteil gegenüber der CML-Methode liegt in der zusätzlich erforderlichen Verteilungsannahme. Bei Fehlspezifikation können die Schätzer verzerrt sein. Vorteile jedoch ergeben sich bei der Schätzung[8] der Personenparameter: Beobachtungen mit extremen Summenwerten r_v besitzen aufgrund der zusätzlich angenommenen Basisverteilung einen endlichen Schätzwert[9].
Die CML-basierte Methode hingegen dient lediglich zur Ermittlung der Itemparameter. Schätzt man anschließend die Personenparameter durch Maximierung von (2.4), wobei die Itemparameter der CML-Schätzung als wahre, feste Werte behandelt werden („Plug-In"-Prinzip), so ergeben sich für die extremen Summenscores ($r_v = 0, r_v = k$) keine endlichen Schätzwerte.
Dieses „Plug-In"-Prinzip ignoriert zudem die Variabilität des $\boldsymbol{\beta}$-Schätzers. Auch für die MML-basierten Schätzer der Fähigkeit ($E(\Theta_v | \, r_v)$) sind „Plug-In"-Schätzer für die Itemparameter erforderlich. Eine Möglichkeit, dieses Problem zu überwinden, bietet die *simultane* Bayes-Schätzung der Item- und Personenparameter (Swaminathan und Gifford, 1982; Tsutakawa und Johnson, 1990; Lee, 2007).

[7] Meistens handelt es sich um die Annahme einer Normalverteilung.

[8] Es wird zwar der Ausdruck „Schätzung" verwendet, eigentlich handelt es sich aber um Prädiktion, da Θ_v als Zufallsvariable angesehen wird.

[9] Basis der Schätzung bildet der bedingte Erwartungswert $\mathrm{E}(\Theta_v | \boldsymbol{x_v})$.

2.5 Parametrische Rasch-Modell-Tests

Die parametrischen Modelltests lassen sich in zwei Klassen unterteilen. Die erste Klasse der „generalized Pearson tests“ fokussiert sich auf die beobachteten Einträge der 2^k-dimensionalen Kontingenztafel, in der die Items kreuzklassifiziert sind. Ein Vergleich mit den erwarteten Einträgen - gemäß des angepassten Rasch Modells - liefert eine asymptotisch χ^2-verteilte Prüfgröße unter der Nullhypothese der Gültigkeit des Rasch-Modells. Spezifischer zeigen Glas und Verhelst (1989), dass die quadratische Form

$$Q(\boldsymbol{U}) := n(\hat{\boldsymbol{\pi}} - \boldsymbol{p})'\boldsymbol{U}(\boldsymbol{U}'\boldsymbol{D}_{\hat{\pi}}\boldsymbol{U})^{-}\boldsymbol{U}'(\hat{\boldsymbol{\pi}} - \boldsymbol{p}) \tag{2.14}$$

unter enstprechenden Bedingungen an die Matrix $\boldsymbol{U}$ asymptotisch χ^2-verteilt ist mit $\text{Rang}(\boldsymbol{U}'\boldsymbol{D}_{\hat{\pi}}\boldsymbol{U}) - q - 1$ Freiheitsgraden. q repräsentiert hierbei die Anzahl geschätzter Modellparameter. In (2.14) bezeichnet $\hat{\boldsymbol{\pi}}$ den Vektor der vorhergesagten Zellwahrscheinlichkeiten der Multinomialverteilung, $\boldsymbol{p}$ den korrespondierenden Vektor der relativen Häufigkeiten sowie $\boldsymbol{D}_{\hat{\pi}}$ eine Diagonalmatrix mit Diagonalelementen $\hat{\boldsymbol{\pi}}$.
Wählt man speziell die Einheitsmatrix für $\boldsymbol{U}$, so ergibt sich die globale Testgröße:

$$Q(\boldsymbol{I}) = n \sum_i \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} \tag{2.15}$$

Prinzipiell reagiert sie auf alle denkbaren Abweichungen von einem Rasch-Modell, da sie mit einem saturierten Multinomial-Modell für die Kontingenztabelle vergleicht. Der große Nachteil ist jedoch, dass die Voraussetzungen für eine gute Annäherung an die asymptotische Verteilung im Regelfall nicht gegeben sind. Geht man etwa von einer üblichen Faustregel von mindestens fünf erwarteten Beobachtungen pro Zelleintrag aus und nimmt einen Stichprobenumfang von $n = 1000$ an, so ist (bei einer zudem unrealistischen Annahme der Gleichverteilung auf die Zellen) die maximal denkbare Testlänge auf sieben Items beschränkt. Typischerweise spielen sich Anwendungen jedoch jenseits dieser Größenordnung ab.
Der Ausweg aus dieser Problematik besteht darin, mittels der Matrix $\boldsymbol{U}$ Häufigkeiten zu gruppieren. Hiermit erreicht man eine bessere Annäherung an die χ^2-Verteilung. Im Gegenzug verliert man allerdings die globale Sensitivität der Größe (2.15).
Beispiele für diesen Ansatz bilden die „first“- und „second-order“-Statistiken (Glas, 1988). Die „first-order“-Statistiken vergleichen $n_{hi} := \sum_{v \in h} X_{vi}$, die Anzahl der Per-

sonen, die Item i lösen und einen Summenscore in Kategorie h erreichen, mit dem Erwartungswert unter dem Rasch-Modell. Dies lässt sich formal mittels einer Matrix $\boldsymbol{U}$ auf die Form (2.14) bringen. Während die „first-order"-Statistiken auf die Lösung eines Items fokussiert sind, orientieren sich die „second-order"-Statistiken an dem simultanen Response auf zwei Items. Verglichen wird $m_{ij} := \sum_v X_{vi}X_{vj}$, die Anzahl der Personen, die sowohl Item i als auch Item j lösen, mit dem modellkonformen Erwartungswert. Auch dies lässt sich in die Form (2.14) überführen (siehe z.B. Glas, 1995).

Abschließend bleibt noch zu klären, wie $\boldsymbol{\pi}$ durch das Rasch-Modell spezifiziert ist. Hierbei muss man zwischen MML-Modell (d.h. ein um eine Verteilungsannahme für θ erweitertes Modell) und CML-Modell unterscheiden. Im MML-Modell wird durch den Term $P(\boldsymbol{X_v} = \boldsymbol{x_v})$ aus (2.13) $\boldsymbol{\pi}$ direkt festgelegt.
Im CML-Kontext hingegen liegt eine Produktmultinomialverteilung vor. Vektoren mit gleicher Randsumme folgen einer Multinomialverteilung gemäß (2.11). Die Multinomialverteilungen unterscheiden sich jedoch für die einzelnen Scoregruppen. Um auf die für (2.14) erforderliche Multinomialverteilung zu gelangen, wird daher das bedingte Rasch-Modell um ein saturiertes Modell für die Verteilung des Summenscores erweitert. Das im CML-Kontext betrachtete Modell lautet dann:

$$P(\boldsymbol{X_v} = \boldsymbol{x_v}) = P(\boldsymbol{X_v} = \boldsymbol{x_v} |\, R_v = r_v)P(R_v = r_v)$$

Der erste Term ist durch (2.11) festgelegt. Der zweite Term wird als seperater Modellparameter aufgefasst, d.h. für jede Ausprägung des Summenscores wird ein neuer Parameter eingeführt. Diese Erweiterung des CML-Modells gewährleistet die Anwendbarkeit von (2.14)[10].

Die zweite parametrische Testklasse verwendet ein zweites Modell für die Daten. Das Rasch-Modell stellt ein Submodell dar und lässt sich über „übliche" Likelihood-Quotienten-Tests[11] prüfen.
Bezeichnet $\boldsymbol{\delta_0}$ den Parametervektor des Rasch-Modells, $\boldsymbol{\delta}$ den (erweiterten) Parametervektor des Alternativmodells und $L(\cdot)$ die jeweils korrespondierende Likelihood,

[10] Eine detaillierte Erläuterung dieser Erweiterung geben Glas und Verhelst (1989).
[11] Ebenso denkbar sind Wald- bzw. Score-Tests.

dann ist unter Regularitätsbedingungen die Größe

$$-2\log\left(\frac{\sup_{\boldsymbol{\delta_0}} L(\boldsymbol{\delta_0})}{\sup_{\boldsymbol{\delta}} L(\boldsymbol{\delta})}\right),$$

bei Gültigkeit des Rasch-Modells asymptotisch χ^2-verteilt. Die Freiheitsgrade entsprechen der Parameterdifferenz von Submodell (Rasch-Modell) zu Obermodell.
Ein Test dieser Klasse wird in den Simulationen ab Kapitel 4 näher betrachtet. Es handelt sich um den Likelihood-Quotienten-Test von Andersen (1973). Er dient als parametrische Vergleichsbasis zur Beurteilung der Teststärke der kombinatorischen Testklasse. Diese Wahl erscheint u.a. vor dem Hintergrund, dass er als „Standard“ für das Testen auf Rasch-Modellkonformität vorgeschlagen wird (Kubinger und Draxler, 2007), interessant.
Insofern erfolgt der Vergleich der im Folgenden darzustellenden nonparametrischen Testklasse gegen einen „etablierten“ parametrischen Test.

Kapitel 3

Nonparametrische Rasch-Modell-Tests

3.1 Kombinatorische Tests

Parametrische Tests für das Rasch-Modell sind, wie im vorherigen Abschnitt angedeutet, i.d.R. Score[1]-, Wald- oder Likelihood-Quotienten-Tests gegenüber bestimmten Alternativmodellen. Ihre Verteilung besitzt nur asymptotische Gültigkeit. Das Verhalten für endlichen Stichprobenumfang kann mitunter fragwürdig ausfallen.
So ergeben sich etwa für die R_0-Statistik (Glas und Verhelst, 1989) - eine Größe der Form (2.14) - bei einem formalen α-Fehler-Niveau von 5% tatsächliche Ablehnwahrscheinlichkeiten von teilweise[2] über 30% (Suárez-Falcón und Glas, 2003).
Neben dieser Form von überhöhtem α-Fehler kann jedoch auch das Gegenteil, ein stark konservatives Testniveau, auftreten. In einer Simulationsstudie berichten Verguts und De Boeck (2000) von einem tatsächlichen α-Fehler des Martin-Löf-Tests von 0.2% bei Matrizen der Größenordnung 1000×24 und einem formalen Niveau von 5%. Somit wird ein großer Teil an potentieller Teststärke „verschenkt“. Auch bei einer Erhöhung auf 5000 Personen bleibt ein konservatives Fehlerniveau von 1.2%.
Die folgende nonparametrische Testklasse hat gegenüber den parametrischen Tests

[1] Die „generalized Pearson tests“ können in direkte Beziehung zu speziellen Score-Tests gesetzt werden (Glas, 2007).

[2] Das Fehlerniveau variiert stark nach der Dimension der Datenmatrix.

zwei Vorteile. Zum Einen werden keine Parameterschätzungen benötigt. Als Basis der Teststatistiken dient direkt die Datenmatrix $\boldsymbol{X}$. Zum Anderen besitzen Verteilungsaussagen für endlichen Stichprobenumfang Gültigkeit. Es handelt sich um exakte[3] Tests für prinzipiell beliebigen Stichprobenumfang.
Grundlegend für diese Testklasse sind die suffizienten Statistiken des Rasch-Modells. Gemäß (2.5) gilt:

$$P(\mathbf{X} \,|\, \theta_1, \ldots, \theta_n,\, \beta_1, \ldots, \beta_k) = v(\boldsymbol{\theta}, \boldsymbol{\beta}) \exp\left(\textstyle\sum_v r_v \theta_v - \sum_i c_i \beta_i\right) \tag{3.1}$$

Die Summenscores r_v bilden suffiziente Statistiken für die Fähigkeitsparameter θ_v. Die Spaltensummen c_i fungieren als suffiziente Statistiken der Itemparameter β_i.
Zwei Matrizen mit gleichen Zeilen- sowie Spaltensummen besitzen gemäß (3.1) die gleiche Wahrscheinlichkeit, unter dem Rasch-Modell aufzutreten. Die bedingte Verteilung der Menge aller Matrizen mit gegebenen Randsummen folgt somit einer Gleichverteilung. Dies bedeutet aber nichts anderes, als dass unter der Nullhypothese des Rasch-Modells die (auf Randsummen bedingte) Verteilung jeder Statistik bekannt ist. Somit lässt sich für eine konkrete Statistik $T(\boldsymbol{X})$, die bezüglich des bedingten Stichprobenraums definiert ist, ein Niveau-α-Test konstruieren: Bezeichnet etwa $\boldsymbol{\Sigma_{rc}}$ die Menge der Matrizen mit gleichen Zeilensummen $\boldsymbol{r}$ und gleichen Spaltensummen $\boldsymbol{c}$ wie die beobachtete Datenmatrix, so besteht der Ablehnbereich aus den $(100 \cdot \alpha)\%$ Matrizen aus $\boldsymbol{\Sigma_{rc}}$, die die höchsten (bzw. extremsten) Werte der Teststatistik $T(\cdot)$ aufweisen[4].
Dieser theoretischen Darstellung stehen jedoch praktische Probleme gegenüber. Die Auflistung aller Matrizen mit gleichen Randsummen ist bereits für die überschaubare Größenordnung einer 100×30 Matrix unmöglich. Schon für eine 8×8 Matrix lassen sich Randsummen angeben, für die die Mächtigkeit der korrespondierenden Matrizenmenge im Bereich 10^9 liegt (Ponocny, 2001).
Da die explizite Angabe des Ablehnbereichs somit misslingt, geht man zu stochastischen Approximationen des p-Werts über. Aus der Literatur sind mehrere Algorithmen zur Simulation des p-Werts bekannt. Während Chen und Small (2005) mit Im-

[3]Das Wort „exakt“ wird im Folgenden jedoch eine Relativierung erfahren, da die Tests zur Angabe des p-Werts auf einer Monte-Carlo-Approximation beruhen.

[4]Diese *bedingte* Formulierung der Teststatistik lässt sich leicht zu einer korrespondierenden *unbedingten* Formulierung ausweiten. Der Ablehnbereich des *unbedingten* Tests besteht aus der Vereinigung aller *bedingten* Ablehnbereiche. Die Vereinigung erfolgt dabei über alle möglichen Ausprägungen der suffizienten Statistik.

portance Sampling arbeiten, benutzen Ponocny (2001) und Verhelst (2008) MCMC-Methoden. Die Grundlage der MCMC-Methoden besteht in der Generierung von Matrizen aus $\boldsymbol{\Sigma_{rc}}$ gemäß einer einfach zu simulierenden und theoretisch geeigneten Verteilung. Ein Satz bezüglich des asymptotischen Verhaltens einer Markov-Kette erlaubt dann die Approximation des p-Werts anhand der generierten Matrizen.
Im Folgenden wird der MCMC-Algorithmus von Verhelst (2008) näher erläutert. Nach der Darlegung eines zentralen, zur Approximation relevanten Satzes der MCMC-Theorie erfolgt im ersten Schritt die Generierung einer irreduziblen Markov-Kette[5] mit stationärer Verteilung. Eine Korrektur - mittels Metropolis-Hastings-Algorithmus - ermöglicht im zweiten Schritt die Modifikation der Markov-Kette, so dass sich eine Gleichverteilung als stationäre Verteilung ergibt.

MCMC-Theorie

Der Zustandsraum besteht aus der Menge $\boldsymbol{\Sigma_{rc}}$ aller (binären) Matrizen mit gleichen Randsummen wie die beobachtete Matrix $\boldsymbol{A_0}$. Die konkrete Übergangsmatrix wird im ersten Schritt (siehe nächster Abschnitt) erläutert. Zentral für die Approximation des p-Werts ist der folgende Satz (Robert und Casella, 2004, S.241):

Satz 1. *Besitzt eine Markov-Kette $(X_n)_{n\in N}$ ein invariantes, σ-finites Maß π, dann sind die folgenden Aussagen äquivalent:*

- *Konvergenz der empirischen Mittelwerte, d.h.*

$$f, g \in L^1(\pi),\ \int g \,\mathrm{d}\pi \neq 0 \quad \Rightarrow \quad \lim_{n\to\infty} \frac{S_n(f)}{S_n(g)} = \frac{\int f \mathrm{d}\pi}{\int g \mathrm{d}\pi}$$

- *Die Markov-Kette ist (Harris-)rekurrent.*

$S_n(f)$ bezeichnet hierbei den empirischen Mittelwert $n^{-1}\sum_i f(X_i)$.
Eine rekurrente Markov-Kette mit invarianter Verteilung π erlaubt folglich die Approximation des bezüglich π gebildeten Erwartungswertes $\int f \mathrm{d}\pi$ anhand des Mittelwerts $S_n(f)$.

[5] Für eine Erläuterung der generellen Theorie einer Markov-Kette sei auf Robert und Casella (2004, Kapitel 6) verwiesen.

Im vorliegenden Fall handelt es sich um einen diskreten, endlichen Zustandsraum. Ein Element des Zustandsraumes ist eine Matrix $\boldsymbol{A} \in \boldsymbol{\Sigma_{rc}}$. Die (angestrebte) stationäre Verteilung π entspricht der Gleichverteilung auf $\boldsymbol{\Sigma_{rc}}$. f steht für die Indikatorfunktion I_M, welche angibt, ob der Wert der Teststatistik größer als der beobachtete Wert $T(\boldsymbol{A_0})$ der Ausgangsmatrix $\boldsymbol{A_0}$ ausfällt. g nimmt den konstanten Wert 1 an.
Der Satz impliziert somit: Wenn die Markov-Kette rekurrent ist und die Gleichverteilung als stationäre Wahrscheinlichkeitsverteilung besitzt, dann approximiert der empirische Mittelwert

$$S_n(I_M) = n^{-1} \sum_i I_M(\boldsymbol{X_i}), \quad M := \{\boldsymbol{A} \in \boldsymbol{\Sigma_{rc}} \,|\, T(\boldsymbol{A}) \geq T(\boldsymbol{A_0})\},$$

den p-Wert (d.h. den Anteil aller Matrizen $\boldsymbol{A} \in \boldsymbol{\Sigma_{rc}}$ mit nicht geringerem Wert der Teststatistik $T(\boldsymbol{A})$ als die Ausgangsmatrix $\boldsymbol{A_0}$)

$$p = \int I_M \mathrm{d}\pi = \sum_{\boldsymbol{A} \in M} \frac{1}{|\boldsymbol{\Sigma_{rc}}|},$$

mit wachsendem n beliebig genau.

Erster Schritt: Konstruktion einer rekurrenten Markov-Kette

Um die Aussagen des Satzes anwenden zu können, verbleibt noch die Konstruktion einer rekurrenten Markov-Kette, deren stationäre Verteilung durch die Gleichverteilung auf $\boldsymbol{\Sigma_{rc}}$ gegeben ist. Dies erfodert die Angabe einer bedingten Übergangswahrscheinlichkeit. Gegeben, die Markov-Kette ist im Zustand $\boldsymbol{A_k}$, wird der neue Zustand, d.h. eine neue Matrix, gemäß einer bestimmten Wahrscheinlichkeitsverteilung auf $\boldsymbol{\Sigma_{rc}}$ bestimmt. Für die Spezifikation dieser bedingten Übergangswahrscheinlichkeit sind zunächst einige Begriffe sowie Definitionen erforderlich.
Grundlegend für die Markov-Kette ist der Begriff der Binomialtransformation einer Matrix innerhalb eines Spaltenpaares (i, j), genannt $\mathbf{B_{ij}}$-Transformation[6]:

Die **$\mathbf{B_{ij}}$-Transformation** (Verhelst, 2008) ist eine Transformation, die alle Spalten außer die i-te und j-te Spalte unverändert lässt. Innerhalb des Spaltenpaares (i, j) sei m_{ij} die Anzahl Zeilen mit Zeilensumme eins. Davon sei in a_{ij} Fällen die 1 in der i-ten Spalte. Dann besteht die B_{ij}-Transformation darin, die a_{ij} Einsen der i-ten

[6]Im Folgenden sei stets $i < j$, d.h. Spaltenpaare sind geordnet. Das Paar (3, 1) ist z.B. nicht definiert.

Spalte auf die m_{ij} verfügbaren Zeilen neu zu verteilen. Die restlichen $b_{ij} = m_{ij} - a_{ij}$ Einträge der i-ten Spalte erhalten den Wert Null. Die j-te Spalte wird so ergänzt, dass sich jeweils wieder Zeilensumme 1 (bezogen auf das Spaltenpaar (i, j)) ergibt.
Inhaltlich handelt es sich bei m_{ij} um die Anzahl aller Personen, die genau ein Item innerhalb des Itempaares (i, j) lösen. Hiervon lösen a_{ij} das i-te Item und b_{ij} das j-te Item. Durch die Transformation werden somit aus der Menge aller Personen, die genau ein Item innerhalb des Itempaares (i, j) lösen, a_{ij} „neue“ Personen ausgewählt, die Item i lösen. „Neu“ bezieht sich hierbei auf die Forderung, dass die Auswahl *mindestens eine* neue Person beinhalten soll. Nach einer B_{ij}-Transformation wurde folglich für mindestens zwei der m_{ij} Personen mit Zeilensumme eins in dem Spaltenpaar (i, j) das Itempattern vertauscht. Insbesondere ändern sich die Randsummen nicht, d.h. die resultierende Matrix ist wiederum ein Element von $\boldsymbol{\Sigma_{rc}}$.
Anhand einer 5×3-Matrix $\boldsymbol{A_1}$ soll ein kurzes Beispiel für eine B_{ij}-Transformation gegeben werden:

$$\boldsymbol{A_1} := \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \quad \xrightarrow{\boldsymbol{B_{13}}} \quad \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} =: \boldsymbol{A_2}$$

Betrachtet man etwa das Spaltenpaar (1, 3), so gibt es $m_{13} = 4$ Personen mit Zeilensumme eins. In zwei Fällen (Personen 2 und 5) wurde hiervon das erste Item gelöst. An Stelle der „alten“ Personenwahl kann nun eine „neue“ Auswahl erfolgen. Eine Möglichkeit - als Matrix $\boldsymbol{A_2}$ realisiert - besteht in der Wahl von Person 3 und 5. Insgesamt existieren in diesem Beispiel $\binom{4}{2} - 1$ zulässige B_{13}-Transformationen.
Die Menge aller Matrizen, die durch die B_{ij}-Transformation aus der Matrix $\boldsymbol{A}$ entstehen können, wird im Folgenden mit $B_{ij}(\boldsymbol{A})$ bezeichnet. Im obigen Beispiel gilt somit $\boldsymbol{A_2} \in B_{13}(\boldsymbol{A_1})$. Für das Spaltenpaar (1, 2) gilt hingegen $B_{12}(\boldsymbol{A_1}) = \emptyset$. Als allgemeine Eigenschaften der sogenannten „Nachbarschaft“ $B_{ij}(\boldsymbol{A})$ lassen sich leicht die folgenden Aussagen - für beliebige $\boldsymbol{A}, \boldsymbol{A_1}, \boldsymbol{A_2} \in \boldsymbol{\Sigma_{rc}}$ - nachprüfen:

i) $\forall (i,j) : \ \boldsymbol{A} \notin B_{ij}(\boldsymbol{A})$

ii) $\forall\, (i,j) \neq (k,l) : \ B_{ij}(\boldsymbol{A}) \cap B_{kl}(\boldsymbol{A}) = \emptyset$

iii) $\boldsymbol{A}_1 \in B_{ij}(\boldsymbol{A}_2) \Leftrightarrow \boldsymbol{A}_2 \in B_{ij}(\boldsymbol{A}_1)$

iv) $|B_{ij}(\boldsymbol{A})| = \binom{m_{ij}}{a_{ij}} - 1$

v) $\boldsymbol{A_1} \in B_{ij}(\boldsymbol{A_2}) \Rightarrow |B_{ij}(\boldsymbol{A_1})| = |B_{ij}(\boldsymbol{A_2})|$

Weiterhin werden zwei Definitionen als Basis für den Algorithmus, der die Übergangswahrscheinlichkeit der Markov-Kette festlegt, benötigt:

Definition 1. Das Spaltenpaar (i, j) heißt regulär, wenn $a_{ij} \cdot b_{ij} > 0$

Nur bei regulären Spaltenpaaren ist es möglich, Binomialtransformationen durchzuführen. Für die Beispielsmatrix $\boldsymbol{A_1}$ ist das Paar (1, 3) das einzige reguläre Spaltenpaar.
Basierend auf dem Begriff eines regulären Spaltenpaares wird ein „Maß“ für die Menge aller Matrizen mit gleichen Randsummen eingeführt. Es entspricht der Anzahl regulärer Spaltenpaare der betreffenden Matrix.

Definition 2. Das k-Maß, definiert für $\boldsymbol{A} \in \boldsymbol{\Sigma_{rc}}$, ist festgelegt durch:
$k(\boldsymbol{A}) := |\{(i,j) \,|\, (i,j) \text{ ist reguläres Spaltenpaar von } \boldsymbol{A}\}|$

Mit Hilfe des Begriffs der Binomialtransformation sowie des k-Maßes ist es nun möglich, die Entwicklung der Markov-Kette zu kennzeichnen. Der Übergang vom Zustand zum Zeitpunkt t (Matrix $\boldsymbol{A}_k$) zu einem Zustand zum Zeitpunkt $t+1$ (Matrix $\boldsymbol{A}_l$) wird durch folgenden Algorithmus determiniert:

- Wähle ein reguläres Spaltenpaar (i, j) von $\boldsymbol{A}_k$ gemäß einer Gleichverteilung auf der Menge der regulären Spaltenpaare von $\boldsymbol{A_k}$.
- Führe eine zufällige Binomialtransformation innerhalb des gewählten Spaltenpaares durch. Dies liefert den neuen Zustand $\boldsymbol{A_l}$.

Damit ergibt sich die einschrittige Übergangsmatrix zu:

$$p_{\boldsymbol{A}_k \boldsymbol{A}_l} = \frac{w_{\boldsymbol{A}_k \boldsymbol{A}_l}}{k(\boldsymbol{A}_k)}$$

$$w_{\boldsymbol{A}_k\boldsymbol{A}_l} = \begin{cases} \left(\binom{m_{ij}(\boldsymbol{A}_k)}{a_{ij}(\boldsymbol{A}_k)} - 1\right)^{-1} & \text{falls} \quad \boldsymbol{A}_l \in B_{ij}(\boldsymbol{A}_k) \\ 0 & \text{sonst} \end{cases}$$

Die Definition von $w_{\boldsymbol{A}_k\boldsymbol{A}_l}$ ist stets eindeutig aufgrund der Eigenschaft $ii)$ der Binomial-Nachbarschaft.
Zur Berechnung der stationären Verteilung bezeichne im Folgenden $\boldsymbol{k}$ den Vektor mit Komponenten $k_u := k(\boldsymbol{A}_u)$. Ferner wird die Notation von Matrizen entkoppelt, so dass „u“ nun für die Matrix „$\boldsymbol{A}_u$“ steht. $\boldsymbol{P}$ bezeichne die oben definierte Übergangsmatrix der Markov-Kette.

Lemma 1 (Invariantes Maß)**.** $\boldsymbol{k'P} = \boldsymbol{k'}$

Beweis. Die v-te Spalte der linken Seite ist:

$$\sum_u p_{uv}k_u = \sum_u \frac{w_{uv}k_u}{k_u} = \sum_u w_{uv} \overset{iii)\ v)}{=} \sum_u w_{vu} \overset{ii)}{=} \sum_{(i,j)} \sum_{\boldsymbol{A_u} \in B_{ij}(\boldsymbol{A_v})} \left(\binom{m_{ij}}{a_{ij}} - 1\right)^{-1} \overset{iv)}{=} k_v$$

□

Dies entspricht der v-ten Spalte der rechten Seite. Somit besitzt die Markov-Kette eine zum k-Maß proportionale stationäre Verteilung.
Des Weiteren liegt Irreduzibilität vor. Jede Matrix $\boldsymbol{A} \in \boldsymbol{\Sigma_{rc}}$ lässt sich durch eine endliche Folge von Binomialtransformationen aus einer Matrix $\boldsymbol{B} \in \boldsymbol{\Sigma_{rc}}$ gewinnen. Dies kann durch Widerspruchsbeweis gezeigt werden (Ponocny, 2001). Da der Zustandsraum zudem endlich ist, folgt die Rekurrenz der Markov-Kette (Grimmett und Stirzaker, 2001a, S.225). Damit sind die Voraussetzungen des Satzes erfüllt.
Jedoch liefert dies noch kein zufriedenstellendes Resultat. Das (normierte) k-Maß kann mitunter deutlich von einer Gleichverteilung abweichen. Um dies zu korrigieren, d.h. um eine Markov-Kette mit der Gleichverteilung als invariantes Maß zu erreichen, wird im zweiten Schritt die Übergangsmatrix $\boldsymbol{P}$ als Vorschlagsdichte in einem Metropolis-Hastings-Algorithmus verwendet.

Zweiter Schritt: Konstruktion einer stationären Gleichverteilung

Wie sich anhand des Lemmas erkennen lässt, weisen Matrizen mit größerem k-Maß eine höhere Wahrscheinlichkeit - gemäß der stationären Verteilung - auf. Die Anwendung von Satz 1 liefert einen p-Wert bezüglich der durch das k-Maß gegebenen Verteilung. Variiert das k-Maß beträchtlich, so ergeben sich deutliche Abweichungen von der Gleichverteilung, die in eine verzerrte Schätzung des p-Werts münden.
Benutzt man die Übergangsmatrix $\boldsymbol{P}$ hingegen lediglich als „Vorschlagsdichte“ in einem Metropolis-Hastings-Algorithmus, so erreicht man die stationäre Gleichverteilung mittels des inhärenten Akzeptanzschrittes (Robert und Casella, 2004, S.272).

Gegeben, die Markov-Kette befindet sich zum Zeitpunkt t im Zustand $\boldsymbol{A_k}$, lautet der endgültige Algorithmus somit:

- Ziehe ein $\boldsymbol{A} \in \boldsymbol{\Sigma_{rc}}$ gemäß der durch die k-te Zeile von $\boldsymbol{P}$ gegebenen Verteilung (d.h. unter Benutzung des Algorithmus aus Schritt 1).
- Akzeptiere $\boldsymbol{A}$ mit Wahrscheinlichkeit

$$\alpha(\boldsymbol{A_k}, \boldsymbol{A}) := \min\left(1, \frac{p_{\boldsymbol{A}\boldsymbol{A_k}}}{p_{\boldsymbol{A_k}\boldsymbol{A}}}\right) = \min\left(1, \frac{k(\boldsymbol{A_k})}{k(\boldsymbol{A})}\right)$$

- Falls $\boldsymbol{A}$ akzeptiert wurde, geht die Markov-Kette zum Zustand $\boldsymbol{A}$ über. Ansonsten verbleibt sie in $\boldsymbol{A_k}$.

Die auf diese Weise generierte Markov-Kette wird nach einem gewissen Zeitpunkt abgebrochen. Der Anteil an Zuständen der Kette mit nicht geringerem Wert der Teststatistik als der Ausgangswert der beobachteten Matrix dient als Schätzung des p-Werts. Meist wird zuvor der erste Teil (burn-in) der simulierten Matrizen aus Gründen der Stationarität ausgeschlossen. Um die Abhängigkeiten zwischen den Matrizen der Markov-Kette zu reduzieren, erfolgt zudem oft ein Sampling bezogen auf jede l-te Matrix (step $= l$), d.h. nur jede l-te simulierte Matrix wird betrachtet. Die $l - 1$ dazwischen liegenden Matrizen bleiben ungeachtet.

Bemerkungen zur Effizienz

Auch wenn der Metropolis-Hastings-Algorithmus unter relativ schwachen Bedingungen (siehe z.B. Robert und Casella, 2004, Kapitel 7) die Konvergenz der empirischen Mittelwerte sichert, impliziert dies nicht, dass ein praktisch brauchbarer Algorithmus resultiert. Die Konvergenz ist zunächst rein formal zu sehen. Der Konvergenzsatz liefert keine anwendbare Abbruchregel, nach der die Markov-Kette zu stoppen wäre. Ferner hängt die Konvergenzgeschwindigkeit von der konkret gewählten Teststatistik $T(.)$ ab. Insofern lässt sich nur durch Simulation ein Einblick in die Effizienz des Algorithmus geben.
Verhelst (2008) variiert die Länge des burn-in sowie die Schrittweite der Kette und stellt für vier verschiedene Teststatistiken - bei Rasch-homogenen 300×30 Matrizen - einen deutlichen Effizienzgewinn (gemessen an der Standardabweichung des geschätzten p-Werts bei wiederholtem Ablauf des Algorithmus) gegenüber dem Importance Sampling-Schätzer von Chen und Small (2005) fest. Die durchschnittlichen p-Werte der beiden Methoden weisen zudem keine nennenswerten Abweichungen auf. In Anhang A ist der Einfluss der Schrittweite sowie der Anzahl simulierter Matrizen näher beschrieben. Hier finden sich Angaben zum Monte-Carlo-Fehler am Beispiel einer 1000×10 Rasch-homogenen Datenmatrix.

Umsetzung des Algorithmus

Der Algorithmus ist im Paket *RaschSampler* (Verhelst u.a., 2007) der Software R implementiert. Das Paket bildet die Grundlage für die Simulationen des fünften Kapitels. Die konkrete Wahl der Tuningparameter (step, burn-in, Anzahl simulierter Matrizen) wird im fünften Kapitel sowie in Anhang A näher behandelt. Hier seien abschließend nur die verschiedenen Parameter und ihre Bedeutung aufgeführt:

- **burn-in**: Anzahl Matrizen, die nicht in den Mittelwert eingehen. Erst ab $t = \text{burn-in} + 1$ wird das empirische Mittel gebildet[7].
- **step**: Schrittweite der Markov-Kette. Von den simulierten Matrizen wird nur jede l-te Matrix (falls step $= l$) betrachtet. Man hofft, so die Abhängigkeiten

[7]Dies gilt nur für step $= 1$. Falls step ≥ 2 gilt, muss der burn-in-Parameter mit der Schrittweite multipliziert werden, um den tatsächlichen burn-in zu erhalten. Siehe auch Verhelst u.a. (2007).

der Markov-Kette zu reduzieren. Dies kann mitunter näher an den Fall unabhängiger Ziehungen aus $\boldsymbol{\Sigma_{rc}}$ gelangen.

- $\boldsymbol{n}_{\text{eff}}$: Anzahl Matrizen, die zur Mittelwertsbildung verwendet werden.

Teststatistiken

In den bisherigen Ausführungen wurde die spezifische Form der Teststatistik bewusst offen gelassen. Der Algorithmus ist von der konkreten Gestalt der Statistik unabhängig[8]. Um einen kurzen Überblick über die zahlreich möglichen Testgrößen $T(\boldsymbol{X})$ zu geben, werden im Folgenden - orientiert an Ponocny (2001) - einige Teststatistiken im Kontext spezieller Annahmeverletzungen erläutert:

- **Erhöhte Itemdiskrimination** lässt sich anhand des klassischen Maßes der Item-Test-Korrelation (Trennschärfe) erfassen. Items mit höherer Diskrimination repräsentieren das Konstrukt „besser“ und weisen dahingehend eine erhöhte Korrelation mit dem Gesamttestwert (Summenscore) auf.
  Eine adäquate Teststatistik ist daher durch $T(\boldsymbol{X}) = \text{Cor}(\boldsymbol{x_{(i)}}, \boldsymbol{r})$ gegeben. Der p-Wert ergibt sich als Anteil aller generierten Matrizen, die bezüglich des i-ten Items eine nicht geringere Trennschärfe aufweisen als die beobachtete Matrix. Sollte die Richtung, in der die Itemdiskrimination abweicht, unbekannt sein, ist auch ein zweiseitiger Test denkbar. Im Ablehnbereich befinden sich alle Matrizen, für die der Trennschärfewert extreme Werte annimmt.
  Globale Teststatistiken bezüglich zum Rasch-Modell diskrepanter Itemdiskrimination basieren auf allen Item-Test-Korrelationen. Ein Beispiel ist der in Kapitel 4 näher beschriebene Test von Chen und Small (2005).

- **Personenspezifische Tests** erfüllen den Zweck, Tendenzen, die die Gültigkeit eines Testergebnisses beeinträchtigen, wie z.B. Rateverhalten, aufzudecken. Eine mögliche Prüfgröße beruht auf einer Gegenüberstellung der Summenwerte zweier Testhälften, die sich von der Schwierigkeit stark unterscheiden: $T(\boldsymbol{X}) = \sum_{i \in A} X_{vi} - \sum_{i \notin A} X_{vi}$[9]. „$A$“ bezeichnet hierbei die $k/2$ leichtesten

[8] Das Konvergenzverhalten kann sich jedoch je nach Teststatistik unterscheiden. Siehe auch Anhang A.

[9] Da die Randsumme konstant ist, lässt sich diese Statistik auch durch $-\sum_{i \notin A} X_{vi}$ ersetzen.

Items. Reines Rateverhalten führt zu einem niedrigen Wert der Teststatistik, weil Items unabhängig von ihrer „wirklichen“ Schwierigkeit beantwortet werden. Für Personen, deren Antwortmuster hingegen dem Rasch-Modell folgt, ergeben sich, da leichtere Items auch häufiger gelöst werden als schwerere Items, höhere Werte der Teststatistik. Folglich lehnt der Test auf „Rateverhalten“ ab, wenn höchstens $(100 \cdot \alpha)\%$ der simulierten Matrizen einen nicht höheren Wert $T(\boldsymbol{X})$ aufweisen als die beobachtete Matrix.

Personen-bezogene Tests dieser Form besitzen einen Vorteil gegenüber den „üblichen“ Personen-orientierten Tests. Diese basieren auf der Annahme, dass es sich bei den geschätzten Itemparametern $\hat{\beta}_i$ um die wahren Größen β_i handelt, d.h. zur Beurteilung des „Personen-Fits“ werden bekannte Itemparameter vorausgesetzt (siehe z.B. Klauer, 1995). Somit erfolgt keine Korrektur der Teststatistik gemäß der Variabilität der Itemparameterschätzung. Der kombinatorische Test hingegen ist auf die Kenntnis der wahren β_i nicht angewiesen.

- Mehrdimensionalität in Form von **Differential-Item-Functioning** (DIF) liegt vor, wenn für Personen gleicher Fähigkeit die Wahrscheinlichkeit, das Item zu lösen, nach Gruppenzugehörigkeit (z.B. Geschlecht) variiert. Ein Test auf DIF bei Item i verwendet $\sum_{v \in H_1} X_{vi}$, die Anzahl der Personen aus der ersten Gruppe (H_1), die das i-te Item lösen. Ein zu hoher Wert spricht für Gruppe eins bevorzugendes DIF.

  Der globale DIF-Test beruht auf Odds-Ratios. Bei Gültigkeit des Rasch-Modells sowie Abwesenheit von DIF resultiert für jede Personengruppe bezüglich des Tests, der nur aus den Items i und j besteht:

$$P(X_{vi} = 1 |\, R_v = 1, H = h(v)) = \frac{\exp(\beta_j - \beta_i)}{1 + \exp(\beta_j - \beta_i)}$$

  bzw:

$$\frac{P(X_{vi} = 1 |\, R_v = 1, H = h(v))}{P(X_{vi} = 0 |\, R_v = 1, H = h(v))} = \exp(\beta_j - \beta_i) \tag{3.2}$$

  Der Quotient n_{ijh}/n_{jih} (siehe Notation) liefert eine konsistente Schätzung der linken Seite von (3.2) in Gruppe h. Bei Modellgültigkeit, d.h. wenn (3.2) gilt, sollten die Schätzungen in den Gruppen ähnlich ausfallen, da (3.2) unabhängig von h ist. Als globale Teststatistik - am Beispiel zweier Gruppen - dient somit:

$$T(\boldsymbol{X}) = \sum_{i,j} |n_{ij1} n_{ji2} - n_{ij2} n_{ji1}| \tag{3.3}$$

- **Suffizienz des Summenscores** lässt sich durch leichte Modifikation der obigen Teststatistik prüfen. An Stelle einer externen Gruppierungsvariable wird lediglich die interne Variable „Summenscore“ zur Gruppenbildung verwendet.

- **Unregelmäßige Verläufe der Item-Response-Funktionen**, wie z.B. Items mit maximaler Lösungswahrscheinlichkeit bei durchschnittlicher Fähigkeit, lassen sich z.B. anhand $T(\boldsymbol{X}) = \sum_{v \in A} X_{vi}$ erkennen. Die Summation verläuft über alle Personen, deren Summenscore nahe dem Median verläuft.

- **Unabhängigkeit des Antwortverhaltens** zweier Personen lässt sich anhand der Anzahl übereinstimmender Antworten $T(\boldsymbol{X}) = \sum_i I_{\{X_{vi}=X_{wi}\}}$ evaluieren.

- **Lokale stochastische Abhängigkeit** manifestiert sich in dem Korrelationsmuster der Items. Eine globale Teststatistik, die in Kapitel 4 näher betrachtet wird, ist $T(\boldsymbol{X}) = \sum_{i,j} |r_{ij} - \rho_{ij}|$. Die Abweichungen der paarweisen Itemkorrelationen $r_{ij} := \mathrm{Cor}(\boldsymbol{x_{(i)}}, \boldsymbol{x_{(j)}})$ von ihrem Erwartungswert ρ_{ij} bezüglich der Gleichverteilung auf $\boldsymbol{\Sigma_{rc}}$ werden aufsummiert. Einzelne Itempaare sind analog über $T(\boldsymbol{X}) = (r_{ij} - \rho_{ij})^2$ prüfbar.

Diese Aufzählung ließe sich beliebig fortsetzen. Jede Funktion $T(\boldsymbol{X})$ der Datenmatrix bildet eine „gültige“ Teststatistik. Denkbar sind Statistiken, die auf den Einträgen nur einer Person basieren (Personen-orientierte Tests), ebenso wie komplexe Funktionen - z.B. Eigenwerte der Korrelationsmatrix - der Datenmatrix. Entscheidend für die Auswahl einer Statistik ist dabei - neben einer aussagekräftigen Interpretationsfähigkeit - die Teststärke. Trotz der vielfältigen Möglichkeiten, eine Statistik zu formulieren, liegen bisher nur wenige Ergebnisse hierzu vor. Einige Resultate finden sich z.B. bei Ponocny (2001), Suárez-Falcón und Glas (2003), Chen und Small (2005) sowie Christensen und Kreiner (2010). Aus diesen Untersuchungen ergeben sich vielversprechende Implikationen. So konnten etwa Chen und Small (2005) sowie Christensen und Kreiner (2010) jeweils verschiedene Teststatistiken bezüglich variabler Itemdiskrimination formulieren, die gegenüber „gängigen“ parametrischen Tests einen deutlichen Zuwachs an Teststärke aufweisen. Dieser Vergleich mit parametrischen Alternativen sowie die weitere Exploration des Potentials einiger kombinatorischer Teststatistiken bilden den Fokus von Kapitel 5. Die schwer zugängliche Verteilung von $T(\boldsymbol{X})$, hervorgerufen u.a. durch die enorme Größe von $\boldsymbol{\Sigma_{rc}}$, macht einen simulativen Zugang zwingend erforderlich.

Im Gegensatz hierzu führen die folgenden beiden Testklassen auf bekannte Teststatistiken für kategoriale Daten. Ferner sind sie in einem breiteren Kontext, der über das (dichotome) Rasch-Modell hinausgeht, gültig.

3.2 Mantel-Haenszel-Tests

Der Mantel-Haenszel-Test ist ein Test auf bedingte Unabhängigkeit in einer dreidimensionalen Kontingenztabelle. Bevor die spezielle Verbindung zum Rasch-Modell mittels des nachfolgenden Satzes hergestellt wird, sei hier zunächst die grundlegende Theorie des Mantel-Haenszel-Tests skizziert.
Für diese Zwecke seien X sowie Y zwei binäre Zufallsvariablen und Z eine mehrkategoriale Zufallsvariable mit K Stufen. Bedingte Unabhängigkeit von X und Y gegeben Z - die Nullhypothese des Tests - liegt vor, wenn für alle i, j, k gilt:

$$P(X = i, Y = j \,|\, Z = k) = P(X = i \,|\, Z = k)P(Y = j \,|\, Z = k)$$

Bzw. wenn für alle k das zur k-ten Schicht korrespondierende Odds-Ratio nicht von eins abweicht. Aus bedingter Unabhängigkeit folgt nicht marginale Unabhängigkeit. Sind X und Y in ihrer gemeinsamen, unbedingten Verteilung abhängig, so impliziert bedingte Unabhängigkeit, dass die Assoziation der beiden Zufallsvariablen bei Kontrolle von Z verschwindet.
Eine Überpüfung der Nullhypothese der bedingten Unabhängigkeit kann für verschiedene Stichprobenschemata (multinomial, produktmultinomial) durch den Mantel-Haenszel-Test erfolgen (Birch, 1964).
n_{11k}[10] bezeichne die Anzahl Beobachtungen für die $X = 1, Y = 1$ und $Z = k$ gilt. $\pi_{j|ik}$ bezeichne des Weiteren die bedingte Wahrscheinlichkeit $P(Y = j \,|\, X = i,\, Z = k)$.
Bei festen Rändern von X und Z, d.h. bei einem produktmultinomialen Design[11], in dem für fest vorgegebene Anzahlen $n_{1.k}$, $n_{0.k}$ lediglich Y als Response erhoben wird, resultieren zwei unabhängige Binomialverteilungen innerhalb der k-ten Schicht. Da Beobachtungen aus unterschiedlichen Schichten ferner unabhängig erhoben werden, genügt es zunächst, die Verteilung innerhalb einer konkret gewählten Schicht k zu

[10] In analoger Form seien n_{01k}, n_{00k}, n_{10k} definiert. Ferner deute ein Punkt die Summation über den entsprechenden Index an.

[11] Durch entsprechendes Bedingen auf X und Z können die übrigen Stichprobenschemata wie z.B. das multinomiale Design auf diesen Fall zurückgeführt werden.

betrachten. Hier ergibt sich für (n_{11k}, n_{01k}) die folgende Produktbinomialverteilung:

$$P(n_{11k}, n_{01k}) = \binom{n_{1.k}}{n_{11k}} \pi_{1|1k}^{n_{11k}} (1-\pi_{1|1k})^{n_{10k}} \binom{n_{0.k}}{n_{01k}} \pi_{1|0k}^{n_{01k}} (1-\pi_{1|0k})^{n_{00k}} \tag{3.4}$$

Wenn gegeben $Z = k$ die Assoziation von X und Y verschwindet, dann unterscheiden sich die beiden bedingten Verteilungen nicht:

$$P(Y = j|\, X = 1,\, Z = k) = P(Y = j|\, X = 0,\, Z = k)$$

Folglich gilt $\pi_{1|1k} = \pi_{1|0k}$ und (3.4) vereinfacht sich zu:

$$P(n_{11k}, n_{01k}) = \binom{n_{1.k}}{n_{11k}} \binom{n_{0.k}}{n_{01k}} \pi_{1|1k}^{n_{.1k}} (1-\pi_{1|1k})^{n_{.0k}}$$

Durch Bedingen auf $n_{.1k}$, welches als Summe zweier unabhängiger Binomialverteilungen (mit dem gleichen Wahrscheinlichkeitsparameter) wiederum einer Binomialverteilung folgt, ergibt sich:

$$P(n_{11k}|\, n_{.1k}) = \frac{\binom{n_{1.k}}{n_{11k}} \binom{n_{0.k}}{n_{01k}} \pi_{1|1k}^{n_{.1k}} (1-\pi_{1|1k})^{n_{.0k}}}{\binom{n_{..k}}{n_{.1k}} \pi_{1|1k}^{n_{.1k}} (1-\pi_{1|1k})^{n_{.0k}}} = \frac{\binom{n_{1.k}}{n_{11k}} \binom{n_{0.k}}{n_{01k}}}{\binom{n_{..k}}{n_{.1k}}}$$

Dies stellt eine hypergeometrische Verteilung mit bekannten Parametern dar. Unter Berücksichtigung der Unabhängigkeit der Daten aus unterschiedlichen Schichten resultiert somit - bedingt auf alle Randsummen der k-ten Schicht $(\forall k)$ - eine produkthypergeometrische Verteilung für den Vektor $\boldsymbol{n} = (n_{111}, n_{112} \ldots n_{11K})^T$.

Die Teststatistik des Mantel-Haenszel-Tests lautet[12]:

$$MH = \frac{\left(\sum_k (n_{11k} - \mu_{11k})\right)^2}{\sum_k \sigma_{11k}^2} \tag{3.5}$$

Die Prüfgröße lehnt ab, wenn der beobachtete Wert größer ist als das $(1-\alpha)$-Quantil einer χ^2-Verteilung mit einem Freiheitsgrad. Die Asymptotik dieser Prüfgröße ist auch bei geringen Besetzungen innerhalb der Schichten gültig, solange die Anzahl an Schichten „genügend groß“ ausfällt (Agresti, 2002, S.233). Alternativ ist auch eine exakte Bestimmung des p-Werts über die hypergeometrische Verteilung möglich.
Wie anhand (3.5) ersichtlich, fällt die Prüfgröße hoch aus, wenn in jeder Schicht $n_{11k} > \mu_{11k}$ (oder in jeder Schicht $n_{11k} < \mu_{11k}$) gilt. Abweichungen $(n_{11k} - \mu_{11k})$

[12] μ_{11k}, σ_{11k}^2 stellen hierbei Erwartungswert bzw. Varianz von n_{11k} gemäß der hypergeometrischen Verteilung dar.

unterschiedlichen Vorzeichens in den Schichten kompensieren sich dagegen. Folglich besitzt die Teststatistik nur dann Power, wenn das schichtspezifische Odds-Ratio stets in gleicher Richtung von eins abweicht. Bei stark schichtspezifisch variierendem Zusammenhang stellt (3.5) keine geeignete Prüfgröße dar. Liegen jedoch annähernd gleichförmige bedingte Assoziationen vor, so kann der Mantel-Haenszel-Test angewandt werden. Insbesondere ergibt sich für die Alternative eines bezüglich der Schicht konstanten Odds-Ratios die Möglichkeit, einen UMP-Test zu konstruieren[13] (Birch, 1964). Fällt die Assoziation somit nicht nur richtungskonform, sondern auch in ihrer Stärke gleichmäßig aus (homogene Assoziation, d.h. in jeder Schicht k nimmt das Odds-Ratio zwischen X und Y den gleichen Wert an), so bildet (3.5) eine besonders geeignete Teststatistik.

Um die Verbindung zum Rasch-Modell herzustellen, ist es zunächst erforderlich, die Notation zu erweitern. Der Responsevektor $\boldsymbol{X}$, der das Testverhalten einer Person beschreibt, wird hierzu partitioniert in $\boldsymbol{X} = (\boldsymbol{W}, \boldsymbol{U})$. Ferner bezeichne H eine binäre Funktion von $\boldsymbol{W}$, d.h. $H = 1$ wenn ein bestimmtes Responseverhalten in dem durch $\boldsymbol{W}$ gegebenen Subtest gezeigt wird. Eine einfache Möglichkeit bildet z.B. die Vorschrift:

$$H = W_i$$

In diesem Fall deutet H lediglich an, ob das i-te Item des Subtests $\boldsymbol{W}$ gelöst wurde. Während H in der „Mantel-Haenszel-Notation“ der binären Variable X entspricht, ist für die Schichtklassifikation, d.h. für die Variable Z, der Summenscore im Subtest $\boldsymbol{U}$ (im Folgenden mit r_2 bezeichnet) zuständig. Alle Personen mit gleichem Summenscore r_2 bilden eine Schicht. Schließlich wird für die Spaltenklassifikation aus der Menge aller Subtestvektoren $\boldsymbol{u}$ mit Summenscore r_2 eine bestimmte Teilmenge $M(r_2)$ ausgewählt. Eine mögliche Wahl wäre beispielsweise:

$$\boldsymbol{u} \in M(r_2) \Leftrightarrow u_j = 1 \text{ und } \textstyle\sum_k u_k = r_2$$

In diesem Fall erfolgt die Klassifikation in Abhängigkeit davon, ob das j-te Item des Subtests $\boldsymbol{U}$ gelöst wurde.
Bezogen auf diese Notation resultiert der folgende Satz (Verguts und De Boeck, 2001):

[13]Der Ablehnbereich wird dabei nicht über die asymptotische Verteilung, sondern über das Neyman-Pearson-Lemma, bezogen auf die Prüfgröße $\sum_k n_{11k}$, bestimmt. Es handelt sich somit um einen exakten Test auf bedingte Unabhängigkeit (Lehmann und Romano, 2005, S.133ff). Der MH-Test kann als gute Approximation dieses exakten Tests betrachtet werden.

Satz 2. *Unter einem Rasch-Modell gilt:*

$$P(\boldsymbol{U} \in M(r_2)|\, r_2, H = 1) = P(\boldsymbol{U} \in M(r_2)|\, r_2, H = 0) \tag{3.6}$$

Beweis. Die linke Seite von (3.6) ergibt sich zu:

$$P(\boldsymbol{U} \in M(r_2)|\, r_2, H = 1) = \sum_{\boldsymbol{u} \in M(r_2)} P(\boldsymbol{u}|\, r_2, H = 1) = \sum_{\boldsymbol{u} \in M(r_2)} \frac{P(\boldsymbol{u}|\, H = 1)}{P(r_2|\, H = 1)} \tag{3.7}$$

Durch Einführen der latenten Variable wird (3.7) zu:

$$\sum_{\boldsymbol{u} \in M(r_2)} \frac{P(\boldsymbol{u}|\, H = 1)}{P(r_2|\, H = 1)} = \sum_{\boldsymbol{u} \in M(r_2)} \frac{\int_{-\infty}^{\infty} P(\boldsymbol{u}|\, \theta, H = 1) \mathrm{d}G(\theta|\, H = 1)}{\int_{-\infty}^{\infty} P(r_2|\, \theta, H = 1) \mathrm{d}G(\theta|\, H = 1)} \tag{3.8}$$

Betrachtet man für einen festen Summanden den Zähler in (3.8) und berücksichtigt man ferner die bedingte Unabhängigkeit von $\boldsymbol{U}$ und $H = H(\boldsymbol{W})$ gegeben θ, so führt dies auf:

$$\int_{-\infty}^{\infty} P(\boldsymbol{u}|\, \theta, H = 1) \mathrm{d}G(\theta|\, H = 1) = \int_{-\infty}^{\infty} P(\boldsymbol{u}|\, \theta) \mathrm{d}G(\theta|\, H = 1) \tag{3.9}$$

Unter Verwendung der konkreten Form der Item-Response-Funktion des Rasch-Modells ergibt sich:

$$\int_{-\infty}^{\infty} P(\boldsymbol{u}|\, \theta) \mathrm{d}G(\theta|\, H = 1) = \int_{-\infty}^{\infty} \frac{\exp(r_2\theta - \sum_k u_k\beta_k)}{\prod_k (1 + \exp(\theta - \beta_k))} \mathrm{d}G(\theta|\, H = 1) \tag{3.10}$$

Somit besitzt der Zähler (eines bestimmten Summanden) aus (3.8) die Gestalt:

$$\exp(-\textstyle\sum_k u_k\beta_k) \displaystyle\int_{-\infty}^{\infty} \frac{\exp(r_2\theta)}{\prod_k (1 + \exp(\theta - \beta_k))} \mathrm{d}G(\theta|\, H = 1) \tag{3.11}$$

Andererseits ergibt sich der Nenner aus (3.8) per Definition - $P(r_2|\, \theta, H = 1) = \sum_{\boldsymbol{u}'\mathbf{1} = r_2} P(\boldsymbol{u}|\, \theta, H = 1)$ - sowie unter Verwendung von (3.11) zu:

$$\sum_{\boldsymbol{u}'\mathbf{1} = r_2} \left(\exp(-\textstyle\sum_k u_k\beta_k) \displaystyle\int_{-\infty}^{\infty} \frac{\exp(r_2\theta)}{\prod_k (1 + \exp(\theta - \beta_k))} \mathrm{d}G(\theta|\, H = 1) \right) \tag{3.12}$$

Nutzt man (3.12) und (3.11) in (3.8), so resultiert[14]:

$$P(\boldsymbol{U} \in M(r_2)|\, r_2, H = 1) = \frac{\sum_{\boldsymbol{u} \in M(r_2)} \exp(-\sum_k u_k\beta_k)}{\sum_{\boldsymbol{u}'\mathbf{1} = r_2} \exp(-\sum_k u_k\beta_k)}$$

Dieser Ausdruck ist unabhängig von H. □

[14] Der Integrand in (3.12) ist für jeden in Betracht kommenden Vektor $\boldsymbol{u}$ identisch.

Tabelle 3.1: *Kreuzklassifikation von X_7 und X_8 innerhalb jeder Stufe des Subtestsummenscores $R_2 = \sum_{k \neq 7} X_k$. Zur Verdeutlichung sind lediglich zwei Stufen dargestellt. Die Indizes beziehen sich auf die Itemposition innerhalb des Würfelaufgaben-Tests des IST 2000 R.*

	$R_2 = 8$				$R_2 = 14$		
	$X_8 = 0$	$X_8 = 1$	Gesamt		$X_8 = 0$	$X_8 = 1$	Gesamt
$X_7 = 0$	5	3	8	$X_7 = 0$	2	3	5
$X_7 = 1$	5	4	9	$X_7 = 1$	5	13	18
Gesamt	10	7	17	Gesamt	7	16	23

Wenn A das Ereignis „$\boldsymbol{U} \in M(r_2)$“ bezeichnet, dann impliziert Satz 2 die bedingte Unabhängigkeit von I_A (Indikatorfunktion des Ereignisses A) und H gegeben den Subtestsummenscore R_2. Für die Vorhersage des Verhaltens im Subtest $\boldsymbol{U}$ ist, bei Kenntnis des Summenscores r_2, das Verhalten im Subtest $\boldsymbol{W}$ irrelevant.
Am Beispiel des Würfelaufgaben-Untertests des IST 2000 R[15] sei eine mögliche Anwendung des Satzes - mit Fokus auf bedingte stochastische Abhängigkeit zweier Items - kurz skizziert. Der Subtest $\boldsymbol{W}$ besteht hierbei aus dem siebten Item (X_7), so dass die Zeilenklassifikation nach dem Antwortverhalten auf X_7 erfolgt. Die übrigen Items bilden den Subtest $\boldsymbol{U}$ und damit stellt der Summenscore dieser Items $\sum_{k \neq 7} X_k$ die Schichtungsvariable dar. Ferner wird die Spaltenklassifikation gemäß dem achten Item vorgenommen. Gleichung (3.9) impliziert, dass der hierdurch festgelegte Test besonders sensitiv gegenüber lokalen stochastischen Abhängigkeiten zwischen Item X_7 und den Items des „restlichen“ Tests ist. Die spezielle Wahl von $M(r_2)$ legt ferner nahe, dass dieser Test besonders auf bedingte Abhängigkeiten zwischen X_7 und X_8 reagiert.
Tabelle 3.1 enthält die Ergebnisse der Kreuzklassifikation für zwei spezifische, zur Verdeutlichung gewählte Schichten ($r_2 = 8$, $r_2 = 14$). Für die Daten der ersten abgebildeten Schicht ($r_2 = 8$) liegt das Odds-Ratio ($\hat{\omega} = 4/3$) nahe dem Unabhängigkeitskonformen Wert von eins. In der zweiten Schicht ergibt sich ein Wert von $\hat{\omega} \approx 1.73$. In beiden dargestellten Schichten sind die geschätzten Wahrscheinlichkeiten, das achte Item zu lösen, für Personen, die Item X_7 lösen, trotz Kontrolle des Subtestscores erhöht. Insgesamt liegen 20 Schichten vor. Diese sind vorwiegend gering besetzt. Der

[15] Analyse und Beschreibung des IST 2000 R erfolgen in Kapitel 6.

p-Wert des Mantel-Haenszel-χ^2-Tests beträgt 0.15, während der korrespondierende p-Wert des exakten Tests - mittels der hypergeometrischen Verteilung - etwas geringer ausfällt (0.12). Somit kann die Nullhypothese der Gültigkeit eines Rasch-Modells basierend auf diesem Test nicht verworfen werden.
Während dieses Beispiel auf lokale Abhängigkeiten zielte, sind durch geeignete Wahl der Zeilen- bzw. Spaltenklassifikation auch andere Verletzungen des Rasch-Modells prüfbar. Einen Überblick bieten Verguts und De Boeck (2001)[16]. Für die *allgemeinen* Eigenschaften des Mantel-Haenszel-Tests sei ferner auf Agresti (2002, Kapitel 6.3 und 7.5) sowie die dort angegebene Literatur verwiesen.

Erweiterungen

Auch wenn das Beispiel sowie die bisherigen Erläuterungen auf $2 \times 2 \times K$-Tabellen beschränkt waren, so kann doch die Aussage des Satzes problemlos auf $I \times J \times K$-Kontingenztafeln übertragen werden. Lässt man z.B. für H mehr als zwei verschiedene Werte zu, so bleibt die Herleitung der Unabhängigkeitsaussage in (3.6) unverändert. Die Theorie des Mantel-Haenszel-Tests für diesen mehrkategorialen Fall wird bei Landis u.a. (1978) besprochen.
Neben dieser Erweiterung auf $I \times J \times K$-Tabellen ist ferner eine Verallgemeinerung des Satzes auf zum dichotomen Rasch-Modell verwandte Modelle möglich.
Ersetzt man die Item-Response-Funktion des Rasch-Modells durch die eines „one parameter logistic model“ (OPLM)

$$f_i(\theta) = \frac{\exp(a_i(\theta - \beta_i))}{1 + \exp(a_i(\theta - \beta_i))}$$

und wählt man für die Schichtvariable an Stelle des einfachen Summenscores den gewichteten Score[17] $\sum_k a_k U_{vk}$, so gelten die Aussagen des Satzes unverändert. Gleichermaßen lässt sich Satz 2 für ordinale Rasch-Modelle ausdehnen.
Nach einer kurzen Bemerkung über die Anwendung der Mantel-Haenszel-Testgröße im Kontext einer weiteren relevanten Erweiterung (*Differential Item Functioning*) soll abschließend ein allgemeiner Satz formuliert werden, der die skizzierten Erweiterungen als Spezialfall umfasst.

[16] Hier finden sich auch Angaben bezüglich der Power eines Tests auf Mehrdimensionalität.
[17] Der Diskriminationsparameter a_i wird im OPLM als *bekannte* Größe vorausgesetzt.

Bemerkung:
Das sogenannte *Differential Item Functioning* (DIF) kann ebenfalls mit der geschilderten Methodik behandelt werden. Dies soll nachfolgend, da es eine Hauptanwendung des Mantel-Haenszel-Tests darstellt, kurz erläutert werden.
Hierzu sei Q eine Zufallsvariable, die die Zugehörigkeit einer Person zu einer bestimmten Gruppe andeutet. Im Kontext der Testfairness findet man häufig die Wahl von Geschlecht, Altersgruppe oder Ethnizität.
Mittels dieser Bezeichnung lautet die „NODIF"-Annahme im Rahmen des Rasch-Modells:

$$P(\boldsymbol{x}|\,\theta, Q = q) = P(\boldsymbol{x}|\,\theta) = \prod_i \frac{\exp(x_i(\theta - \beta_i))}{1 + \exp(\theta - \beta_i)} \tag{3.13}$$

Für ein einzelnes Item führt dies auf:

$$P(X_i = 1|\,\theta,\, Q = q) = P(X_i = 1|\,\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}$$

Bedingt auf die Fähigkeit, ist das Antwortverhalten auf dem i-ten Item unabhängig von der Gruppenzugehörigkeit. Mit anderen Worten: Die Gruppen dürfen zwar über differierende Fähigkeitsverteilungen verfügen, das Verhalten bei gleicher Fähigkeit ist jedoch unbeeinflusst von der Gruppenzugehörigkeit. In allen Gruppen gilt - bedingt auf θ - ein Rasch-Modell mit *denselben* β-Parametern. Unterscheidet sich hingegen für ein Item i die Wahrscheinlichkeit $P(X_i = 1|\,\theta,\, Q = q)$ zwischen den Gruppen, so liegt DIF für Item i vor.
Da (3.13) eine Aussage analog zur bedingten Unabhängigkeit der Items gegeben θ darstellt, lässt sich leicht zeigen, dass Satz 2 Gültigkeit behält, wenn an Stelle der Variable H die externe Zufallsgröße Q verwendet wird. Alle Gleichungen des Beweises behalten - dank (3.13) - Gültigkeit. Die Zeilenklassifikation erfolgt nun nach der externen Größe Q. Das Antwortverhalten auf Item i bildet die Spaltenklassifikation und die Schichtvariable ist der Summenscore des gesamten Tests[18].
Bezüglich der Anwendung der Mantel-Haenszel-Prüfgröße im DIF-Kontext existieren zahlreiche Ergebnisse. Eine Gegenüberstellung der Mantel-Haenszel-Methodik mit anderen Testmöglichkeiten findet sich z.B. bei Waldherr (2001). Holland und Wainer (1993) befassen sich ferner mit DIF in einem größeren Kontext als den des Rasch-Modells. Literaturangaben über bekannte Resultate sowie einen Abriss der aktuellen Fragestellungen für mehrkategoriale bzw. ordinale Items geben Fidalgo und Madeira (2008).

[18]Die ursprüngliche Aufteilung des Tests entfällt, d.h. es gilt: $\boldsymbol{W} = \emptyset$, $\boldsymbol{U}$=$\boldsymbol{X}$.

Verallgemeinerung der Mantel-Haenszel-Testklasse

Das Ziel dieses Abschnitt ist die Reformulierung von Satz 2 unter möglichst allgemeinen Bedingungen. Insbesonders sollen die skizzierten Erweiterungen (OPLM, ordinales Rasch-Modell, DIF) sich in dieser neuen Formulierung wiederfinden.

Im Kontext der einführenden Bemerkungen aus Kapitel 2.1 (Definition eines IRT-Modells) werden für einen partitionierten Test[19] $\boldsymbol{X} = (\boldsymbol{W}, \boldsymbol{U})$ folgende Annahmen vorausgesetzt:

(1) θ ist eindimensional;

(2) $\boldsymbol{U}$ und $\boldsymbol{W}$ sind, gegeben θ, unabhängig;

(3) Die Items des Subtests $\boldsymbol{U}$ sind, gegeben θ, lokal unabhängig;

(4) $P(U_i = u_i|\,\theta) = v_i(\theta, \boldsymbol{\beta_i}) \exp(\theta f_i(u_i) - g_i(\boldsymbol{\beta_i}, u_i))$

Analog zu Satz 2 wird eine Funktion $H = H(\boldsymbol{W})$ sowie die suffiziente Statistik (des Subtests $\boldsymbol{U}$) $R_2 := \sum_i f_i(u_i)$ definiert. $M(r_2)$ bezeichne wiederum eine bestimmte Teilmenge der Menge aller Subtestvektoren $\boldsymbol{U}$, für die $\sum_i f_i(u_i) = r_2$ gilt.

Mit diesen Bezeichnungen gilt dann:

$$P(\boldsymbol{U} \in M(r_2)|\, r_2, H = h) = P(\boldsymbol{U} \in M(r_2)|\, r_2) \tag{3.14}$$

Beweis. Die linke Seite von (3.14) ergibt sich zu:

$$P(\boldsymbol{U} \in M(r_2)|\, r_2, H = h) = \sum_{\boldsymbol{u} \in M(r_2)} P(\boldsymbol{u}|\, r_2, H = h) = \sum_{\boldsymbol{u} \in M(r_2)} \frac{P(\boldsymbol{u}|\, H = h)}{P(r_2|\, H = h)} \tag{3.15}$$

Durch Einführen der latenten Variable wird (3.15) zu:

$$\sum_{\boldsymbol{u} \in M(r_2)} \frac{P(\boldsymbol{u}|\, H = h)}{P(r_2|\, H = h)} = \sum_{\boldsymbol{u} \in M(r_2)} \frac{\int_{-\infty}^{\infty} P(\boldsymbol{u}|\, \theta, H = h) \mathrm{d}G(\theta|\, H = h)}{\int_{-\infty}^{\infty} P(r_2|\, \theta, H = h) \mathrm{d}G(\theta|\, H = h)} \tag{3.16}$$

Betrachtet man für einen festen Summanden den Zähler in (3.16) und berücksichtigt man ferner die Annahme (2), so führt dies auf:

$$\int_{-\infty}^{\infty} P(\boldsymbol{u}|\, \theta, H = h) \mathrm{d}G(\theta|\, H = h) = \int_{-\infty}^{\infty} P(\boldsymbol{u}|\, \theta) \mathrm{d}G(\theta|\, H = h)$$

[19] Der Index v wird im Folgenden unterdrückt, d.h. $\boldsymbol{X}$ steht für den Responsevektor $\boldsymbol{X_v}$.

Unter Verwendung der Annahmen (3) und (4) ergibt sich:

$$\int_{-\infty}^{\infty} P(\boldsymbol{u}|\,\theta)\mathrm{d}G(\theta|\,H=h) = \int_{-\infty}^{\infty} v(\theta,\boldsymbol{\beta})\exp(r_2\theta - \textstyle\sum_i g_i(\boldsymbol{\beta_i},u_i))\mathrm{d}G(\theta|\,H=h)$$

Somit besitzt der Zähler (eines bestimmten Summanden) aus (3.16) die Gestalt:

$$\exp(-\textstyle\sum_i g_i(\boldsymbol{\beta_i},u_i))\displaystyle\int_{-\infty}^{\infty} v(\theta,\boldsymbol{\beta})\exp(r_2\theta)\mathrm{d}G(\theta|\,H=h) \tag{3.17}$$

Andererseits ergibt sich der Nenner aus (3.16) per Definition - $P(r_2|\,\theta,H=h) = \sum_{\boldsymbol{f(u)'1}=r_2} P(\boldsymbol{u}|\,\theta,H=h)$ - sowie unter Verwendung von (3.17) zu:

$$\sum_{\boldsymbol{f(u)'1}=r_2}\left(\exp(-\textstyle\sum_i g_i(\boldsymbol{\beta_i},u_i))\displaystyle\int_{-\infty}^{\infty} v(\theta,\boldsymbol{\beta})\exp(r_2\theta)\mathrm{d}G(\theta|\,H=h)\right) \tag{3.18}$$

Nutzt man (3.18) und (3.17) in (3.16), so resultiert[20]:

$$P(\boldsymbol{U}\in M(r_2)|\,r_2,H=h) = \frac{\sum_{\boldsymbol{u}\in M(r_2)}\exp(-\sum_i g_i(\boldsymbol{\beta_i},u_i))}{\sum_{\boldsymbol{f(u)'1}=r_2}\exp(-\sum_i g_i(\boldsymbol{\beta_i},u_i))}$$

Dieser Ausdruck ist unabhängig von h. □

Anmerkungen zur Notation:

„$\sum_{\boldsymbol{f(u)'1}=r_2}$“ bezeichnet eine Summe über alle Vektoren $\boldsymbol{u}$, die den gleichen Wert bezüglich der suffizienten Statistik r_2 aufweisen;
$v(\theta,\boldsymbol{\beta}) := \prod_i v_i(\theta,\boldsymbol{\beta_i})$;

Abschließend seien stichpunktartig einige Bemerkungen gegeben, die die Allgemeinheit des Satzes nochmals herausstellen:

- Bezüglich des Subtests $\boldsymbol{W}$ bestehen außer der Annahme der bedingten Unabhängigkeit keine restriktiven Forderungen. Insbesondere müssen die Wahrscheinlichkeitsfunktionen der Items des Subtests $\boldsymbol{W}$ *nicht* der Annahme (4) folgen.
- An Stelle eines Subtests kann W auch für eine externe Variable (z.B. Geschlecht) stehen. In diesem Fall entspricht die Annahme (2) genau der „NO-DIF“-Annahme.

[20] Der Integrand in (3.18) ist für jeden in Betracht kommenden Vektor $\boldsymbol{u}$ identisch.

- Der Beweis erfordert keine restriktiven Annahmen bezüglich der Anzahl der Ausprägungen der Zufallsvariable H. Er gilt folglich auch für mehrkategorial definiertes $H = H(\boldsymbol{W})$.

- Die Items dürfen - wie in Kapitel 2.1 ursprünglich eingeführt - mehr als zwei Kategorien aufweisen. Ferner kann die Anzahl an Kategorien über die Items variieren.

- Das dichotome Rasch-Modell ergibt sich als Spezialfall, wenn die Items binär sind und wenn $f_i(u_i) = u_i$, $g_i(\beta_i, u_i) = \beta_i u_i$ gewählt wird.

- Das OPLM ergibt sich als Spezialfall, wenn die Items binär sind und wenn $f_i(u_i) = a_i u_i$, $g_i(\beta_i, u_i) = a_i \beta_i u_i$ gewählt wird.

- Das ordinale Rasch-Modell ergibt sich als Spezialfall, wenn $f_i(u_i) = u_i$ und $g_i(\boldsymbol{\beta_i}, u_i) = \sum_l \beta_{il} I_{u_i=l}$ gewählt wird.

- Da die Items unterschiedliche Kategorienanzahlen besitzen können, sind auch Kombinationen der obigen Modelle möglich. Der erste Teil der Items des Subtests $\boldsymbol{U}$ folgt dann z.B. einem dichotomen Rasch-Modell, während der zweite Teil des Subtests $\boldsymbol{U}$ Items aus einem ordinalen Rasch-Modell beinhaltet.

3.3 Nonparametrische Tests in Obermodellen

Während die Testklassen der vorherigen Abschnitte stets auf die spezifische parametrische Form der Item-Response-Funktion im Rasch-Modell zurückgreifen, bauen die Tests dieses Abschnitts lediglich auf globaleren Eigenschaften wie „Monotonie der Item-Response-Funktion“ oder „lokale stochastische Unabhängigkeit“ auf. Aus der Gültigkeit dieser relativ allgemeinen Attribute folgen Aussagen über die Assoziation bestimmter Zufallsvariablen. Die Existenz dieser Assoziationen ist wiederum über „herkömmliche“ Teststatistiken für kategoriale Daten prüfbar.
In Abhängigkeit der gewählten Teststatistik ergibt sich ein asymptotischer oder ein exakter Test. Die Aussage zur Assoziationsstruktur ist jedoch für jeden beliebigen Stichprobenumfang gültig und erfordert insofern keine Asymptotik[21].
Bevor die Prüfgrößen dargestellt werden können, ist es notwendig, einige grundlegende Definitionen sowie Schlussfolgerungen bezüglich der Assoziation von Zufallsvariablen zu geben[22] (Esary u.a., 1967):

Definition 3. Ein Zufallsvektor $\boldsymbol{Y}$ heißt *assoziiert* genau dann, wenn für jedes Paar monoton wachsender, reellwertiger Funktionen g_1, g_2 gilt:

$$\mathrm{Cov}(g_1(\boldsymbol{Y}), g_2(\boldsymbol{Y})) \geq 0$$

„Monoton wachsend“ bezieht sich hierbei auf jede einzelne Komponente, bei festen Werten der anderen Komponenten. Implizit wird ferner die Existenz der Kovarianz vorausgesetzt.

Eine erste Schlussfolgerung ergibt sich aus der Betrachtung eines eindimensionalen Zufallsvektors:

Lemma 2. *Ein eindimensionaler Zufallsvektor Y ist assoziiert.*

Beweis. Für beliebige y_1, y_2 gilt aufgrund der gleichen Monotonie von g_1 und g_2:

$$(g_1(y_1) - g_1(y_2))(g_2(y_1) - g_2(y_2)) \geq 0$$

[21] Nonparametrische Teststatistiken, die auf asymptotischen Aussagen beruhen, finden sich z.B. bei Stout (1987) oder bei Douglas u.a. (1998).

[22] Beweise werden stets wiedergegeben, insofern sie das Zusammenwirken der Annahmen des jeweiligen Item-Response-Modells verdeutlichen oder für spätere Aussagen relevant sind.

Für[23]

$$Y_1, Y_2 \overset{i.i.d}{\sim} F_Y$$

gilt somit:

$$(g_1(Y_1) - g_1(Y_2))(g_2(Y_1) - g_2(Y_2)) \geq 0$$

Durch Bildung des Erwartungswerts auf beiden Seiten sowie Umstellung von Termen folgt unmittelbar die Behauptung. □

Eine weitere zentrale Folgerung betrifft das Verhalten zweier unabhängiger Zufallsvektoren (Esary u.a., 1967):

Lemma 3. *Wenn $\boldsymbol{W}$ und $\boldsymbol{Z}$ jeweils assoziierte Zufallsvektoren sind, dann ist auch $\boldsymbol{Y} := (\boldsymbol{W}, \boldsymbol{Z})$ assoziiert, falls $\boldsymbol{W}$ von $\boldsymbol{Z}$ unabhängig ist.*

Beweis.

$$\mathrm{Cov}(g_1(\boldsymbol{Y}), g_2(\boldsymbol{Y})) = \mathrm{Cov}\left(E(g_1(\boldsymbol{Y})|\boldsymbol{W}), E(g_2(\boldsymbol{Y})|\boldsymbol{W})\right) + E\left(\mathrm{Cov}(g_1(\boldsymbol{Y}), g_2(\boldsymbol{Y})|\boldsymbol{W})\right)$$

Die Zufallsvariablen $E(g_1(\boldsymbol{Y})|\boldsymbol{W})$ sowie $E(g_2(\boldsymbol{Y})|\boldsymbol{W})$ innerhalb des ersten Summanden sind aufgrund der Monotonie von g_1 (bzw. g_2) und der Unabhängigkeit von $\boldsymbol{Z}$ und $\boldsymbol{W}$ (die bedingte Verteilung über die der entsprechende Erwartungswert gebildet wird, ist für jeden Wert von $\boldsymbol{W}$ die gleiche) monoton wachsend in $\boldsymbol{W}$. Da $\boldsymbol{W}$ assoziiert ist, folgt unmittelbar die Nichtnegativität des ersten Summanden. Bezüglich des zweiten Terms gilt:

$$\mathrm{Cov}(g_1(\boldsymbol{Y}), g_2(\boldsymbol{Y})|\boldsymbol{W}) \geq 0$$

Dies folgt, da g_1 und g_2 für festes $\boldsymbol{W} = \boldsymbol{w}$ monoton wachsende Funktionen in $\boldsymbol{Z}$ sind. Da die Zufallsvektoren $\boldsymbol{W}$ und $\boldsymbol{Z}$ unabhängig sind, wird die Kovarianz von zwei monoton wachsenden Funktionen in $\boldsymbol{Z}$ somit über die marginale Verteilung von $\boldsymbol{Z}$ gebildet. Aus der Assoziation von $\boldsymbol{Z}$ folgt die Behauptung. □

[23] F_Y bezeichnet hierbei die Verteilungsfunktion der Zufallsvariable Y.

Assoziation im „Monotone Homogeneity Model“

Die bisherigen Ergebnisse stellen direkte Schlussfolgerungen aus Definition 3 dar. Ein erster Bezug zur Item-Response-Theorie kann über folgendes Lemma (Rosenbaum, 1984) hergestellt werden[24]:

Lemma 4. *In einem „Monotone Homogeneity Model“ ist für beliebiges, monoton wachsendes g die Funktion $E(g(\boldsymbol{X})|\,\theta)$ monoton wachsend in θ.*

Beweis. Sei $\theta_1 \leq \theta_2$, es gilt:

$$\begin{aligned} E(g(\boldsymbol{X})|\,\theta_2) - E(g(\boldsymbol{X})|\,\theta_1) &= \sum_{\boldsymbol{x}} g(\boldsymbol{x})(P(\boldsymbol{X}=\boldsymbol{x}|\,\theta_2) - P(\boldsymbol{X}=\boldsymbol{x}|\,\theta_1)) \\ &= \sum_{\boldsymbol{x}} g(\boldsymbol{x}) \left(\frac{P(\boldsymbol{X}=\boldsymbol{x}|\,\theta_2)}{P(\boldsymbol{X}=\boldsymbol{x}|\,\theta_1)} - 1 \right) P(\boldsymbol{X}=\boldsymbol{x}|\,\theta_1) \end{aligned}$$

Der Quotient

$$q(\boldsymbol{x}) := \frac{P(\boldsymbol{X}=\boldsymbol{x}|\,\theta_2)}{P(\boldsymbol{X}=\boldsymbol{x}|\,\theta_1)} = \frac{\prod_i f_i(\theta_2)^{x_i}(1-f_i(\theta_2))^{1-x_i}}{\prod_i f_i(\theta_1)^{x_i}(1-f_i(\theta_1))^{1-x_i}}$$

ist aufgrund der Monotonie-Annahme des MHM monoton wachsend in $\boldsymbol{x}$. Somit bilden q und g ein Funktionenpaar mit identischem Monotonieverhalten und es gilt:

$$\sum_{\boldsymbol{x}} g(\boldsymbol{x})(q(\boldsymbol{x})-1)P(\boldsymbol{X}=\boldsymbol{x}|\,\theta_1) = \mathrm{Cov}(g(\boldsymbol{X}), q(\boldsymbol{X})|\,\Theta=\theta_1) \geq 0 \tag{3.19}$$

Der letzte Übergang in (3.19) folgt, da - gegeben $\Theta = \theta_1$ - die Variablen in $\boldsymbol{X}$ unabhängig sind. Die Kombination von Lemma 2 mit Lemma 3 („Vereinigungen“ eindimensionaler, unabhängiger Zufallsvektoren sind assoziiert!) ergibt somit, dass $\boldsymbol{X}$ gegeben θ assoziiert ist. Berücksichtigt man die Monotonie von g und q, so resultiert mittels Definition 3 die Behauptung. □

Basierend auf diesem Resultat für das MHM und den „technischen“ Vorüberlegungen (Lemma 2 und 3) kann die erste „Testklasse“ von Rosenbaum (1984) näher beschrieben werden. Hierfür wird der Zufallsvektor $\boldsymbol{X}$, der das Antwortverhalten einer Person auf die unterschiedlichen Items kennzeichnet, in zwei disjunkte Vektoren $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z})$ unterteilt. Die zentrale Aussage lässt sich nun in folgendem Satz formulieren:

[24] Zur Vereinfachung der Notation wird der Index v unterdrückt. $\boldsymbol{X}$ steht somit stellvertretend für den Responsevektor $\boldsymbol{X_v}$.

Satz 3. *Bezeichnet $h(\cdot)$ eine beliebige Funktion von $\boldsymbol{Z}$ und sind die Annahmen des „Monotone Homogeneity Model“ erfüllt, dann ist $\boldsymbol{Y}$ gegeben $h(\boldsymbol{Z}) = c$ assoziiert, d.h. für monoton wachsende Funktionen g_1 und g_2 gilt für beliebiges c:*

$$Cov(g_1(\boldsymbol{Y}), g_2(\boldsymbol{Y})|h(\boldsymbol{Z}) = c) \geq 0$$

Beweis. Nach der „tower property“ (Grimmett und Stirzaker 2001a, S.69) der bedingten Erwartung gilt:

$$E\{g_1(\boldsymbol{Y})g_2(\boldsymbol{Y})|\, h(\boldsymbol{Z})\} = E\{E[g_1(\boldsymbol{Y})g_2(\boldsymbol{Y})|h(\boldsymbol{Z}),\, \Theta]|\, h(\boldsymbol{Z})\} \tag{3.20}$$

Aufgrund der Unabhängigkeit (zweite Annahme des MHM) von $\boldsymbol{Y}$ und $\boldsymbol{Z}$ unter der Bedingung $\Theta = \theta$ folgt:

$$E\{E[g_1(\boldsymbol{Y})g_2(\boldsymbol{Y})|\, h(\boldsymbol{Z}),\, \Theta]|\, h(\boldsymbol{Z})\} = E\{E[g_1(\boldsymbol{Y})g_2(\boldsymbol{Y})|\, \Theta]|\, h(\boldsymbol{Z})\} \tag{3.21}$$

Weiterhin ist $\boldsymbol{Y}$ - gegeben θ - ein Vektor, dessen Komponenten voneinander unabhängig sind. Mittels Lemma 2 und Lemma 3 folgt, dass $\boldsymbol{Y}$ gegeben die Fähigkeit θ assoziiert ist. Somit wird (3.20) in Kombination mit (3.21) zu:

$$E\{g_1(\boldsymbol{Y})g_2(\boldsymbol{Y})|\, h(\boldsymbol{Z})\} \geq E\{E[g_1(\boldsymbol{Y})|\, \Theta]E[g_2(\boldsymbol{Y})|\, \Theta]|\, h(\boldsymbol{Z})\} \tag{3.22}$$

Betrachtet man die Größe $\psi_i(\Theta) := E(g_i(\boldsymbol{Y})|\Theta)$, so stellt diese aufgrund der ersten Annahme des MHM eine eindimensionale Zufallsvariable (in Θ) dar. Gemäß Lemma 2 ist eine eindimensionale Zufallsvariable assoziiert. Da ψ_1 und ψ_2 monoton wachsende Funktionen von θ sind (nach Lemma 4), folgt somit für die rechte Seite von (3.22):

$$E\{E[g_1(\boldsymbol{Y})|\, \Theta]E[g_2(\boldsymbol{Y})|\, \Theta]|\, h(\boldsymbol{Z})\} \geq E\{E[g_1(\boldsymbol{Y})|\, \Theta]|\, h(\boldsymbol{Z})\}E\{E[g_2(\boldsymbol{Y})|\, \Theta]|\, h(\boldsymbol{Z})\} \tag{3.23}$$

Aufgrund der bedingten Unabhängigkeit und der „tower property“ gilt aber:

$$E\{E[g_i(\boldsymbol{Y})|\, \Theta]|\, h(\boldsymbol{Z})\} = E\{E[g_i(\boldsymbol{Y})|\, h(\boldsymbol{Z})\,, \Theta]|\, h(\boldsymbol{Z})\} = E\{g_i(\boldsymbol{Y})|\, h(\boldsymbol{Z})\} \tag{3.24}$$

In Kombination mit (3.22) und (3.23) ergibt sich die Behauptung:

$$E\{g_1(\boldsymbol{Y})g_2(\boldsymbol{Y})|\, h(\boldsymbol{Z})\} \geq E\{g_1(\boldsymbol{Y})|\, h(\boldsymbol{Z})\}E\{g_2(\boldsymbol{Y})|\, h(\boldsymbol{Z})\}$$

□

Ein erstes Beispiel liefert $\boldsymbol{Y} = (X_i, X_j)$, $Z = X_k$, $h(Z) = Z$. Hier wird für einen festen Wert auf Item k die Assoziation zweier Items i und j betrachtet. Die Wahl von $g_1(X_i, X_j) = X_i$ und $g_2(X_i, X_j) = X_j$ führt mittels des Satzes auf eine nicht negative Korrelation zweier Items i und j, gegeben jedes Level auf Item k. Existiert hingegen ein Niveau ($X_k = 1$ oder $X_k = 0$), auf dem das Itempaar negativ korreliert, so kann daraus der Schluss gezogen werden, dass diese drei Items unabhängig von der Gestaltung der restlichen Testitems in kein MHM „passen".
Lässt man im obigen Beispiel $\boldsymbol{Z}$ leer, so resultiert eine nicht negative *marginale* Korrelation für das Itempaar (i, j). In einem MHM sind sämtliche Itempaare nicht negativ korreliert. Satz 3 beinhaltet jedoch noch eine wesentlich stärkere Aussage, da er sich auf *bedingte Assoziation bezüglich jeder beliebigen Partition* des Vektors $\boldsymbol{X}$ bezieht. Die Nichtnegativität der marginalen Korrelationen hingegen gilt bereits unter abgeschwächten Bedingungen (Holland, 1981).

Genau genommen ist durch Satz 3 noch kein spezifischer Test festgelegt. Zunächst impliziert er eine Vielfalt an Assoziationsaussagen, die z.B. bei der Itemauswahl bzw. Testbeurteilung deskriptiv eingesetzt werden können (siehe hierzu auch die Analyse des IST 2000 R in Kapitel 6). Das Überprüfen der paarweisen Itemkorrelationen bezüglich ihres Vorzeichens sowie die eventuelle Ausdehnung dieser Prozedur auf bedingte Assoziationen erlauben eine leicht durchzuführende Vorab-Evaluation der Qualität des jeweiligen Datensatzes. Entdeckt man auf diese Art Verstöße gegen Satz 3, so zieht dies die starke Implikation mit sich, dass keines der „gängigen" eindimensionalen Item-Response-Modelle - d.h. kein MHM - die Daten beschreiben kann. Holland (1981, S.79) bemerkt in einem ähnlichen Kontext hierzu:

„*The main value of the conditions derived here is to identify data sets which are inconsistent with any item response model that assumes local independence.*"

An Stelle der rein deskriptiven Prüfung kann jedoch auch die „strengere" Form des Hypothesentests angewendet werden. Als Teststatistiken kommen je nach gewähltem Funktionenpaar (g_1, g_2) unterschiedliche Prüfgrößen in Frage.
Eine Möglichkeit, insbesondere für mehrkategoriale, ordinale Funktionenpaare, bildet Kendalls τ. Falls für zwei Zufallsvariablen X, Y die Bedingungen („positive quadrant dependence", Lehmann, 1966)

$$P(X \geq x, Y \geq y) \geq P(X \geq x)P(Y \geq y) \quad \forall x, \forall y \tag{3.25}$$

gelten, nimmt Kendalls τ einen nicht negativen Wert an (Lehmann, 1966).

Die Verbindung zu Satz 3 kann wie folgt hergestellt werden:
Man wähle

$$g_1^* := I_{g_1 \geq a},\ g_2^* := I_{g_2 \geq b},$$

dann besitzen g_1^*, g_2^* die gleiche Monotonie (bezüglich $\boldsymbol{Y}$) wie g_1, g_2 und die Anwendung von Satz 3 auf das neue Funktionenpaar (g_1^*, g_2^*) ergibt, bedingt auf $h(\boldsymbol{Z})$:

$$0 \leq \mathrm{Cov}(g_1^*(\boldsymbol{Y}), g_2^*(\boldsymbol{Y})) = P(g_1(\boldsymbol{Y}) \geq a, g_2(\boldsymbol{Y}) \geq b) - P(g_1(\boldsymbol{Y}) \geq a)P(g_2(\boldsymbol{Y}) \geq b)$$

Das Funktionenpaar (g_1, g_2) erfüllt somit (3.25). Aus Assoziation (Definition 3) folgt somit „positive quadrant dependence“ bzw. ein nicht negativer Wert für Kendalls τ. Dies stellt eine einfache Testmöglichkeit - realisierbar anhand der korrespondierenden U-Statistik $\hat{\tau}$ (Puri und Sen, 1971, Kapitel 3; Lehmann und Romano, 2005, Kapitel 6) - im Falle eines nicht binären Funktionenpaares dar. Es sei einschränkend bemerkt, dass diese Prüfgröße ($\hat{\tau}$) auf Asymptotik beruht.
Für *binäre* Funktionen hingegen ist ein Test bezüglich des Odds-Ratios durchführbar. Am Beispiel des Zahlenreihen-Untertests des IST 2000 R sei dies verdeutlicht. Die 20 Items des Subtests seien durchnummeriert und mit $X_1, \ldots, X_{20}$ bezeichnet. Unter der Annahme der Gültigkeit eines MHM - eine notwendige Bedingung für ein Rasch-Modell - folgt aus Satz 3 eine nicht negative Assoziation zwischen den Items X_{14} und X_{16} für jede Stufe von X_{15}. Die entsprechenden Kontingenztafeln sind in Tabelle 3.2 aufgeführt.
Für jene Personen, die an Item X_{15} „scheitern“, befindet sich das Odds-Ratio $\hat{\omega}$ in zu Satz 3 konformer Richtung ($\hat{\omega} = 7.56 \geq 1$). Hingegen zeigt sich für die zweite Bedingung ($X_{15} = 1$) eine zu Satz 3 widersprüchliche Richtung ($\hat{\omega} = 0.72$). Der zugehörige einseitige Test (Agresti 2002, S.99) auf „$H_0 : \omega \geq 1$“ versus „$H_1 : \omega < 1$“

Tabelle 3.2: *Kreuzklassifikation von X_{14} und X_{16} innerhalb jeder Stufe des Items X_{15}. Die Indizes beziehen sich auf die Itemposition innerhalb des Zahlenreihen-Tests des IST 2000 R.*

	$X_{15} = 0$				$X_{15} = 1$		
	$X_{16} = 0$	$X_{16} = 1$	Gesamt		$X_{16} = 0$	$X_{16} = 1$	Gesamt
$X_{14} = 0$	51	6	57	$X_{14} = 0$	1	14	15
$X_{14} = 1$	18	16	34	$X_{14} = 1$	15	152	167
Gesamt	69	22	91	Gesamt	16	166	182

liefert allerdings einen nicht signifikanten p-Wert von 0.61. Somit können auf der Basis dieses Tests keine Abweichungen von einem MHM festgestellt werden. Es sei jedoch darauf hingewiesen, dass sich bei einer anderen Auswahl der bedingenden Variable (z.B. X_{14} „tauscht die Rolle“ mit X_{15}) andere Ergebnisse bezüglich der MHM-Konformität dieser drei Items ergeben könnten.

Wie das Beispiel zeigt, ist es erforderlich für jede Schicht, d.h. für jede Stufe der bedingenden Variable, einen seperaten Test durchzuführen[25]. Dies könnte des Weiteren für jede Auswahl einer bedingenden Variable aus den drei zu prüfenden Items erfolgen. Ferner ließen sich auch neben der Assoziation zwischen X_{14} und X_{16} andere Itempaare betrachten. Die Vielzahl an potentiellen Tests führt zwangsläufig auf ein multiples Testproblem. Die notwendige Korrektur des α-Niveaus behandelt der abschließende Teil dieses Kapitels. Zuvor wird jedoch eine weitere nonparametrische Testklasse präsentiert, die ebenso die Korrekturen im Rahmen des multiplen Testens erforderlich macht.

Latente proportionale Odds

Die zweite Testklasse, die im folgenden Satz indirekt formuliert wird, geht ebenfalls auf Rosenbaum (1987) zurück. Für die Gültigkeit der Aussagen dieses Satzes sind jedoch im Gegensatz zu der vorhergehenden Klasse, die auf den drei Eigenschaften des MHM beruht, andere Annahmen erforderlich. Die Monotonie der Item-Response-Funktion wird ebenso wie die Eindimensionalität der latenten Variable aufgegeben. Somit bleibt als Basisannahme lediglich die lokale Unabhängigkeit gegeben ein (potentiell) mehrdimensionales $\boldsymbol{\theta}$. Ergänzt wird diese Eigenschaft durch die Annahme einer bestimmten Relation zwischen den Item-Response-Funktionen zweier Items i und j.

Definition 4. Zwei Items i und j besitzen *latente proportionale Odds* (LPO) genau dann, wenn eine Konstante κ existiert, so dass gilt:

$$\frac{f_i(\boldsymbol{\theta})}{1 - f_i(\boldsymbol{\theta})} = \kappa \frac{f_j(\boldsymbol{\theta})}{1 - f_j(\boldsymbol{\theta})} \tag{3.26}$$

[25] Es existieren auch Möglichkeiten die Daten der unterschiedlichen Schichten simultan zu nutzen. Erscheint z.B. eine „gleichmäßige“ Abweichung des Odds-Ratios in den Schichten plausibel, so kann mit dem (einseitigen) Mantel-Haenszel-Test eine gemeinsame Prüfgröße formuliert werden.

Jedes Itempaar eines Rasch-Modells besitzt LPO. Man wähle $\kappa = \exp(\beta_j - \beta_i)$, dann gilt:

$$\frac{f_i(\theta)}{1 - f_i(\theta)} = \exp(\theta - \beta_i) = \kappa \exp(\theta - \beta_j) = \kappa \frac{f_j(\theta)}{1 - f_j(\theta)}$$

Basierend auf (3.26) kann nun der zentrale Satz formuliert werden. Als Grundlage dient die Zerlegung des Vektors $\boldsymbol{X_v}$ in einen Teil X_{vi}, der den Response auf Item i angibt, einen Teil X_{vj} sowie einen Vektor $\boldsymbol{Z_v}$, der den Response auf den restlichen Items umfasst[26].

Satz 4. *Sind gegeben ein (potentiell) mehrdimensionales $\boldsymbol{\theta}$ die Items lokal stochastisch unabhängig und besitzt das Itempaar (i, j) LPO, so gilt für jede beliebige Funktion $h(\boldsymbol{Z})$ sowie jedes c:*

$$P(X_i = 1 | \, X_i + X_j = 1, \, h(\boldsymbol{Z}) = c) = \frac{\kappa}{\kappa + 1}$$

Beweis. Zunächst wird die latente Variable $\boldsymbol{\theta}$ mit der zugehörigen bedingten Dichte $l(\boldsymbol{\theta}) := g(\boldsymbol{\theta} | \, X_i + X_j = 1, \, h(\boldsymbol{Z}) = c)$ eingeführt:

$$P(X_i = 1 | \, X_i + X_j = 1, \, h(\boldsymbol{Z}) = c) = \int P(X_i = 1 | \, X_i + X_j = 1, \, h(\boldsymbol{Z}) = c, \, \boldsymbol{\theta}) l(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{3.27}$$

Aufgrund der bedingten Unabhängigkeit von $\boldsymbol{Z}$ und (X_i, X_j) gilt:

$$\begin{aligned} P(X_i = 1 | \, X_i + X_j = 1, \, h(\boldsymbol{Z}) = c, \, \boldsymbol{\theta}) &= P(X_i = 1 | \, X_i + X_j = 1, \, \boldsymbol{\theta}) \\ &= \frac{f_i(\boldsymbol{\theta})(1 - f_j(\boldsymbol{\theta}))}{f_i(\boldsymbol{\theta})(1 - f_j(\boldsymbol{\theta})) + (1 - f_i(\boldsymbol{\theta}))f_j(\boldsymbol{\theta})} \\ &= \frac{\kappa}{\kappa + 1} \end{aligned}$$

Nutzt man dieses Resultat in (3.27), so folgt unmittelbar die Behauptung. □

Wenn $h(\boldsymbol{Z})$ C verschiedene Ausprägungen aufweist, erhält man somit die Aussage, dass für jede der C-Stufen, gegeben, genau eines der Items i und j wurde gelöst, die Wahrscheinlichkeit, dass Item i gelöst wurde gleich ist. Das Antwortverhalten auf den restlichen Items liefert somit zur Prädiktion, welches der beiden Items gelöst wurde, keinen Beitrag. Mit diesem Resultat können „gewöhnliche“ Tests auf Verteilungshomogenität für eine $C \times 2$ Kreuztabelle angewandt werden.

[26] Im Folgenden wird aus Gründen der Übersichtlichkeit erneut der Index v unterdrückt.

Am Beispiel des IST 2000 R sei dies verdeutlicht. Der Satzergänzungstest des IST 2000 R beinhaltet 20 Items. Diese seien im Folgenden mit $X_1, \ldots, X_{20}$ bezeichnet. Falls dieser Subtest Rasch-skalierbar ist, gilt für jedes beliebige Itempaar die Aussage des Satzes. Betrachtet man z.B. die beiden letzten Items des Subtests sowie $\boldsymbol{Z} = (X_1, \ldots X_{18})$ und wählt man für h die Funktion des „Restsummenscores“

$$h(\boldsymbol{Z}) = \begin{cases} 0 & \text{falls } \sum_{l=1}^{18} X_l \leq 10 \\ 1 & \text{falls } \sum_{l=1}^{18} X_l \geq 11 \end{cases},$$

so impliziert Satz 4 - bezogen auf jene Personen, die genau eines der letzten beiden Items lösen - die Verteilungshomogenität der Zufallsvariable X_{19} über die zwei Ausprägungen von $h(\boldsymbol{Z})$. Gegeben eine „gute Performance“ auf den restlichen Items ($h(\boldsymbol{Z}) = 1$), ist die bedingte Wahrscheinlichkeit, Item 19 richtig beantwortet zu haben (gegeben, genau eines der letzten beiden Items wurde gelöst) identisch mit der entsprechenden Wahrscheinlichkeit bei „schlechter Performance“.
Tabelle 3.3 beinhaltet die korrespondierende Kreuzklassifikation. Das Zeilenmerkmal ist die binäre Performance im restlichen Test. Das Spaltenmerkmal wird durch Item 19 gebildet. Obwohl Item 20 nicht direkt erscheint, ist es implizit anhand der Personenvorauswahl gegenwärtig, da nur Personen, die genau eines der Items X_{19}, X_{20} gelöst haben, in die Tabelle eingehen. Die Odds, Item 19 zu lösen sind für die erste Gruppe ($Z \geq 11$) 2.46 mal so hoch wie die Odds der zweiten Gruppe. Bei Gültigkeit von Satz 4 wäre ein (Populations-)Odds-Ratio von eins zu erwarten. Der beobachtete Wert weicht nicht signifikant von dieser Erwartung ab. Der exakte Test von Fisher liefert einen nicht signifikanten p-Wert ($p = 0.15$).

Tabelle 3.3: *Kreuzklassifikation von $Z = \sum_{l=1}^{18} X_l$ und X_{19} für Personen, die genau eines der zwei Items X_{19} und X_{20} lösen. Die Indizes beziehen sich auf die Itemposition innerhalb des Satzergänzungs-Tests des IST 2000 R.*

	Häufigkeiten				Relative Häufigkeiten		
	$X_{19} = 0$	$X_{19} = 1$	Gesamt		$X_{19} = 0$	$X_{19} = 1$	Gesamt
$Z \leq 10$	5	16	21	$Z \leq 10$	0.03	0.10	0.13
$Z \geq 11$	16	126	142	$Z \geq 11$	0.10	0.77	0.87
Gesamt	21	142	163	Gesamt	0.13	0.87	1

Multiples Testen

Das Beispiel bezog sich auf ein konkret gewähltes Itempaar. Da in der Praxis davon auszugehen ist, dass nur selten Hypothesen bezüglich eines spezifischen Itempaares vorliegen und da ferner, nutzt man Satz 4 zur Falsifikation des Rasch-Modells, alle Itempaare von Interesse wären, stellt sich die Frage nach der simultanen Betrachtung aller $s := \frac{k(k-1)}{2}$ Teststatistiken.
Hierfür stehen Prozeduren für multiple Testprobleme zur Verfügung. Das einfachste Verfahren führt jeden individuellen Test auf dem α/s-Niveau durch. Jene Bonferroni-Korrektur hat jedoch den Nachteil eines starken Power-Verlusts mit steigender Anzahl zu testender Hypothesen. Eine leichte Verbesserung bietet die Holm-Prozedur, die auf den geordneten p-Werten beruht (Lehmann und Romano, 2006, S.350f).
Beide bisher erwähnten „Korrekturen“ gründen auf dem Konzept der FWER („family wise error rate“). Wenn $H_1, \ldots, H_s$ die zu testenden Nullhypothesen bezeichnen und wenn J die Indexmenge der wahren Nullhypothesen darstellt (d.h. $j \in J \Leftrightarrow H_j$ wahr), dann kontrolliert eine multiple Testprozedur die FWER, wenn für jede beliebige Konstellation falscher und richtiger Nullhypothesen gilt:

$$P(\exists\, j \in J :\ H_j \text{ wird verworfen}) \leq \alpha$$

Sowohl Bonferroni- als auch Holm-Prozedur kontrollieren die FWER. Für jede Kombination wahrer und falscher Nullhypothesen beträgt die Wahrscheinlichkeit mindestens eine wahre Nullhypothese zu verwerfen höchstens α.
Im Gegensatz zu den vorherigen Verfahren bietet die Methode von Benjamini und Hochberg (1995) eine höhere Testpower. Sie basiert jedoch auf einem anderen Fehlerkonzept. Die FDR („false discovery rate“), definiert als erwarteter Anteil wahrer Nullhypothesen unter den verworfenen Nullhypothesen, stellt hier die zu kontrollierende Fehlergröße dar $\left(\frac{0}{0} := 0\right)$:

$$FDR := E(Q), \quad Q := \frac{\sum_{i \in J} I_{\{H_i \text{ wird abgelehnt}\}}}{\sum_{l=1}^{s} I_{\{H_l \text{ wird abgelehnt}\}}}$$

Eine Prozedur kontrolliert die FDR, wenn für jede Konstellation wahrer und falscher Nullhypothesen gilt:

$$E(Q) \leq \alpha$$

Das Verfahren von Benjamini und Hochberg (1995) kontrolliert die FDR. Basierend auf den geordneten p-Werten $p_{(1)}, \ldots, p_{(s)}$ sowie den korrespondierenden Nullhypothesen $H_{(1)}, \ldots, H_{(s)}$ lautet die Vorschrift:

- Bestimme den maximalen Index k, für den gilt:
$$p_{(k)} \leq \frac{k}{s}\,\alpha$$
- Lehne $H_{(1)}, \ldots, H_{(k)}$ ab.

Dieser Ablauf garantiert für unabhängige Teststatistiken (Benjamini und Hochberg, 1995) sowie für gewisse Formen der Abhängigkeit („positive regression dependence“) die Einhaltung der FDR zum Niveau α. Eine leichte Modifikation erlaubt auch die Kontrolle der FDR für beliebige Formen der Abhängigkeit zwischen den einzelnen Teststatistiken[27].
Gegenüber der Bonferroni-Methode ermöglicht dieses Verfahren eine erhebliche Verbesserung der Teststärke (Benjamini und Hochberg, 1995). Im Kontext des Rasch-Modells wurde es z.B. von Christensen und Kreiner (2010) zur Identifikation von Items mit abweichender Diskrimination verwendet.
Der Einsatz dieser multiplen Testprozedur ist nicht auf die Überprüfung der LPO-Eigenschaft für alle Itempaare beschränkt. So wurde in Kapitel 3.2 ein Mantel-Haenszel-Test auf lokale stochastische Abhängigkeit zweier Items vorgestellt. Auch hier kann es von Interesse sein, diesen Test für eine größere Menge an Itempaaren durchzuführen. Auf die so resultierende Menge an Teststatistiken kann ebenfalls eine multiple Testprozedur angewandt werden, um einer Kumulation des α-Fehlers vorzubeugen. Die gleiche Bemerkung gilt im Kontext der kombinatorischen Testklasse. Hier lässt sich für ein Itempaar (i, j) über $T(\boldsymbol{X}) = (r_{ij} - \rho_{ij})^2$ lokale stochastische Abhängigkeit testen. Für mehrere Itempaare resultiert auch hier wiederum ein multiples Testproblem.
Bei Ablehnung der Nullhypothese des Rasch-Modells, d.h. falls mindestens eine Teststatistik das entsprechende, korrigierte Signifikanzniveau unterschreitet, sind unmittelbar kritische Itempaare ersichtlich. Die zu den abgelehnten Hypothesen korrespondierenden Teststatistiken bieten aufgrund ihrer Ausrichtung an spezifischen Itempaaren einen direkten Ansatzpunkt für eine Verbesserung der Skala. Die Teststatistiken des nächsten Kapitels stellen hingegen globale Statistiken dar. Der Verlust an

[27] Für den Beweis sei auf Benjamini und Yekutieli (2001) verwiesen.

Power, der durch multiple Testprozeduren entsteht, wird hier umgangen, da eine einzelne Teststatistik zur globalen Beurteilung dient. Anstatt für jedes Itempaar einen Test auf lokale Abhängigkeit zu formulieren und auf die so entstandene Menge an Prüfgrößen eine multiple Testprozedur anzuwenden, erfolgt die Evaluation anhand *einer* allgemeinen Statistik (siehe hierzu Kapitel 4.4 für einen globalen Test auf lokale Abhängigkeit), in die alle Itempaare eingehen. Der so erzielte Gewinn an Teststärke wird durch eine „diffuse“ Interpretation erkauft. Bei signifikantem Ergebnis ist - im Gegensatz zur multiplen Testprozedur, die auf Prüfung spezifischer Itempaare basiert - durch den globalen Charakter der Prüfgröße noch kein spezielles Itempaar als „Verursacher des Misfits“ ausgezeichnet.

Kapitel 4

Globale Alternativmodelle

4.1 Vorbemerkungen zu den Alternativmodellen

Nachdem im vorherigen Kapitel drei grundlegende nonparametrische Testklassen vorgestellt wurden, von denen zwei auf bekannte Prüfgrößen für kategoriale Daten führen, dienen die folgenden zwei Kapitel zur Ergründung der Teststärke der kombinatorischen Tests. Da deren Verhalten aufgrund der Mächtigkeit von $\boldsymbol{\Sigma_{rc}}$ analytisch kaum zugänglich ist, soll es anhand der Simulation bestimmter Alternativmodelle untersucht werden.

Während die Simulation in Kapitel 5 stattfindet, stellt dieses Kapitel die Alternativmodelle vor, bezüglich deren die Power der kombinatorischen Tests beurteilt werden soll. Die Auswahl der Alternativen[1] orientiert sich dabei an „gängigen“ Item-Response-Modellen und korrespondiert weitestgehend mit der Simulationsstudie von Suárez-Falcón und Glas (2003). Es handelt sich um globale Alternativen, d.h. die Verletzung des Rasch-Modells betrifft nicht wenige, vereinzelte Items, nach deren Elimination das Rasch-Modell wieder Gültigkeit besäße, sondern es sind prinzipiell mehrere Items von einer Abweichung betroffen.

Die Präsentation der Modelle erfolgt nach abnehmender Kompatibilität mit den Annahmen eines MHM. Während die ersten beiden Alternativmodelle noch alle Forderungen des MHM erfüllen und somit stochastische Ordnung (siehe Kapitel 2.3)

[1] Da die Menge an Alternativmodellen für die zufällige Datenmatrix $\boldsymbol{X}$ beliebig groß ist, erfolgt eine Beschränkung auf in der Literatur häufig vorkommende Alternativmodelle.

garantieren, weisen die letzten beiden Modelle Verstöße gegen die MHM-Annahmen auf.

Jedes Alternativmodell wird in Kombination mit einer oder mehreren Teststatistiken dargestellt. Diese Prüfgrößen sollten sensitiv gegenüber der von dieser Alternative erzeugten Abweichung vom Rasch-Modell reagieren. Eine „gängige“ Unterteilung ist durch die „first-order“/„second-order“-Klassifikation gegeben. „First-order“-Statistiken fokussieren sich auf den Zusammenhang zwischen dem Summenwert und dem Antwortverhalten auf einem Item, während „second-order“-Statistiken auf den Zusammenhang zweier Items ausgerichtet sind. Erstere sollten diskrepante Verläufe in der Item-Response-Funktion entdecken. Letztere sollten gegenüber lokalen Abhängigkeiten sensitiv ausfallen. Es sei jedoch erwähnt, dass eine strikte Zuteilung von Teststatistiken zu bestimmten Alternativmodellen i.A. kaum haltbar ist. Viele Teststatistiken im Kontext des Rasch-Modells weisen mitunter beträchtliche Power gegen „fremde“ Alternativmodelle auf (siehe hierzu die Simulationsstudie von Suárez-Falcón und Glas (2003)). So besitzt beispielsweise der R_{2c}-Test, ein „generalized Pearson test“, der speziell für Verletzungen der lokalen Unabhängigkeit konstruiert wurde, ebenso bei Modellen, die eine abweichende Form der Item-Response-Funktion aufweisen, beträchtliche Power. Auch der R_0-Test („generalized Pearson test“) verfügt über Power bei abweichenden Item-Response-Funktionen. Dies erscheint besonders problematisch, da er ursprünglich zur Evaluation der im MML-Kontext postulierten Verteilungsannahme konstruiert wurde (Glas und Verhelst, 1989).

Die nachfolgende Präsentation einer Teststatistik anhand eines Alternativmodells ist insofern eher als Motivation zu verstehen denn als strikte Zuweisung. Gegenüber welchen Alternativen eine Prüfgröße Power aufweist und wie spezifisch ihr Verhalten ausfällt, ist letztendlich Gegenstand der Simulationsstudie.

Abschließend sei noch darauf hingewiesen, dass die Motivation der Teststatistiken anhand eines Alternativmodells, d.h. die Begründung, warum eine Prüfgröße für ein spezielles Alternativmodell geeignet erscheint, häufig eher informellen Charakter besitzt. Eine „exakte“ Argumentation mittels der Verteilung auf $\boldsymbol{\Sigma_{rc}}$ unter der Alternative ist sowohl aufgrund der Mächtigkeit der Menge als auch wegen analytischer Schwierigkeiten - nur unter dem Rasch-Modell ergibt sich eine „einfache“ Verteilung - nicht möglich.

4.2 Variable Itemdiskrimination

Modell

Das Alternativmodell für die erste Simulation ist das „two parameter logistic model“ (2PL-Modell). In diesem Szenario folgt das Antwortverhalten einem MHM. Die Item-Response-Funktionen lauten:

$$f_i(\theta) = \frac{\exp(a_i(\theta - \beta_i))}{1 + \exp(a_i(\theta - \beta_i))} \tag{4.1}$$

Ferner seien mindestens zwei a_i-Parameter (Diskriminationsparameter) verschieden. Ist der Träger der latenten Variable durch die Menge der reellen Zahlen gegeben, so liegt kein DMM vor. Die Item-Response-Funktionen zweier Items mit verschiedenem Diskriminationsparameter schneiden sich und die Ordnung der Items gemäß ihrer Schwierigkeit fällt für unterschiedliche Fähigkeitsbereiche verschieden aus. Des Weiteren ist durch die Einführung des Parameters a_i die Eigenschaft „LPO“ für Itempaare mit nicht identischen Diskriminationsparametern nicht mehr erfüllt.
Die Interpretation des Parameters a_i als Diskriminationsparameter (und damit als IRT-Analogon zum klassischen Maß der Trennschärfe) ergibt sich unmittelbar aus der Steigung der Item-Response-Funktion:

$$\frac{\partial f_i(\theta)}{\partial \theta} = a_i f_i(\theta)(1 - f_i(\theta)) \tag{4.2}$$

Der maximale Anstieg liegt vor, wenn $f_i(\theta) = \frac{1}{2}$, d.h. an der Stelle des Schwierigkeitsparameters. Der Wert des maximalen Anstiegs ist proportional zu a_i. In einem 2PL-Modell variiert folglich die maximale Diskriminationsfähigkeit[2]. Items mit hohem Diskriminationsparameter ermöglichen - lokal gesehen - eine bessere Unterscheidung zwischen den einzelnen Fähigkeitsstufen.
Eine weitere, alternative Interpretationsmöglichkeit von a_i orientiert sich an der Wirkung der latenten Variable auf das Antwortverhalten bezüglich Item i. Sieht man (4.1) als Regressionsmodell mit einer latenten erklärenden Variable, so ist die Wirkung der latenten Variable itemspezifisch. Der Zusammenhang zwischen latenter Variable und Item - gemessen am Odds-Ratio e^{a_i} - ist umso höher, je stärker das

[2]Auch in einem Rasch-Modell variiert - gemessen an der Steigung von $f_i(\theta)$ - die Diskriminationsfähigkeit der Items in Abhängigkeit von θ. Die *maximale* Diskriminationsfähigkeit ist jedoch für jedes Item identisch.

Item diskriminiert. So gesehen repräsentieren Items mit höherem a_i-Parameter das Konstrukt „besser“. Eine Änderung der latenten Variable zeigt hier stärkere Auswirkungen auf das Antwortverhalten.
Neben diesen die Assoziationsstruktur betreffenden Auswirkungen verändert der Parameter a_i auch die suffiziente Statistik für θ. An Stelle eines einfachen Summenscores ist bei bekannten[3] Diskriminationsparametern die gewichtete Summe $\sum_i a_i X_i$ nun suffizient für θ.
Diese durch das 2PL-Modell implizierten Verletzungen besitzen große praktische Relevanz. Datensätze mit variierender Itemdiskrimination dürften eher der Regelfall als die Ausnahme sein. In einer Diskussion des Rasch-Modells schließt Reckase (2009, S.21) etwa mit den Worten:

„*Item analysis results of this kind provide strong evidence that test items are not equal in discriminating power. These results imply that a model that has the same value for the maximum slope of the ICCs for all test items does not realistically describe the interaction between persons and items*“

Anmerkung: Der Ausruck „ICC“ steht synonym für die Item-Response-Funktion.

Tests

Ein parametrischer Test, der auf abweichende Formen der Item-Response-Funktionen reagiert, ist durch den Likelihood-Quotienten-Test von Andersen (1973) gegeben. Er beruht auf der Tatsache, dass bei Gültigkeit des Rasch-Modells die Itemparameter anhand der Daten jeder Scoregruppe - via CML-Methode - konsistent geschätzt werden können. Darauf basierend vergleicht der Likelihood-Quotienten-Test die Parameterschätzungen in verschiedenen Scoregruppen (d.h. für jede Scoregruppe wird ein seperates Rasch-Modell angepasst) mit der Schätzung eines gemeinsam gültigen Rasch-Modells. Die Teststatistik lautet:

$$\lambda(\mathbf{X}) = -2\left(l(\mathbf{X}, \hat{\boldsymbol{\beta}}) - \sum_g l_g(\mathbf{X_g}, \hat{\boldsymbol{\beta}}_{\boldsymbol{g}})\right) \tag{4.3}$$

[3]Im Fall von unbekannten Diskriminationsparametern ist $\sum_i a_i X_i$ keine „reine“ Funktion der Daten mehr, sondern beinhaltet unbekannte Größen. Folglich stellt der Ausdruck $\sum_i a_i X_i$ - definitionsgemäß - keine Statistik dar.

$\mathbf{X_g}$ bezeichnet hierbei die Zeilen der Datenmatrix $\boldsymbol{X}$, deren Summenscore in Scoregruppe g fällt. $\boldsymbol{\beta_g}$ ist der jeweils gruppenspezifische Itemparametervektor des Alternativmodells (Rasch-Modell für jede Gruppe). Die Log-Likelihood $l_g()$ entspricht der logarithmierten CML-Likelihood gemäß (2.12), beschränkt auf alle Personen aus der jeweiligen Scoregruppe. $l()$ entspricht dem Logarithmus von (2.12), wobei alle Personen zur Likelihood beitragen. Es werden folglich zwei genestete Modell verglichen: Das größere Modell spezifiziert in jeder Scoregruppe ein eigenes Rasch-Modell. Das Teilmodell hingegen legt ein globales Rasch-Modell fest. Die Prüfgröße (4.3) ist bei Gültigkeit des Teilmodells asymptotisch χ^2-verteilt. Die Freiheitsgrade entsprechen der Differenz an Parametern der beiden Modelle.
Voraussetzung für die Wohldefiniertheit von (4.3) ist die Existenz des Schätzers $\hat{\boldsymbol{\beta}}_{\boldsymbol{g}}$ in jeder Gruppe g. Wird ein Item von allen Personen einer Gruppe richtig beantwortet, so existiert der Schätzer $\hat{\boldsymbol{\beta}}_{\boldsymbol{g}}$ nicht. Analoges gilt für den Fall, dass ein Item von keiner Person gelöst wird. Der Ausschluss dieser beiden Szenarien stellt nur eine notwendige - jedoch keine hinreichende - Bedingung für die Existenz des CML-Schätzers $\hat{\boldsymbol{\beta}}_{\boldsymbol{g}}$ dar[4]. Im Falle der Nichtexistenz erfolgt i.d.R. eine Elimination passender Items, bis die resultierende Statistik wohldefiniert ist. Dieses Vorgehen findet auch in den Simulationen von Kapitel 5 statt.

Der Test scheint für das Problem der variablen Itemdiskrimination angemessen. Er benutzt als zentrales Element die Suffizienz des Summenscores, die durch das 2PL-Modell verletzt wird. Die Gruppierung nach Summenscore ist auch bei Abwesenheit eines Rasch-Modells eine Näherung für die Gruppierung nach der latenten Variable. Bei Gültigkeit eines MHM ist die latente Variable stochastisch geordnet, bezogen auf den Summenscore. Die Gruppe mit höherem Score besitzt somit tendenziell die höhere Fähigkeit. Insofern lässt sich obige Teststatistik auch als ein Vergleich von Itemparametern (Itemschwierigkeiten) in verschiedenen Bereichen der latenten Variable betrachten. In einem Rasch-Modell ergibt sich wiederum unabhängig von dem betrachteten Bereich die gleiche Ordnung der Items bezüglich ihrer Schwierigkeit. In einem 2PL-Modell hingegen kann für die höhere Scoregruppe (stochastisch gesehen die Gruppe mit der höheren Fähigkeit) eine andere Itemordnung resultieren als in der Gruppe mit geringem Summenscore. Folglich differieren die Itemschwierigkeiten (gemeint sind die relativen Positionen der Itemparameter) zwischen den Scoregrup-

[4]Notwendige und hinreichende Bedingungen für die Existenz lassen sich anhand des Begriffs der Binomialtransformation formulieren (Ponocny, 2001).

pen, und ein Modell mit getrennten Itemparametern ermöglicht einen besseren Fit als ein gemeinsames Rasch-Modell. Dies resultiert wiederum in einem erhöhten Wert der Prüfgröße (4.3).
Eine offene Frage betrifft die Art der Gruppierung. Die Gruppierung nach allen potentiell beobachtbaren Summenscores ermöglicht zwar eine differenzierte Einteilung, ist jedoch nicht praktikabel. Manche Summenscores treten mitunter nicht auf oder weisen sehr geringe Fallzahlen auf. Als Konsequenz können sich starke Abweichungen von der asymptotischen χ^2-Verteilung ergeben[5]. In Simulationsstudien zeigt sich - für eine gewählte Einteilung - ein akzeptables Verhalten insofern eine Gruppengröße von jeweils mindestens 50 Personen gewährleistet ist (Gustafsson, 1977).
Für die in Kapitel 5 folgenden Simulationen wurde stets der Split am Mittelwert $(\frac{1}{n}\sum_v r_v)$, d.h. eine Zwei-Gruppen-Aufteilung, verwendet. Gerade bei einem geringen Stichprobenumfang von 100 Personen scheint dieser - angesichts der obigen Faustregel - gerechtfertigt. Für die höheren Fallzahlen wurde er aus Gründen der Vergleichbarkeit beibehalten.

Eine nonparametrische Prüfgröße der kombinatorischen Testklasse basiert auf der Item-Test-Korrelation (Trennschärfe), dem klassischen Maß für Itemdiskrimination. Die von Chen und Small (2005) vorgeschlagene Prüfgröße lautet:

$$Y := \sum_{i=1}^{k} \frac{(d_i(\boldsymbol{X}) - \mu_i)^2}{\sigma_i^2} \tag{4.4}$$

d_i bezeichnet hierbei die Trennschärfe des i-ten Items. μ_i stellt den Erwartungswert von d_i gemäß der Gleichverteilung auf $\boldsymbol{\Sigma_{rc}}$ dar. Dieser lässt sich direkt - unter Rückgriff auf Satz 1 - anhand der mittleren i-ten Trennschärfe der generierten Matrizen schätzen. Analoges gilt für σ_i^2, der Varianz von d_i. Der Test lehnt ab, wenn höchstens $(100 \cdot \alpha)\%$ der simulierten Matrizen einen nicht geringeren Wert der Statistik aufweisen als die beobachtete Matrix[6].
In einer Simulation von Chen und Small (2005), die sich allerdings auf Tests bestehend aus nur sechs Items beschränkte, zeigte sich für die Größe Y eine deutlich höhere Teststärke - versus 2PL-Modell - verglichen mit dem R_{1c}-Test. Der R_{1c}-Test -

[5]Zudem tritt das Problem der Wohldefiniertheit mit wachsender Gruppenanzahl verstärkt auf. Dies kann somit in einer erhöhten Itemelimination bei zunehmend feinerer Gruppierung resultieren.

[6]Eine zusätzliche Ablehnung bei zu geringen Werten der Teststatistik wäre auch denkbar, wird aber hier nicht weiter verfolgt.

ein Vertreter der „generalized Pearson tests“ - besitzt wiederum eine dem Likelihood-Quotienten-Test ähnliche Gütefunktion bei 2PL-Modellen (Suárez-Falcón und Glas, 2003). Es kann folglich vermutet werden, dass die kombinatorische Prüfgröße Y - zumindest bei kleinen Testlängen - dem parametrischen Likelihood-Quotienten-Test überlegen ist. Die Frage, ob sich diese Beziehung auch im Fall längerer Tests fortsetzt, ist u.a. Gegenstand von Kapitel 5.2.
Abschließend sei noch erwähnt, dass Y und λ als „first-order“-Statistiken klassifiziert werden können. Beide fokussieren sich nicht auf den Zusammenhang der Itempaare, sondern sind auf den Zusammenhang zwischen Summenscore und Item ausgerichtet. Bei Y ist dies direkt anhand der Korrelation zwischen Summenscore und Item ersichtlich. Für λ erkennt man dies mittels der suffizienten Statistiken des Obermodells (=Rasch-Modell in jeder Scoregruppe). Diese sind durch die Anzahl richtiger Lösungen auf einem Item in der jeweiligen Scoregruppe gegeben. Folglich wird hier ebenfalls das Zusammenspiel zwischen Summenscore und Antwortverhalten auf den Items verwendet und es handelt sich somit um eine „first-order“-Statistik.

Bemerkung. An Stelle einer gewichteten Distanz zwischen dem Trennschärfenvektor $\boldsymbol{d}$ und seinem Erwartungswert $\boldsymbol{\mu_d}$ sind auch Modifikationen der Prüfgröße Y denkbar. Zum Einen wäre eine ungewichtete Abwandlung der Größe (4.4) möglich, d.h. nur quadrierte Abweichungen $(\boldsymbol{d}-\boldsymbol{\mu_d})'(\boldsymbol{d}-\boldsymbol{\mu_d})$ (oder Beträge) werden bei der Definition der Teststatistik verwendet. Zum Anderen ließe sich (4.4) zu einer „Mahalanobis-Distanz“ erweitern. Die resultierende Prüfgröße

$$Y_m := (\boldsymbol{d}-\boldsymbol{\mu_d})'\boldsymbol{\Sigma_d^-}(\boldsymbol{d}-\boldsymbol{\mu_d})$$

berücksichtigt somit auch Abhängigkeiten innerhalb des Vektors $\boldsymbol{d}$.
Neben diesen die Gewichtung betreffenden Veränderungen sind auch Modifikationen am Diskriminationsmaß möglich. Der Trennschärfenvektor $\boldsymbol{d}$ wäre z.B. durch den Vektor der Skalierungskoeffizienten[7] $\boldsymbol{h}$ ersetzbar. Eine Diskussion der möglichen Konsequenzen dieser Veränderung findet sich in Anhang B.

[7]Zur Definition des Skalierungskoeffizienten siehe Anhang B bzw. Kapitel 6.2.

4.3 Variable Itemdiskrimination mit Ratetendenzen

Modell

Ein besonders im Falle von Multiple-Choice-Aufgaben auftretendes Problem stellt das zufällige Raten der korrekten Antwort dar. Hieraus ergeben sich nicht nur verminderte Reliabilitäten, da das Raten als eine Form des Messfehlers betrachtet werden kann, sondern auch die Notwendigkeit das Item-Response-Modell adäquat anzupassen. Nach dem Rasch-Modell besitzt eine Person mit niedriger Fähigkeit eine äußerst geringe Lösungswahrscheinlichkeit für ein schweres Item. Dies kann in Konflikt mit der Annahme des zufälligen Ratens stehen, die stets für jedes Fähigkeitslevel eine gewisse, mitunter beträchtliche Mindestwahrscheinlichkeit, ein Item zu lösen, garantiert.
Das Alternativmodell dieses Abschnitts berücksichtigt Ratetendenzen durch die Verwendung eines zusätzlichen Parameters. Die Item-Response-Funktionen dieses „three parameter logistic model“ (3PL-Modell) lauten[8]:

$$f_i(\theta) = c_i + (1 - c_i)\frac{\exp(a_i(\theta - \beta_i))}{1 + \exp(a_i(\theta - \beta_i))} \tag{4.5}$$

Analog zum 2PL-Modell verfügen die Items über (gegebenenfalls) unterschiedliches Diskriminationspotential, modelliert anhand des Parameters a_i. Der Parameter c_i hingegen erlaubt den Einbezug des Rateverhaltens. Eine Person mit sehr niedriger Fähigkeit besitzt stets mindestens eine Wahrscheinlichkeit von c_i, das i-te Item zu lösen. Für hohe c_i-Werte verläuft die Item-Response-Funktion nahezu konstant. Folglich geht die Korrelation mit anderen Items bei wachsendem c_i gegen Null.
Die Steigung

$$\frac{\partial f_i(\theta)}{\partial \theta} = (1 - c_i)a_i f_i^*(\theta)(1 - f_i^*(\theta))$$

ist gegenüber einem 2PL-Modell mit identischen a_i-Größen abgeflacht[9]. Der Parameter c_i beeinflusst somit auch die Diskriminationsfähigkeit eines Items. Bezüglich der maximalen Diskriminationsfähigkeit der Items spielt $1 - c_i$ eine zum Parameter a_i analoge Rolle.

[8] Es gelten weiterhin die Annahmen des MHM.
[9] f_i^* entspricht der Form (4.1) des 2PL-Modells.

Typischerweise wird c_i als Rateparameter bezeichnet. Reckase (2009, S.32) führt jedoch an, dass Schätzungen für c_i in realen Datensätzen häufig unter dem theoretisch zu erwartenden Wert (dieser entspricht dem Inversen der Anzahl an Kategorien des Items) liegen und plädiert daher dafür, den Parameter als „untere Asymptote“ zu bezeichnen.
Fernab der Interpretation dieses Parameters ergeben sich, neben den bereits im Kontext des 2PL-Modells erörterten Konsequenzen in Bezug auf variierende Itemdiskrimination, statistische „Probleme“: Das 3PL-Modell verfügt - auch bei bekannten a_i-Werten - über keine eindimensionale suffiziente Statistik für den Personenparameter. Die Inferenz bezüglich θ_v ist auf den gesamten Responsevektor $\boldsymbol{x_v}$ angewiesen.

Tests

Da die Item-Response-Funktion direkt an den Fall des 2PL-Modells anknüpft, liegt es nahe, die Teststatistiken im Kontext des 2PL-Modells auch auf den Fall des 3PL-Modells anzuwenden. Eine weitere Motivation für diese Parallele liefert eine Taylorentwicklung erster Ordnung um β_i:

$$f_i(\theta) \approx \frac{1}{2}(1 + c_i) + \frac{(1 - c_i)a_i}{4}(\theta - \beta_i) \tag{4.6}$$

Falls die Verteilung der Fähigkeit stark um β_i konzentriert ist, verhält sich das Item nahezu wie ein Item mit linearer Response-Funktion. Betreffend der Steigung dieser linearen Näherung stehen $(1-c_i)$ und a_i in einem austauschbaren Verhältnis. Folglich kann in der linearen Approximation das 3PL-Item wie ein Item eines 2PL-Modells mit Diskriminationsparameter $\alpha_i := (1 - c_i)a_i$ behandelt werden.
Dies bildet eine informelle Grundlage für die Übertragung der Teststatistiken des 2PL-Modells auf das um Rateeffekte erweiterte 3PL-Modell.

4.4 Lokale stochastische Abhängigkeit

Modell

Für die Simulation lokal abhängiger Daten wird die gemeinsame Wahrscheinlichkeitsfunktion faktorisiert:

$$P(\boldsymbol{X_v} = \boldsymbol{x_v}|\, \theta_v) = P(X_{v1} = x_{v1}|\, \theta_v) \prod_{l>1} P(X_{vl} = x_{vl}|\, \theta_v,\, x_{v(l-1)}, \ldots, x_{v1})$$

Unter der Annahme der bedingten Unabhängigkeit, gegeben die Antwort auf das vorherige Item (und θ), gilt:

$$P(X_{vl} = x_{vl}|\, \theta_v,\, x_{v(l-1)}, \ldots, x_{v1}) = P(X_{vl} = x_{vl}|\, \theta_v,\, x_{v(l-1)}) \tag{4.7}$$

Die bedingte Wahrscheinlichkeit in (4.7) wird nun parametrisiert[10].

$$P(X_{vl} = x_{vl}|\, \theta_v,\, x_{v(l-1)}) = \frac{\exp(x_{vl}(\theta_v - \beta_l + (x_{v(l-1)} - 0.5)\delta_{(l-1)l}))}{1 + \exp(\theta_v - \beta_l + (x_{v(l-1)} - 0.5)\delta_{(l-1)l})} \tag{4.8}$$

Der Parameter $e^{\delta_{(l-1)l}}$ entspricht dem bedingten Odds-Ratio zwischen Item l und Item $l-1$, gegeben θ. Falls $\delta_{(l-1)l} > 0$ gilt, erhöhen sich die Odds, Item l zu lösen, um den Faktor $(e^{\delta_{(l-1)l}} - 1)$, wenn das vorherige Item $l-1$ gelöst wurde (relativ zur „Nichtlösung“ des Items). Dies könnte inhaltlich bedeuten, dass die Aufgabe $l-1$ wichtige Informationen zur Lösung der nächsten Aufgabe beinhaltet, so dass eine korrekte Antwort die Lösungswahrscheinlichkeit des nachfolgenden Items erhöht.

Als Wahrscheinlichkeitsfunktion der gesamten Datenmatrix ergibt sich unter der weiterhin gültigen Annahme der Unabhängigkeit des Antwortverhaltens der verschiedenen Personen:

$$P(\boldsymbol{X}|\, \boldsymbol{\theta},\, \boldsymbol{\beta},\, \boldsymbol{\delta}) = \prod_v \left(P(X_{v1} = x_{v1}|\, \theta_v) \prod_{l>1} P(X_{vl} = x_{vl}|\, \theta_v,\, x_{v(l-1)}) \right) \tag{4.9}$$

Das Rasch-Modell resultiert, wenn:

$$\delta_{12} = \delta_{23} = \ldots \delta_{(k-1)k} = 0$$

[10] Für die Wahrscheinlichkeitsfunktion von X_{v1} gegeben θ_v gelte die „übliche“ Form (2.3) des Rasch-Modells.

In diesem Fall sind alle Abhängigkeiten gegeben θ_v eliminiert. Als wesentlicher Unterschied zum Rasch-Modell lässt sich somit das Vorhandensein von bedingten Itemassoziationen ausmachen. Tragen im Rasch-Modell die restlichen Items - bei bekannter Fähigkeit - nicht zur Vorhersage des Antwortverhaltens auf ein spezifisch gewähltes Item bei, so gilt dies nicht für lokal abhängige Daten.

Tests

Sind zwei Items (gegeben θ) nicht unabhängig, so besteht zwischen ihnen eine (bedingte) Korrelation[11]. Die Korrelation bildet zugleich den natürlichen Ansatzpunkt für Teststatistiken, die auf Itemabhängigkeiten zielen, da sich lokale Abhängigkeiten in der Assoziationsstruktur der Items widerspiegeln (siehe z.B. Satz 3 aus Kapitel 3.3).
Ein Beispiel ist der von Ponocny (2001) vorgeschlagene „overall test“ auf lokale Unabhängigkeit, für den Ponocny anmerkt, dass er seiner Erfahrung nach zu den teststärksten Prüfgrößen innerhalb der von ihm vorgeschlagenen kombinatorischen Tests gehört. Die Testgröße basiert auf den Itemkorrelationen $r_{ij} := \mathrm{Cor}(\boldsymbol{x_{(i)}}, \boldsymbol{x_{(j)}})$ und lautet:

$$T_{11} := \sum_{i,j} |r_{ij}(\boldsymbol{X}) - \rho_{ij}| \tag{4.10}$$

ρ_{ij} stellt den mittels Satz 1 approximierten Erwartungswert von r_{ij} gemäß der Gleichverteilung auf $\boldsymbol{\Sigma_{rc}}$ dar. Der Test lehnt bei zu hohen Werten der Prüfgröße T_{11} ab.
Prinzipiell ist es auch denkbar, die Summation in (4.10) auf benachbarte Itempaare zu begrenzen. Damit geht ein Gewinn an Teststärke bezüglich den entsprechenden Alternativen, d.h. wenn vorrangig benachbarte Paare in Verletzungen des Rasch-Modells involviert sind, einher, jedoch ebenso ein Verlust an Sensitivität gegenüber komplexeren Itemabhängigkeiten.
An Stelle des Korrelationskoeffizienten wären auch andere „Assoziationsmaße“ verwendbar. Ein ähnlicher Test resultiert z.B., wenn die Korrelation r_{ij} durch die Anzahl simultaner Lösungen $\sum_v X_{vi}X_{vj}$ ersetzt wird. Dies folgt direkt aufgrund der Konstanz der Randsummen der Matrizen, da das einzige variierende Element der

[11] Für zwei binäre Variablen sind Unabhängigkeit und Unkorreliertheit äquivalent (Grimmett und Stirzaker, 2001b, S.174).

Korrelation r_{ij} somit durch den Ausdruck $\sum_v X_{vi}X_{vj}$ gegeben ist. Die Anzahl simultaner Lösungen wiederum stellt die zentrale Größe für Tests auf lokale stochastische Abhängigkeit („second-order"-Tests) innerhalb der Klasse der „generalized Pearson tests" dar (Glas, 1988). Positive Abhängigkeit ($\delta > 0$) führt zu einer erhöhten Anzahl simultaner Lösungen gegenüber einem Rasch-Modell.
Die Beziehungen zwischen den einzelnen Assoziationsmaßen und ihre Implikationen bezüglich der Anwendung sind in Anhang B näher dargelegt.

4.5 Mehrdimensionalität

Modell

Unter Beibehaltung der lokalen stochastischen Unabhängigkeit ist das Alternativmodell dieses Abschnitts durch die folgende Item-Response-Funktion spezifiziert:

$$f_i(\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{a_i'\theta} - \beta_i)}{1 + \exp(\boldsymbol{a_i'\theta} - \beta_i)} \tag{4.11}$$

$\boldsymbol{\theta}$ bezeichnet eine mehrdimensionale, latente Variable, die (üblicherweise) einer Normalverteilung folgt:

$$\boldsymbol{\theta} \sim N_p(\boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta})$$

Die Kovarianzmatrix $\boldsymbol{\Sigma_\theta}$ des latenten Vektors kann dabei o.B.d.A. als Einheitsmatrix betrachtet werden. Gilt nämlich $\boldsymbol{\Sigma_\theta} \neq \boldsymbol{I}$, so folgt für

$$\boldsymbol{\theta^*} := \boldsymbol{\Sigma_\theta^{-1/2}\theta} \quad [\boldsymbol{\theta^*} \sim N(\boldsymbol{\mu_\theta^*}, \boldsymbol{I})]$$

ein empirisch ununterscheidbares Modell mit neuer Ladungsmatrix (die Ladungsmatrix ist die Matrix $\boldsymbol{A}$, deren i-te Zeile den Diskriminationsvektor $\boldsymbol{a_i}$ des i-ten Items beinhaltet) $\boldsymbol{A^*} := \boldsymbol{A\Sigma_\theta^{1/2}}$:

$$f_i(\boldsymbol{\theta^*}) = \frac{\exp(\boldsymbol{a_i^{*\prime}\theta^*} - \beta_i)}{1 + \exp(\boldsymbol{a_i^{*\prime}\theta^*} - \beta_i)}$$

Sämtliche Antwortwahrscheinlichkeiten bleiben durch diese Transformation erhalten.

Die Interpretation der Ladungsmatrix kann auf zwei Arten vorgenommen werden: Anhand von (4.11) lässt sich erkennen, dass $e^{a_{ij}}$ genau dem Odds Ratio, Item i zu

lösen, bei einer Erhöhung des Wertes der j-ten latenten Variable um eine Einheit entspricht[12]. Je größer die Ladung, desto stärker fällt somit diese bedingte Assoziation zwischen Item und j-ter latenter Variable aus.
Ein zweiter Zugang verläuft über die Faktorenanalyse. Nimmt man an, dass die dichotomen Antworten als Folge eines Schwellenwertmodells (Lee, 2007, S.178) entstehen, so lässt sich die Interpretation der Ladungsmatrix analog der Ladungsmatrix einer Faktorenanalyse vornehmen. Genauer: Postuliert man einen k-dimensionalen Zufallsvektor $\boldsymbol{Y_v}$, der einem faktorenanalytischen Modell folgt ($\boldsymbol{\Lambda}$ feste $k \times p$ Matrix, $\boldsymbol{U_v}$ und $\boldsymbol{\theta_v}$ unabhängig, $U_{v1}, \ldots, U_{vk}$ unabhängig),

$$\boldsymbol{Y_v} = \boldsymbol{\Lambda\theta_v} + \boldsymbol{U_v},$$

sowie einen Zusammenhang zwischen $\boldsymbol{Y_v}$ und dem beobachtbaren Responsevektor $\boldsymbol{X_v}$ mittels eines Schwellenwertes,

$$X_{vi} = 1 \quad \Leftrightarrow \quad Y_{vi} < -\beta_i,$$

so ergibt sich ($\boldsymbol{\lambda_i}$ bezeichnet hierbei die i-te Zeile der Matrix $\boldsymbol{\Lambda}$):

$$P(X_{vi} = 1|\,\boldsymbol{\theta_v}) = P(Y_{vi} < -\beta_i|\,\boldsymbol{\theta_v}) = P(U_{vi} < -\boldsymbol{\lambda_i'\theta_v} - \beta_i) = \Phi(-\boldsymbol{\lambda_i'\theta_v} - \beta_i)$$

Wählt man für Φ die logistische Verteilungsfunktion, so folgt ein Modell analog zu (4.11). Somit stehen potentiell zwei verschiedene Interpretationsarten für die Diskriminationsmatrix $\boldsymbol{A}$ zur Verfügung.

Ferner liefert die Matrix $\boldsymbol{A}$ eine Basis zur Klassifikation der durch (4.11) gegebenen Modelle. Lädt jedes Item auf genau einer Dimension, d.h. in jeder Zeile der Ladungsmatrix befindet sich nur ein Eintrag ungleich Null, so spricht man von „Between-Item-Multidimensionality“. Kann ein Item hingegen gegenüber mehreren Dimensionen sensitiv sein, so liegt „Within-Item-Multidimensionality“ vor (Adams u.a., 1997; Adams und Wu, 2007). Die in Kapitel 5 folgende Simulation befasst sich mit dem Fall der „Between-Item-Multidimensionality“. Items können somit immer klar (genau) einer Dimension zugeordnet werden.
Das Rasch-Modell ergibt sich als Spezialfall, wenn alle Items den gleichen Diskriminationsvektor ($\boldsymbol{a_1} = \boldsymbol{a_2} = \cdots = \boldsymbol{a_k}$) besitzen.

[12] Dies gilt nur bei festen Werten der übrigen latenten Größen.

Tests

Aufgrund des indirekten Bezugs zur Faktorenanalyse stellt die Korrelationsmatrix - als zentrales Element der Faktorenanalyse - einen naheliegenden Zugang zur Prüfung auf Mehrdimensionalität dar.
Ebenso lässt sich über die Verknüpfung zur lokalen Abhängigkeit argumentieren: Mehrdimensionalität impliziert, dass eine einzelne Variable nicht ausreichend ist, um die (bedingte) Korrelation der Items zu „beseitigen“. Folglich besteht nach Bedingen auf *eine* latente Größe noch lokale Abhängigkeit zwischen den Items. Dies wiederum bildet eine Verbindung zur Prüfung auf lokale Abhängigkeit. Betrachtet man z.B. den Fall eines zweidimensionalen Item-Response-Modells gemäß (4.11) mit nichtnegativer Ladungsmatrix[13] und reduziert dieses auf ein eindimensionales Modell, indem lediglich die erste Komponente θ_1 Berücksichtigung findet, so gilt für die bedingte Kovarianz zweier Items:

$$\mathrm{Cov}(X_i, X_j \,|\, \theta_1) = \mathrm{Cov}(f_i(\boldsymbol{\theta}), f_j(\boldsymbol{\theta}) \,|\, \theta_1) \geq 0$$

Die erste Gleichung folgt dabei aus der bedingten Unabhängigkeit gegeben $\boldsymbol{\theta}$. Die Relation „$\geq$“ ergibt sich, da f_i, f_j monoton wachsende Funktionen in θ_2 darstellen und da die Kovarianz bezüglich der (bedingten) Verteilung einer eindimensionalen Zufallsvariable (θ_2) gebildet wird. Nach Lemma 2 (Kapitel 3.3) ist eine eindimensionale Zufallsvariable assoziiert. Folglich resultiert ein eindimensionales Item-Response-Modell, in dem die Items eine nichtnegative bedingte Korrelation aufweisen. Ferner lässt sich anhand des Beweises von Lemma 2 erkennen, dass für *streng* monoton wachsende Funktionen die Relation „$\geq$“ durch die Relation „$>$“ substituiert werden kann. Folglich liegt bei (in θ_2) streng monotonen Item-Response-Funktionen positive lokale Abhängigkeit vor. Das mehrdimensionale Modell wurde somit auf ein eindimensionales Modell mit lokal abhängigen Items reduziert. Aus diesem Blickwinkel betrachtet, erscheint eine strikte Trennung dieser beiden eng verzahnten Modellverletzungen nicht haltbar. Lokale Abhängigkeit und Mehrdimensionalität sind verbundene Eigenschaften eines Item-Response-Modells. Es scheint daher auch nicht ungewöhnlich, dass Tests, die auf lokale Abhängigkeiten sensitiv sind, auch gegenüber Mehrdimensionalität empfindlich reagieren (Suárez-Falcón und Glas, 2003).

[13] „Nichtnegative Ladungsmatrix“ bedeutet, dass jeder Diskriminationsvektor aus nichtnegativen Elementen besteht. Dies bildet häufig eine weitere Standardannahme im Kontext des Modells (4.11) und wird auch als kompensatorisches Modell bezeichnet (Reckase, 2009).

Die Separierung der folgenden Teststatistiken von der Prüfgröße T_{11} ist unter diesen Gesichtspunkten als eher willkürlich anzusehen.

Ein erster, im Paket *RaschSampler* (Verhelst u.a., 2007) realisierter Test der kombinatorischen Klasse bezieht sich - analog der T_{11}-Prüfgröße - auf Itemkorrelationen.

$$\phi(\boldsymbol{X}) := \max_{i,j} r_{ij}(\boldsymbol{X}) - \min_{i,j} r_{ij}(\boldsymbol{X})$$

Die Spannweite der Interitemkorrelationen dient als Teststatistik. Der Test lehnt ab, wenn der beobachtete Wert das $(1-\alpha)/2$-Quantil überschreitet bzw. das $(\alpha/2)$-Quantil unterschreitet.

Eine weitere Möglichkeit bietet ein von Ponocny (2001) vorgeschlagener, kombinatorischer Test auf Eindimensionalität. Er basiert auf der Varianzformel

$$\mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i), \tag{4.12}$$

deren Gültigkeit mit der Annahme der Unabhängigkeit verbunden ist. Gilt Eindimensionalität, so ist obige Formel auf die bedingte Varianz des Summenscores anwendbar. Bei Anwesenheit einer zweiten Dimension hingegen gilt (4.12) nur bei zusätzlichem Bedingen auf die zweite latente Größe[14]. Bestehen erhöhte (relativ zum Rasch-Modell) Abhängigkeiten zwischen gewissen Itempaaren, so spiegeln sich diese in der Streuung des Summenwerts einer geeignet gewählten Itemteilmenge (Subtest) wider.
Die Prüfgröße des Tests lautet:

$$\sigma_r^2(\boldsymbol{X}) = \mathrm{Var}(\boldsymbol{r}_S) \tag{4.13}$$

$\boldsymbol{r}_S$ stellt hierbei den Vektor der Summenwerte, gebildet über eine ausgewählte Item-Teilmenge S, dar. Der Test lehnt bei zu großen Werten der Prüfgröße ab[15].
Problematisch erscheint die Wahl der entsprechenden Subskala. Nur in wenigen Fällen dürften konkrete Vermutungen vorliegen, die eine adäquate Spezifikation von S gewährleisten. Für die Simulationsstudien soll daher eine konstante Spezifikation von S, bestehend aus den Items der ersten Testhälfte, erfolgen. Der so definierte Test sollte gegenüber einem Modell, in dem die erste Testhälfte eine eigene Skala bildet (ein anderes Merkmal misst), sensitiv reagieren.

[14]Dieser Überlegung bildet gleichfalls die Basis für einen recht allgemeinen nonparametrischen, asymptotischen Test auf Eindimensionalität (Stout, 1987).

[15]Ein „zweiseitiger“ Test, der zusätzlich bei zu geringen Werten ablehnt, wäre ebenso durchführbar (siehe auch die Analyse des IST 2000 R in Kapitel 6.3).

Kapitel 5

Simulationsstudie

5.1 Simulationsdesign und Fehler erster Art

Design

Für die folgenden Simulationen wurden jeweils vier Stufen des Stichprobenumfangs, entsprechend den Fällen von kleinen (n=100, n=250), moderaten (n=500) sowie größeren (n=1000) Stichprobenumfängen, mit jeweils drei Testlängen (k=10, 20, 30) kombiniert. In jedem Alternativmodell (4.2 - 4.5) wurden für jede der 12 Kombinationen von Item- und Personenanzahlen $m = 500$ Datensätze simuliert. Dies erfolgte für jedes Alternativmodell in drei Ausführungen, die sich jeweils in der Stärke der Abweichung vom Rasch-Modell unterschieden. Pro Alternativmodell ergaben sich somit $4 \times 3 \times 3$ Simulationszenarien (Stichprobenumfang$\times$Testlänge$\times$Abweichungsgrad) mit je 500 Datenmatrizen.
Die konkrete Generierung einer Matrix $\boldsymbol{X}$ richtet sich nach dem Alternativmodell und wird in dem entsprechendem Abschnitt behandelt. Einige allgemeine Vorbemerkungen können dennoch getroffen werden:

- Itemparameter β_i entstammten stets unabhängig einer Standardnormalverteilung. Die Ziehungen erfolgten für jede der $m = 500$ Matrizen von Neuem.
- Analog zu den Itemparametern stammten Personenparameter, sofern es sich um ein eindimensionales Alternativmodell handelt, aus einer Standardnormal-

verteilung. Es gelten die gleichen Bemerkungen wie im Fall der Itemparameter $\boldsymbol{\beta}$.

- Um den Anteil der Matrizen, für die der Likelihood-Quotienten-Test wohldefiniert ist, zu erhöhen, wurde eine generierte Datenmatrix nur dann akzeptiert, wenn sie kein Item beinhaltete, welches von allen oder keinen Personen gelöst wurde[1]. Ansonsten erfolgte die Generierung einer neuen Datenmatrix. Da trotz dieser Maßnahme die Existenz des Likelihood-Quotienten-Tests nicht gewährleistet sein muss, wurde in den übrigen Fällen durch Elimination der verantwortlichen Items der Likelihood-Quotienten-Test bezüglich des verkürzten Tests berechnet[2]. Als Gruppierungskriterium diente stets die Aufteilung am Mittelwert.

- Die Parameter der Markov-Kette lagen bei: burn-in = 100, step = 16 und $n_{eff} = 1200$. Dieselben 1200 Matrizen bildeten die Grundlage für jede der vier kombinatorischen Prüfgrößen.

- Das Signifikanzniveau betrug - konstant über alle Simulationen - 5%.

Des Weiteren erfolgte keine spezifische Betrachtung, d.h. jede in Kapitel 4 dargelegte Teststatistik wurde unter jedem Alternativmodell untersucht. Dies ermöglicht ein Urteil über die Eindeutigkeit der Modelldiagnostik. Erbringt eine Statistik z.B. nur unter dem 2PL-Modell eine hohe Teststärke, so ist der Rückschluss auf variierende Itemdiskrimination bei Vorlage eines signifikanten Wertes weniger problematisch als bei einer Teststatistik, die auf Modellverletzungen mehrerer Art reagiert. Zur systematischen Fehlersuche liefert eine solche, spezifische Teststatistik einen deutlich höheren Beitrag.

Fehler erster Art

Zunächst wurden zur Beurteilung des Fehlers erster Art für jede der 12 Dimensionierungen $m = 500$ Datenmatrizen gemäß dem Rasch-Modell erzeugt. Für die Simulation gelten die obigen Vorbemerkungen zum Design.

[1] Spalten mit lediglich einem Response besitzen zudem eine Varianz der Trennschärfe (bezüglich $\boldsymbol{\Sigma_{rc}}$) von Null. Dies würde zur Nichtexistenz der Prüfgröße Y führen.

[2] Dies entspricht der Vorgehensweise der Funktion LRtest() im *eRm*-Paket der Software R.

Tabelle 5.1: *Anteil (in %) verworfener Nullhypothesen - getrennt nach Prüfgröße - bei Rasch-homogenen Datenmatrizen in Abhängigkeit von n und k.*

n	k	Y	T_{11}	ϕ	σ_r^2	λ
100	10	6.0	4.8	4.4	4.8	4.0
	20	5.0	5.8	3.4	5.6	4.4
	30	6.0	4.8	6.2	4.6	6.2
250	10	5.2	4.2	6.0	5.4	5.8
	20	3.0	4.2	5.2	4.2	6.0
	30	6.2	5.4	4.2	6.2	5.8
500	10	5.8	7.0	4.0	5.0	5.0
	20	6.4	5.8	3.6	5.2	6.4
	30	5.0	5.4	6.4	4.8	6.2
1000	10	6.0	4.8	4.4	4.8	4.2
	20	7.8	6.0	6.4	5.0	6.0
	30	5.0	6.2	6.0	5.6	5.4

Ergebnisse. Tabelle 5.1 gibt für jede der vier kombinatorischen sowie für den parametrischen Test - getrennt nach der Dimension der Datenmatrix - den Anteil verworfener Nullhypothesen (aus $m = 500$ Datenmatrizen) wieder. Sowohl die nonparametrischen Tests als auch λ zeigen nahezu keine auffälligen Abweichungen vom nominellen Signifikanzniveau. Pro Teststatistik vollzogene, einseitige Binomialtests mit Normalverteilungsapproximation und Bonferroni-Korrektur ($s = 12$ simultane Tests) resultieren lediglich für die Testgröße Y in einem signifikanten Wert. Die Diskrepanz der Größe Y tritt in der Bedingung ($n = 1000, k = 20$) auf. Bei den benachbarten Dimensionen ($n = 1000, k = 10$) sowie ($n = 1000, k = 30$) zeigen sich jedoch keine auffälligen Werte, so dass ein einfacher Zusammenhang dieser Abweichung mit der Dimensionierung der Matrix unwahrscheinlich erscheint.

Bemerkung. Eine angemessene Methode zur Beurteilung der *Parameterwahl* der Markov-Kette ist durch eine simultane Betrachtung aller 4×12 Bedingungen gegeben. Die zugehörigen Binomialtests mit Bonferroni-Korrektur lehnen keine Hypothese ab. Die Parameterwahl scheint akzeptabel auszufallen. Es sei an dieser Stelle jedoch bemerkt, dass für wesentlich geringere Längen der Markov-Kette (z.B. $n_{eff} = 100$) erhöhte Abweichungen vom 5%-Niveau resultieren (siehe auch Anhang A).
Insgesamt kann festgehalten werden, dass alle Werte der Teststatistiken in einem akzeptablen Bereich liegen. Y weist vereinzelt ein leicht erhöhtes Fehlerniveau auf.

5.2 Variable Itemdiskrimination

Die Datenmatrizen zur Simulation variabler Itemdiskrimination entstammten dem 2PL-Modell aus Kapitel 4.2. Für jede der $m = 500$ Datenmatrizen wurden die Diskriminationsparameter der Items unabhängig gemäß einer Lognormalverteilung ($a_i = \exp(Z_i)$, $Z_i \overset{i.i.d}{\sim} N(0, \sigma^2)$) gezogen.
Mittels des Parameters σ, der auf drei Stufen variierte, wurde die Stärke der Abweichung festgelegt. Insgesamt erfolgten drei verschiedene Simulationen, mit wachsender Diskrepanz zum Rasch-Modell, die anhand zunehmender Werte des Parameters σ realisiert wurden.

Ergebnisse. Tabelle 5.2 gibt für jede Dimensionierung (n, k) und jeden Abweichungsgrad (σ) den Anteil signifikanter Matrizen - hinsichtlich jeder Prüfgröße - wieder. Betrachtet man vorerst die nonparametrischen Tests unter der Bedingung $\sigma = 0.12$, so ergeben sich auf den ersten Blick zwei Gruppen. Zum Einen die deutlich sensitiven Tests Y und T_{11}. Zum Anderen die Gruppe der Tests auf Eindimensionalität (ϕ, σ_r^2), die selbst bei hohem Stichprobenumfang nur sehr geringe Teststärken aufweisen. Innerhalb der ersten Gruppe erkennt man ferner, dass der Y-Test beständig höhere Werte als der T_{11}-Test besitzt. Für beide Gruppen gilt, dass die Teststärke mit steigendem Stichprobenumfang wächst. Gleiches gilt für eine wachsende Itemzahl[3]. Geht man zu stärkeren Abweichungen über $(\sigma = 0.25, \sigma = 0.5)$, so lassen sich obige Interpretationen mit einer Ausnahme beibehalten. Diese betrifft das Verhalten des ϕ-Tests, der sich bei moderatem bis großem Stichprobenumfang als ebenfalls sensitiv darstellt. In der Bedingung $\sigma = 0.5$ besteht für einen Stichprobenumfang von 250 Personen „bereits“ eine hohe Ablehnwahrscheinlichkeit. Die Güte des σ_r^2-Test hingegen liegt selbst bei hohem Stichprobenumfang deutlich unter 35%. Für $\sigma = 0.5$ erreicht σ_r^2 bei maximaler Personen- sowie Itemanzahl $(n = 1000, k = 30)$ einen Wert, der sogar unter den Werten der sensitiven Tests Y und T_{11} in der Bedingung mit minimaler Personen- und Itemanzahl liegt. Letztere weisen dort Teststärken von ca. 50% auf.
Die Monotonie-Aussagen betreffend n und k lassen sich auch auf den parametrischen Likelihood-Quotienten-Test übertragen. λ reagiert auf variable Itemdiskrimination, jedoch liegen fast alle Teststärken unterhalb der korrespondierenden Power des T_{11}-Tests.

[3]Wenn gleich dieser Effekt für die erste Gruppe wesentlich deutlicher erkennbar ist.

Tabelle 5.2: *Anteil als signifikant erklärter Datenmatrizen (in %) von $m = 500$ nach dem 2PL-Modell simulierten Datenmatrizen in Abhängigkeit von σ, n und k.*

σ	n	k	Y	T_{11}	ϕ	σ_r^2	λ
0.12	100	10	9.6	7.0	5.8	5.8	6.0
		20	11.8	7.2	6.0	3.2	9.8
		30	16.6	13.4	5.2	4.6	11.8
	250	10	14.2	12.0	6.2	6.8	9.8
		20	21.6	20.2	4.8	7.4	15.0
		30	35.4	28.6	7.6	6.0	22.0
	500	10	19.4	18.4	7.4	8.0	13.4
		20	50.6	39.8	10.0	8.0	31.8
		30	70.2	52.4	8.6	9.8	40.0
	1000	10	43.8	33.0	10.2	8.0	27.6
		20	81.8	70.6	14.0	11.0	65.6
		30	94.6	89.2	19.8	13.2	78.6
0.25	100	10	20.0	16.0	7.6	5.8	15.0
		20	37.8	28.8	6.0	8.0	24.8
		30	56.2	44.8	7.8	11.2	31.0
	250	10	45.2	33.8	13.4	10.8	31.8
		20	83.4	69.0	19.4	10.0	64.6
		30	96.6	91.4	22.6	15.2	81.2
	500	10	76.2	65.0	26.2	12.8	58.4
		20	98.2	94.2	44.2	18.8	90.6
		30	99.8	99.6	51.0	16.4	98.8
	1000	10	94.4	87.2	50.4	21.2	84.4
		20	100.0	99.6	67.2	23.4	99.4
		30	100.0	100.0	81.2	25.0	100.0
0.5	100	10	56.6	43.6	21.2	10.8	35.6
		20	89.8	83.8	32.0	13.8	70.8
		30	98.8	96.2	43.6	17.0	91.8
	250	10	92.8	85.0	53.0	17.4	77.6
		20	99.8	99.4	79.4	25.4	98.6
		30	100.0	100.0	90.4	26.0	100.0
	500	10	98.0	96.6	78.4	21.4	94.2
		20	100.0	100.0	95.4	28.0	100.0
		30	100.0	100.0	98.8	29.0	100.0
	1000	10	100.0	100.0	92.0	25.4	99.2
		20	100.0	100.0	99.4	32.0	100.0
		30	100.0	100.0	100.0	31.6	100.0

Bemerkung. Prinzipiell besitzt jede der aufgeführten Teststatistiken eine gewisse Sensitivität gegenüber variierender Itemdiskrimination. Für vier der fünf Prüfgrößen kann unter entsprechenden Bedingungen eine Power von eins erreicht werden. Alle Statistiken - auch die schwach reagierende Prüfgröße σ_r^2 - weisen einen Anstieg der Teststärke mit wachsendem Abweichungsgrad auf. Während diese Beobachtung im Fall von λ und Y gemäß den Erläuterungen aus Kapitel 4.2 plausibel scheint, soll im Folgenden eine *mögliche*, informelle Begründung für die Sensitivität der drei nonparametrischen Größen ϕ, T_{11} und σ_r^2 gegeben werden.

Das sensitive Verhalten dieser Gruppe könnte darauf zurückzuführen sein, dass stets in direkter oder indirekter Form die Kovarianz der Itempaare involviert ist. Bei den Teststatistiken T_{11} sowie ϕ in nahezu expliziter Form, aber auch für σ_r^2 lässt sich eine Beziehung zur Kovarianz - anhand der Varianz einer Linearkombination - herstellen. Bezeichnet etwa $\boldsymbol{X_1}$ die entsprechende Submatrix von $\boldsymbol{X}$, welche die für σ_r^2 relevanten Items beinhaltet[4] und bezeichnet $\boldsymbol{r_1}$ den Summenscorevektor bezüglich dieser Submatrix, so gilt:

$$\boldsymbol{r_1} = \boldsymbol{X_1 1}$$

Für die empirische Varianz folgt somit ($\boldsymbol{C} := \boldsymbol{I} - \frac{1}{n}\boldsymbol{11'}$ entspricht der Zentrierungsmatrix):

$$\mathrm{Var}(\boldsymbol{r_1}) = \frac{1}{n}\boldsymbol{r_1' C r_1} = \frac{1}{n}\boldsymbol{1' X_1' C X_1 1} = \boldsymbol{1' S_1 1}$$

Da die empirischen Varianzen der Items (Diagonalelemente von $\boldsymbol{S_1}$) bei festen Randsummen konstant sind, gehen in die Prüfgröße σ_r^2 lediglich Kovarianzen zwischen Itempaaren ein. Folglich lehnt der Test für den vorliegenden Fall ab, wenn die paarweisen Itemkovarianzen der ersten Testhälfte relativ zum Rasch-Modell erhöht ausfallen.

Die Beziehung zwischen der Kovarianz und den Diskriminationsparametern kann wiederum heuristisch durch ein Taylor-Polynom erster Ordnung hergestellt werden. Nach (2.1) gilt:

$$\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(f_i(\Theta), f_j(\Theta)) \tag{5.1}$$

Entwickelt man die Item-Response-Funktionen um den Schwierigkeitsparameter, so ergibt sich:

$$f_i(\theta) \approx \frac{1}{2} + \frac{a_i}{4}(\theta - \beta_i)$$

[4] Für den hier zu diskutierenden Fall entspricht dies der ersten Testhälfte.

Nutzt man diese Linearisierung in (5.1), so resultiert:

$$\mathrm{Cov}(X_i, X_j) \approx \mathrm{Cov}\left(\frac{a_i}{4}\Theta, \frac{a_j}{4}\Theta\right) = \frac{a_i a_j}{16}\,\mathrm{Var}(\Theta) \tag{5.2}$$

Diese Approximation impliziert einen monotonen Zusammenhang zwischen Diskriminationsparameter und *marginaler* Kovarianz.
Nimmt man als Arbeitshypothese an, dass sich diese marginale Beziehung auch in der relevanten bedingten Verteilung zeigt, so ergibt dieser Zusammenhang zwischen Diskriminationsparameter und bedingter Kovarianz eine plausible Erklärung für die Sensitivität der Teststatistiken T_{11} und ϕ (a_i-Variationen sind in der bedingten Kovarianz reflektiert und Abweichungen in der bedingten Kovarianz werden von T_{11} und ϕ registriert).
Für σ_r^2 gilt es hingegen zusätzlich die Richtung der Abweichung der Kovarianz zu berücksichtigen. Wegen der einseitigen Formulierung lehnt σ_r^2 bei *erhöhter* Varianz des Summenwerts der ersten Testhälfte ab. Itempaare mit erhöhter Kovarianz tragen hierzu positiv bei, Paare mit verringerter Kovarianz dagegen negativ. Gemäß obiger Heuristik sind Erstere durch Itempaare mit hohem Diskriminationswert und Letztere durch Itempaare mit niedriger Diskrimination gegeben. Eine hohe Power resultiert somit, wenn die erste Testhälfte primär aus den trennschärfsten Items zusammengesetzt ist. Besteht die erste Testhälfte hingegen aus den gering diskriminierenden Items, so resultiert ein - verglichen mit der Erwartung gemäß der Gleichverteilung auf $\boldsymbol{\Sigma_{rc}}$ - niedriger Wert für die Varianz der ersten Testhälfte und σ_r^2 lehnt nicht ab. Folglich könnte die Sensitivität von σ_r^2 darauf beruhen, dass die erste Testhälfte aus trennschärferen Items, d.h. Items mit höheren a_i-Werten, besteht. Da in den Simulationen die Diskriminationsparameter zufällig gezogen wurden, sind einige Fälle denkbar, in denen die trennschärfsten Items sich per Zufall in der ersten Testhälfte befinden. In diesem Fall käme es gemäß der Heuristik zu erhöhten Itemkovarianzen innerhalb der Items der ersten Testhälfte und demnach zu einem hohen Wert der Prüfgröße σ_r^2. Dieser Hypothese soll nun neben anderen durch die bisherigen Ergebnisse aufgeworfenen Fragen durch ergänzende Simulationen nachgegangen werden.

Ergänzende Simulationen

Dieser Abschnitt[5] befasst sich mit einigen im Kontext der vorherigen Simulation resultierenden Fragen, die im Folgenden anhand zusätzlicher Simulationen ergründet werden sollen. Er ist anhand einzelner, isolierter Fragen gegliedert und kann ohne signifikante Beeinträchtigung des Verständisses späterer Kapitel ausgelassen werden.

Simulation 1: *Wie kann die (geringe) Teststärke von σ_r^2 erklärt werden? Ist gegebenenfalls eine Verbesserung möglich?*

Zur Beantwortung dieser Fragen wurden zwei weitere Simulationen durchgeführt. Beiden Simulationen lag ein Test mit $n = 250$ Personen und $k = 10$ Items zugrunde. Im Gegensatz zur bisherigen Vorgehensweise erfolgte eine *feste* Wahl der Diskriminationsparameter[6].
Für die erste Simulation betrug der Diskriminationsparameter der ersten fünf Items eins. Die zweite Testhälfte dagegen bestand aus Items mit doppelter Diskriminationsfähigkeit ($a_i = 2$). Gemäß den Überlegungen sollte dieser Fall zu einer sehr geringen Teststärke für σ_r^2 führen. Die Items der ersten Testhälfte verfügen über die geringste Diskriminationsfähigkeit. Folglich fallen die Kovarianzen der Itempaare des ersten Testteils (relativ zu einem Rasch-Modell) niedrig aus und die Prüfgröße σ_r^2 nimmt geringe Werte an. Die Varianz des Summenscores der ersten Testhälfte fällt durch die niedrigere Trennschärfe der Items geringer aus.
In der zweiten Simulation wurden die Rollen vertauscht. Die erste Testhälfte verfügte nun über die besser diskriminierenden Items ($a_i = 2$). In dieser Situation wird eine hohe Teststärke erwartet, da in der ersten Testhälfte aufgrund der größeren a_i-Parameter erhöhte Kovarianzen auftreten. Von den $m = 100$ generierten Matrizen der ersten Simulation wurde keine Matrix von σ_r^2 abgelehnt. In der zweiten Simulation hingegen erreichte σ_r^2 eine Power von 99%. Dieser Wert lag sogar über dem entsprechenden Wert (95%) des Y-Tests.
Mit anderen Worten: Nimmt man eine Aufteilung des Tests anhand der Diskriminationsfähigkeit der Items vor, so dass in σ_r^2 die trennschärfsten Items eingehen, so

[5] Analoges gilt für die späteren Fortsetzungen dieses Abschnitts (in den Kapiteln 5.3, 5.4 und 5.5), die sich an den anderen Alternativmodellen („3PL“, „lokale Abhängigkeit“ und „Mehrdimensionalität“) orientieren.

[6] Die Ziehung der β-Parameter geschah unverändert gemäß der Normalverteilung.

resultiert ein mächtiger Test gegen variierende Itemdiskrimination. Dieser Vorteil von σ_r^2 gegenüber Y beruht allerdings auf einer Vorkenntnis der Diskriminationsfähigkeit der Items. Ist diese Vorkenntnis nicht gegeben, so liefern die Tabellenwerte ein gutes Indiz für die schwache Power des Tests. Insofern ist in Situationen ohne Vorkenntnis von dem Einsatz der Prüfgröße σ_r^2 im Hinblick auf Erkennung variierender Itemdiskrimination abzuraten.

Simulation 2: *Kann die höhere Teststärke von Y gegenüber T_{11} durch die zusätzliche Skalierung der Abweichungen erklärt werden?*

Im Folgenden wird zur Beantwortung dieser Frage die skalierte Version von T_{11} mit der Prüfgröße Y verglichen. Die neue zu betrachtende Statistik lautet:

$$T(\boldsymbol{X}) := \sum_{i,j} \frac{(r_{ij} - \rho_{ij})^2}{\sigma_{ij}^2} \tag{5.3}$$

σ_{ij}^2 stellt hierbei die mittels Satz 1 approximierte Varianz von r_{ij} bezüglich der Gleichverteilung auf $\boldsymbol{\Sigma_{rc}}$ dar. Die so resultierende Prüfgröße besitzt die skalierte Form des Y-Tests und verwendet an Stelle der Trennschärfe d_j die paarweisen Itemkorrelationen r_{ij}.
Wenn (5.3) eine ähnliche Power wie Y aufweist, dann kann die Überlegenheit von Y gegenüber T_{11} (bei 2PL-Daten) auf die zusätzlich vorgenommene Skalierung zurückgeführt werden. Unterschiede von Y und T_{11} in der Teststärke würden dann lediglich auf der Form (mit bzw. ohne Skalierung) beruhen und nicht auf der verwendeten Grundgröße d_j bzw. r_{ij}.
Zum Vergleich wurden für drei willkürlich gewählte Bedingungen aus Tabelle 5.2 die korrespondierenden Werte für die Prüfgröße (5.3) ermittelt. In der Bedingung ($n = 100, k = 30, \sigma = 0.25$) ergab sich ein Wert von 45.8%, bei ($n = 250, k = 20, \sigma = 0.25$) resultierte ein Wert von 73.2% und für ($n = 1000, k = 10, \sigma = 0.25$) lag der Wert bei 88%. Diese Ergebnisse liegen zwar beständig über den entsprechenden Werten von T_{11}, jedoch immer noch deutlich unter den Teststärken des Y-Tests. Folglich kann durch eine zusätzliche Skalierung u.U. eine geringe Verbesserung erzielt werden. Trotzdem erweist sich ein Maß, das auf der Trennschärfe basiert als effektiveres Mittel zum Testen auf variierende Itemdiskrimination.

Simulation 3: *Kann ϕ durch eine einseitige Formulierung Teststärke gewinnen?*

Dies scheint angesichts der im Kontext des σ_r^2-Tests aufgestellten Heuristik (5.2) plausibel. Da der Diskriminationsparameter im Rasch-Modell keiner Variation unterliegt, ist es denkbar, dass sich diese mangelnde Variabilität auf die Itemkorrelationen überträgt. Eine Variation der Diskriminationsparameter, wie sie im 2PL-Modell erfolgt, führt dann zu einer *gesteigerten* Varianz der Itemkorrelationen gegenüber dem Rasch-Modell. Folglich könnte ein einseitiges ϕ, welches nur bei zu hohen Werten ablehnt, eine höhere Power versus 2PL-Modell aufweisen als das zweiseitig formulierte ϕ, dass auch bei zu geringer Variabilität ablehnt.
Um dies zu überprüfen, wurde das einseitig formulierte ϕ in zwei willkürlich gewählten Bedingungen der Tabelle 5.2 betrachtet. Für $(n = 250, k = 20, \sigma = 0.25)$ ergab sich ein deutlich höherer Wert (27.8%) gegenüber der ursprünglichen ϕ-Statistik (19.4%). Die zweite Bedingung $(n = 100, k = 20, \sigma = 0.5)$ bestätigte diese Tendenz (42.8% versus 32.0%).
Durch leichte Modifikation können somit höhere Teststärken im Fall variierender Itemdiskrimination erlangt werden. Auch die veränderte ϕ-Statistik bleibt jedoch deutlich den Prüfgrößen Y, T_{11} und λ unterlegen.

Anmerkung:

Erfolgt - ähnlich wie im Falle des einseitigen ϕ-Tests - ein Vergleich einer neu definierten nonparametrischen Prüfgröße mit einer bereits existierenden Statistik (d.h. mit Y, T_{11}, ϕ oder mit σ_r^2), so werden stets die gleichen 1200 generierten Matrizen als Grundlage verwendet (siehe auch die Vorbemerkungen in Kapitel 5.1). Ferner beziehen sich alle Schätzungen der Power - solange nicht anders bemerkt - auf $m = 500$ simulierte Datenmatrizen.

5.3 Variable Itemdiskrimination mit Ratetendenzen

Sämtliche Bemerkungen zum Simulationsdesign des 2PL-Modells lassen sich auf das 3PL-Szenario übertragen. Zusätzlich zu den Spezifikationen des vorherigen Abschnitts ist lediglich die Angabe der Rateparameter erforderlich. Diese wurden für jedes Item unabhängig gemäß einer Betaverteilung, $c_i \overset{i.i.d}{\sim} \text{Be}(5, 17)$, gezogen. Die Ziehungen erfolgten für jede der 500 Datenmatrizen jeweils von Neuem.

Ergebnisse. Die Resultate gleichen größtenteils denen des vorherigen Abschnitts. Sowohl die Monotonieaussagen bezüglich der Testlänge und der Personenanzahl als auch die Aufteilung der Teststatistiken bezüglich ihrer Sensitivität lassen sich übertragen. Y stellt auch hier die Größe mit der höchsten Power dar. Im Gegensatz zum 2PL-Modell liegt die Power des T_{11}-Tests jedoch nicht mehr konstant über der des parametrischen Tests: In manchen Bedingungen weist λ eine ähnliche Teststärke auf. Ein Vergleich mit Tabelle 5.2 zeigt ferner, dass die Teststärken der Prüfgrößen bei 3PL-Daten mit Ausnahme der schwachen Bedingung $\sigma = 0.12$ geringer gegenüber den korrespondierenden Teststärken im 2PL-Modell ausfallen.

Bemerkung. Die veringerten Teststärken relativ zu einem 2PL-Modell könnten *informell* mit der in Kapitel 4.3 gegebenen Näherung (4.6) erläutert werden. Approximiert man wie in (4.6) die Diskrimination eines 3PL-Items mit $\alpha_i = a_i(1 - c_i)$ und vergleicht $\sigma_\alpha^2 := \text{Var}(\alpha_i)$ mit $\sigma_a^2 := \text{Var}(a_i)$, so ergibt sich:

$$\begin{aligned} \sigma_\alpha^2 < \sigma_a^2 \qquad &\text{für} \quad \sigma = 0.25, 0.5 \\ \sigma_\alpha^2 > \sigma_a^2 \qquad &\text{für} \quad \sigma = 0.12 \end{aligned}$$

Somit gilt genau für die beiden Bedingungen ($\sigma = 0.25, 0.5$) mit verringerter Power des 3PL-Modells eine verringerte Streuung der Diskrimination.
Alternativ lässt sich das 3PL-Modell auch als Mischverteilungsmodell betrachten. Mit Wahrscheinlichkeit $(1 - c_i)$ stammt der Response aus einem 2PL-Modell und mit Wahrscheinlichkeit c_i aus einer degenerierten (Rasch-konformen) Verteilung. Je höher der Parameter c_i, umso stärker geht die konforme Komponente in das Modell ein. Dies bietet einen zweiten möglichen Zugang zur Erklärung verringerter Teststärken, der jedoch nicht den Ausnahmefall $\sigma = 0.12$ erklären kann.

Tabelle 5.3: *Anteil als signifikant erklärter Datenmatrizen (in %) von $m = 500$ nach dem 3PL-Modell simulierten Datenmatrizen in Abhängigkeit von σ, n und k.*

σ	n	k	Y	T_{11}	ϕ	σ_r^2	λ
0.12	100	10	9.8	10.2	6.0	6.2	8.2
		20	25.8	16.2	5.8	6.6	14.4
		30	32.4	26.8	7.0	8.2	18.6
	250	10	26.0	19.8	8.4	7.4	17.4
		20	60.2	44.0	8.8	9.4	42.4
		30	77.0	60.4	9.0	10.8	54.0
	500	10	49.6	34.6	12.2	11.8	34.6
		20	87.6	75.0	12.4	13.4	72.4
		30	98.0	91.4	15.2	14.2	91.2
	1000	10	76.2	61.8	16.0	14.6	60.4
		20	99.4	95.8	22.6	16.6	96.0
		30	100.0	99.8	28.8	19.4	100.0
0.25	100	10	15.8	10.0	7.0	7.0	13.0
		20	37.0	23.8	7.0	5.6	24.0
		30	54.0	39.4	7.6	8.8	35.6
	250	10	40.6	28.0	10.8	13.0	26.6
		20	79.8	61.2	13.6	14.2	64.2
		30	96.0	84.2	17.0	14.8	80.6
	500	10	68.2	54.2	20.0	13.8	50.8
		20	98.0	91.8	24.6	14.2	88.8
		30	99.6	98.8	31.4	17.8	98.8
	1000	10	90.8	80.6	33.8	18.2	81.8
		20	99.8	100.0	50.8	20.6	99.2
		30	100.0	100.0	62.6	25.4	100.0
0.5	100	10	36.0	26.0	8.8	7.6	23.6
		20	68.4	52.4	19.6	10.6	46.8
		30	87.8	73.6	19.6	11.6	68.6
	250	10	69.4	53.6	27.0	14.8	51.0
		20	97.6	91.8	43.0	17.2	90.6
		30	99.8	99.0	52.0	19.8	98.4
	500	10	91.4	83.6	46.6	15.6	80.4
		20	100.0	99.6	73.2	22.6	99.6
		30	100.0	100.0	87.6	25.6	100.0
	1000	10	98.2	95.0	71.2	20.4	94.0
		20	100.0	100.0	94.4	27.2	100.0
		30	100.0	100.0	98.4	29.0	100.0

Ergänzende Simulationen

Simulation 4: *Wie ändern sich die Resultate bei konstanter Diskrimination?*

Mit dieser Frage soll die Power der Tests gegen ein Alternativmodell, in dem alle Items über das gleiche Diskriminationspotential ($a_i = 1$) verfügen und sich lediglich im Wert der unteren Asymptote von einem Rasch-Modell unterscheiden, betrachtet werden. Der isolierte Effekt des Rateparameters bildet somit den Gegenstand der Untersuchung.

Für die vier möglichen Kombinationen von zwei Testlängen ($k = 10, k = 20$) mit zwei Stichprobengrößen ($n = 250, n = 500$) wurden $m = 500$-Datenmatrizen gemäß eines 3PL-Modells mit konstanter Diskrimination ($a_i = 1$) simuliert. Die Ziehung der Rateparameter erfolgte wiederum gemäß der angegebenen Betaverteilung.

Die Ergebnisse sind in Tabelle 5.4. dargestellt. Das Verhalten der Tests gleicht - wie zu erwarten war - dem 3PL-Szenario mit geringer Streuung ($\sigma = 0.12$). Die Teststärken fallen lediglich noch geringer aus.

λ weist nun eine nahezu identische Power wie der T_{11}-Test auf. Beide Teststärken liegen unterhalb der korrespondierenden Werte des Y-Tests. Somit erweist sich Y auch bei Vorlage eines reinen (variierenden) Rateeffekts als die beste Prüfgröße. Allerdings erreicht diese Teststatistik bei geringem k selbst für 500 Personen noch keine hohe Teststärke.

Die Prüfgrößen ϕ und σ_r^2 reagieren kaum auf das Vorliegen eines Rateeffekts. Dies ähnelt dem bereits in Tabelle 5.3 festgestellten Verhalten in der Bedingung $\sigma = 0.12$.

Tabelle 5.4: *Geschätzte Teststärken für das 3PL-Modell mit konstanter Diskrimination der Items ($a_i = 1$). Die Rateparameter entstammen einer Be$(5, 17)$-Verteilung.*

n	k	Y	T_{11}	ϕ	σ_r^2	λ
250	10	18.8	10.8	4.0	6.0	13.8
250	20	51.2	34.8	6.2	8.8	33.8
500	10	37.0	26.0	7.4	10.2	26.0
500	20	83.6	60.6	11.8	12.0	60.6

5.4 Lokale stochastische Abhängigkeit

Die Simulation lokal abhängiger Daten erfolgte gemäß Modell (4.9) aus Kapitel 4.4. Für die Dimensionen der Datenmatrix sowie die Ziehung der Schwierigkeits- und Fähigkeitsparameter gelten die in den Vorbemerkungen erwähnten Ausführungen. Zur vollen Spezifikation bedarf es somit lediglich der Angabe von $\delta_{(l-1)l}$. Von den $k-1$ vorhandenen δ-Parametern waren fünf mit Werten ungleich Null besetzt. Die Parameter der fünf betroffenen Itempaare, deren Positionen innerhalb des Tests in Tabelle 5.5 dargestellt sind, nahmen den identischen Wert δ an. Dieser variierte - analog zum Parameter σ in der 2PL-Bedingung - auf drei Stufen (Abweichungsgrade).

Tabelle 5.5: *Indizes l, für die $\delta_{(l-1)l} \neq 0$ gilt (in Abhängigkeit der Testlänge k).*

Testlänge $k =$	Indizes l: $\delta_{(l-1)l} \neq 0$
10	2, 4, 6, 8, 10
20	2, 6, 10, 14, 18
30	2, 8, 14, 20, 26

Ergebnisse. Alle Teststatistiken weisen zunehmende Power bei einem Anstieg des Parameters δ auf (Tabelle 5.6). Somit sind, wie auch bei den Modellverletzungen der vorherigen Abschnitte, zunächst alle Prüfgrößen - in einer absoluten Betrachtungsweise - sensitiv[7] gegenüber lokalen Abhängigkeiten.

Ein Vergleich der Prüfgrößen legt ferner eine Gruppierung in zwei Klassen nahe. Die erste Klasse besteht aus den - relativ gesehen - stärker sensitiven Tests ϕ und T_{11}, welche beide auf Assoziationen zwischen Itempaaren basieren. In der zweiten Klasse befinden sich die auf variable Itemdiskrimination ansprechenden Tests Y und λ sowie die Prüfgröße σ_r^2. Für diese Gruppe zeigt sich generell eine geringere Sensitivität. Außer in den Bedingungen mit maximalem Stichprobenumfang und maximalem δ-Parameter liegen hier alle Teststärken konstant unter 20%.

In beiden Klassen wächst die Power mit steigendem Stichprobenumfang. Bezüglich der Testlänge[8] k zeigt sich hingegen ein unterschiedliches Monotonieverhalten.

[7] „Sensitiv“ meint hier lediglich einen Anstieg an Power bei zunehmender Modellverletzung.

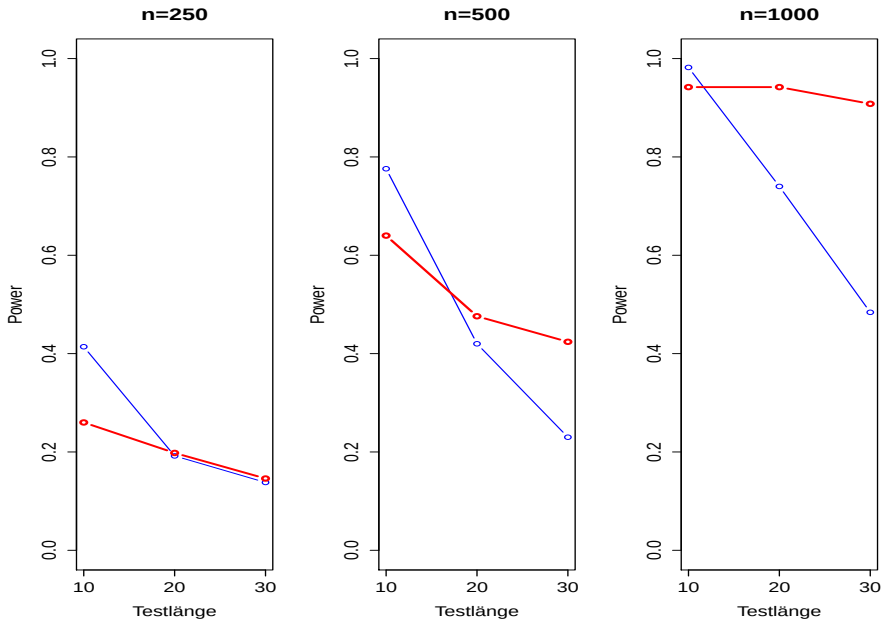
[8] Die folgenden Betrachtungen zum Einfluss der Testlänge beziehen sich auf Bedingungen, in denen die Teststärke deutlich über 5% liegt. Unter anderen Bedingungen ist der Einfluss schwerer auszumachen.

Tabelle 5.6: *Anteil als signifikant erklärter Datenmatrizen (in %) von $m = 500$ nach dem Modell (4.9) simulierten Datenmatrizen in Abhängigkeit von δ, n und k.*

δ	n	k	Y	T_{11}	ϕ	σ_r^2	λ
0.25	100	10	6.2	5.8	4.8	5.4	5.4
		20	6.6	6.2	4.2	5.2	5.2
		30	6.0	6.4	5.0	5.8	5.6
	250	10	6.8	6.8	5.8	7.0	5.2
		20	6.4	5.8	4.2	8.2	6.0
		30	5.0	6.6	5.0	6.6	6.0
	500	10	5.0	9.2	7.0	8.0	4.2
		20	5.6	6.8	3.8	7.4	6.6
		30	6.4	6.4	6.4	6.2	5.8
	1000	10	6.0	17.0	11.0	7.8	4.4
		20	7.4	9.2	7.8	6.6	8.2
		30	8.6	11.4	5.0	9.2	6.4
0.5	100	10	5.2	8.6	5.2	5.4	5.8
		20	7.6	6.8	7.0	7.8	6.2
		30	7.0	5.8	5.0	6.2	6.0
	250	10	7.8	20.4	10.2	11.2	6.4
		20	7.2	10.6	7.4	10.2	6.8
		30	7.6	9.6	4.6	9.2	6.4
	500	10	7.8	37.8	22.2	11.8	7.4
		20	9.8	16.6	12.4	13.6	7.8
		30	12.6	17.4	14.0	11.4	7.8
	1000	10	12.6	74.2	52.6	13.6	7.8
		20	18.6	36.6	45.8	18.8	12.6
		30	17.4	26.2	39.8	11.4	13.2
0.75	100	10	7.0	18.0	8.0	8.0	5.6
		20	7.8	11.4	9.4	10.6	6.0
		30	8.0	8.2	5.0	9.6	6.8
	250	10	8.4	41.4	26.0	9.2	7.4
		20	11.4	19.2	19.8	13.6	7.8
		30	9.0	13.8	14.6	12.0	9.4
	500	10	9.2	77.6	64.0	14.4	8.2
		20	20.0	42.0	47.6	19.0	13.6
		30	17.4	23.0	42.4	17.8	11.8
	1000	10	19.6	98.2	94.2	21.2	13.8
		20	42.2	74.0	94.2	28.4	25.8
		30	32.4	48.4	90.8	17.0	22.2

Für die erste Gruppe fällt die Teststärke mit steigender Itemzahl. Dies mag aufgrund der Tatsache, dass der prozentuale Anteil an zusätzlichen Parametern (5/k) gegenüber dem Rasch-Modell mit steigendem k sinkt, plausibel erscheinen. Mit zunehmendem k verringert sich der Anteil lokal abhängiger Itempaare bezogen auf die Testlänge.
Die Tests Y und λ der zweiten Gruppe zeigen dagegen ein überraschendes Verhalten. Unter den Bedingungen $k = 20, k = 30$ ergeben sich höhere Teststärken als unter der Bedingung mit geringem k. Ferner weist in einigen Fällen der 20 Item-Test eine höhere Power auf als der 30 Item-Test. Ein ähnliches nicht monotones Verhalten der sogenannten „first-order"-Statistiken in Abhängigkeit von der Testlänge (bei lokal abhängigen Daten) findet sich auch in der Simulationsstudie von Suárez-Falcón und Glas (2003). Eine Erklärung dieses Sachverhalts steht jedoch noch aus.
Abschließend lässt sich noch bezüglich der ersten Gruppe ein Unterschied in der Abnahme der Power mit steigender Itemzahl konstatieren: ϕ verliert für wachsendes k wesentlich langsamer an Teststärke als die T_{11}-Statistik. Dies könnte darin begründet sein, dass in T_{11} eine Reihe „wenig relevanter" Itempaare eingehen, während ϕ stets nur auf den zwei extremsten Itempaaren beruht. Da T_{11} über alle Itempaare verläuft, gehen in T_{11} auch konforme Itempaare, d.h. unkorrelierte (gegeben θ) Itempaare mit ein. Diese konformen Itempaare überwiegen mit steigender Testlänge, da die Anzahl diskrepanter Itempaare konstant bleibt. In ϕ hingegen geht ein Itempaar nur dann ein, wenn es ein Extremum bezüglich der Korrelation repräsentiert. Zwei weitere Rasch-konforme Items tragen stets zu einem veränderten Wert von T_{11} bei. Bezüglich ϕ dürften diese jedoch, gerade bei einem bereits vorhandenem Itempaar mit extremer Korrelation, keine Auswirkungen haben.
Liegt folglich ein geringer Anteil lokal abhängiger Itempaare vor, so ist ϕ gegenüber T_{11} vorzuziehen. Dies gilt allerdings nur für Bedingungen, unter denen die Teststärke nicht zu gering ausfällt (z.B. $\delta = 0.75$).
Für $\delta = 0.75$ und $n > 100$ sind die entsprechenden Werte aus Tabelle 5.6 grafisch in Abbildung 5.1 dargestellt. Unter jeder Bedingung des Stichprobenumfangs fällt die Power des T_{11}-Tests (blaue Linie) deutlich erkennbar mit wachsender Testlänge. Der - relativ betrachtet - wesentlich geringere Abfall der Teststärke von ϕ (rote Linie) führt dazu, dass ϕ bei kleinem Anteil verstoßender Itempaare (d.h. bei hoher Testlänge) überlegen ist.

Abbildung 5.1: *Teststärken von ϕ (rot) und T_{11} (blau) in der Bedingung $\delta = 0.75$.*



Bemerkung. Es ist zu beachten, dass die Teststärken der Tabelle 5.6 nicht unmittelbar mit den Teststärken der 2PL- bzw. 3PL-Modelle vergleichbar sind, da sie auf stark unterschiedlichen Modellen beruhen[9]. Die Schlussfolgerung einer - aufgrund der tendenziell niedrigen Werte in Tabelle 5.6 - geringeren Sensitivität der Prüfgrößen gegenüber lokal abhängigen Daten verglichen mit dem Szenario variabler Itemdiskrimination ist unzulässig. Eine Modellverletzung, realisiert anhand des Parameters δ, lässt sich nicht unmittelbar mit einer durch den Parameter σ induzierten Modellverletzung vergleichen.

Ergänzende Simulationen

Simulation 5: *Sind λ und Y gegen lokal abhängige Daten sensitiv?*

Auch wenn diese Frage anhand der Daten der Tabelle 5.6 auf den ersten Blick positiv beantwortet werden könnte, so ergeben sich doch bei näherem Hinsehen Probleme. Anders als für die vorherigen Alternativmodelle (2PL/3PL-Modell), in denen der Verlauf der Item-Response-Funktionen (Faktor 1) unter Beibehaltung der lokalen

[9]Diese Problematik wird am Ende von Kapitel 5.6 indirekt wieder aufgegriffen.

Unabhängigkeit (Faktor 2) variiert wurde, erfolgte in dem Alternativmodell dieses Abschnitts keine systematische Variation *eines* Faktors bei Konstanthaltung des anderen Faktors, d.h. es wurde zwar die lokale Unabhängigkeit der Items verletzt, jedoch *nicht* unter Beibehaltung des Verlaufs der Item-Response-Funktion des Rasch-Modells.
Dies sei an einem Beispiel der Testlänge $k = 10$ demonstriert. Während für das erste Item per Definition (siehe Abschnitt 4.4) die Item-Response-Funktion dem Rasch-Modell folgt, gilt für das zweite Item ($\delta_{12} \neq 0$ nach Tabelle 5.5) gemäß dem iterierten Erwartungswert:

$$f_2(\theta) = E(X_2|\,\theta) = E(E(X_2|\,X_1, \theta)|\,\theta) \tag{5.4}$$

Verwendet man die Abkürzungen

$$f_{12}(0, \theta) := \frac{\exp(\theta - \beta_2 - 0.5\delta_{12})}{1 + \exp(\theta - \beta_2 - 0.5\delta_{12})}$$

sowie

$$f_{12}(1, \theta) := \frac{\exp(\theta - \beta_2 + 0.5\delta_{12})}{1 + \exp(\theta - \beta_2 + 0.5\delta_{12})},$$

so wird (5.4) - nach kurzer Umformung - zu:

$$f_2(\theta) = f_{12}(0, \theta) + f_1(\theta)(f_{12}(1, \theta) - f_{12}(0, \theta)) \tag{5.5}$$

Dies besitzt nicht die Gestalt einer logistischen Verteilungsfunktion und repräsentiert somit eine vom Rasch-Modell abweichende Item-Response-Funktion.
Um zu klären, ob die Teststärke von Y und λ allein auf den diskrepanten Verlauf von (5.5) zurückführbar ist, scheint eine weitere Simulation angebracht, in der lediglich der zweite Faktor - bei Konstanthaltung des ersten Faktors - variiert wird. Dies erfordert ein Alternativmodell, in dem $f_i(\theta)$ einem Rasch-Modell folgt, die Items jedoch lokal abhängig sind. Eine einfache Methode hierfür stellt das „Kopieren“ gewisser Spalten einer Rasch-homogenen Datenmatrix dar. Wie anhand obiger Gleichungen leicht ersichtlich, erzielt man denselben Effekt für den Grenzübergang $\delta_{12} \to \infty$. In diesem Fall ist die zweite Spalte lediglich eine Kopie der ersten Spalte und die Item-Response-Funktion (5.5) nimmt die logistische Gestalt an:

$$f_2(\theta) = f_1(\theta)$$

Folglich sollte, falls die beiden Prüfgrößen tatsächlich nur auf einen diskrepanten Verlauf von $f_i(\theta)$ reagieren, bei sehr groß gewähltem δ die Teststärke sehr gering ausfal-

len. Eine entsprechende Simulation[10] (Tabelle 5.7) für $\delta = 100$ ergab jedoch eine hohe Power der beiden Tests. Folglich sind die beiden Tests - wenn auch in einem deutlich geringeren Umfang als die anderen Prüfgrößen - sensitiv gegenüber lokal abhängigen Daten. Man beachte, dass die Wahl $\delta = 100$ den Extremfall („kopierte“ Spalten) lokaler Abhängigkeit zweier Items darstellt. Insofern fallen die Teststärken von Y und insbesondere von λ gemessen an dem Verletzungsgrad gering aus. Gleichzeitig stellt Tabelle 5.7 noch einmal den bereits erwähnten, nicht monotonen Zusammenhang mit der Testlänge k heraus.

Tabelle 5.7: *Prüfgrößen-spezifischer Anteil (in %) als signifikant erklärter Datenmatrizen von* $m = 500$ *gemäß dem Modell (4.9) simulierten Datenmatrizen.*

δ	n	k	Y	T_{11}	ϕ	σ_r^2	λ
100	500	10	14.4	100	100	99.6	12.8
		20	98.0	100	100	98.2	69.8
		30	70.6	100	100	71.0	44.0

Simulation 6: *Sind durch geeignete Modifikationen höhere Teststärken erzielbar?*

Eine Verbesserung der Teststärken lässt sich durch entsprechendes Vorwissen über relevante Itempaare erzielen: Modifiziert man etwa die Größe σ_r^2, so dass an Stelle der ersten Testhälfte die Summation über alle Items erfolgt, die mit einem positiven δ-Parameter assoziiert sind[11], ergibt sich in der $(\delta = 0.5, n = 500, k = 20)$-Bedingung eine Teststärke von 34.4%. Dies stellt eine Verdoppelung an Power gegenüber dem T_{11}-Test dar.
Wenn $S \subset \{1 \dots k\}$ eine Menge an Items bezeichnet, dann ist die entsprechende Modifikation von σ_r^2 formal gegeben durch:

$$T(\boldsymbol{X}) = \mathrm{Var}(\boldsymbol{r}_S), \quad \boldsymbol{r}_S := \boldsymbol{X}\boldsymbol{1}_S$$

[10] Es gelten weiterhin die Bedingungen, unter denen die Werte aus Tabelle 5.6 simuliert wurden. An Stelle der vorherigen Werte (0.25, 0.5 und 0.75) nimmt δ nun einen extrem hohen Wert (100) an. Die „verletzten“ Itempaare sind unverändert aus Tabelle 5.5 zu entnehmen.

[11] Für k=20 - z.B. - betrifft dies die Items 1, 2, 5, 6, 9, 10, 13, 14, 17 und 18. Für k=10 ist diese Prozedur nicht anwendbar.

$\mathbf{1}_S$ bezeichnet hierbei einen Vektor, für dessen i-tes Element gilt:

$$(\mathbf{1}_S)_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases}$$

In der ursprünglichen Form entspricht die Menge S der ersten Testhälfte. Die „verbesserte“ Version hingegen beinhaltet in S alle Items, die in eine lokale Abhängigkeit involviert sind. Da in $\mathrm{Var}(\boldsymbol{r}_S)$ alle paarweisen Kovarianzen zweier Items aus der Menge S eingehen und da diese Menge genau aus den abweichenden Itempaaren (jene mit positivem δ) besteht, dürfte die resultierende Zunahme der Power intuitiv nahe liegen.

Eine weitere Steigerung der Teststärke ist durch eine analoge Modifikation der T_{11}-Größe möglich. Hier erfolgt die Summation in (4.10) lediglich über alle Itempaare mit positivem δ-Parameter. Aus der Beschränkung auf Itempaare, die von der Modellabweichung betroffen sind, ergibt sich eine Teststärke von 92% unter der ($\delta = 0.5, n = 500, k = 20$)-Bedingung.

Anhand dieser Modifikationen wird die Rolle des Vorwissens akzentuiert. Auch wenn keine der in Tabelle 5.6 aufgeführten Prüfgrößen eine hohe Teststärke aufweist, so ist es dennoch möglich durch Vorwissen über relevante Itempaare eine Statistik mit ausreichender Power zu konstruieren. Ähnliches deutete sich bereits in Abschnitt 5.2 (Simulation 1) an, wo eine entsprechende, an den Trennschärfen der Items orientierte Anpassung der Prüfgröße σ_r^2 eine hohe Teststärke ermöglichte.

5.5 Mehrdimensionalität

Das Modell (4.11) des Kapitels 4.5 diente zur Simulation zweidimensionaler Daten. Die Ziehung der Personenparameter[12] erfolgte - für jede der $m = 500$ Matrizen von Neuem - gemäß einer bivariaten Normalverteilung mit Einheitsvarianzen. Die Korrelation der beiden Dimensionen variierte, analog zu den Parametern δ und σ der vorherigen Alternativmodelle, in drei Stufen bzw. Abweichungsgraden ($\rho = 0.75, 0.5, 0.25$). Alle Diskriminationsvektoren wurden des Weiteren so gewählt, dass jedes Item auf genau einer Dimension lud und dessen zu jener Dimension korrespondierende Diskriminationswert eins betrug, d.h. jedes $\boldsymbol{a_i}$ entsprach einem Einheitsvektor.
In einem ersten Szenario erfolgte die Zuteilung der Items zu den Dimensionen symmetrisch (1:1), d.h. die erste Hälfte der Items lud auf der ersten Dimension und die zweite Testhälfte auf der zweiten Dimension. Für das zweite Szenario kam eine asymmetrische Aufteilung (4:1) zur Anwendung. Hier waren bis auf das letzte Fünftel alle Items der ersten Dimension zugeordnet.

Ergebnisse. Bezüglich der *symmetrischen* Aufteilung (Tabelle 5.8), welche zuerst dargestellt werden soll, ergeben sich zwei nicht bzw. äußerst schwach sensitive Statistiken (Y und λ). Die Power dieser Tests liegt stets unter 12%.
Die drei sensitiven Teststatistiken lassen sich - nahezu unter jeder Bedingung - gleichmäßig anordnen ($\phi << T_{11} < \sigma_r^2$). Da σ_r^2 quasi spezifisch für diesen Fall (erste Testhälfte misst anderes Konstrukt) konstruiert wurde, ist dieses Ergebnis plausibel. Sowohl σ_r^2 als auch T_{11} ermöglichen für $\rho = 0.25$ und für $\rho = 0.5$ bereits ab einem relativ geringen Stichprobenumfang hohe Teststärken. Für niedrige und moderate Korrelationen reichen in den meisten Fällen $n = 100$ bzw. $n = 250$ Personen aus, um Teststärken über 50% zu erlangen. Die Bedingung hoher Korrelation setzt sich dagegen deutlich ab. Hier garantiert lediglich der maximale Stichprobenumfang (oder ein hoher Stichprobenumfang in Kombination mit maximaler Testlänge) Teststärken über 50%.
Betreffend der Monotonie ergeben sich die „üblichen“ Resultate. Alle drei sensitiven Tests weisen steigende Teststärke bezüglich eines Anstiegs des Stichprobenumfangs, der Testlänge oder des Grades der Modellverletzung (d.h. abnehmender Korrelation ρ) auf.

[12] Für den Schwierigkeitsparameter gelten die Vorbemerkungen.

Tabelle 5.8: *Anteil als signifikant erklärter Datenmatrizen (in %) von $m = 500$ nach dem Modell (4.11) simulierten Datenmatrizen (1:1-Aufteilung) in Abhängigkeit von ρ, n und k.*

ρ	n	k	Y	T_{11}	ϕ	σ_r^2	λ
0.75	100	10	6.4	8.0	6.8	14.8	7.2
		20	5.2	10.4	5.8	34.0	5.6
		30	5.6	15.0	4.0	44.8	5.6
	250	10	5.0	9.2	4.4	29.6	5.2
		20	5.0	23.2	8.2	59.6	5.4
		30	5.2	31.8	6.2	73.6	4.0
	500	10	6.2	24.4	7.0	47.6	6.0
		20	5.2	46.2	8.6	78.0	6.0
		30	6.0	71.0	8.8	91.4	6.8
	1000	10	3.8	50.0	14.4	72.0	3.6
		20	6.2	87.0	16.4	96.6	6.0
		30	6.6	99.4	16.0	99.0	4.2
0.5	100	10	6.2	23.0	6.2	42.2	4.6
		20	6.8	34.0	7.0	63.6	7.0
		30	4.6	54.2	9.4	84.6	5.2
	250	10	5.0	47.4	15.4	68.0	5.6
		20	5.2	83.2	18.0	94.8	6.6
		30	6.4	97.6	19.2	98.8	7.4
	500	10	4.6	89.4	31.6	92.4	4.8
		20	8.4	99.8	42.8	99.8	6.4
		30	8.0	100.0	46.6	100.0	6.6
	1000	10	5.8	99.6	63.4	99.2	7.6
		20	6.6	100.0	76.4	100.0	5.6
		30	6.0	100.0	84.2	100.0	7.2
0.25	100	10	3.8	38.2	13.8	60.0	3.6
		20	6.2	77.6	17.0	90.2	7.6
		30	6.6	92.8	21.2	97.8	4.8
	250	10	6.8	89.6	35.2	89.8	6.8
		20	9.4	99.4	49.0	99.6	10.0
		30	7.4	100.0	57.8	100.0	7.6
	500	10	7.6	100.0	75.6	99.6	6.6
		20	6.0	100.0	87.6	100.0	7.6
		30	6.0	100.0	93.2	100.0	6.6
	1000	10	7.6	100.0	97.8	100.0	11.0
		20	6.8	100.0	99.6	100.0	8.0
		30	10.6	100.0	100.0	100.0	8.0

Tabelle 5.9: *Anteil als signifikant erklärter Datenmatrizen (in %) von $m = 500$ nach dem Modell (4.11) simulierten Datenmatrizen (4:1-Aufteilung) in Abhängigkeit von ρ, n und k.*

ρ	n	k	Y	T_{11}	ϕ	σ_r^2	λ
0.75	100	10	7.6	9.0	6.4	9.2	7.2
		20	10.0	10.0	4.2	20.2	7.0
		30	11.6	12.6	5.4	25.2	8.4
	250	10	14.2	10.4	6.8	16.4	9.6
		20	21.2	21.8	6.0	32.0	14.6
		30	23.6	26.4	5.0	46.0	17.2
	500	10	18.4	19.2	9.6	24.2	12.4
		20	34.4	43.8	6.4	45.0	20.6
		30	44.6	58.4	8.8	66.4	25.6
	1000	10	45.0	46.2	10.8	39.2	28.0
		20	67.4	78.8	13.2	75.2	44.6
		30	81.8	94.4	15.0	90.6	55.0
0.5	100	10	15.6	14.8	7.0	17.8	11.6
		20	24.8	27.2	7.6	41.4	17.6
		30	36.6	46.2	8.0	56.6	23.6
	250	10	38.4	39.0	14.2	35.0	23.0
		20	65.0	80.2	16.6	74.2	45.2
		30	74.8	90.2	16.8	89.4	54.8
	500	10	74.2	78.4	30.4	63.6	57.2
		20	93.4	98.4	33.8	94.4	74.4
		30	97.6	99.8	38.0	99.0	87.8
	1000	10	96.0	97.4	56.2	87.2	83.2
		20	100.0	100.0	68.2	99.8	98.0
		30	100.0	100.0	72.2	100.0	100.0
0.25	100	10	38.6	37.4	14.2	32.0	25.6
		20	55.4	65.8	16.0	65.6	37.0
		30	68.8	84.6	15.6	83.2	51.8
	250	10	76.2	82.0	34.0	66.8	55.6
		20	95.8	98.6	46.4	96.0	84.0
		30	98.8	100	49.2	99.4	92.4
	500	10	98.0	99.4	67.6	89.2	88.0
		20	100.0	100.0	81.2	100.0	97.2
		30	100.0	100.0	90.0	100.0	100.0
	1000	10	100.0	100.0	95.6	98.8	99.8
		20	100.0	100.0	98.0	100.0	100.0
		30	100.0	100.0	99.4	100.0	100.0

Das gleiche Monotonieverhalten zeigt sich für die *asymmetrische* Aufteilung (Tabelle 5.9). Die Teststärken der drei sensitiven Tests fallen hier gegenüber der 1:1-Aufteilung geringer aus. Dies erscheint aufgrund der geringeren Verletzung (der Test besteht quasi - bis auf wenige abweichende Items - aus einer Dimension) plausibel. Weiterhin kann der stärkere Verlust an Power für σ_r^2 auf die nicht mehr optimale Testaufteilung zurückgeführt werden. Die erste Hälfte besteht zwar vollständig aus den Items einer Dimension, jedoch ist auch ein gewisser Anteil an Items der zweiten Testhälfte mit der ersten Dimension assoziiert. Der daraus resultierende Verlust an Power führt zu einer Überlegenheit des T_{11}-Tests in der asymmetrischen Aufteilung. Für $\rho \neq 0.75$ und $n > 100$ liegen die Teststärken des T_{11}-Tests konsistent über den entsprechenden Werten von σ_r^2.
Die größten Änderungen gegenüber der symmetrischen Situation ergeben sich für die „first-order"-Statistiken Y und λ, deren Teststärken nun wesentlich deutlicher mit dem Grad der Modellverletzung, dem Stichprobenumfang und der Testlänge zunehmen. Beide Prüfgrößen weisen durchgehend höhere Teststärken auf als der ϕ-Test. Diese Zunahme lässt sich möglicherweise informell über den Zusammenhang zwischen Item- und Summenscore verdeutlichen. Im asymmetrischen Fall misst der Summenwert primär die erste Dimension. Folglich besitzen die Items der ersten Dimension eine höhere Trennschärfe als die Items der zweiten Dimension. Diese Variation der Diskriminationsfähigkeit schlägt sich in den Statistiken Y und λ nieder. λ weist hierbei - übereinstimmend mit den Resultaten bezüglich variierender Diskrimination - konstant eine geringere Teststärke als Y auf. Die Teststärke von Y erreicht bei Bedingungen geringer Testlänge ($k = 10$) in den meisten Fällen sogar das Niveau der T_{11}-Prüfgröße.

Bemerkung. Dem mehrdimensionalen Modell dieses Abschnitts lässt sich, wie bereits in Kapitel 4.5 skizziert, ein empirisch ununterscheidbares eindimensionales, lokal abhängiges Modell zuordnen. Wählt man o.B.d.A. θ_1 als latente Variable, so resultiert ein eindimensionales Modell mit bedingten Abhängigkeiten zwischen gewissen Itempaaren. Für ein Itempaar (X_i, X_j) gilt dann:

$$\mathrm{Cov}(X_i, X_j | \, \theta_1) = \mathrm{Cov}(E(X_i | \, \boldsymbol{\Theta}), E(X_j | \, \boldsymbol{\Theta}) | \, \theta_1) = \mathrm{Cov}(f_i(\boldsymbol{\Theta}), f_j(\boldsymbol{\Theta}) | \, \theta_1) \qquad (5.6)$$

Die rechte Seite stellt eine Kovarianz zwischen zwei Funktionen einer skalaren Zufallsvariable (θ_2 variiert, θ_1 ist fest) dar. Da f_i und f_j monoton wachsend in θ_2 sind,

folgt gemäß Lemma 2:

$$\mathrm{Cov}(X_i, X_j | \, \theta_1) \geq 0$$

Es ist jedoch noch eine weitere Differenzierung für die konkret vorliegende Situation möglich. Für zwei Items, die auf der ersten Dimension laden[13], sind die Item-Response-Funktionen nicht von θ_2 abhängig. Folglich gilt $\mathrm{Cov}(X_i, X_j | \, \theta_1) = 0$. Da eine Zufallsvariable mit einer konstanten Zufallsvariable keine Kovarianz aufweist, besteht dieselbe Gleichung für ein Itempaar, in dem genau ein Item der ersten Dimension enthalten ist. Nur für zwei Items, die beide zur zweiten Dimension korrespondieren, resultiert eine positive Kovarianz.
Insgesamt ergibt sich somit ein lokal abhängiges, eindimensionales Modell mit positiven bedingten Kovarianzen, die sich in dem Teil des Tests, der die zweite Dimension misst, manifestieren. Es ist daher nicht verwunderlich, dass Tests für lokale Abhängigkeiten (z.B. T_{11} und ϕ) auch bezüglich des mehrdimensionalen Modells sensitiv reagieren.
Eine analoge Vorgehensweise lässt sich für die Item-Response-Funktionen des zugehörigen eindimensionalen Modells durchführen. Es gilt:

$$f_i(\theta_1) = P(X_i = 1 | \, \theta_1) = E(E(X_i | \, \boldsymbol{\Theta}) | \, \theta_1) = E(f_i(\boldsymbol{\Theta}) | \, \theta_1) \tag{5.7}$$

Für ein Item der ersten Dimension ergibt sich, da $f_i(\boldsymbol{\theta})$ unabhängig von θ_2 ist:

$$f_i(\theta_1) = f_i(\boldsymbol{\Theta})$$

Dies folgt der „üblichen“ logistischen Form.
Für ein Item der *zweiten* Dimension gilt gemäß (5.7) hingegen:

$$f_i(\theta_1) = \int \frac{\exp(a_{i2}\theta_2 - \beta_i)}{1 + \exp(a_{i2}\theta_2 - \beta_i)} \mathrm{d}\Phi_{\theta_1}(\theta_2) \tag{5.8}$$

$\Phi_{\theta_1}(\theta_2)$ bezeichnet hierbei die bedingte Verteilungsfunktion von θ_2 gegeben θ_1. Folglich ergibt sich i.A. keine logistische Form der Item-Response-Funktion. Für den Fall der hier vorliegenden Normalverteilung lässt sich (5.8) jedoch gut durch eine logistische Funktion approximieren. Es resultiert dann aus (5.8) für die gegebene Spezifikation (siehe Ip (2009)):

$$f_i(\theta_1) \approx \frac{\exp(a_i^*\theta_1 - \beta_i^*)}{1 + \exp(a_i^*\theta_1 - \beta_i^*)}$$

[13]Im Fall der 1:1-Aufteilung gilt dies z.B. für jedes Itempaar der ersten Testhälfte.

Hierbei ist

$$a_i^* := \frac{a_{i2}\rho}{\sqrt{1 + k^2 a_{i2}^2 (1 - \rho^2)}}, \quad \beta_i^* := \frac{\beta_i}{\sqrt{1 + k^2 a_{i2}^2 (1 - \rho^2)}}$$

und

$$k := 16\sqrt{3}/(15\pi) \approx 0.588.$$

Setzt man den Wert $a_{i2} = 1$ sowie exemplarisch den Wert $\rho = 0.5$ ein, so folgt ein Diskriminationsparameter für ein Item der zweiten Dimension von $a_i^* \approx 0.45$, d.h. das Item besitzt ein nur halb so großes Diskriminationspotential wie ein Item der ersten Dimension ($a_{i1} = 1$). Diese Approximation bietet somit eine weitere, grobe Verdeutlichung anhand variierender Diskrimination, warum auch die Statistiken λ und Y sensitiv gegenüber dem Modell (4.11) reagieren.

Anhand der Form von a_i^* erkennt man ferner die Rolle der Korrelation. Mit wachsender Korrelation fällt der Effekt der verringerten Diskrimination geringer aus. Folglich fallen die Teststärken mit steigendem Korrelationswert.
Eine alternative Erklärung für den Einfluss der Korrelation orientiert sich an der ursprünglichen, zweidimensionalen Formulierung. Wie bereits in den einleitenden Bemerkungen von Kapitel 4.5 erwähnt, lässt sich ein Modell mit korrelierten Dimensionen in ein äquivalentes, unkorreliertes Modell umformulieren.
Die resultierenden neuen Ladungsmatrizen (zur Definition siehe Kapitel 4.5) sind für die drei verschiedenen Szenarien $\rho = 0.25, 0.5, 0.75$ unten aufgeführt. Aus Wiederholungsgründen wurde ferner die Ladungsmatrix exemplarisch auf zwei Items reduziert.

$$\rho = 0.25: \qquad \boldsymbol{A} := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \longrightarrow \quad \begin{pmatrix} 0.99 & 0.13 \\ 0.13 & 0.99 \end{pmatrix} =: \boldsymbol{A}^*$$

$$\rho = 0.5 \ : \qquad \boldsymbol{A} := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \longrightarrow \quad \begin{pmatrix} 0.97 & 0.26 \\ 0.26 & 0.97 \end{pmatrix} := \boldsymbol{A}^*$$

$$\rho = 0.75: \qquad \boldsymbol{A} := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \longrightarrow \quad \begin{pmatrix} 0.91 & 0.41 \\ 0.41 & 0.91 \end{pmatrix} := \boldsymbol{A}^*$$

Dies führt auf den Fall der „Within-Item-Multdimensionality“. Jedes Item ist betreffend dieser unkorrelierten Darstellung mit beiden Dimensionen verbunden. Je höher

die Korrelation, desto ähnlicher (gemessen z.B. anhand des Winkels) fallen die Vektoren korrespondierend zu den zwei verschiedenen Itemtypen aus und desto geringer sind folglich die Unterschiede im Diskriminationsvermögen. Ferner gilt es zu beachten, dass eine nahezu äquidistante Zunahme der Ladungen $(0.13, 0.26, 0.41)$ sich aufgrund des rapiden Anstiegs der Exponentialfunktion nicht in einer gleichmäßigen Erhöhung des Odds-Ratio niederschlägt. Dies bildet eine mögliche Verdeutlichung für die bereits festgestellte deutliche Abgrenzung des Falles $\rho = 0.75$ von den Fällen geringerer Korrelation.

Ergänzende Simulationen

Simulation 7: *Ist die einseitige Modifikation von ϕ - ähnlich wie in Abschnitt 5.2 - der ursprünglichen ϕ-Prüfgröße überlegen?*

Aufgrund der Darstellung des mehrdimensionalen Modells als (lokal abhängiges) eindimensionales Modell mit variierender Itemdiskrimination ließe sich - angesichts der Ergebnisse der ergänzenden Simulation 3 aus Kapitel 5.2 - vermuten, dass ϕ auch im Fall mehrdimensionaler Daten von einer einseitigen Formulierung profitieren könnte. Die Resultate einer hierzu korrespondierenden Simulation mit jeweils $m = 500$ Wiederholungen deuten in diese Richtung. Für die symmetrische Aufteilung ergab sich sowohl in der $(n = 250, k = 10, \rho = 0.5)$-Bedingung eine höhere Teststärke (23.6% gegenüber 15.4%), als auch in der $(n = 500, k = 30, \rho = 0.5)$-Bedingung (61.4% gegenüber 46.6%). Im Fall der asymmetrischen Aufteilung beliefen sich die Werte für die $(n = 250, k = 10, \rho = 0.5)$-Bedingung auf 20.4% (vs. 14.2%) und für $n = 500, k = 30, \rho = 0.5$ auf 52% (vs. 38%).
Folglich besitzt die einseitig formulierte ϕ-Statistik für alle hier betrachteten Szenarien eine höhere Teststärke als die zweiseitig definierte ϕ-Größe.

5.6 Diskussion und Ausblick

Die Resultate der Simulationen sollen in diesem Kapitel mit Bezug auf die drei abstrahierten Themenkomplexe „Spezifität“, „Globalität“ und „Vergleich parametrischer und nonparametrischer Tests“ diskutiert werden.
Das Kapitel endet mit einem theoretischen Resultat über die Optimalität der kombinatorischen Testklasse sowie einem Ausblick auf nicht behandelte Alternativmodelle.

Spezifität der Teststatistiken

Die vier in diesem Kapitel näher untersuchten Alternativmodelle lassen sich grob in zwei Klassen gruppieren. In der ersten Klasse (2PL bzw. 3PL-Modell) verlaufen die Item-Response-Funktionen diskrepant zu einem Rasch-Modell. Die Items sind jedoch, gegeben ein eindimensionales θ, unabhängig. Dies gilt nicht für die Modelle der zweiten Klasse (lokale stochastische Abhängigkeit, Mehrdimensionalität). Hier resultieren, gegeben eine eindimensionale latente Variable, Kovarianzen zwischen den Items.
Eine korrespondierende Einteilung von Prüfgrößen kann anhand der Klassifikation in „first-order“/„second order“-Statistiken vorgenommen werden. „First-order“-Größen orientieren sich an dem Zusammenhang zwischen dem Summenscore und dem Antwortverhalten auf einem Item, während „second-order“-Größen auf das simultane Antwortverhalten auf ein Itempaar zielen (siehe auch Kapitel 2.5). Erstere sollten vorrangig auf Verletzungen des Verlaufs der Item-Response-Funktionen reagieren, während Letztere vorrangig auf lokal abhängige Daten sensitiv ausfallen sollten (Glas, 1988; Suárez-Falcón und Glas, 2003).
Wie die Ergebnisse der Simulationen aufzeigen, ist eine solche strikte Einteilung nicht möglich. Es existieren sowohl „second-order“-Statistiken (T_{11}), die beträchtliche Power gegen einen diskrepanten Verlauf von $f_i(\theta)$ besitzen, als auch „first-order“-Größen (λ, Y), die Teststärke gegen mehrdimensionale (bzw. lokal abhängige) Daten aufweisen. Diese Überlappung wurde auch in einer Simulationsstudie von Suárez-Falcón und Glas (2003) gefunden. Der Effekt der Überlappung scheint ferner für die „second-order“-Tests wesentlich stärker auszufallen. Der T_{11}-Test ist bezüglich des 2PL-Modells sogar dem Likelihood-Quotienten-Test („first-order“-Test) überlegen. Die beiden „first-order“-Tests besitzen hingegen deutlich niedrigere Power bei Al-

ternativmodellen der zweiten Klasse (siehe z.B. die Tabellen 5.6 und 5.7) gegenüber dem T_{11}-Test.
Insgesamt gesehen kann diese Klassifizierung der Teststatistiken somit nicht ausreichend dem Umstand ihrer geringen Spezifität gerecht werden. Ein signifikantes Ergebnis eines „second-order"-Tests impliziert *nicht zwingend* einen Verstoß gegen die lokale stochastische Unabhängigkeit bzw. Eindimensionalität. Die Klassifikation in „first"- und „second-order"-Größen scheint insofern eher eine *einseitige* Implikation darzustellen: Ein „first-order"-Test besitzt Power gegenüber variierender Itemdiskrimination. Die Umkehrung ist hingegen nicht gegeben, d.h. ein Test, der Power gegenüber variierender Itemdiskrimination aufweist, muss nicht zwangsläufig ein „first-order"-Test sein (eine analoge Bemerkung gilt für „second-order"-Tests und lokal abhängige Daten).
Die Möglichkeit einer detaillierten Diagnose des mangelnden Modellfits ist folglich nicht gegeben. Eine Missachtung dieser mangelnden Spezifität kann zu falschen Interpretationen führen. Unterscheiden sich etwa die Diskriminationsfähigkeiten der beiden Testhälften eines eindimensionalen Tests stark, so führt dies zu einem signifikanten σ_r^2-Wert. Für den Betrachter könnte diese Signifikanz - in Unkenntnis der mangelnden Spezifität - fälschlicherweise als Indiz für einen mehrdimensionalen Test angesehen werden.

Das Problem der geringen Spezifität kann ebenso bei Item-fokussierten Tests auftreten. Beurteilt man z.B. lokale Abhängigkeit zweier Items anhand der Statistik $T(\boldsymbol{X}) = (r_{ij} - \rho_{ij})^2$, so können auch in Abwesenheit lokaler Abhängigkeit erhöhte Signifikanzen resultieren. Dies sei an einem kurzen Simulationsbeispiel demonstriert. Für $n = 500$ Personen und $k = 10$ Items wurden Daten gemäß einem 2PL-Modell simuliert[14]. Insgesamt fanden 500 Durchgänge statt. Die *festen* Diskriminationsparameter a_i betrugen:

$$a_i = \begin{cases} 1.5 & i = 1, 2 \\ 1.0 & i \geq 3 \end{cases}$$

Trotz Abwesenheit lokaler Abhängigkeit lehnte die Prüfgröße $T(\boldsymbol{X}) = (r_{12} - \rho_{12})^2$ 49.4% der simulierten Matrizen ab. Die Statistik reagiert somit *ebenfalls* auf Diskriminationsunterschiede in dem betrachteten Itempaar. Die Problematik der geringen

[14] Analog zu den Simulationen des fünften Kapitels erfolgen die Ziehungen der Schwierigkeits- und Personenparameter - wenn nicht anders erwähnt - stets gemäß der Standardnormalverteilung.

Spezifität zeigt sich folglich auch bei einer Itempaar-zentrierten Testgröße.
Eine deutliche Verbesserung in Hinblick auf eine eindeutige Modelldiagnose scheint die korrespondierende Mantel-Haenszel-Prüfgröße zu ermöglichen. Nutzt man für die gleichen Daten die in Kapitel 3.2 skizzierte Statistik[15] zum Testen auf lokale Abhängigkeit zweier Items, so erfolgt lediglich in 11.2% der Fälle eine Ablehnung der Nullhypothese. Das Analogon der Mantel-Haenszel-Klasse besitzt folglich eine wesentlich höhere Spezifität.
Dies scheint außerdem nur in geringem Maße zu Lasten der Power bei abhängigen Daten zu gehen. In einer Simulation (gemäß Modell (4.9) aus Kapitel 4.4) mit lokalen Abhängigkeiten bezüglich des ersten Itempaares ($n = 500, k = 10, \delta_{12} = 0.75, \delta_{(l-1)l} = 0$ für $l > 2$) zeigte sich verglichen mit der obigen kombinatorischen Prüfgröße lediglich ein Unterschied in der Teststärke von 1% (82.4% gegenüber 81.4%).
Die Teststatistik der Mantel-Haenszel-Testklasse könnte somit wertvolle Informationen im Hinblick auf eine spezifischere Modelldiagnose liefern[16].

„Overall"-Tests

Sucht man dagegen nicht nach spezifischen, sondern nach globalen Statistiken, d.h. nach Prüfgrößen, die im Sinne der χ^2-Statistik (2.15) auf jede Form der Modellverletzung reagieren, so scheint der T_{11}-Test dieser Forderung nahe zu kommen. Sowohl für lokal abhängige Modelle als auch für Modelle mit variierender Diskrimination weist er - verglichen mit anderen dargestellten Tests - hohe Teststärken auf. Es sei jedoch betont, dass der unterstellte globale Charakter - streng betrachtet - nicht gegeben sein kann. Eine Alternative, bezüglich derer T_{11} geringe Power besitzt, ist leicht konstruierbar. Man ordne genau den Matrizen aus $\boldsymbol{\Sigma_{rc}}$ mit geringem T_{11}-Wert eine hohe Wahrscheinlichkeit unter dem Alternativmodell zu. Gegenüber dem so enstandenen Modell besitzt T_{11} eine (beliebig) geringe Teststärke. Allerdings dürfte dieses Modell inhaltlich nur schwer interpretierbar und evtl. lediglich eine abstrakte Konstruktion sein, die realen Datensätzen nicht zugrundeliegt. Dies ist letztendlich eine empirische

[15] Das erste Item dient zur Zeilenklassifikation. Spaltenzuordnungen sind durch das zweite Item gegeben, und für die Schichtbildung ist der Summenscore $\sum_{i \neq 1} X_i$ verantwortlich.

[16] Eine Begründung für diese höhere Spezifität bietet die verallgemeinerte Formulierung der Mantel-Haenszel-Testklasse in Kapitel 3.2. Dort wurde implizit gezeigt, dass 2PL-Items - solange sie im Subtest $\boldsymbol{W}$ vorkommen - die Gültigkeit der Nullhypothese nicht beeinflussen.

Frage und wird hier offen gelassen.
Somit muss die Aussage über den globalen Charakter von T_{11} eingeschränkt werden. Wenn das datengenerierende Modell durch eines der in Kapitel 4 diskutierten Modelle näherungsweise dargestellt werden kann und wenn ferner kein spezifisches Vorwissen über die Items vorhanden ist, dann erscheint die Wahl von T_{11} vernünftig. Bei Vorwissen über relevante Modell-diskrepante Itempaare kann hingegen - wie in Simulation 6 aus Kapitel 5.4 angedeutet - eine Beschränkung der Summation auf diese abweichenden Itempaare zu deutlich höheren Teststärken führen.

Vergleich parametrischer und nonparametrischer Statistiken

Die Koppelung an das Alternativmodell sollte auch für den Vergleich von nonparametrischen mit parametrischen Prüfgrößen berücksichtigt werden. So erscheint angesichts der Simulationen das nonparametrische Y als eine Art „bessere Variante“ des parametrischen Likelihood-Quotienten-Tests. Diese Aussage gilt wiederum nur für die hier realisierten Alternativen. Es sind durchaus Szenarien denkbar - z.B. das Alternativmodell des Likelihood-Quotienten-Tests (seperates Rasch-Modell in den Scoregruppen) - in denen λ dem nonparametrischen Analogon überlegen sein könnte.
Aber auch bezüglich der in Kapitel 4 realisierten Modelle ließe sich einwenden, dass λ durch eine feinere Einteilung der Scoregruppen an Power gewinnen könnte und die Unterlegenheit gegenüber Y somit auf die grobe Zwei-Klassen-Gruppierung am Mittelwert zurückzuführen wäre. Eine Untersuchung respektive einer Drei-Klassen-Gruppierung ergab jedoch keine höheren Teststärken gegenüber der ursprünglichen Gruppierung. Die Überlegenheit des nonparametrischen Tests scheint folglich auch einer feineren Gruppierung standzuhalten[17].
Es ist aber kritisch zu hinterfragen, ob diese „globale“ Überlegenheit wünschenswert ist. Die deutlich geringere Power von λ gegenüber Y in den Fällen lokal abhängiger Daten zeigt z.B. eine höhere Spezifität der parametrischen Größe auf. Im Hinblick auf eine eindeutige Modelldiagnose könnte dies von Vorteil sein.

[17]Eine Modifikation von Y der Form Y_m (siehe die abschließende Bemerkung aus Kapitel 4.2) könnte zudem einen weiteren Zugewinn an Power gegenüber den 2PL/3PL-Modellen ermöglichen. Unter fünf willkürlich ausgewählten Bedingungen aus Tabelle 5.2 ergaben sich jeweils um 1-2% höhere Teststärken gegenüber Y.

Optimalität der kombinatorischen Tests

Die Suche nach Alternativmodellen, bezüglich derer ein gegebener nonparametrischer Test $T(\boldsymbol{X})$ Optimalitätseigenschaften besitzt, lässt sich - wie bereits von Ponocny (2001) skizziert - anhand des Lemmas von Neyman und Pearson durchführen. Betrachtet man die durch einen weiteren skalaren Parameter μ gekennzeichnete Klasse an Alternativen

$$P(\boldsymbol{X}|\,\boldsymbol{\theta},\,\boldsymbol{\beta},\,\mu) = v(\boldsymbol{\theta},\boldsymbol{\beta},\mu)\exp(\mu T(\boldsymbol{X}) + \textstyle\sum_v r_v\theta_v - \sum_i c_i\beta_i),$$

so resultiert nach Elimination der Störparameter $(\boldsymbol{\theta},\boldsymbol{\beta})$ durch Bedingen auf die suffizienten Statistiken $\boldsymbol{r}$ und $\boldsymbol{c}$:

$$P(\boldsymbol{X}|\,\boldsymbol{\theta},\,\boldsymbol{\beta},\,\mu,\,\boldsymbol{r},\,\boldsymbol{c}) = P(\boldsymbol{X}|\,\mu,\,\boldsymbol{r},\,\boldsymbol{c}) = \frac{\exp(\mu T(\boldsymbol{X}))}{\sum_{\boldsymbol{A}\in\boldsymbol{\Sigma}_{\boldsymbol{rc}}}\exp(\mu T(\boldsymbol{A}))}$$

Diese Familie an Wahrscheinlichkeitsfunktionen besitzt monotone Dichtequotienten („monotone likelihood ratio“) in $T(\boldsymbol{X})$. Ferner ergibt sich das Rasch-Modell als Spezialfall für $\mu = 0$. Folglich lässt sich ein UMP-Test nach dem Neyman-Pearson-Lemma für einseitge Fragestellungen ($H_0 : \mu \leq 0$) bzw. ein UMPU-Test für zweiseitige Probleme ($H_0 : \mu = 0$) konstruieren (Lehmann und Romano, 2005). Die korrespondierenden Tests lehnen bei zu hohen Werten (einseitig) bzw. bei zu hohen und zu niedrigen Werten (zweiseitig) von $T(\boldsymbol{X})$ ab.
Auch wenn somit bezüglich einer formal definierten Modellklasse optimale Eigenschaften resultieren, bleibt die Frage der „Realitätsnähe“ der konstruierten Modellklasse offen. Für die Modelle aus Kapitel 4 hingegen lassen sich reale Datensätze angeben, die näherungsweise durch eines der Modelle beschreibar sind.

Ausblick

Es sei jedoch an dieser Stelle bemerkt, dass die Auswahl der Alternativmodelle des vierten Kapitels sich zwar auf häufig in der Literatur vorkommende Modelle bezieht, aber bei weitem keine ausreichende Repräsentation der möglichen Item-Response-Modelle bietet. Die Analyse des mehrdimensionalen Modells beschränkte sich z.B. auf den idealisierten Fall der „Between-Item-Multidimensionality“ (siehe Kapitel 4.5). „Within-Item-Multidimensionality“ beinhaltet nicht eindeutig zu einer Dimension

zuteilbare Items. Dieser Fall wurde ebenso wenig behandelt wie der Fall eines partiell kompensatorischen mehrdimensionalen Modells. Ein Vertreter dieser Klasse ist z.B. gegeben durch

$$f_i(\boldsymbol{\theta}) = \prod_l \frac{\exp(\theta_l - \beta_{il})}{1 + \exp(\theta_l - \beta_{il})}.$$

Es gilt

$$f_i(\boldsymbol{\theta}) \leq \frac{\exp(\theta_l - \beta_{il})}{1 + \exp(\theta_l - \beta_{il})},$$

so dass eine niedrige Fähigkeit auf einer Dimension auch durch hohe Werte auf den anderen Dimensionen nicht ausgeglichen werden kann.
Jede dieser bisher beschriebenen Erweiterungen (Itemquerladungen, partiell kompensatorische Modelle) ist zudem mit der Wahl einer anderen Verteilungsfunktion (z.B. die asymmetrische Gompertzverteilung an Stelle der logistischen Funktion) kombinierbar. Insofern ergibt sich eine weitere große Modellklasse, bezüglich der die Untersuchung der Prüfgrößen in Hinblick auf Mehrdimensionalität ausweitbar wäre. Aber auch im Fall eines eindimensionalen Modells ist die getroffene Auswahl keineswegs erschöpfend. Es sind sowohl unstetige Item-Response-Funktionen als auch Item-Response-Funktionen mit nicht monotonem Verlauf denkbar. Eine „etablierte“ Möglichkeit für Letzteres liefert das Modell (Andrich und Luo, 1993):

$$f_i(\theta) = \frac{\exp \lambda_i}{\exp \lambda_i + \exp(\theta - \delta_i) + \exp(\delta_i - \theta)}$$

Die Item-Response-Funktion ist monoton wachsend für $\theta \leq \delta_i$ und monoton fallend für $\theta \geq \delta_i$. Die maximale Lösungswahrscheinlichkeit ergibt sich für $\theta = \delta_i$:

$$f_i(\delta_i) = \frac{\exp \lambda_i}{\exp \lambda_i + 2} < 1$$

Somit realisiert dieses Modell neben einer nicht monotonen Funktion auch eine obere Schranke für die Lösungswahrscheinlichkeit. Diese beiden Formen der Modellverletzungen stellen mitunter durchaus realistische Abweichungen dar. Verstöße gegen die Monotonie können sich z.B. bei ambivalenten Fragestellungen ergeben (Andrich, 1997), und obere Schranken/Asymptoten repräsentieren im psychopathologischen Anwendungsbereich ein gängiges Phänomen (Meijer und Baneke, 2004).

Insofern sollte eine weitergehende Analyse der kombinatorischen Modelltests diese Alternativen[18] einbeziehen und gegebenenfalls spezifische, für diese Modelle geeignete Prüfgrößen entwickeln. Diese Untersuchung könnte dann auch den Effekt der Itemschwierigkeiten, den dieses Kapitel aufgrund des zufälligen Ziehens der $\boldsymbol{\beta}$-Parameter unberücksichtigt ließ, thematisieren. Sowohl theoretisch unter Betrachtung des Grenzfalls $\beta_i \to \infty$ als auch praktisch anhand von Simulationsstudien lassen sich Effekte des Parameters β_i auf Teststärken nachweisen (Christensen und Kreiner, 2010).

Neben der Untersuchung der nonparametrischen Teststatistiken in Bezug auf diese erweiterte Auswahl an Alternativmodellen erscheint aber vor allem - gerade aus entscheidungstheoretischen Gründen - die Quantifizierung des Verlusts bei Verwendung eines falschen Modells relevant. Die Entscheidung für eine bestimmte Teststatistik der kombinatorischen Testklasse ist, wie in der Diskussion kenntlich gemacht, abhängig von dem unterstellten Alternativmodell. Bei einer eng umrissenen Alternativmodellklasse (z.B. 2PL-Modell) lässt sich zwar relativ leicht und ohne Bezugnahme auf Verlustfunktionen eine geeignete Prüfgröße finden (Y). In Anbetracht stark verschiedener, plausibler Alternativen (z.B. 2PL-Modell und ein lokal abhängiges Modell), in welchem Fall mehrere Statistiken zur Prüfung der Nullhypothese geeignet scheinen, müsste die Wahl einer Teststatistik aber auch die Verluste einbeziehen, die aus einer Fehlentscheidung resultieren. Für den Fall einer fälschlich angenommenen Nullhypothese mündet dies in der Frage, welche statistischen Eigenschaften die aus einem Rasch-Modell resultierenden Schätzwerte $\hat{\theta}$ unter dem von einem Alternativmodell (z.B. 2PL-Modell) induzierten Wahrscheinlichkeitsmaß besitzen. Ein erster Hinweis für Modelle aus der MHM-Klasse bietet die stochastische Ordnung der latenten Größe anhand des Summenscores. Abgesehen davon bedarf es jedoch weiterer Untersuchungen, um einer adäquaten, entscheidungstheoretisch-basierten Auswahl einer Teststatistik näher zu kommen.

[18] Natürlich stellt dies nur eine unvollständige Auswahl der vorhandenen Möglichkeiten dar.

Kapitel 6

Nonparametrische Analyse des IST 2000 R

6.1 Vorbemerkungen

Der IST 2000 R ist ein in der Praxis weit verbreiteter (Schorr, 1995), nach der klassischen Testtheorie konstruierter Test zur Erfassung unterschiedlicher Dimensionen des Intelligenz-Konstrukts, der in verschiedenen Versionen realisiert ist. Die folgende Auswertung orientiert sich an der Kurzform A, die sich aus neun, jeweils 20 Items umfassenden Subskalen zusammensetzt:

- Satzergänzung
- Analogien
- Gemeinsamkeiten
- Rechenaufgaben
- Zahlenreihen
- Rechenzeichen
- Figurenauswahl
- Würfelaufgaben
- Matrizen

Eine Analyse von Bühner u.a. (2006) zeigte unzureichende Reliabilitäten[1] der verbalen Skalen (Satzergänzung, Analogien, Gemeinsamkeiten) auf. Die Frage der Rasch-Modell-Konformität wurde mittels des Bootstrap-Tests (auf der globalen χ^2-Statistik

[1] Die geringen Reliabilitäten sind auf eine Studentenpopulation bezogen und können (daher) teilweise durch eine geringe Streuung der latenten Variable erklärt werden.

basierend) für die meisten Skalen positiv beantwortet. Vor dem Hintergund, dass eine kürzliche Untersuchung (Mang, 2009) der Teststärke des Bootstrap-Tests jedoch „alarmierende“ Befunde bezüglich dessen Gütefunktion lieferte (die Power des Bootstrap-Tests liegt für die in dieser Arbeit betrachteten Dimensionen meistens in der Nähe des 5%-Niveaus), erscheint eine Reanalyse dieser Skalen angebracht.

Bevor die Überprüfung auf Rasch-Skalierbarkeit der Subskalen erfolgt, soll jedoch zunächst mittels nonparametrischer deskriptiver Methoden - orientiert an Kapitel 3.3 - ein Einblick in die psychometrischen Eigenschaften der Skalen gegeben werden. Der Einsatz von nonparametrischen Methoden zur (Vor-)Analyse eines Datensatzes wurde bereits an mehreren Stellen (Holland, 1981; De Koning u.a., 2002, 2003) vorgeschlagen und gewinnbringend eingesetzt. Nonparametrische deskriptive Methoden ermöglichen nicht nur einen schnellen Überblick mittels relativ einfacher Statistiken, sondern können auch zur Skalenkonstruktion eingesetzt werden. Dieser Aspekt wird in Kapitel 6.4 näher behandelt.

Die Vorgehensweise der deskriptiven Analyse orientiert sich dabei an den Obermodellen des Rasch-Modells. Zunächst werden über eine Analyse der Itemkorrelationen Eigenschaften des MHM überprüft. Die Erfüllung dieser Eigenschaften stellt eine notwendige Bedingung für ein Rasch-Modell dar und dient gleichzeitig als Indiz, ob die Skalen einer ordinalen Messintention im Sinne der stochastischen Ordnung (siehe Kapitel 2.3) genügen können.
Neben dieser die MHM-Konformität betreffenden Frage steht aber auch die Messgenauigkeit der Skalen im Vordergrund. Selbst bei Vorlage eines Rasch-Modells kann eine Skala sich als „nutzlos“ erweisen. Dies lässt sich gut am Fall einer Matrix verdeutlichen, deren Einträge i.i.d. Bernoulli-verteilte Zufallsvariablen sind. Es liegt zwar ein Rasch-Modell vor (z.B. $\theta_v = \beta_i = 0$), die Messung besteht jedoch zu 100% aus Messfehlern.
Es erfolgt somit im *ersten Teil* der Analyse eine deskriptive Überprüfung der MHM-Implikationen unter gleichzeitiger Berücksichtigung der Forderung nach ausreichender Reliabiltät.
Der *zweite Teil* befasst sich abschließend mit der Frage der invarianten Itemordnung (DMM), die eine über die Untersuchung der MHM-Axiome hinausgehende, weitere notwendige Bedingung für ein Rasch-Modell darstellt.

6.2 Deskriptive Analyse

Stichprobe

Detaillierte Angaben zur Stichprobe finden sich bei Bühner u.a. (2006). Hier sei nur erwähnt, dass es sich um 176 weibliche sowie 97 männliche Studenten der Philipps-Universität Marburg handelt. Ferner liegt keine Zufallsstichprobe vor, so dass die nachfolgenden Analysen vor diesem Hintergrund zu relativieren sind. Eine weitere Einschränkung ergibt sich durch die Diskrepanz zwischen Studienpopulation (Studenten der Philipps-Universität Marburg) und der tatsächlich intendierten Zielpopulation[2] des IST 2000 R. Die folgende Auswertung bezieht sich stets auf die psychometrischen Eigenschaften des IST 2000 R in der hypothetischen Studienpopulation.

Analyse der Korrelationsmatrizen

Einen ersten Eindruck über die Qualität einer Skala ermöglicht die Inspektion der Inter-Item-Korrelationsmatrix. Aus Satz 3 (Kapitel 3.3) ist bekannt, dass bei Vorlage eines MHM die Korrelationen der Items nichtnegativ ausfallen. Folglich kann eine Untersuchung der Vorzeichen innerhalb der Korrelationsmatrix ein erstes Indiz für die Messeigenschaften der Skala liefern.
Tabelle 6.1 beinhaltet in der zweiten Spalte für jede Subskala den Anteil negativer Korrelationen gemessen an allen paarweisen Itemkorrelationen. Für nahezu die Hälfte der Skalen liegt dieser Wert über 20%. Besonders betroffen sind die verbalen Skalen. Als unbedenklich hingegen kann die Skala „Würfelaufgaben“ angesehen werden.
Es sollte an dieser Stelle betont werden, dass eine nichtnegative marginale Itemkorrelation eine relativ schwache Mindestanforderung darstellt. Sie kann bereits unter milderen Annahmen als denen eines MHM hergeleitet werden (Mokken, 1971; Holland, 1981). Eine Verletzung dieser basalen Forderung impliziert, dass kein „gängiges“ eindimensionales Item-Response-Modell (z.B. 3PL-Modell) die Daten beschreiben kann.

[2]Der IST 2000 R ist für den Einsatz in mehreren Populationen gedacht. Die Population bestehend aus „jungen“ Personen mit Gymnasialabschluss erscheint hier - für den Vergleich - noch am besten geeignet.

Tabelle 6.1: *Beurteilung der Skalenqualität anhand der marginalen Korrelationen (Cor_m), bedingten Korrelationen (Cor_b) sowie diverser Skalierungskoeffizienten (Erläuterung siehe Text).*

Skala	$\text{Cor}_m < 0$ (in %)	$\text{Cor}_b < 0$ (in %)	H	$\min H_i$	$\max H_i$
Satzergänzung	24	33	0.11	0.02	0.28
Analogien	23	31	0.12	-0.01	0.26
Gemeinsamkeiten	30	36	0.09	-0.02	0.21
Rechenaufgaben	9	18	0.31	0.07	0.65
Zahlenreihen	13	17	0.49	-0.25	0.73
Rechenzeichen	8	18	0.35	0.07	0.59
Figurenauswahl	14	23	0.11	0.03	0.22
Würfelaufgaben	4	13	0.31	0.08	0.44
Matrizen	23	31	0.11	0.01	0.20

Marginale nicht negative Korrelation folgt - als Spezialfall - aus marginaler Assoziation (siehe Definition 3 in Kapitel 3.3). Satz 3 bezieht sich jedoch auf die stärkere Form der *bedingten* Assoziation[3]. *Eine*[4] Möglichkeit, diese bedingte Assoziation zu „prüfen“, bildet die Untersuchung der bedingten Korrelation zweier Items auf jeder Stufe (0 oder 1) eines dritten Items. Aus Satz 3 folgt, dass negative bedingte Korrelationen im Widerspruch zu einem MHM stehen.

Die dritte Spalte der Tabelle 6.1 gibt den Anteil negativer bedingter Itemkorrelationen an allen bedingten Korrelationen innerhalb der Skala wieder. Es wurden hierfür alle möglichen Kombinationen an „Dreier-Items“ gebildet (für jede feste Wahl dreier Items (i, j, k) existieren drei Möglichkeiten, eine bedingende Variable zu bilden. Da die bedingende Variable jeweils zwei Ausprägungen besitzt, sind folglich insgesamt $3 \cdot 2 = 6$ Korrelationen für diese feste Auswahl zu untersuchen). Wie anhand der Werte deutlich erkennbar, zeigt sich ein zum Fall der marginalen Korrelation ähnliches Bild. Bei den verbalen Skalen sowie der Skala „Matrizen“ steht nahezu jede dritte bedingte Korrelation in Diskrepanz zu einem MHM. Die geringste prozentuale Abweichung weist auch hier der Würfelaufgabentest auf. Insgesamt sind aber für jede Skala noch stärkere Abweichungen - verglichen mit dem Fall der mar-

[3] Aus bedingter Assoziation folgt u.a. marginale Assoziation. Eine Darstellung und Diskussion der Zusammenhänge verschiedener Assoziationskonzepte geben Holland und Rosenbaum (1986).

[4] Satz 3 offeriert eine große Anzahl potentiell untersuchbarer Zusammenhänge. Hier wird lediglich *eine* konkrete Wahl diskutiert.

ginalen Korrelation - festzustellen.
Somit bestehen - basierend auf marginalen und bedingten Korrelationen - für *mindestens* vier der neun Skalen erhebliche Zweifel an der Gültigkeit eines MHM und damit insbesonders an der Rasch-Skalierbarkeit.

Bisher wurde lediglich das Vorzeichen der Inter-Item-Korrelationen zur Beurteilung einer Skala verwendet. Der naheliegende Gedanke, den Fit eines Itempaares zusätzlich anhand der *Höhe* der Korrelation zu evaluieren, kann jedoch zu Trugschlüssen führen: Bei dichotomen Items ist der Korrelationskoeffizient aufgrund der Randverteilungen des Itempaares mitunter starken Restriktionen unterworfen. Große Schwierigkeitsunterschiede der Items führen zu geringen (maximal möglichen) Korrelationen. Folglich würde eine Evaluation der Items anhand der Höhe der Korrelation fehlleiten. Die Beurteilung, in welchem Maße zwei Items die gleiche Fähigkeit messen, ist konfundiert mit der Itemschwierigkeit. Dies ist zuweilen auch ein Grund, warum lineare Faktorenanalysen zu falschen Aussagen bezüglich der Dimensionalität eines Datensatzes gelangen können (McDonald und Ahlawat, 1974).
Ein Maß, welches diese Problematik umgeht - d.h. dieses Maß kann unabhängig von den Randverteilungen seinen Maximalwert erreichen - ist durch den H_{ij}-Koeffizienten gegeben (Mokken, 1971). Nach einer kurzen Darlegung des Koeffizienten zu Beginn des folgenden Abschnitts wird dieser anschließend zur Identifikation auffälliger Itempaare im IST eingesetzt.

Skalierungskoeffizienten

Bezeichne[5] π_i die Wahrscheinlichkeit, das i-te Item zu lösen - d.h. in Termini der Item-Response-Theorie ausgedrückt

$$\pi_i = \int f_i(\theta) \mathrm{d}G(\theta)$$

und bezeichne ferner π_{ij} die Wahrscheinlichkeit, sowohl Item i als auch Item j zu lösen, dann ist der H_{ij}-Koeffizient definiert als (o.B.d.A. sei $\pi_i < \pi_j$):

$$H_{ij} := \frac{\mathrm{Cov}(X_i, X_j)}{\mathrm{Cov}_{\max}(X_i, X_j)} = \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i(1-\pi_j)} = 1 - \frac{\pi_i - \pi_{ij}}{\pi_i(1-\pi_j)} \tag{6.1}$$

[5]Zur *Schätzung* der nachfolgenden Größen ersetze man lediglich die Populationsgrößen durch die korrespondierenden relativen Häufigkeiten.

Tabelle 6.2: *Itempaare - pro Subskala - mit „stark“ negativen Skalierungskoeffizienten.*

Skala	Itempaare (i, j) mit der Eigenschaft: $H_{ij} \leq -0.3$
Satzergänzung	(06,20), (10,15)
Analogien	(01,10), (01,11), (01,17), (02,16), (04,16), (05,13), (05,16)
Gemeinsamkeiten	(02,14), (02,19), (02,20), (03,20), (04,20), (12,20), (17,20)
Rechenaufgaben	(01,12)
Zahlenreihen	(01,14), (01,15), (01,16), (01,17), (01,18), (01,19), (01,20)
Rechenzeichen	
Figurenauswahl	
Würfelaufgaben	(02,20)
Matrizen	(01,08), (02,16), (02,20), (04,19), (09,20)

$\mathrm{Cov}_{\max}(X_i, X_j)$ bezeichnet hierbei die maximal mögliche Kovarianz zweier Items, die gleiche Randverteilungen wie Item i und Item j aufweisen. Da explizit durch die maximale Kovarianz dividiert wurde, kann der H_{ij}-Koeffizient unabhängig von den Randverteilungen der Items den Maximalwert eins erreichen. Ferner sieht man unmittelbar, dass das Vorzeichen des H_{ij}-Koeffizienten identisch mit dem Vorzeichen der Kovarianz ist. Folglich stehen negative Werte im Widerspruch zu einem MHM.
Eine alternative Interpretation des H_{ij}-Koeffizienten bietet die rechte Seite von (6.1). Der Zähler des Bruchs der rechten Seite von (6.1) entspricht der Wahrscheinlichkeit eines Guttman-Fehlers (d.h. das schwerere Item i wird gelöst, nicht jedoch das leichtere Item j). Der Nenner beschreibt die Wahrscheinlichkeit eines Guttman-Fehlers unter Annahme der marginalen Unabhängigkeit der Items. Somit ist ein H_{ij}-Koeffizient von 0.3 gleichzusetzen mit einer Reduktion des Guttman-Fehlers um 30% gegenüber dem Fall der Unabhängigkeit.
Tabelle 6.2 listet für jede Skala die Itempaare mit „stark“ negativem H_{ij}-Koeffizienten (nach der Heuristik $H_{ij} \leq -0.3$) auf. Diese bilden einen ersten Ansatzpunkt zu einer potentiellen Verbesserung der Skala.
Auffallend ist vor allem der Zahlenreihentest. Jedes „kritische“ Itempaar beinhaltet Item 1. Eine nähere Betrachtung dieses Items zeigt, dass es sich um ein extrem leichtes Item handelt. Es wurde lediglich von einer Person falsch beantwortet. Dieses Item stellt sich somit - unabhängig von dem Wert des Skalierungskoeffizienten - aus Gründen mangelnder Diskrimination als kritisch dar.
In der Skala „Gemeinsamkeiten“ kommt des Weiteren das letzte Item fünfmal in

einem kritischen Itempaar vor. Auch im Analogientest finden sich mehrfach vorkommende Items. Das wirft die Frage auf, ob diese Items adäquat für die jeweilige Skala sind.

Eine Größe, die für die Beantwortung dieser Frage geeignet scheint, ist der H_i-Koeffizient. Abstrahiert von den bisherigen paarweisen Betrachtungen, quantifiziert er den Fit eines Items in eine Skala.

$$H_i := \frac{\sum_{j \neq i} \mathrm{Cov}(X_i, X_j)}{\sum_{j \neq i} \mathrm{Cov}_{\max}(X_i, X_j)} = 1 - \frac{\sum_{j \neq i} O_{ij}}{\sum_{j \neq i} E_{ij}} \tag{6.2}$$

E_{ij} bezeichnet hierbei die Wahrscheinlichkeit eines Guttman-Fehlers bei Unabhängigkeit der Items (siehe (6.1)). Analog steht O_{ij} für die tatsächliche Wahrscheinlichkeit eines Guttman-Fehlers.
Der H_i-Koeffizient stellt gewissermaßen ein zur Trennschärfe bzw. Itemdiskrimination analoges Maß dar, welches jedoch die Restriktionen bezüglich der Randverteilungen berücksichtigt. Ein hoher[6] Koeffizient spricht für einen guten Fit eines Items in eine Skala.
Bezüglich des Rasch-Modells ist jedoch Vorsicht geboten. Eine hohe Variabilität der H_i-Koeffizienten einer Skala deutet auf unterschiedliche Itemdiskrimination hin, da der H_i-Koeffizient als ein nonparametrisches Analogon zum a_i-Parameter des 2PL-Modells interpretiert werden kann (Meijer und Baneke, 2004). Im Hinblick auf ein Rasch-Modell ist folglich eine *geringe Variation* der Koeffizienten erwünscht. Das Rasch-Modell stellt keine Forderungen bezüglich der Höhe des H_i-Koeffizienten (nur die Implikation $H_i \geq 0$, die für ein MHM gilt). Gleichwohl ist der Aspekt der Höhe in Bezug auf die Messgenauigkeit einer Skala nicht auszuklammern. Somit ergibt sich die Forderung nach hohen H_i-Koeffizienten (Reliabilität) in Kombination mit geringer Variation (gleiche Itemdiskrimination).
Die extremsten H_i-Koeffizienten innerhalb der einzelnen Subskalen des IST sind in Tabelle 6.1 aufgeführt. Das aufgrund der negativen H_{ij}-Koeffizienten als „kritisch“ bezeichnete Item 20 der Skala „Gemeinsamkeiten“ besitzt den niedrigsten H_i-Koeffizienten innerhalb dieser Skala. Der starke Ausreißer von $H_i = -0.25$ innerhalb des Zahlenreihentests ist ferner auf Item 1 zurückzuführen. Somit können die

[6] $H_i > 0.3$ bildet eine grobe Orientierung für einen akzeptablen Wert gemäß einer Faustregel von Mokken (1971).

Ergebnisse, die aus der Untersuchung kritischer Itempaare resultierten, anhand der H_i-Koeffizienten bestärkt werden.
Bezüglich der Forderung nach geringer Variabilität, beurteilt über die Spannweite der Skalierungskoeffizienten H_i, sind vor allem die nicht verbalen Skalen „Rechenaufgaben“, „Zahlenreihen“, „Rechenzeichen“ sowie „Würfelaufgaben“ als kritisch anzusehen.

Betrachtet man hingegen den Reliabilitätsaspekt der Skalen, so fallen vorrangig die verbalen Skalen negativ auf. Der H-Koeffizient (Loevinger, 1948) ist ein Maß für die Gesamtqualität einer Skala. Er ergibt sich als gewichtetes Mittel der H_i-Koeffizienten (Sijtsma und Molenaar, 2002, S.58). Respektive dieser Größe erscheinen - gemäß der Faustregel von Mokken ($H \geq 0.3$) - lediglich die Skalen „Rechenaufgaben“, „Zahlenreihen“, „Rechenzeichen“ sowie „Würfelaufgaben“ als akzeptabel (Tabelle 6.1). Ein Resultat, welches im Einklang mit den korrespondierenden Reliabilitätsschätzungen von Bühner u.a. (2006) steht.

Insgesamt lässt sich aus der Untersuchung der Skalierungskoeffizienten festhalten, dass insbesondere die verbalen Skalen eine unzureichende Messgenauigkeit aufweisen. Für zwei der drei verbalen Skalen konnte ferner ein größere Anzahl an stark negativen H_{ij}-Koeffizienten festgestellt werden. Ebenso ergab eine Analyse der marginalen und bedingten Korrelationen Zweifel an der Gültigkeit eines MHM für die drei verbalen Skalen[7] und den Matrizentest. Die Skala „Würfelaufgaben“ verfügte über den geringsten Anteil an negativen Korrelationen. Innerhalb der Skalen mit höherer Reliabilität fanden sich gemessen an der Spannweite der H_i-Koeffizienten stark variierende Itemdiskriminationen. Somit bestehen, was die *simultane* Forderung nach Rasch-Skalierbarkeit und Messgenauigkeit betrifft, an allen Subskalen Zweifel.

Invariante Itemordnung

Die bisherige Analyse beschränkte sich im Wesentlichen auf eine deskriptive Prüfung diverser Implikationen eines „Monotone-Homogeneity-Model“ sowie einer Inspektion der Messgenauigkeit anhand des H-Koeffizienten. In diesem Abschnitt soll die

[7] Auch für die numerischen Skalen sowie den Figurenauswahltest fanden sich einige inkonsistente Korrelationen.

zusätzliche Eigenschaft des „Double-Monotonicity-Model“ betrachtet werden.
Die Forderung, welche im Folgenden anhand einfacher deskriptiver Methoden untersucht werden soll, lautet (o.B.d.A. sei $\pi_i < \pi_j$):

$$f_i(\theta) \leq f_j(\theta) \quad \forall\theta \tag{6.3}$$

Dies entspricht der Eigenschaft (NI) aus Kapitel 2.3, die eine von der spezifischen Fähigkeitstufe unabhängige Ordnung der Items postuliert. Für jedes beliebige Fähigkeitsniveau θ ist Item i schwieriger als Item j.

Um eine empirisch überprüfbare Aussage der Eigenschaft (NI) abzuleiten, ist es notwendig, die latente Variable zu „entfernen“ und eine Schätzung der Itemordnung vorzunehmen. Letzteres kann auf naheliegende Art geschehen. Die Populationsgrößen π_i werden durch ihre relativen Häufigkeiten p_i ersetzt. Hierdurch ist eine Ordnung der Items gemäß der relativen Häufigkeiten gegeben, die eine gute Näherung der tatsächlichen Ordnung bezüglich π_i widerspiegelt. Wenn die Eigenschaft (NI) gilt, dann besteht die gleiche Ordnung innerhalb jeder Fähigkeitsstufe.
Betrachtet man ein festes Item i sowie die simultane Lösungswahrscheinlichkeit mit Item j $\pi_{ij} := P(X_i = 1, X_j = 1)$ und einem relativ zu j leichterem Item l (π_{il}), so gilt unter der Annahme (NI) sowie lokaler stochastischer Unabhängigkeit:

$$\begin{aligned}
\pi_{ij} - \pi_{il} &= \int (P(X_i = 1, X_j = 1|\,\theta) - P(X_i = 1, X_l = 1|\,\theta))\,\mathrm{d}G(\theta) && (6.4)\\
&= \int f_i(\theta)(f_j(\theta) - f_l(\theta))\mathrm{d}G(\theta) && (6.5)\\
&\leq 0 && (6.6)
\end{aligned}$$

Folglich ist nach einer Umordnung bzw. Neunummerierung der Items gemäß ihrer Schwierigkeit ($\pi_1 \leq \pi_2 \leq \cdots \leq \pi_{k-1} \leq \pi_k$) die Funktion $f_i^1(j) := \pi_{ij}$ (definiert für $j \neq i$) monoton *wachsend* in j für jedes i. Ersetzt man die entsprechenden Populationsgrößen π_{ij} durch ihre beobachteten relativen Häufigkeiten p_{ij}, ergibt sich eine empirisch prüfbare Aussage.
Eine zu (6.4)-(6.6) analoge Gleichungskette führt ferner zur Monotonie der Funktion $f_i^0(j) := \pi_{ij}^0 := P(X_i = 0, X_j = 0)$. Unter der Annahme (NI) liegt hier - nach obiger Neunummerierung - eine monoton *fallende* Funktion in j vor.

Es sei hierbei bemerkt, dass die Schlussfolgerungen (6.4)-(6.6) ihre Gültigkeit allein aus der Annahme (NI) und der lokalen Unabhängigkeit beziehen. Auch bei nicht

montonen Item-Response-Funktionen bleiben die Aussagen bestehen. Gleiches gilt - solange lokale stochastische Unabhängigkeit vorliegt - für eine *mehrdimensionale* latente Variable[8]. Insofern liegt in der Überprüfung der Monotonie-Eigenschaft von $f_i^1(j)$ „eigenständige“ Information. Auch Modelle außerhalb der Klasse des MHM können die Eigenschaft (NI) erfüllen. (NI) stellt lediglich Forderungen an die Beziehung zweier Item-Response-Funktionen, nicht jedoch an den Verlauf einer einzelnen Item-Response-Funktion.

Für den vorliegenden Fall wurde für jedes Item i einer Skala die Anzahl an Verstößen gegen die Monotonie bezüglich $f_i^1(j)$ und $f_i^0(j)$ gezählt. Ein Verstoß ist dabei durch ein Itempaar (j, l) definiert, dessen Differenz $f_i^1(j) - f_i^1(l)$ (bzw. $f_i^0(j) - f_i^0(l)$) nicht konform ausfällt. Ferner wurden Verstöße nur gezählt, wenn die korrespondierende Wahrscheinlichkeitsdifferenz mindestens 3% betrug.
Die drei verbalen Skalen zeigen ein insgesamt unauffälliges Verhalten. Es ergeben sich weder hervorstechende Items noch stellt die Skala als Ganzes betrachtet, d.h. die Summation der Verletzungen über alle Items der Skala, einen erkennbaren Widerspruch zu (NI) dar. Der Satzergänzungstest z.B. weist ingesamt fünf Verletzungen auf. Dies ist angesichts der hohen Anzahl potentiell möglicher Verletzungen ein sehr geringer Wert. Analoge Resultate zeigen sich für die übrigen verbalen Skalen.
Bei den Skalen „Rechenaufgaben“, „Rechenzeichen“ sowie „Matrizen“ finden sich trotz größerer Werte als bei den Verbalskalen ebenfalls keine auffällig starken Verletzungen.
Ähnliches gilt für die Skalen „Zahlenreihen“ sowie „Figurenauswahl“. Allerdings existiert im Figurenauswahltest ein Item (Item 20) mit 8 Verstößen bezüglich $f_{20}^1(j)$. Die restlichen Items sind hingegen unauffällig.
Die mit Abstand größten Verletzungen der Annahme (NI) ergeben sich für die Skala „Würfelaufgaben“. Tabelle 6.3 beinhaltet für jedes Item i dieses Tests die Anzahl der Verstöße bezogen auf $f_i^1(j)$ und $f_i^0(j)$. Die meisten Verletzungen resultieren bei Items der zweiten Testhälfte, insbesondere den Itemindizes 13 bis 16. Ferner ist $f_i^1(j)$ wesentlich stärker betroffen als $f_i^0(j)$.
Betrachtet man den Integranden in (6.5) und nimmt des Weiteren monoton wachsendes $f_i(\theta)$ an, so wird ersichtlich, dass das Gewicht $f_i(\theta)$, mit dem eine Verletzung $(f_j(\theta) > f_l(\theta))$ in das Integral eingeht, umso größer ist, je höher die Fähigkeit θ (bzw. $f_i(\theta)$) ausfällt.

[8]Die Gleichungen bleiben erhalten, falls (6.3) durch $f_i(\boldsymbol{\theta}) \leq f_j(\boldsymbol{\theta})\ \forall \boldsymbol{\theta}$ ersetzt wird.

Tabelle 6.3: *Anzahl Verstöße gegen die gemäß (NI) erwartete Monotonie von $f_i^1(j)$ und $f_i^0(j)$ für die Items der Skala „Würfelaufgaben“.*

Item $\boldsymbol{i}=$	$f_i^1(j)$	$f_i^0(j)$	Item $\boldsymbol{i}=$	$f_i^1(j)$	$f_i^0(j)$
1	2	1	**11**	7	1
2	1	3	**12**	4	3
3	2	5	**13**	11	5
4	1	1	**14**	12	4
5	1	3	**15**	15	1
6	0	0	**16**	16	6
7	0	0	**17**	7	0
8	0	0	**18**	4	0
9	0	4	**19**	11	0
10	0	0	**20**	0	0

Im Fall von $f_i^0(j)$ gilt hingegen:

$$\begin{aligned} \pi_{ij}^0 - \pi_{il}^0 &= \int (P(X_i=0, X_j=0|\,\theta) - P(X_i=0, X_l=0|\,\theta))\,\mathrm{d}G(\theta) \\ &= \int (1-f_i(\theta))(f_l(\theta)-f_j(\theta))\mathrm{d}G(\theta) \qquad (6.7) \\ &\geq 0 \end{aligned}$$

In (6.7) fällt das Gewicht einer Verletzung umso höher aus, je geringer die Fähigkeit bzw. je höher $1-f_i(\theta)$ ist. Folglich kann aufgrund der Dominanz von $f_i^1(j)$- gegenüber $f_i^0(j)$-Verstößen vermutet werden, dass die Abweichungen von der invarianten Itemordnung verstärkt bei hohen Fähigkeiten auftreten.
Ingesamt kann somit für die Skala „Würfelaufgaben“ die Existenz eines lokal unabhängigen (potentiell mehrdimensionalen) Modells mit der Eigenschaft (NI) stark angezweifelt werden.

Bemerkung. Die Methode dieses Abschnitts ist in der Literatur unter dem Namen „Item-Splitting“ (Sijtsma und Molenaar, 2002) bekannt. Eine ähnliche Methode namens „Restscore-Splitting“ führte angewandt auf die Subskalen zu identischen Resultaten.

6.3 Prüfung auf Rasch-Skalierbarkeit

Aus den bisherigen Analysen konnte zwar ein erster Eindruck bezüglich der Konformität der Skalen mit einem MHM bzw. DMM (und damit implizit mit einem Rasch-Modell) gewonnen werden, die Überprüfung der Rasch-Modell-Konformität anhand einer Teststatistik wird dadurch jedoch nicht obsolet.
Die Prüfung auf Rasch-Skalierbarkeit erfolgt anhand der kombinatorischen Modelltests. Als Entscheidungsgrundlage dient der T_{11}-Test, der in den Simulationsstudien gegen jedes Alternativmodell sensitiv reagierte. Die übrigen drei Teststatistiken - ebenso wie der parametrische Likelihood-Quotienten-Test - werden lediglich als zusätzliche Orientierung präsentiert. Die Entscheidung, ob die Nullhypothese des Rasch-Modells zu verwerfen ist, erfolgt jedoch allein aufgrund des p-Werts der T_{11}-Statistik ($\alpha = 0.05$).
Die p-Werte der kombinatorischen Tests sind in Tabelle 6.4 aufgeführt. Sie basieren auf jeweils $n_{eff} = 5000$ generierten Matrizen, um einen möglichst geringen Monte-Carlo-Fehler zu gewährleisten. Alle p-Werte des T_{11}-Tests führen zur Ablehnung der Rasch-Skalierbarkeit der jeweiligen Subskala. Identische Schlussfolgerungen erlaubt die Prüfgröße Y. Die Teststatistik ϕ hingegen lehnt die Rasch-Skalierbarkeit der verbalen Skalen nicht ab. Wie aus dem Simulationsabschnitt 5.4 bekannt, besitzt bei einem - relativ zur Testlänge - hohen Anteil lokal abhängiger Itempaare T_{11} gegenüber ϕ eine deutlich höhere Power. Nimmt man die starken Verletzungen bezüglich des Vorzeichens der bedingten Korrelation (Tabelle 6.1) als Anzeichen für eine hohe An-

Tabelle 6.4: *p-Werte der Tests auf Rasch-Skalierbarkeit für die Subskalen des IST 2000 R.*

Skala	Y	T_{11}	ϕ	σ_r^2	λ
Satzergänzung	0	0	0.224	0.989	0
Analogien	0	0	0.411	0.755	0.002
Gemeinsamkeiten	0.008	0.024	0.055	0.078	0.006
Rechenaufgaben	0	0	0	1	0
Zahlenreihen	0	0	0	1	0
Rechenzeichen	0	0	0	0.988	0.003
Figurenauswahl	0	0.001	0	0	0
Würfelaufgaben	0	0	0	0.989	0
Matrizen	0.038	0	0.009	0.176	0.590

zahl lokal abhängiger Itempaare, so liegt genau diese Situation für die verbalen Skalen vor.
Bezüglich der Beurteilung der Werte des σ_r^2-Tests in Tabelle 6.4 ist zu beachten, dass der p-Wert sich auf einen einseitigen Test bezieht. Ein p-Wert von 98.9% bedeutet, dass 98.9% der generierten Matrizen eine höhere Varianz bezüglich des Summenscores der ersten Testhälfte aufweisen. Dies impliziert aber zugleich ein signifikantes Resultat bei einer zweiseitigen Formulierung des Tests. Somit lehnt der *zweiseitige* σ_r^2-Test lediglich bei den Skalen „Analogien“, „Gemeinsamkeiten“ und „Matrizen“ die Nullhypothese nicht ab. Trotz ähnlicher Eigenschaften bei der deskriptiven Analyse wird hingegen die verbale Skala „Satzergänzung“ abgelehnt. Ein möglicher Grund liegt - in Anlehnung an die ergänzende Simulation 1 aus Kapitel 5.2 - in dem Vergleich der Diskriminationsfähigkeit der Items der ersten Testhälfte mit den Items der zweiten Testhälfte. Betrachtet man für beide Testhälften jeweils den Mittelwert der Skalierungskoeffizienten H_i als Diskriminationsmaß, so fällt auf, dass sowohl für den Analogien- als auch den Gemeinsamkeiten-Test beide Testhälften über nahezu den gleichen Mittelwert (Differenz der Mittelwerte ≈ 0.01) verfügen. Im Gegensatz dazu liegt der mittlere H_i-Koeffizient der Items der ersten Testhälfte des Satzergänzungstests ($\bar{H}_i = 0.07$) deutlicher unter dem entsprechenden Wert der zweiten Testhälfte ($\bar{H}_i = 0.15$). Unter den zehn Items mit den geringsten H_i-Koeffizienten befinden sich acht Items der ersten Testhälfte. Die erste Testhälfte weist somit weniger trennscharfe Items auf als die zweite Testhälfte. Dies erklärt[9] möglicherweise die verminderte Varianz des Summenscores der ersten Testhälfte und damit das signifikante Resultat für die Skala „Satzergänzung“.

Der parametrische Test lehnt für acht der neun Subskalen die Skalierbarkeit ab. Für den Matrizentest ergibt sich jedoch ein nicht signifikanter p-Wert. Dieses Resultat steht im starken Widerspruch zu den Ergebnissen der deskriptiven Analyse. Die Skala „Matrizen“ weist beträchtliche Verstöße gegen ein MHM auf (siehe Tabelle 6.1). Nahezu jede vierte marginale und jede dritte bedingte Korrelation fällt negativ aus. Trotz dieser starken Verletzungen lehnt λ die Nullhypothese nicht ab. Eine mögliche Erklärung bietet die geringe Streuung der H_i-Koeffizienten (Tabelle 6.1). Die Skala „Matrizen“ verfügt - verglichen mit den anderen Subskalen - über die geringsten Unterschiede im Diskriminationspotential der Items.

[9] Siehe hierzu auch die Bemerkungen zur „Verbesserung“ der Prüfgröße σ_r^2 in Kapitel 5.2.

Insgesamt zeigt sich für die Analyse des Datensatzes somit eine ähnliche Ordnung der Teststatistiken wie in den Simulationsabschnitten. Die beiden stärksten Prüfgrößen der Simulationen Y und T_{11} lehnen das Rasch-Modell für jede Subskala des IST ab. ϕ und σ_r^2 führen - verglichen mit den anderen Statistiken - zu weniger Ablehnungen. Es mag zudem auf den ersten Blick erstaunlich wirken, dass gerade die aufgrund geringerer Reliabilität und stärkeren Verstößen gegen die MHM-Annahmen psychometrisch schlechteren Skalen (verbale Skalen) weniger deutlich abgelehnt werden. Bei näherer Betrachtung zeichnet sich jedoch ab, dass der Aspekt der Reliabilität unabhängig von der Gültigkeit eines Rasch-Modells ist[10]. Des Weiteren führt die zentrale Eigenschaft des Rasch-Modells, d.h. die Austauschbarkeit der Items bezüglich ihres Beitrags zur suffizienten Statistik, zu einer (besonderen) Anfälligkeit von Teststatistiken gegenüber variierender Diskrimination. Unterschiede in den Trennschärfen stehen jedoch in keinem unmittelbaren Verhältnis zu den Annahmen eines MHM.
Mit anderen Worten: Eine Eigenschaft, die in „keiner“ Beziehung zu den Annahmen eines MHM steht, fließt maßgeblich in die Power einer Prüfgröße ein.
Die verbalen Skalen weisen trotz stärkerer Verstöße gegen MHM-Annahmen nur gering variierende H_i-Koeffizienten (nonparametrische Trennschärfen) auf. Ebenso ergeben sich bei der Analyse der invarianten Itemordnung nur äußerst geringe Verletzungen für die verbalen Skalen. Die bezüglich MHM-Diskrepanz am wenigsten auffällige Skala „Würfelaufgaben“ weist hingegen starke Verstöße gegen die invariante Itemordnung auf und die numerischen Skalen fallen besonders in Bezug auf variierende Trennschärfen auf. Es sind genau diese vier Skalen (und die Skala „Figurenauswahl“), die konsistent von allen Teststatistiken in Tabelle 6.4 abgelehnt werden.

[10]Das bereits erwähnte Beispiel einer reinen Zufallsmatrix ($\theta_v = \beta_i = 0 \quad \forall v, \forall i$) verdeutlicht dies.

6.4 Itemselektion

Nachdem die bisherigen Ergebnisse bezüglich Eindimensionalität (MHM), Reliabilität sowie Rasch-Skalierbarkeit überwiegend negativ ausfallen, soll in diesem Abschnitt der Versuch unternommen werden, durch Elimination ausgewählter Items die Skalen zu „verbessern“. Im ersten Schritt sollen weitestgehend eindimensionale[11] Skalen erzielt werden.

Faktorenanalysen stellen die gängige Methode für diesen Zweck dar. Allerdings wurde von mehreren Autoren Kritik an diesem Vorgehen geäußert. Lineare Faktorenanalysen, d.h. Faktorenanalysen angewandt auf die „normale“ Korrelationsmatrix der binären Items, können aufgrund der bereits erwähnten Restriktionen an Korrelationen zwischen dichotomen Variablen zu Artefakten führen (McDonald und Ahlawat, 1974; Sijtsma und Molenaar, 2002, Kapitel 5). Die Dimensionalitätsbestimmung ist hier konfundiert mit der Itemschwierigkeit.
Die zweite Art der Faktorenanalyse basiert daher auf einer „korrigierten“ Korrelationsmatrix („tetrachoric correlation matrix“). Auch wenn diese Methode die Mängel der linearen Faktorenanalyse behebt, so kann auch sie zu „zweifelhaften“ Ergebnissen führen. Dies tritt verstärkt auf, wenn die latenten Größen keiner Normalverteilung folgen (siehe Roussos u.a. (1998) sowie die dort angegebenen Referenzen). Hinzu kommt, dass diese Prozedur nur bei Vorlage einer bestimmten Klasse von Alternativmodellen („two-parameter normal-ogive“) theoretisch gerechtfertigt ist.
Aufgrund der angedeuteten Problematik und da diese Arbeit auf nonparametrische Methoden fokussiert ist, wird hier eine Vorgehensweise zur Erzielung eindimensionaler Skalen betrachtet, die mit nonparametrischen Größen arbeitet. Die nonparametrische Art der Dimensionalitätsbestimmung basiert auf Satz 3 (Kapitel 3.3). Unter Verwendung eines an diesen Satz angelehnten Ähnlichkeitsmaßes (bezüglich zweier Items) erfolgt die Gruppierung in eindimensionale Skalen anhand einer hierarchischen Clusteranalyse. Dieses Vorgehen bildet den ersten Schritt der Itemselektion und wurde bereits bei Roussos u.a. (1998) sowie bei Van Abswoude u.a. (2004) vielversprechend angewandt.

Während die Clusteranalyse des ersten Schrittes der Bildung MHM-konformer Skalen dient, vollzieht der zweite Schritt eine weitere Itemauswahl, die schließlich in

[11] Genauer formuliert: Die Bildung MHM-konformer Skalen ist das primäre Ziel im ersten Schritt.

ein Rasch-Modell münden soll. Die Items einer Skala aus Schritt eins besitzen aufgrund der (angenommenen) MHM-Konformität monoton wachsende Item-Response-Funktionen[12]. Für ein Rasch-Modell bedarf es jedoch eines spezifischen Verlaufs dieser Funktionen. Insbesonders sollten die Items über gleiches Diskriminationspotential verfügen. Um dies zu erreichen, wird daher im zweiten Schritt mittels eines kombinatorischen Tests Item-spezifisch geprüft, ob die Diskriminationsfähigkeit eines Items zu einem Rasch-Modell „passt“. Da jedes Item einen seperaten Test generiert, findet eine Korrektur des Signifikanzniveaus mit Hilfe der am Ende von Kapitel 3.3 dargestellten Methoden zum multiplen Testen statt.
Im Folgenden werden die beiden skizzierten Schritte näher erläutert.

Schritt 1: MHM-Konformität

Die Formung der Subskalen basiert auf einer agglomerativ-hierarchischen Clusteranalyse. Das zugrundeliegende Distanzmaß orientiert sich dabei an Satz 3. Wie bereits in der deskriptiven Analyse geschehen, wird von der Tatsache Gebrauch gemacht, dass negative marginale Korrelationen in Konflikt zu einem MHM stehen. Es erscheint somit rational, dass eine Clusteranalyse positiv korrelierende Items zu gemeinsamen Objekten (d.h. Skalen) gruppiert. Itempaare mit negativer Korrelation sollten hingegen in getrennten Clustern auftreten. Diese Überlegung legt den Korrelationskoeffizienten als Ähnlichkeitsmaß nahe. Aus bereits erwähnten Gründen bezüglich der Randrestriktionen erfolgt die Quantifizierung der Ähnlichkeit zweier Items jedoch nicht anhand des Korrelationskoeffizienten, sondern mit Hilfe des Skalierungskoeffizienten H_{ij}. Dies gewährleistet, dass die Beurteilung der Ähnlichkeit nicht mit den Itemschwierigkeiten konfundiert ist.
Somit lautet das *Ähnlichkeitsmaß* für die einelementigen Ausgangsobjekte (Items):

$$s_{ij} = s(X_i, X_j) := H_{ij}$$

Als Distanzmaß[13] dient:

$$d_{ij} = d(X_i, X_j) := -s_{ij} = -H_{ij}$$

[12]Dies ist jedenfalls die Intention des ersten Auswahlschritts. Ob dies tatsächlich erreicht wurde, kann nur durch Prüfung an einem zweiten Datensatz evaluiert werden.

[13]Streng betrachtet stellt dies kein Distanzmaß dar ($d_{ij} < 0$ ist möglich). Dies ist jedoch für die folgenden Schritte unerheblich. Jede monoton fallende Transformation des Ähnlichkeitsmaßes führt zur gleichen Clusterbildung.

Nach dem ersten Schritt der Clusteranalyse wurden zwei Items mit minimaler Distanz zu einem neuen Objekt vereint. Folglich bedarf es für die weiteren Schritte eines Distanzmaßes zwischen zwei Objekten, die aus mehreren Items bestehen können. Hierfür wurde das „Complete Linkage"-Verfahren gewählt. Es besitzt nicht nur allgemein „sinnvolle" Eigenschaften wie z.B. die Invarianz gegenüber monotonen Transformationen der Distanzen d_{ij} (Mardia u.a., 1979), sondern scheint gerade für den vorliegenden spezifischen Kontext (Erzielung eindimensionaler Skalen) geeignet. In einer Simulationsstudie erbrachte das „Complete-Linkage"-Verfahren deutlich bessere Ergebnisse gegenüber zwei alternativen Methoden: Von generierten mehrdimensionalen Daten konnte die korrekte Dimensionalität eher festgestellt und die Zuordnung der Items zu den Dimensionen mit geringerem Fehler durchgeführt werden als mit konkurrierenden Verfahren (Van Abswoude u.a., 2004).

Die Distanz zweier Objekte O_i und O_j im „Complete-Linkage"-Verfahren lautet[14]:

$$d(O_i, O_j) := \max(A_{ij}),\ A_{ij} := \{d(X_k, X_l) |\ X_k \in O_i, X_l \in O_j\} \tag{6.8}$$

Mittels (6.8) werden sukzessive zwei Objekte mit minimaler Distanz zu einem neuen Objekt vereint, bis ein bestimmtes Abbruchkriterium erreicht ist. Die Objektkonstellation zu diesem Zeitpunkt entspricht dann der vorzunehmenden Skalenbildung. Als Abbruchkriterium dient ein festgesetzter Cut-Off-Wert bezüglich $d(O_i, O_j)$. Überschreitet der minimal mögliche Wert $\min_{i,j} d(O_i, O_j)$ den Cut-Off ψ, so endet der Cluster-Algorithmus. Für die Wahl des Cut-Off-Wertes erscheint angesichts Satz 3 die Forderung von nichtnegativen Itemkorrelationen innerhalb eines Clusters relevant. Zwei Objekte sollten folglich nur dann vereint werden, wenn sie zu einem neuen Objekt führen, in dem alle paarweisen Itemkorrelationen nichtnegativ ausfallen. Somit lautet der Cut-Off-Wert, der aus der Forderung $H_{ij} \geq 0$ resultiert:

$$\psi = 0$$

Die Ergebnisse der Clusteranalyse sind - getrennt nach Subskala des IST 2000 R - in den Abbildungen 6.1-6.5 (siehe übernächste Seite) dargestellt. Der Cut-Off-Wert ist durch eine rote Linie markiert. Während sich für die verbalen Skalen 4-6-dimensionale Lösungen[15] ergeben, liegen die Resultate bei den psychometrisch besseren, numerischen Skalen im Bereich von 3-4 Dimensionen. Bei den figuralen Skalen bestehen

[14] Ein Objekt ist in der folgenden Definition als Menge bestehend aus Items zu betrachten.

[15] Die Lösung ist durch die kleinstmögliche Clusterbildung, die mit dem Cut-Off-Wert verträglich ist, gegeben.

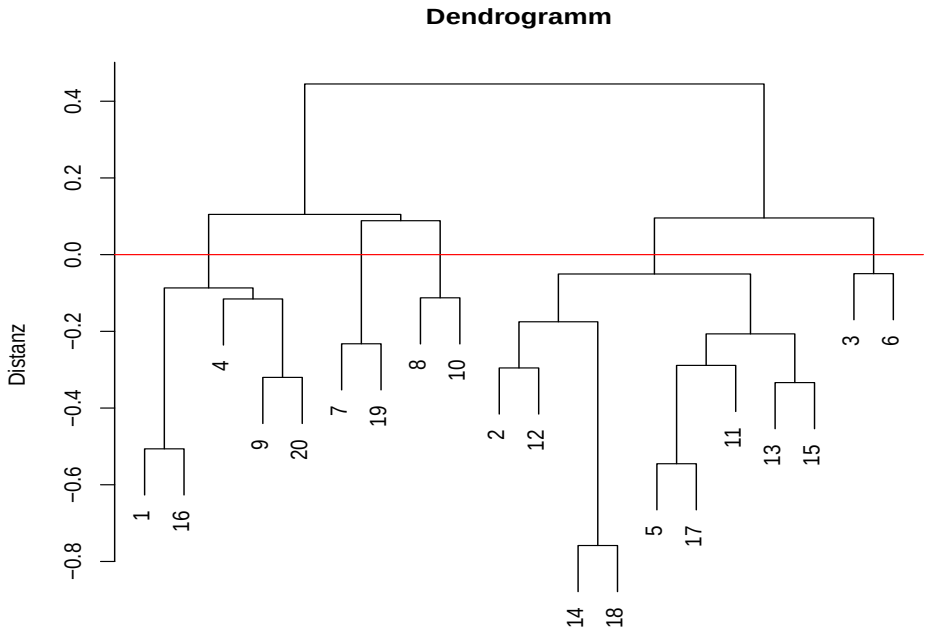
außer für den Matrizentest dreidimensionale Lösungen. Diese hohen Dimensionen stehen in Einklang mit den tendenziell starken Verstößen (Tabelle 6.1) gegen direkte Implikationen der MHM-Annahmen und führen zu relativ gering besetzten Skalenvorschlägen (Clustern).
Für manche Skalen lassen sich ferner Effekte der Reihenfolge vermuten. Am deutlichsten zeigt dies die Skala „Würfelaufgaben“. Die vorgeschlagene Dimensionalität korrespondiert hier nahezu mit der Aufteilung gemäß den Testhälften.
Insgesamt stehen die Ergebnisse bezüglich der Dimensionalität einer Skala in Einklang mit Tabelle 6.1. Für Skalen mit stärkeren Verstößen bezüglich des Vorzeichens der marginalen/bedingten Korrelationen ergeben sich sehr hochdimensionale Lösungen. Die Skalen mit - relativ betrachtet - geringeren Verletzungen führen dagegen zu dreidimensionalen Strukturen.

Im Folgenden wird ein Cluster - aus Gründen der Messgenauigkeit - nur dann einer weiteren Betrachtung unterzogen, wenn es aus mindestens acht Items besteht. Die nach dieser Forderung verbleibenden (hypothetisch) MHM-konformen Cluster (Tabelle 6.5) werden anschließend in einem zweiten Schritt (siehe nächster Abschnitt) bezüglich variierender Itemdiskrimination beurteilt.
Wie anhand Tabelle 6.5 ersichtlich, ergeben sich für die Skalen „Analogien“, „Gemeinsamkeiten“ und „Matrizen“ jeweils keine Cluster, die der minimal geforderten Testlänge genügen.

Tabelle 6.5: *Nach Schritt 1 verbleibende Cluster unter Berücksichtigung einer Mindestgröße von acht Items.*

Ursprüngliche Skala	Itemindizes der neuen Skala
Satzergänzung	2, 5, 11-15, 17, 18
Rechenaufgaben	1, 8, 10, 13-15, 17, 18, 20
Zahlenreihen	5-7, 9-12, 14-20
Rechenzeichen	5, 6, 13, 15-19
Figurenauswahl	3-8, 11, 13, 16
Würfelaufgaben	10-20

Abbildung 6.1: *Dendrogramme des Satzergänzungs- sowie des Analogientests.*



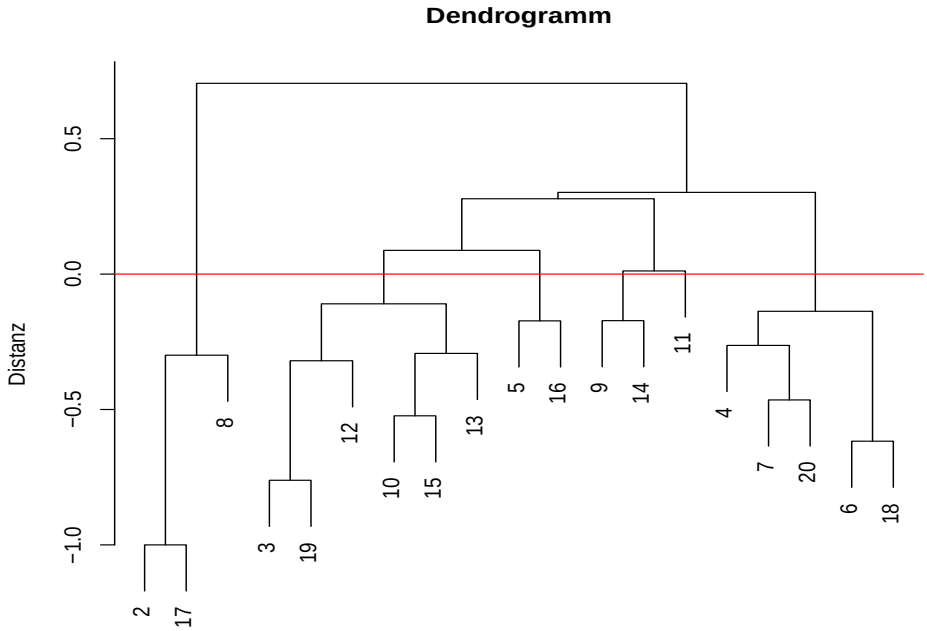
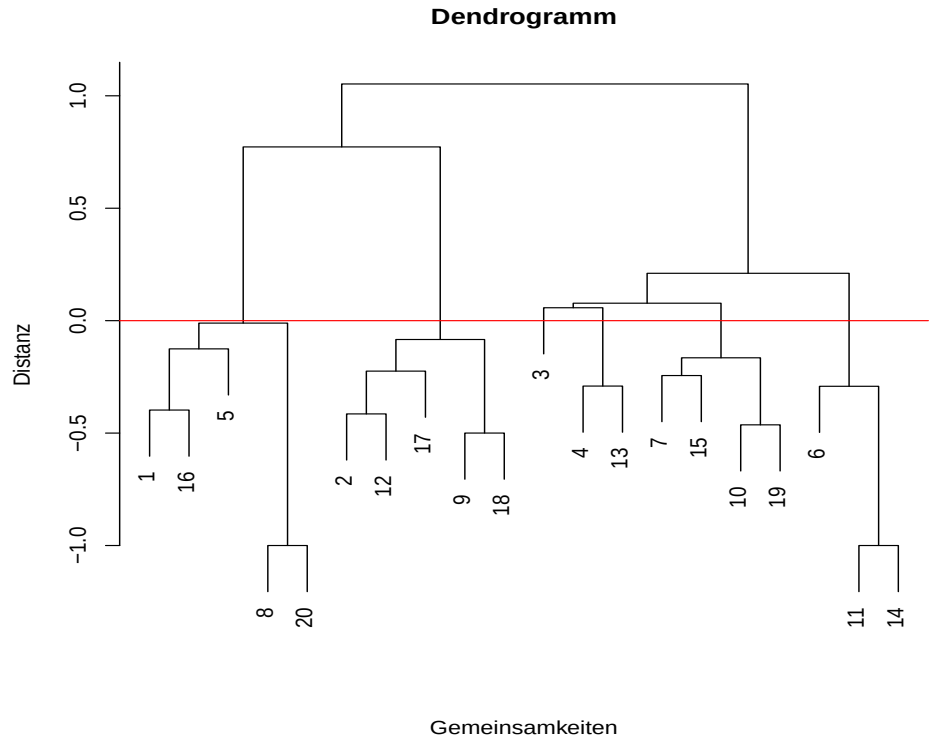


Abbildung 6.2: *Dendrogramme des Gemeinsamkeiten- sowie des Rechenaufgabentests.*



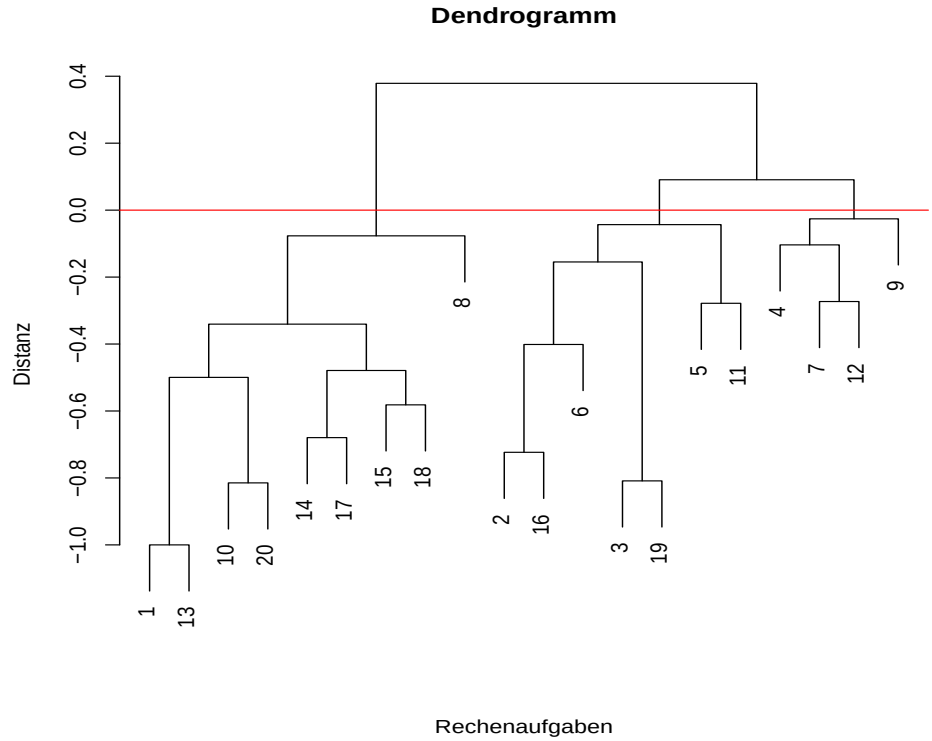


Abbildung 6.3: *Dendrogramme des Zahlenreihen- sowie des Rechenzeichentests.*

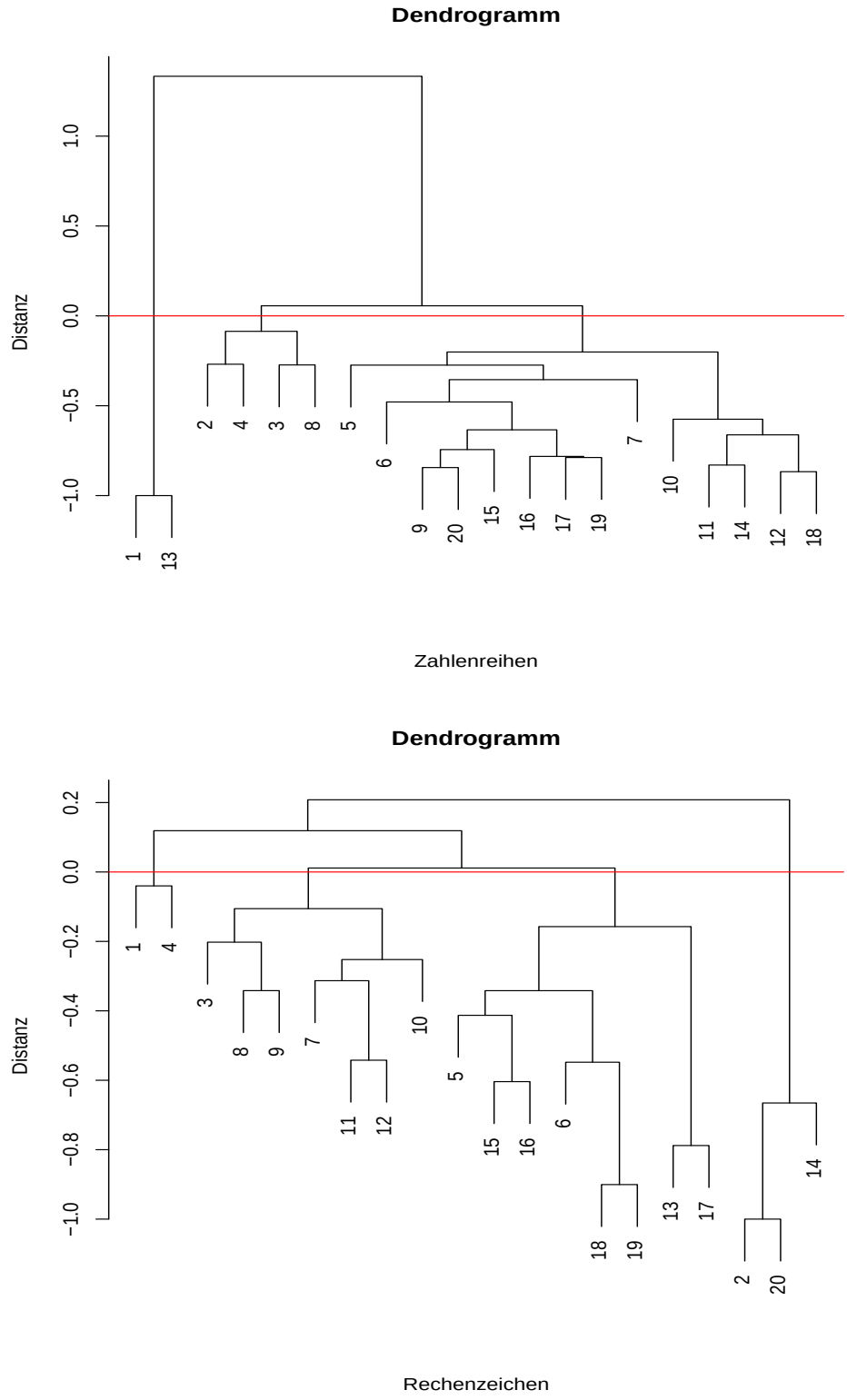
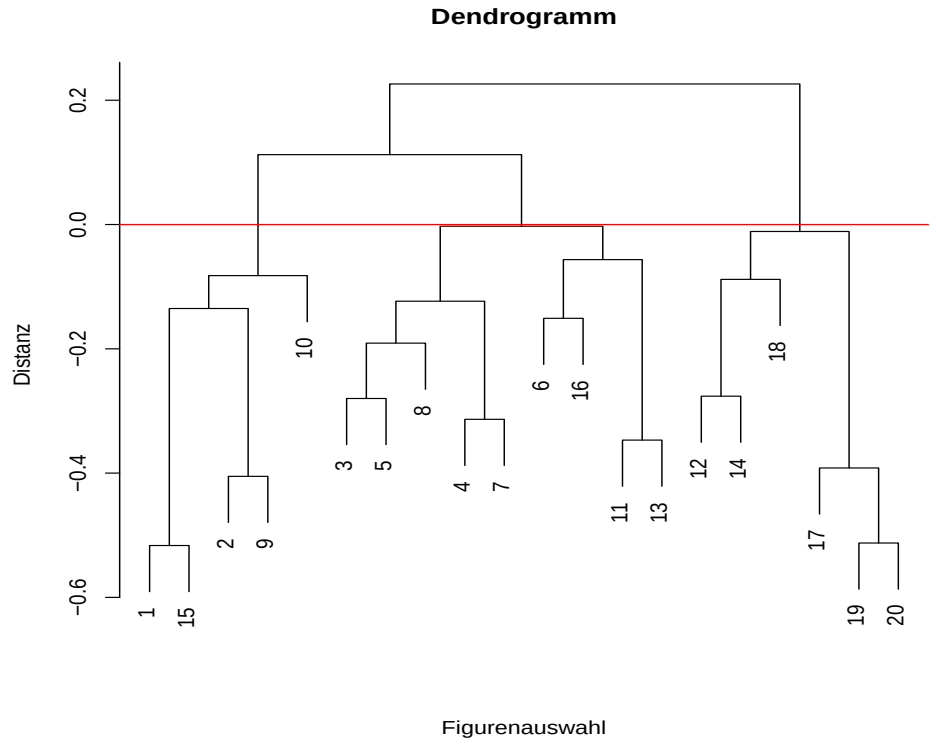


Abbildung 6.4: *Dendrogramme des Figurenauswahl- sowie des Würfelaufgabentests.*



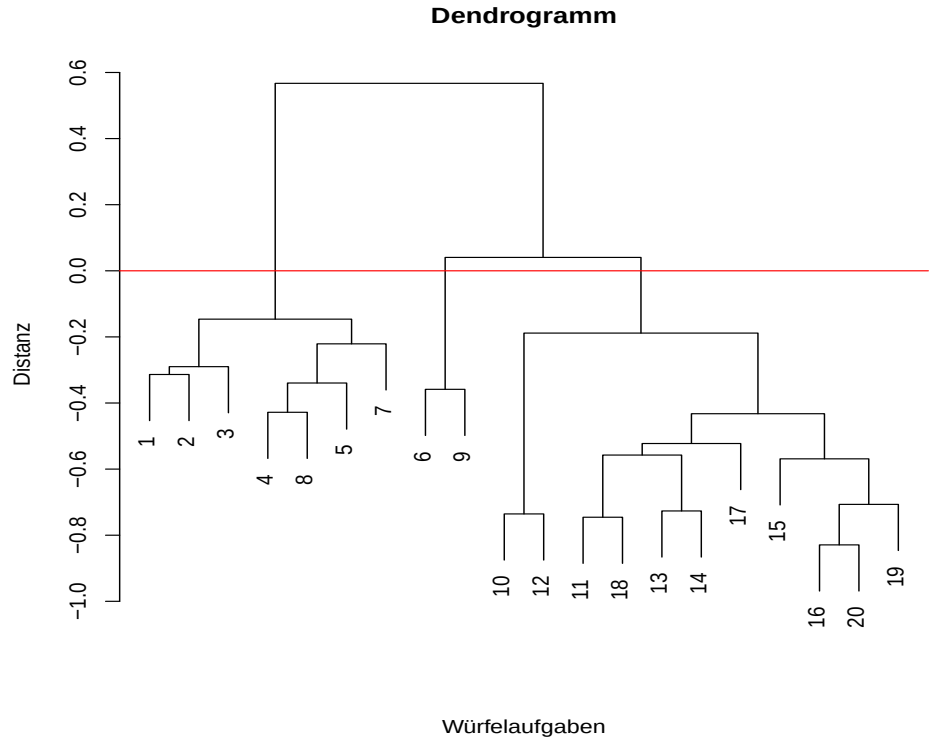
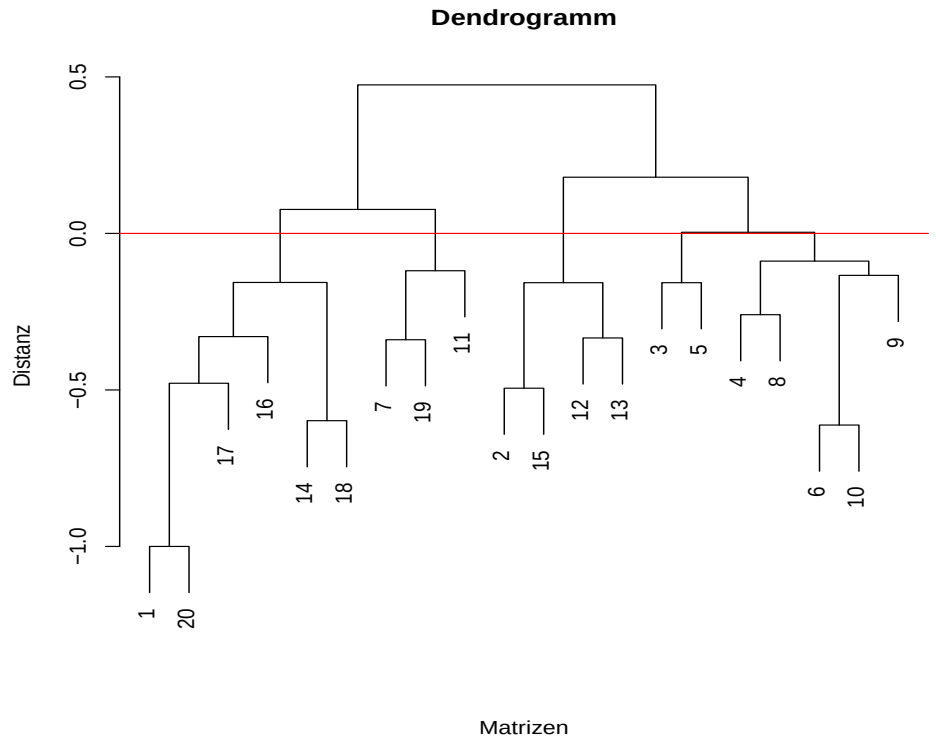


Abbildung 6.5: *Dendrogramm des Matrizentests.*



Anmerkung: Item 1 der Skala „Analogien“ ist, da es von allen Personen gelöst wurde, von der Selektionsprozedur ausgeschlossen.

Schritt 2: Testen auf variierende Itemdiskrimination

Das (idealisiert angenommene) Ergebnis des ersten Auswahlschritts sind MHM-konforme Skalen, die jedoch im monoton wachsenden Verlauf der Item-Response-Funktion beliebige Formen aufweisen können. An dieser Stelle setzt der zweite Auswahlschritt an. Er prüft die Konformität der Item-Response-Funktionen mit einem Rasch-Modell anhand eines kombinatorischen Modelltests. Dieser Test basiert auf einer Item-spezifischen Form der Prüfgröße Y. Für jedes der s ($\leq k$) Items eines Clusters aus Tabelle 6.5 wird folgende, auf dem Skalierungskoeffizienten H_i beruhende Statistik formuliert:

$$T_i(\boldsymbol{X}) := (H_i(\boldsymbol{X}) - \mu_i)^2, \qquad \mu_i := E(H_i) \tag{6.9}$$

Der p-Wert ist durch den Anteil der simulierten Matrizen mit nicht geringerem Wert der Teststatistik gegeben. Wie „üblich“ bezeichnet μ_i den gemäß der Gleichverteilung auf $\boldsymbol{\Sigma_{rc}}$ zu erwartenden Skalierungskoeffizienten des i-ten Items.
Die Gesamtheit der Prüfgrößen $T_1, \ldots, T_s$ bildet ein multiples Testproblem. Demnach erfolgt eine „Korrektur“ nach der Methode von Benjamini und Hochberg (1995). Ohne Korrektur könnte sich - gegeben den Fall eines Rasch-konformen Datensatzes - ein relativ hoher Anteil eliminierter, Rasch-homogener Items ergeben. Dies gilt es zu vermeiden und zugleich in Anbetracht des relativ geringen Stichprobenumfangs trotzdem eine gewisse Teststärke zu sichern. Die Kombination der Prüfgröße (6.9) mit der Hochberg-Korrektur scheint, wie Christensen und Kreiner (2010) zeigen, in dieser Hinsicht vielversprechend. Die Statistik (6.9) ist dem parametrischen Analogon in Form der U_i-Statistik (Molenaar, 1983) überlegen.

Nachdem für jedes der in Tabelle 6.5 definierten Cluster die obige Testprozedur angewandt wurde, erfolgte bei signifikanten Ergebnissen eine Elimination der entsprechenden Items. Die so resultierenden Cluster sind in Tabelle 6.6 aufgeführt. Sie bilden, insofern die Forderung bezüglich der Mindestgröße von acht Items erfüllt ist, den endgültigen Skalenvorschlag dieser zweischrittigen Prozedur.
Für drei Cluster („Satzergänzung“, „Rechenzeichen“ und „Figurenauswahl“) konnte die nach dem ersten Schritt vorhandene Itemauswahl beibehalten werden. Hingegen ergaben sich vier signifikante Resultate für das Cluster „Zahlenreihen“ sowie zwei signifikante Statistiken für die Skala „Würfelaufgaben“. Nach Elimination verfügen jedoch beide Cluster noch über die vorgegebene Mindestgröße. Dies gilt nicht im Fall der Skala „Rechenaufgaben“. Hier führten zwei signifikante Prüfgrößen zum Ausschluss der gesamten Skala.

Tabelle 6.6: *Nach Schritt 2 verbleibende Cluster unter Berücksichtigung einer Mindestgröße von acht Items.*

Ursprüngliche Skala	Itemindizes der endgültigen Skala
Satzergänzung	2, 5, 11-15, 17, 18
Zahlenreihen	6, 9-12, 14-17, 19
Rechenzeichen	5, 6, 13, 15-19
Figurenauswahl	3-8, 11, 13, 16
Würfelaufgaben	11-15, 17-20

Eine Betrachtung der Eigenschaften der ausgeschlossenen Items zeigt ein im Kontext der Itemselektion gemäß dem Rasch-Modell häufig zu beobachtendes Muster. Es werden (häufig) genau jene Items ausgeschlossen, die ein Extremum bezüglich der Diskriminationsfähigkeit repräsentieren. So befinden sich z.B. unter den vier ausgeschlossenen Items der Skala „Zahlenreihen“ die beiden Items mit den geringsten H_i-Koeffizienten sowie die beiden Items mit den höchsten H_i-Koeffizienten. Ähnliches gilt für die Skala „Würfelaufgaben“. Hier wurde das Item mit geringstem H_i-Wert und das am besten diskriminierende Item (höchster H_i-Wert) ausgeschlossen. Auch wenn die Elimination des am besten diskriminierenden Items aus Gründen der Messgenauigkeit absurd erscheint, so ist dies eine Notwendigkeit, um die Suffizienz des Summenscores, d.h. die „Austauschbarkeit“ der Items, zu gewährleisten.

Überprüfung der gebildeten Skalen

Ein zweiter Datensatz bestehend aus $n = 341$ Beobachtungen, die an Studenten der Ludwig-Maximilians-Universität München im Zeitraum zwischen 2005 und 2008 erhoben wurden, dient zur Prüfung der neu geformten Skalen (Tabelle 6.6) auf Rasch-Skalierbarkeit. Die Entscheidung bezüglich der Nullhypothese erfolgt dabei gemäß der T_{11}-Statistik. Für vier der fünf in Frage kommenden Skalen wird die Konformität mit einem Rasch-Modell (Tabelle 6.7) verworfen. Die Modifikation der Skala „Figurenauswahl“ hält hingegen allen Prüfgrößen stand.
Um zu klären, ob diese gute Anpassung tatsächlich auf den Effekt der Selektionsprozedur zurückzuführen ist und nicht lediglich durch die verringerte Testlänge (d.h. verminderte Teststärken) erklärt werden kann, wurden zehn zufällig gewählte neun-

Tabelle 6.7: *p-Werte der Tests auf Rasch-Skalierbarkeit für die nach Itemselektion gebildeten Skalen.*

Skala	Y	T_{11}	ϕ	σ_r^2	λ
Satzergänzung	0	0	0	0.807	0
Zahlenreihen	0	0	0	0.519	0
Rechenzeichen	0	0	0	0.466	0.001
Figurenauswahl	0.200	0.276	0.907	0.313	0.769
Würfelaufgaben	0.001	0	0.270	0.564	0.001

Tabelle 6.8: *Beurteilung der Skalenqualität anhand der marginalen Korrelationen (Cor_m), bedingten Korrelationen (Cor_b) sowie diverser Skalierungskoeffizienten.*

Skala	$\mathrm{Cor}_m < 0$ (in %)	$\mathrm{Cor}_b < 0$ (in %)	H	$\min H_i$	$\max H_i$
Satzergänzung	6	13	0.21	0.02	0.35
Zahlenreihen	0	3	0.60	0.45	0.76
Rechenzeichen	0	12	0.50	0.26	0.60
Figurenauswahl	3	12	0.16	0.11	0.25
Würfelaufgaben	0	7	0.61	0.54	0.71

elementige Teilmengen der ursprünglichen Skala „Figurenauswahl“ auf Skalierbarkeit geprüft. In neun von zehn Fällen konnte die Nullhypothese abgelehnt werden. Dies spricht für einen echten Effekt der Auswahlprozedur, auch wenn der Einfluss der verringerten Testlänge „spürbar“ ist.
Trotz dieses Resultats muss jedoch festgehalten werden, dass die Selektionsprozedur größtenteils negativ - in Hinblick auf die Bildung Rasch-konformer Skalen - verlief[16]. Bezüglich der Evaluation der MHM-Konformität (Tabelle 6.8) ergeben sich hingegen deutlich bessere Befunde. Die Verstöße bezüglich des Vorzeichens der marginalen/bedingten Korrelationen sind gegenüber den ursprünglichen Skalen (Tabelle 6.1) in einem starken Umfang verringert. Für drei der fünf Skalen liegt zudem der H-Koeffizient - als Maßstab für die Qualität der Gesamtskala (Mokken, 1971; Sijtsma und Molenaar, 2002) - in einem hohen Bereich. Die Rasch-konforme Skala „Figurenauswahl“ kann andererseits aufgrund des äußerst geringen H-Koeffizienten als wenig messgenau betrachtet werden.

Die detaillierte Ergründung der Effekte dieser Auswahlprozedur und ein potentieller Vergleich mit konkurrierenden Prozeduren würde den Rahmen dieser Arbeit sprengen. Abschließend sei daher lediglich in kurzer Form eine hierzu korrespondierende Bemerkung gemacht.
Die Skalenbildung einer anderen Selektionsmethode („backward elimination“ mit Hilfe der T_{11}-Statistik) führte - ähnlich wie die dargestellte Clustermethode - zu *einer* Rasch-konformen Skala („Analogien“). Aus den übrigen Skalen konnte keine skalier-

[16]Man beachte jedoch, dass die beiden Datensätze nicht aus Zufallsstichproben aus der gleichen Population stammen.

bare Subskala gebildet werden. Ein unangemessener Nebeneffekt dieser Auswahlmethode bestand in der Formung von Skalen, die z.T. deutliche (deskriptive) Verstöße gegen MHM-Implikationen aufwiesen (Vorzeichen der marginalen/bedingten Korrelationen). Dies kann als ein Vorteil der Clustermethode gesehen werden, da diese Methode positive Korrelationen (und damit implizit höhere Messgenauigkeiten) „garantiert“, während die „backward elimination“ aufgrund mangelnder Teststärke z.T. negative Korrelationen in der Skala „duldet“.

Insgesamt weisen die Resultate beider Selektionsprozeduren darauf hin, dass die Bildung Rasch-konformer Skalen (von einer gewissen Mindestgröße) im IST 2000 R kaum möglich ist. In den wenigen erfolgreichen Fällen („Figurenauswahl“ und „Analogien“) bestehen außerdem erhebliche Zweifel an der „Brauchbarkeit“ der entstandenen Rasch-konformen[17] Subskalen. Die H-Koeffizienten liegen in beiden Fällen deutlich unter 0.2. Dies impliziert, dass eine beträchtliche Anzahl zusätzlicher Rasch-konformer Items erforderlich wäre, um eine akzeptable Reliabilität zu gewährleisten[18]. In Kombination mit den Ergebnissen der deskriptiven Analyse aus Kapitel 6.2 lässt sich somit basierend auf den vorliegenden Daten festhalten, dass der IST 2000 R psychometrische Mängel[19] aufweist. Mit Hilfe nonparametrischer deskriptiver Methoden konnten Verstöße gegen grundlegende Implikationen eines MHM/DMM aufgezeigt werden. Der anschließende Einsatz kombinatorischer Rasch-Modell-Tests betonte ferner anhand extremer p-Werte die große Diskrepanz der Skalen zu einem Rasch-Modell. Trotz eines relativ geringen Stichprobenumfangs konnten (extreme) Signifikanzen erreicht werden. Dieses Resultat und die Fehleinschätzung der Skala „Matrizen“ durch den parametrischen Test stellen abschließend die Nützlichkeit der kombinatorischen Testklasse zur praktischen Analyse eines Datensatzes nochmals heraus.

[17] Da ein nicht signifikantes Ergebnis eines Hypothesentests nicht die Angemessenheit der Nullhypothese impliziert, ist der Ausdruck „Rasch-konform“ entsprechend relativiert zu lesen.

[18] In Populationen, die eine deutlich größere Varianz bezüglich der latenten Variable aufweisen als die (fiktive) Studentenpopulation, könnte sich dieses Problem jedoch auflösen.

[19] Man beachte jedoch die relativierenden Bemerkungen zur Stichprobe sowie zur Population aus Kapitel 6.2.

Kapitel 7

Diskussion und Ausblick

Ein großer Vorteil der kombinatorischen (bzw. nonparametrischen) Testklasse liegt in ihrer Anwendbarkeit für geringen Stichprobenumfang. Aus theoretischer Sicht ist die Einhaltung des Fehlerniveaus garantiert. Wie anhand der Simulationen zum Fehler erster Art ersichtlich, dürfte sich dieser Umstand auch in Anbetracht der praktisch erforderlichen, approximativen Monte-Carlo-Prozedur nicht ändern.
Folglich kann sich die Suche nach einer adäquaten Statistik innerhalb dieser Klasse im Wesentlichen auf die Auswahl einer geeigneten Funktion $T(\boldsymbol{X})$ beschränken. Im parametrischen Fall steht dagegen stets neben der angemessenen Formulierung einer Statistik auch die Überprüfung des Verhaltens bei endlichem bzw. geringem Stichprobenumfang im Vordergrund. Zudem können sich bei auf den MML-Kontext bezogenen Statistiken wegen der erforderlichen Angabe einer Verteilungsfunktion $G(\theta)$ Probleme bezüglich der Fehlspezifikation dieser Verteilung ergeben. Die Gültigkeit der kombinatorischen Testklasse ist dagegen - ähnlich wie bei CML-basierten parametrischen Prüfgrößen - aufgrund des Bedingens auf die suffizienten Statistiken gesichert. Unabhängig von der tatsächlich vorliegenden Verteilung $G(\theta)$ hält die kombinatorische Prüfgröße $T(\boldsymbol{X})$ das vorgegebene Fehlerniveau ein.

Für den Vergleich von nonparametrischen mit parametrischen Prüfgrößen bezüglich der Teststärke sollte, wie bereits in den Bemerkungen von Kapitel 5.6 erwähnt, immer die Abhängigkeit von dem konkret „gewählten“ Alternativmodell berücksichtigt

werden[1]. Dies gilt ebenso beim Vergleich verschiedener Prüfgrößen innerhalb der kombinatorischen Klasse. Jede der vier kombinatorischen Größen stellt sich unter spezifischen Bedingungen als „optimal“ heraus. Y besitzt bei 2PL/3PL-Daten konstant die höchste Teststärke, σ_r^2 erweist sich im Falle mehrdimensionaler Daten mit symmetrischer Aufteilung als herrausragend, ϕ ergibt bei einem geringen Anteil lokal abhängiger Itempaare die größte Power und T_{11} ist bei einem moderaten/hohen Anteil an lokal abhängigen Itempaaren „optimal“.

Global, d.h. über alle in dieser Arbeit realisierten Alternativen betrachtet, erwies sich vorrangig die T_{11}-Statistik als vielversprechend. Dies deckt sich zugleich mit der Bemerkung von Ponocny (2001), dass es sich hierbei um einen der stärksten Tests innerhalb der kombinatorischen Testklasse handele. In die gleiche Richtung weisen die Ergebnisse der Analyse des IST 2000 R. Hier führte die T_{11}-Statistik in Einklang mit deskriptiven Voranalysen zur konsistenten Ablehnung aller Skalen.

Die Ergründung der Ursachen dieser Ablehnung scheint jedoch mit den dargestellten Statistiken nur schwer möglich. Gerade die T_{11}-Statistik („second-order“-Statistik) besitzt, neben ihrer eigentlichen Intention betreffend der Erkennung lokal abhängiger Daten, beträchtliche Power gegenüber Modellen mit variierender Itemdiskrimination. Eine ähnliche Problematik[2] ergibt sich für Itempaar-zentrierte Statistiken[3] der Gestalt $T(\boldsymbol{X}) = (r_{ij} - \rho_{ij})^2$.

An dieser Stelle können nonparametrische (deskriptive) Methoden, orientiert an Kapitel 3.3, wertvolle Zusatzinformationen liefern, die es erlauben, den Misfit näher zu charakterisieren. So kann ein Verstoß bezüglich der durch Satz 3 implizierten Assoziationsrichtung unmittelbar auf eine Verletzung der MHM-Annahmen zurückgeführt werden. Variable Itemdiskrimination (in Form eines 2PL-Modells) scheidet dagegen als Ursache dieses Verstoßes aus. Analog sind Diskrepanzen zu Satz 4 direkt mit der Eigenschaft LPO verbunden und nicht mit Mehrdimensionalität zu erklären. Ähnlich gut interpretierbare Implikationen liefert die Untersuchung invarianter Itemordnung, wie sie in Kapitel 6.2 durchgeführt wurde.

Da einige dieser deskriptiven Verfahren zudem allein auf einzelnen Itempaaren bzw. Dreier-Tupeln basieren, ist es möglich, weitergehende Schlussfolgerungen aufzustel-

[1] Auch wenn diese Aussage trivial erscheint, so soll mit ihr nochmals die Koppelung der Ergebnisse dieser Arbeit an die verwendeten Alternativmodelle herausgestellt werden.

[2] Siehe die abschließende Diskussion zur Spezifität aus Kapitel 5.6.

[3] Für diesen Fall scheint ein Ausweichen auf das wesentlich spezifischer reagierende Analogon der Mantel-Haenszel-Testklasse empfehlenswert.

len. Ein Verstoß bezüglich des Vorzeichens einer bedingten Korrelation (wie z.B. in Tabelle 6.1 betrachtet) zieht nicht nur die Ablehnung eines MHM-konformen Modells für die Skala mit sich, sondern schließt gleichzeitig die Konformität für jede beliebige Skala, die jene drei Items beinhaltet, aus[4].

Diese an einzelnen Itemtupeln ausgerichtete Diagnose bildet auch die Grundlage der clusteranalytischen Methoden von Roussos u.a (1998) und Van Abswoude u.a. (2004), die der Itemselektion aus Kapitel 6.4 zugrundeliegen. Durch die Wahl von „Complete Linkage“ in Kombination mit dem Cut-Off-Wert $\psi = 0$ verfügt die Lösung über eine plausible Interpretation. Die Items eines Clusters sind wechselseitig nichtnegativ korreliert und die Cluster „maximal besetzt“. Durch eine weitergehende Zusammenlegung zweier Cluster würde angesichts negativer Itemkorrelationen eine Skala resultieren, die gegen grundlegende Implikationen eines eindimensionalen Modells (MHM) verstößt.
Als kritisch gegenüber dieser Art der Dimensionalitätsbestimmug ließen sich sowohl die uneindeutige Wahl des Cut-Offs als auch die Abhängigkeit der Ergebnisse von dem gewählten Distanzmaß anführen. Stärkere Forderungen bezüglich des Cut-Offs wie z.B. $\psi = -0.3$ münden (verglichen mit $\psi = 0$) in vielen schwach besetzten Skalen, die relativ hohe H-Koeffizienten aufweisen. Zwecks der Identifikation der korrekten Dimensionalität kann diese stärkere Forderung mitunter zu besseren Resultaten führen (Van Abswoude u.a., 2004).
Eine detaillierte Untersuchung der Auswirkungen des Cut-Offs sowie des gewählten Distanzmaßes in Kombination mit einem Vergleich bezüglich alternativer Methoden zur Dimensionalitätsbestimmung (z.B. Faktorenanalysen) wären an dieser Stelle angebracht. Ebenso von großem Interesse wären Studien, die unterschiedliche Vorgehensweisen der Itemselektion in Hinblick auf ein Rasch-Modell thematisieren. Eine solche Studie könnte diverse Möglichkeiten wie etwa eine „backward elimination“ mit Hilfe unterschiedlicher Prüfgrößen (z.B. T_{11}-Statistik) oder eine Clusteranalyse basierend auf Itempaar-zentrierten p-Werten (z.B. p-Werte bezüglich der Statistik $T(\boldsymbol{X}) = (r_{ij} - \mu_{ij})^2$) mit der zweistufigen Methodik dieser Arbeit vergleichen. Besondere Beachtung verdient hierbei auch der „Ausschuss“ dieser Methoden im Fall eines bereits zu Beginn der Itemselektion vorliegenden Rasch-Modells.

[4]Natürlich ist hierbei, insbesondere bei geringem Stichprobenumfang, der Stichprobenfehler bei der Schätzung der relevanten Größen zu berücksichtigen.

Zum Abschluss soll noch ein Aspekt thematisiert werden, der bei der stark statistischen Ausrichtung dieser Arbeit leicht unterzugehen droht:
Auch wenn in jedem Kapitel stets ein Item-Response-Modell mit latenter Variable gegenwärtig war, so wurde die Rolle der latenten Variable in einem inhaltlichen Sinn nie thematisiert. Die unbeobachtbare Variable hatte lediglich die Funktion einer mathematischen Größe, die das Zustandekommen der Itemassoziationen formalisiert.
Gleichwohl könnte im Kontext der konkreten Anwendung (siehe z.B. das Eingangsbeispiel des Intelligenztests aus Kapitel 1) diese abstrakte Größe θ mit substantieller Bedeutung „versehen“ werden. θ stellt dann nicht mehr nur eine abstrakte mathematische Größe dar, sondern steht für eine Eigenschaft wie „Intelligenz“, „Leistungsmotivation“, „Extraversion“ etc.
Diese Übertragung ist unzulässig. Mit anderen Worten: Auch die Vorlage eines Rasch-Modells (bzw. eines anderen eindimensionalen Modells) gewährleistet nicht, dass eine „reale“ Größe gemessen wird. Die mathematische Größe kann lediglich eine abstrakte Konstruktion ohne reale Manifestation darstellen:

„*One should realize the tentative nature of the latent variable. Be careful not to make the error of reification - treating an abstract construction as if it has actual existence [. . .]*“(Agresti, 2002, S.544)

Anhang A

Parameterwahl der Markov-Kette

Die in Kapitel 3.1 dargelegte Markov-Kette liefert zwar formal (d.h. für $n_{eff} \to \infty$) einen korrekten p-Wert, jedoch ist das Verhalten nach Abbruch bei einer gewissen Anzahl n_{eff} simulierter Matrizen unbekannt. Dieser Anhang soll einen Eindruck über die Präzision der Schätzung in Abhängigkeit der Länge der Markov-Kette geben. Ferner werden die Auswirkungen des „Tuning"-Parameters *step* thematisiert.
Anhand *einer* 1000×10 Datenmatrix $\boldsymbol{X}$, die in Analogie zur Simulation des Fehlers erster Art (Kapitel 5.1) aus einem Rasch-Modell mit zufällig gezogenen $N(0,1)$-verteilten Schwierigkeits- und Fähigkeitsparametern stammt, erfolgt die Evaluation des Monte-Carlo-Fehlers mittels $N = 500$ generierten Markov-Ketten. Jede dieser Markov-Ketten[1] liefert n_{eff}-Matrizen, die in eine Schätzung des p-Werts münden. Die Standardabweichung der $N = 500$ (geschätzten) p-Werte gibt einen Anhaltspunkt für die Monte-Carlo-Variation. Ferner ergibt der Durchschnitt der geschätzten p-Werte eine Schätzung des zu erwartenden p-Werts.
Diese beiden Größen (durchschnittlicher p-Wert und Standardabweichung) können dann für verschiedene Einstellungen der Parameter n_{eff} und *step* verglichen werden. n_{eff} wird auf drei Stufen ($n_{eff} = 100, 1200, 1700$) und *step* auf fünf Stufen ($step{=}2^j, j = 1, \ldots, 5$) variiert. Als Teststatistiken dienen die in Kapitel 4 dargestellten Prüfgrößen.

Betrachtet man zunächst den Mittelwert der 500 geschätzten p-Werte (Tabelle A.1), so fällt vorerst die Abhängigkeit von der gewählten Teststatistik auf. Während für

[1] Die Ausgangsmatrix jeder Markov-Kette ist stets die beobachtete Datenmatrix $\boldsymbol{X}$.

Tabelle A.1: *Durchschnittliche p-Wert-Schätzungen für verschiedene Einstellungen der „Tuning"-Parameter - getrennt nach Teststatistik - bei einer Rasch-homogenen* 1000×10 *Datenmatrix. Die Mittelwerte beruhen jeweils auf* 500 *Schätzungen.*

n_{eff}	*step*	Y	T_{11}	ϕ	σ_r^2
100	2	0.79888	0.67952	0.45346	0.55960
	4	0.83776	0.79408	0.58966	0.55574
	8	0.85256	0.83380	0.62644	0.55822
	16	0.86024	0.85082	0.65352	0.55954
	32	0.86096	0.85772	0.65880	0.55972
1200	2	0.87311	0.86970	0.67145	0.56004
	4	0.87239	0.87587	0.67704	0.56296
	8	0.87129	0.87892	0.68426	0.56266
	16	0.87321	0.87990	0.68560	0.56125
	32	0.87659	0.88041	0.68766	0.56121
1700	2	0.86908	0.87029	0.67301	0.56383
	4	0.87405	0.87568	0.68064	0.56122
	8	0.87096	0.87984	0.68673	0.56179
	16	0.87391	0.88111	0.68660	0.56187
	32	0.87140	0.87991	0.68779	0.56077

σ_r^2 die Mittelwerte kaum nach der Länge der Markov-Kette bzw. der Schrittweite variieren, zeigen sich für die anderen Prüfgrößen bei geringem n_{eff} in Kombination mit geringen bis mittleren Schrittweiten inakzeptable Werte. Ab $n_{eff} = 1200$ finden sich hingegen für alle Prüfgrößen „nur" noch Unterschiede - bei einer weiteren Erhöhung eines Parameters - von maximal 1-2%.

Der Einfluss der „Tuning"-Parameter manifestiert sich deutlicher in den Standardabweichungen. Die Standardabweichungen der 500 geschätzten p-Werte sind in Tabelle A.2 aufgeführt. Sowohl eine Vergrößerung der Schrittweite als auch eine Erhöhung der Länge n_{eff} resultieren in einer geringeren Standardabweichung.

Nimmt man ferner den durchschnittlichen p-Wert für $n_{eff} = 1700$ und $step = 32$ als wahren p-Wert (p_w) an, so lassen sich die in der Tabelle aufgeführten Standardabweichungen mit dem Standardfehler $\sigma_w := \sqrt{p_w(1-p_w)/500}$ im Fall von unabhängigen Ziehungen aus $\boldsymbol{\Sigma_{rc}}$ vergleichen.

Tabelle A.2: *Empirische Standardabweichungen der $N = 500$ p-Wert-Schätzungen für verschiedene Einstellungen der „Tuning"-Parameter - getrennt nach Teststatistik - bei einer Rasch-homogenen 1000×10 Datenmatrix.*

n_{eff}	step	Y	T_{11}	ϕ	σ_r^2
100	2	0.14352	0.15733	0.22489	0.12259
	4	0.11393	0.10492	0.17513	0.09240
	8	0.09604	0.07800	0.12907	0.06837
	16	0.08901	0.05840	0.09829	0.05804
	32	0.08595	0.05177	0.08431	0.05258
1200	2	0.06076	0.03551	0.06512	0.03838
	4	0.04738	0.02564	0.04735	0.02717
	8	0.03601	0.01801	0.03435	0.02103
	16	0.02989	0.01532	0.02675	0.01682
	32	0.02850	0.01445	0.02349	0.01565
1700	2	0.04966	0.03042	0.05686	0.03169
	4	0.03814	0.02120	0.03892	0.02260
	8	0.02889	0.01558	0.02823	0.01747
	16	0.02550	0.01293	0.02462	0.01450
	32	0.02483	0.01216	0.01993	0.01206
σ_w	σ_w	0.01497	0.01454	0.02072	0.02219

Die letzte Zeile der Tabelle A.2 beinhaltet diese Referenzwerte. Wie unmittelbar ersichtlich, fallen die Werte der Monte-Carlo-Prozedur im Regelfall höher aus. Mit entsprechend groß gewählter Länge bzw. Schrittweite lassen sich jedoch Werte im Bereich der Referenz (Unabhängige Ziehungen aus $\boldsymbol{\Sigma_{rc}}$) erzielen. Es zeigt sich allerdings auch hier eine Abhängigkeit von der gewählten Statistik. Während für T_{11} beispielsweise die Wahl von $n_{eff} = 1200$ und $step = 16$ zur Erlangung des Referenzwertes nahezu ausreichend ist, fällt die Standardabweichung für Y bei derselben Parameterwahl doppelt so hoch gegenüber dem Referenzwert aus.
Für die Simulationsstudie in Kapitel 5 wurde diese Einstellung ($n_{eff} = 1200$ und $step = 16$) als Kompromiss zwischen Rechenzeit und Präzision gewählt. Wie anhand Tabelle A.1 ersichtlich, ändert sich der zu erwartende p-Wert bei einer Erhöhung

($n_{eff} > 1200$ oder $step > 16$) der Parameter nur geringfügig. Ähnliche Resultate bezüglich der Konstanz des mittleren p-Wertes sowie der abnehmenden, den Referenzwert erreichenden Standardabweichung wurden bereits für den Fall einer 300×30-Datenmatrix von Verhelst (2008) festgestellt.
Abschließend sei noch betont, dass insbesondere im Fall extremer Dimensionen (z.B. 100×50) eine Vorsimulation, wie sie in diesem Abschnitt durchgeführt wurde, für eine adäquate Wahl der Parameter erfolgen sollte. Eine globale (d.h. für beliebige Dimensionierung der Datenmatrix und beliebige Randsummen gültige) Angemessenheit der Parameterwahl $n_{eff} = 1200$ und $step = 16$ ist nicht garantiert.
Für die Analyse eines konkreten Datensatzes erscheint es in Hinblick auf präzise Schätzungen des p-Werts zudem empfehlenswert, eine möglichst hohe Anzahl effektiver Matrizen (z.B. $n_{eff} = 5000$) zu wählen[2].
Im Fall eines „grenzwertigen" p-Werts ($p \approx \alpha$) sollte des Weiteren sichergestellt werden, dass die Ablehnung/Beibehaltung der Nullhypothese nicht auf den Monte-Carlo-Fehler zurückzuführen ist.

Bemerkung. Die systematisch geringeren p-Werte für $n_{eff} = 100$ (Tabelle A.1) könnten eine Erklärung für überhöhte α-Fehler in Simulationsstudien mit geringer Anzahl simulierter Matrizen darstellen.
Suárez-Falcón und Glas (2003) bemerkten bereits im Kontext ihrer Simulationsstudie einen überhöhten Fehler erster Art eines kombinatorischen Rasch-Modell-Tests bei einer geringen Anzahl simulierter Matrizen. Dies steht in Einklang mit dem Ergebnis aus Kapitel 5.1. Auch dort ergab sich zunächst, d.h. für $n_{eff} = 100$, keine Einhaltung des Testniveaus. In beiden Fällen konnte jedoch mit einer größeren Anzahl n_{eff} diese Problematik behoben werden.

[2]Da der *step*-Parameter einen ähnlichen Effekt wie n_{eff} aufweist, ließe sich eine geringe Varianz der Schätzung alternativ auch über eine hoch gewählte Schrittweite realisieren.

Anhang B

Äquivalente Statistiken

Die Besonderheit der kombinatorischen Testklasse liegt in dem Bedingen auf die suffizienten Modellstatistiken. Die resultierende Aussage über eine parameterfreie (Gleich-)Verteilung auf $\boldsymbol{\Sigma_{rc}}$ erlaubt eine große Flexibilität bezüglich der Formulierung einer Teststatistik $T(\boldsymbol{X})$. Aufgrund der konstanten Randsummen kann es jedoch vorkommen, dass Teststatistiken, welche augenscheinlich stark verschiedene Aussagen bergen, in Wahrheit äquivalent sind. Dies soll im Folgenden anhand zweier Beispiele erläutert werden.
Das erste Beispiel widmet sich Teststatistiken, die auf paarweisen Itemassoziationen basieren. Konkret betrachte man die Teststatistiken:

- $T_r(\boldsymbol{X}) := \mathrm{Cor}(\boldsymbol{x_{(i)}}, \boldsymbol{x_{(j)}})$
- $T_c(\boldsymbol{X}) := \mathrm{Cov}(\boldsymbol{x_{(i)}}, \boldsymbol{x_{(j)}})$
- $T_h(\boldsymbol{X}) := h_{ij} = \mathrm{Cov}(\boldsymbol{x_{(i)}}, \boldsymbol{x_{(j)}})\mathrm{Cov}^{-1}_{\max}(\boldsymbol{x_{(i)}}, \boldsymbol{x_{(j)}})$
- $T_s(\boldsymbol{X}) := s_{ij} := \boldsymbol{x}^{\boldsymbol{T}}_{\boldsymbol{(i)}}\boldsymbol{x_{(j)}}$

$\mathrm{Cov}_{\max}(\boldsymbol{x_{(i)}}, \boldsymbol{x_{(j)}})$ bezeichnet hierbei die maximal erreichbare Kovarianz für die gegebenen Randsummen der Items i und j.
Die Relationen zwischen diesen Teststatistiken werden nun im folgenden Lemma deutlich:

Lemma 5. *Bezüglich* $\boldsymbol{\Sigma_{rc}}$[1] *sind für zwei Itempaare (i, j) mit nicht trivialen*[2] *Spaltensummen* T_r*,* T_c*,* T_h *und* T_s *lineare Funktionen voneinander.*

Beweis. Die Verifikation erfolgt leicht anhand der folgenden Bemerkungen:

(1) Aus der Konstanz der Spaltensummen folgt unmittelbar die Invarianz der Itemmittelwerte.

(2) Aufgrund des binären Antwortformats gilt $x_{vi}^2 = x_{vi}$. Hieraus resultiert die Konstanz der Itemvarianz.

(3) Die maximale Kovarianz bei gegebenen Randsummen ist - per Definition - lediglich eine Funktion der Mittelwerte der betroffenen Items.

(4) Der Ausschluss trivialer Spaltensummen dient dem Zweck der Wohldefiniertheit von T_r und T_h.

□

Eine direkte Implikation des Satzes ist die Äquivalenz zweier Statistiken zum Testen auf lokale Abhängigkeit eines Itempaares. Wählt man etwa T_r als Prüfgröße, so befinden sich jene Matrizen im Ablehnbereich, die die größten T_r-Werte aufweisen. Da jedoch T_c eine lineare - und damit insbesondere monotone - Transformation von T_r darstellt, resultiert der gleiche Ablehnbereich für die Prüfgröße T_c.
Diese Äquivalenz ist nicht auf einzelne Itempaare beschränkt. Beurteilt man z.B. mit einer Prüfgröße des Typs

$$\sum_{(i,j)} \frac{(r_{ij}(\boldsymbol{X}) - \rho_{ij})^2}{\sigma_{ij}^2} \tag{B.1}$$

den globalen Fit, so ergeben sich auch hier aufgrund der Linearität identische Ablehnbereiche, falls die Korrelation r_{ij} durch den Skalierungskoeffizienten h_{ij} (mit analoger Substitution der entsprechenden Erwartungswerte), die Kovarianz oder die Anzahl simultaner Lösungen s_{ij} ersetzt wird.

[1] „Bezüglich $\boldsymbol{\Sigma_{rc}}$“ bedeutet, dass sich im Definitionsbereich der Funktionen nur Matrizen mit gleichen Randsummen befinden.

[2] Die Spaltensummen 0 und n werden als triviale Spaltensummen bezeichnet.

Verzichtet man jedoch auf eine Skalierung wie z.B. im Falle des T_{11}-Tests,

$$\sum_{(i,j)} |r_{ij}(\boldsymbol{X}) - \rho_{ij}|\,, \tag{B.2}$$

so resultieren unterschiedliche Ablehnbereiche. Unter Verwendung der linearen Beziehung zwischen r_{ij} und h_{ij} bezüglich jedes Spaltenpaares lässt sich die Teststatistik

$$\sum_{(i,j)} |h_{ij}(\boldsymbol{X}) - E(H_{ij})| \tag{B.3}$$

in eine gewichtete Form von (B.2) bringen:

$$\sum_{(i,j)} |h_{ij}(\boldsymbol{X}) - E(H_{ij})| = \sum_{(i,j)} a_{ij} |r_{ij}(\boldsymbol{X}) - \rho_{ij}|$$

Um die für ein bestimmtes Alternativmodell „bessere“ Form zu wählen, ist es erforderlich, die Koeffizienten a_{ij} näher zu betrachten. Fällt der Koeffizient für ein Itempaar, welches (stark) in eine Modellverletzung involviert ist, relativ zu Itempaaren, die nur geringfügig an einer Verletzung beteiligt sind, hoch aus, so liefert h_{ij} eine bessere Teststärke. Das relevante Itempaar wird stärker gewichtet.
Im vorliegenden Beispiel gilt $a_{ij}^{-1} = \mathrm{Cor}_{\max}(x_{(i)}, x_{(j)})$. Somit fällt das Gewicht für ein Itempaar umso höher aus, je geringer die maximal erreichbare Korrelation ist. Eine geringe (maximal erreichbare) Korrelation ergibt sich bei Itempaaren mit stark unterschiedlichen Randverteilungen (Sijtsma und Molenaar, 2002). Folglich dürfte h_{ij} gegenüber r_{ij} höhere Teststärken erbringen, wenn die Verletzung der lokalen Abhängigkeit sich vorrangig in Itempaaren mit (stark) unterschiedlicher Schwierigkeit manifestiert.

War die bisherige Diskussion auf Maße der lokalen Abhängigkeit beschränkt, so lassen sich analoge Resultate bezüglich Maßen der variablen Itemdiskrimination aufstellen. Als zu vergleichende Größen dienen:

- $T_d(\boldsymbol{X}) := d_i = \mathrm{Cor}(\boldsymbol{x_{(i)}}, \boldsymbol{r})$
- $T_g(\boldsymbol{X}) := h_i = (\sum_{j \neq i} \mathrm{Cov}(\boldsymbol{x_{(i)}}, \boldsymbol{x_{(j)}}))(\sum_{j \neq i} \mathrm{Cov}_{\max}(\boldsymbol{x_{(i)}}, \boldsymbol{x_{(j)}}))^{-1}$
- $T_{r_1}(\boldsymbol{X}) := \sum_{v:x_{vi}=1} r_v$

Für diese Prüfgrößen gilt:

Lemma 6. *Bezüglich* $\boldsymbol{\Sigma_{rc}}$ *sind - für ein beliebig gewähltes Item i -* T_d*,* T_g *und* T_{r_1} *lineare Funktionen*[3] *voneinander.*

Beweis. Die Bemerkungen des vorherigen Beweises können direkt übertragen werden. Ferner ist das Skalarprodukt $\boldsymbol{x_{(i)}^T r}$ der einzige nicht konstante Term in der Korrelation. Dieses entspricht aber genau der Summe der Gesamttestwerte jener Person, die Item i lösen und damit der Prüfgröße T_{r_1}.
Bezüglich des Skalierungskoeffizienten folgt die lineare Beziehung zu T_d aus der Linearität des Zählers zur Trennschärfe (und aus der Konstanz des Nenners). Der Zähler von T_g besitzt die Struktur:

$$\sum_{j \neq i} \mathrm{Cov}(\boldsymbol{x_{(i)}}, \boldsymbol{x_{(j)}}) = \mathrm{Cov}(\boldsymbol{x_{(i)}}, \boldsymbol{r}) - \mathrm{Cov}(\boldsymbol{x_{(i)}}, \boldsymbol{x_{(i)}})$$

Der letzte Term ist lediglich die (konstante) Varianz des i-ten Items. Da des Weiteren $\mathrm{Cov}(\boldsymbol{x_{(i)}}, \boldsymbol{r})$ eine lineare Funktion von $\mathrm{Cor}(\boldsymbol{x_{(i)}}, \boldsymbol{r})$ ist, folgt die Behauptung. □

Zusammenfassend lässt sich somit festhalten, dass gewisse Prüfgrößen bezüglich einzelner Items (bzw. Itempaare) trotz augenscheinlicher Unterschiede gleichwertig sind. Gleiches gilt für skalierte Summen des Typs (B.1). Eine Diskrepanz hingegen resultiert für unskalierte Summen der Form (B.2). Hier lohnt es sich u.U. die mit der jeweiligen Teststatistik assoziierten Gewichte näher zu betrachten, um damit zu einer verbesserten Teststärke zu gelangen.

[3] Die Wohldefiniertheit der Größen wird implizit vorausgesetzt.

Literaturverzeichnis

Adams, R.J., Wilson, M.R. und Wang, W. 1997. The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement, 21, 1-23.

Adams, R.J. und Wu, M.L. 2007. The mixed-coefficients multinomial logit model: A generalized form of the Rasch model. In Von Davier, M. und Carstensen, C.H. 2007, Multivariate and mixture distribution Rasch models - Extensions and applications, 57-75. New York: Springer.

Agresti, A. 2002. Categorical data analysis (2nd ed.). New York: Wiley.

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

Andrich, D. 1997. An hyperbolic cosine IRT model for unfolding direct responses of persons to items. In Van der Linden, W.J. und Hambleton, R.K., Handbook of modern item response theory, 399-414. New York: Springer.

Andrich, D. und Luo, G. 1993. A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. Applied Psychological Measurement, 17, 253-276.

Benjamini, Y. und Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B 57, 289-300.

Benjamini, Y. und Yekuteli, D. 2001. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics, 29, 1165-1188.

Birch, M.W. 1964. The detection of partial association I: The 2 × 2 case. J. Roy. Statist. Soc. Ser. B 26, 313-324.

Bühner, M. 2006. Einführung in die Test- und Fragebogenkonstruktion, 2., aktualisierte Auflage. München: Pearson.

Bühner, M., Ziegler, M., Krumm, S. und Schmidt-Atzert, L. 2006. Ist der I-S-T 2000 R Rasch-skalierbar? Diagnostica, 52 (3), 119-130.

Chen, Y. und Small, D. 2005. Exact tests for the Rasch model via sequential importance sampling. Psychometrika, 70, 11-30.

Christensen, K.B. und Kreiner, S. 2010. Monte Carlo tests of the Rasch model based on scalability coefficients. British Journal of Mathematical and Statistical Psychology, 63, 101-111.

De Koning, E., Sijtsma, K. und Hamers, J.H.M. 2002. Comparison of four IRT models when analyzing two tests for inductive reasoning. Applied Psychological Measurement, 26, 302-320.

De Koning, E., Sijtsma, K. und Hamers, J.H.M. 2003. Construction and validation of a test for inductive reasoning. European Journal of Psychological Assessment, 19, 24-39.

Dorans, N.J., Pommerich, M. und Holland, P.W. 2007. Linking and aligning scores and scales. New York: Springer.

Douglas, J., Kim, H.R., Habing, B. und Gao, F. 1998. Investigating local dependence with conditional covariance functions. Journal of Educational and Behavioral Statistics, 23, 129-151.

Esary, J.D., Proschan, F. und Walkup, D.W. 1967. Association of random variables, with applications. Annals of Mathematical Statistics, 38, 1466-1474.

Fidalgo, A.M. und Madeira, J.M. 2008. Generalized Mantel-Haenszel methods for differential item functioning detection. Educational and Psychological Measurement, 68, 940-958.

Fischer, G.H. 1995. Derivations of the Rasch model. In Fischer, G.H. und Molenaar, I.W., Rasch models: Foundations, recent developments, and applications, 15-38. New York: Springer.

Fischer, G.H. und Molenaar, I.W. 1995. Rasch models - Foundations, recent developments, and applications. New York: Springer.

Glas, C.A.W. 1988. The derivation of some tests for the Rasch model from the multinomial distribution. Psychometrika, 53, 525-546.

Glas, C.A.W. 2007. Testing generalized Rasch models. In Von Davier, M. und Carstensen, C.H. 2007, Multivariate and mixture distribution Rasch models - Extensions and applications, 37-55. New York: Springer.

Glas, C.A.W. und Verhelst, N.D. 1989. Extensions of the partial credit model. Psychometrika, 54, 635-659.

Glas, C.A.W. und Verhelst, N.D. 1995. Testing the Rasch model. In Fischer, G.H. und Molenaar, I.W., Rasch models: Foundations, recent developments, and applications, 69-96. New York: Springer.

Grayson, D.A. 1988. Two group classification in latent trait theory: Scores with monotone likelihood ratio. Psychometrika, 53, 383-392.

Grimmett, G.R. und Stirzaker, D.R. 2001a. Probability and random processes (3rd ed.). Oxford: Oxford University Press.

Grimmett, G.R. und Stirzaker, D.R. 2001b. One thousand exercises in probability. Oxford: Oxford University Press.

Gustafsson, J.E. 1977. The Rasch model of dichotomous items: Theory, applications and a computer program. Göteborg: Institute of Education, University of Göteborg.

Holland, P.W. 1981. When are item response models consistent with observed data? Psychometrika, 46, 79-92.

Holland, P.W. und Hoskens, M. 2003. Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. Psychometrika, 68, 123-149.

Holland, P.W. und Rosenbaum, P.R. 1986. Conditional association and unidimensionality in monotone latent variable models. The Annals of Statistics, 14, 1523-1543.

Holland, P.W und Wainer H. 1993. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum.

Ip, E.H. 2009. Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. British Journal of Mathematical and Statistical Psychology (in press).

Karlin, S. und Rinott, Y. 1980. Classes of orderings of measures and related correlation inequalities, I. Multivariate totally positive distributions. J. Multivariate Anal. 10, 467-498.

Klauer, K.C. 1995. The assessment of person fit. In Fischer, G.H. und Molenaar I.W., Rasch models: Foundations, recent developments, and applications, 97-110. New York: Springer.

Kubinger, K.D. und Draxler, C. 2007. Probleme bei der Testkonstruktion nach dem Rasch-Modell. Diagnostica, 53, 131-143.

Landis, J.R., Heyman, E.R. und Koch, G.G. 1978. Average partial association in three-way contingency tables: A review and discussion of alternative tests. Internat. Statist. Rev. 46, 237-254.

Lee, S.Y. 2007. Structural equation modeling - A bayesian approach. New York: Wiley.

Lehmann, E.L. 1966. Some concepts of dependence. Ann. Math. Statist. 37, 1137-1153.

Lehmann, E.L. und Romano, J.P. 2005. Testing statistical hypotheses (3rd ed.). New York: Springer.

Loevinger, J. 1948. The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. Psychological Bulletin, 45, 507-530.

Mair, P. und Hatzinger, R. 2007. Extended Rasch modeling: The eRm package for the application of IRT models in R. Journal of Statistical Software, 20, 1-20.

Mang, J. 2009. Evaluating goodness-of-fit tests for the dichotomous Rasch model. Diplomarbeit, Ludwig-Maximilians-Universität München.

Mardia, K.V., Kent, J.T. und Bibby, J.M. 1979. Multivariate analysis. London: Academic Press.

McDonald, R.P und Ahlawat, K.S. 1974. Difficulty factors in binary data. British Journal of Mathematical and Statistical Psychology, 27, 82-99.

Meijer, R.R und Baneke J.J 2004. Analyzing psychopathology items: A case for nonparametric item response theory modelling. Psychological Methods, 9, 354-368.

Mokken, R.J. 1971. A theory and procedure of scale analysis. The Hague: Mouton/Berlin: De Gruyter.

Mokken R.J. und Lewis, C. 1982. A nonparametric approach to the analysis of dichotomous item responses. Applied Psychological Measurement, 6, 417-430.

Molenaar, I.W. 1983. Some improved diagnostics for failure in the Rasch model. Psychometrika, 48, 49-72.

Molenaar, I.W. 1995. Some background for item response theory and the Rasch model. In Fischer, G.H. und Molenaar, I.W., Rasch models: Foundations, recent developments, and applications, 3-14. New York: Springer.

Pfanzagl, J. 1994. On item parameter estimation in certain latent trait models. In Fischer, G.H. und Laming, D., Contributions to Mathematical Psychology, Psychometrics, and Methodology, 249-263. New York: Springer-Verlag.

Ponocny, I. 2001. Nonparametric goodness-of-fit tests for the Rasch model. Psychometrika, 66, 437-460.

Puri, M.L. und Sen, P.K. 1971. Nonparametric methods in multivariate analysis. New York: Wiley.

Reckase, M.D. 2009. Multidimensional item response theory. New York: Springer.

Rizopoulos, D. 2006. ltm: An R package for latent variable modeling and item response theory analyses. Journal of Statistical Software, 17(5), 1-25.

Robert, C.P. und Casella, G. 2004. Monte Carlo statistical methods (2nd ed.). New York: Springer.

Rosenbaum, P.R. 1984. Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika, 49, 425-435.

Rosenbaum, P.R. 1987. Comparing item characteristic curves. Psychometrika, 52, 217-233.

Roussos, L.A., Stout, W.F. und Marden, J.I. 1998. Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. Journal of Educational Measurement, 35, 1-30.

Schorr, A. 1995. Stand und Perspektiven diagnostischer Verfahren in der Praxis. Ergebnisse einer repräsentativen Befragung westdeutscher Psychologen. Diagnostica, 41, 3-20.

Sijtsma, K. und Molenaar, I.W. 2002. Introduction to nonparametric item response theory. Thousand Oaks, CA: Sage.

Stout, W.F. 1987. A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Suárez-Falcón, J.C. und Glas, C.A.W. 2003. Evaluation of global testing procedures for item fit to the Rasch model. British Journal of Mathematical and Statistical Psychology, 56. 127-143.

Swaminathan, H. und Gifford, J.A. 1982. Bayesian estimation in the Rasch model. Journal of Educational Statistics, 7, 175-192.

Tsutakawa, R.K. und Johnson, J.C. 1990. The effect of uncertainty of item parameter estimation on ability estimates. Psychometrika, 55, 371-390.

Van Abswoude, A.A.H., Vermunt, J.K., Hemker, B.T. und Van der Ark, L.A. 2004. Mokken scale analysis using hierarchical clustering procedures. Applied Psychological Measurement, 28, 332-354.

Van der Ark, L.A. 2007. Mokken scale analysis in R. Journal of Statistical Software, 20(11), 1-19.

Van der Linden, W.J. und Hambleton, R.K. 1997. Handbook of modern item response theory. New York: Springer.

Verguts, T. und De Boeck, P. 2000. A note on the Martin-Löf test for unidimensionality. Methods of Psychological Research Online, 5, 77-82.

Verguts, T. und De Boeck, P. 2001. Some Mantel-Haenszel tests of Rasch model assumptions. British Journal of Mathematical and Statistical Psychology, 54, 21-37.

Verhelst, N.D. 2008. An efficient MCMC algorithm to sample binary matrices with fixed marginals. Psychometrika, 73, 705-728.

Verhelst, N.D., Hatzinger, R. und Mair, P. 2007. The Rasch sampler. Journal of Statistical Software, 20(4), 1-14.

Von Davier, M. und Carstensen, C.H. 2007. Multivariate and mixture distribution Rasch models - Extensions and applications. New York: Springer.

Waldherr, K. 2001. Differential Item Functioning-Analysen mittels der Familie der Rasch-Modelle. Diss. Univ. Wien, Wien.

Verwendete Hilfsmittel

Diese Diplomarbeit wurde mit LaTeX kompiliert.
Zur Simulation sowie Datenauswertung wurde die Software R verwendet.
Die Pakete *RaschSampler* (Verhelst u.a., 2007), *eRm* (Mair und Hatzinger, 2007) und *ltm* (Rizopoulos, 2006) dienten dabei als Basis der in Kapitel 5 beschriebenen Simulationsstudie. Für die Auswertung des Datensatzes in Kapitel 6 wurde zusätzlich das Paket *mokken* (Van der Ark, 2007) benutzt.

Erklärung

Hiermit bestätige ich, Pascal Jordan, dass ich die vorliegende Diplomarbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

München, den 2. Februar 2010 .....

Pascal Jordan