Anwendung probabilistisch-testtheoretischer Modelle auf Statistikklausuren des Grundstudiums

Applying item response models to undergraduate stastics exams

Bachelorarbeit

zur Erlangung des Grades

Bachelor of Science in Volkswirtschaftslehre

Christian Westermeier

Matrikelnummer: 515934

Humboldt-Universität zu Berlin
Wirtschaftswissenschaftliche Fakultät
Lehrstuhl für Statistik

Betreuer: Dr. Sigbert Klinke
Prüfer: Prof. Dr. Wolfgang Härdle

eingereicht am: 10.12.2009

18. Januar 2010

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung

Klausuren gehören zum Alltag eines jeden Studierenden - seit der Einführung von Bachelor- und Masterstudiengängen noch mehr als je zuvor. Die Kenntnisse werden dabei größtenteils nach alter Tradition abgefragt: Der Professor überlegt sich Aufgaben, die seiner Meinung nach einen guten Überblick über die Leistungen eines jeden Prüflings geben. Die erreichte Gesamtpunktzahl gibt ihm Auskunft genug, um den Studierenden angemessen zu benoten. (So zumindest die herrschende Vorgehensweise.)

Was den Professor interessiert, ist ein ganz bestimmtes Merkmal: die Kompetenz eines Studierenden im jeweiligen Fachgebiet. Diese versucht er durch seine Klausuren abzubilden, die einzelnen Aufgaben sind seine Indikatoren. Eine Bewertung dieses Vorganges, etwa inwiefern die Aufgaben tatsächlich das Gewünschte messen, wie gut die Aufgaben zwischen guten und schlechten Studenten trennen usw., findet im Prüfungsalltag im besten Fall mittels der Methoden der klassischen Testtheorie statt.

Dabei gibt es seit den 70er Jahren alternative Möglichkeiten das Merkmal und seine jeweilige Ausprägung bei den Studierenden zu messen: Die Methoden der probabilistischen Testtheorie, im englischen Sprachraum besser bekannt als Item Response Theory. Für die internationalen Studien TIMSS (Third International Mathematics and Science Study) und PISA (Program for International Student Asessment) wurden die Skalen so konstruiert, dass sie den Annahmen der probabilistischen Testtheorie genügen - und beide Studien hatten weitreichende Folgen für die Bildungspolitik der teilnehmenden Länder. Im Unterschied zur klassischen Testtheorie arbeitet sie mit Lösungswahrscheinlichkeiten; man nimmt also an, dass mit zunehmender Ausprägung der latenten Variablen bei einem Studenten die Wahrscheinlichkeit steigt, dass er die Aufgabe löst. Im dritten Kapitel wird genauer darauf eingegangen und die Modelle von Rasch (1960), erweitert durch Birnbaum (1968), und Mokken (1982) werden vorgestellt.

Als Datengrundlage einer empirischen Analyse dienen vier Klausuren der Kurse Statistik I und II aus den Jahren 2004 und 2005, sie wurden an der Wirtschaftswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin gestellt. Alle Studierenden der Wirtschaftswissenschaften müssen diese beiden Kursen im Grundstudium abschließen, weshalb auf eine relativ große Fallzahl von jeweils über 100 Teilnehmer zurückgegriffen werden kann. Somit handelt es sich um den Versuch, die Aufgaben einzeln sowie die Klausuren als Ganzes mittels den Methoden der probabilistischen Testtheorie zu bewerten. Die Ergebnisse dieses ersten empirischen Versuches für das Fach Statistik werden im vierten Kapitel dargelegt.

Zuerst jedoch wird das zweite Kapitel einen kurzen Überblick über die Daten gewähren und die zur Analye nötige Dichotomisierung des Datensatzes erläutern.

2 Datengrundlage, deskriptive Statistiken und Dichotomisierung

2.1 Datengrundlage

In die vorliegenden Analysen fließen die Ergebnisse der Statistikklausuren des Sommersemesters 2004 und des Wintersemesters 2004/05 ein. Dabei handelt es sich um die Ergebnisse eines Jahrgangs: die Klausur zur Veranstaltung Statistik I findet regelmäßig im Sommersemester und die Klausur zu Statistik II regelmäßig im Wintersemester statt. Zu beiden Veranstaltung gab es jeweils 2 Prüfungstermine, zwischen denen die Studenten frei wählen konnten: ein Termin fand jeweils in den ersten beiden Wochen nach Ende der Vorlesungszeit statt (am 20.07.2004 für Statistik I und am 22.02.2005 für Statisk II), der zweite Prüfungstermin lag jeweils kurz vor Beginn des nächsten Semesters (am 13.10.2004 für Statistik I und am 08.04.2005 für Statistik II).

Die Aufgaben der beiden Prüfungstermine unterscheiden sich inhaltlich, wie an späterer Stelle noch genauer erläutert wird, sehr stark, weshalb in allen Analysen die Klausuren getrennt betrachtet werden. Zu jedem Prüfungstermin wurde die Klausur außerdem in zwei verschiedenen Versionen ausgehändigt, um Betrugsversuchen vorzubeugen. Dabei wurden allerdings nur Zahlen und die Reihenfolge, nicht aber die Struktur der Aufgaben verändert, sodass sich die beiden Versionen nur strukturell und nicht inhaltlich unterscheiden und als identisch betrachtet werden können. Damit stehen 4 verschiedene Klausuren zur Analyse zur Verfügung.

2.2 Deskriptive Statistiken

Um einen ersten Überblick und die Daten zu erhalten, erscheint es lohnenswert sich die Punkteverteilung der betrachteten Klausuren näher anzusehen. Tabelle 2.1

listet Minimum, Maximum, Median, Mittelwert, 25%- und 75%-Quantile sowie die Anzahl N der Studenten, die an den jeweiligen Klausurterminen teilnahmen, auf. Dabei sticht ins Auge, dass die Ergebnisse des 1. Termins für den Kurs Statistik II sich deutlich positiv von den anderen Terminen abhebt. Sowohl Mittelwert als auch Median liegen deutlich über den Werten der anderen Klausuren. Diese Klausur hatte auch die meisten Teilnehmer (N = 210).
In allen Klausuren waren maximal 50 Punkte zu erreichen.

Klausur	Statistik I Termin 1 20.07.2004	Statistik I Termin 2 12.10.2004	Statistik II Termin 1 22.02.2005	Statistik II Termin 2 08.04.2005
Minimum	0	0	4	0
25%-Quantil	17	18	23,25	13,25
Median	26	25	32	21
Mittelwert	25,33	24,95	30,89	21,04
75%-Quantil	33,25	31	38,75	29,50
Maximum	49	50	50	47
Anzahl N	176	181	210	110

Tabelle 2.1: Deskriptive Statistiken: Überblick der Punkteverteilung zu allen betrachteten Klausuren

Abbildung 2.1 verdeutlicht diese Werte anhand von Boxplots. Dies bestätigt den Eindruck der guten Leistungen für den 1. Termin von Statistik II: zwischen den Medianen des 1. und 2. Termins liegt eine Differenz von 11 Punkten vor. Ein anderes Bild zeigen die beiden Klausuren des Kurses Statistik I: Dort wurden zum 2.Termin im Durchschnitt weniger Punkte erzielt. Die interessante Frage wird mit Hinblick auf die Analysen mittels probabilistischer Testmodelle sein, wie die Schwierigkeitsgrade der betreffenden Aufgaben eingeordnet werden, also etwa ob die Aufgaben für den 1. Termin von Statistik II deutlich einfach waren, als für den 2. Termin.

2.3 Dichotomisierung des Datensatzes

Die einzelnen Klausuraufgaben galten für eine notwendige Dichotomisierung dann als gelöst, wenn sie im Sinne der statistischen Kompetenz, die durch die latenete Variable in allen Modellen geschätzt werden sollte, angemessen gelöst wurde. Dabei wurden die Korrekturen aller Klausuraufgaben überprüft und der nötige prozentuale Anteil der Punkte zur Lösung der Aufgabe so gewählt, dass Rechenfehler keine Rolle spielen; die mathematische Kompetenz, die in den Klausuren zwangsläufig mitabgefragt wird, wurde also ausgeblendet. Im Vordergrund stand also nicht die komplette

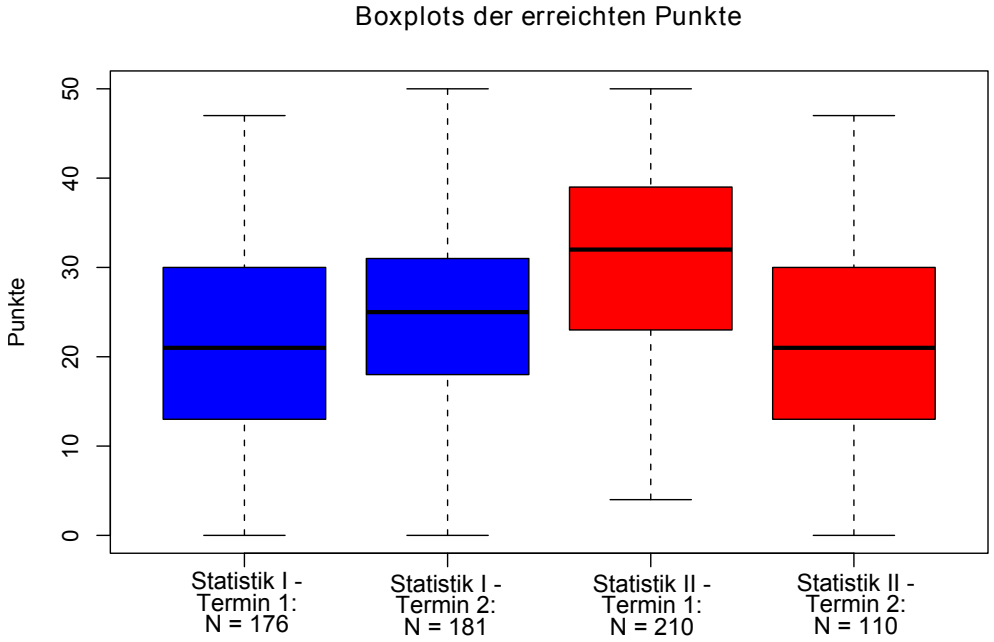

Abbildung 2.1: Boxplots der erreichten Punkte aller betrachteten Klausuren

Lösung der Aufgaben, sondern die Erfassung, wie stark die statistische Kompetenz bei den einzelnen Personen ausgeprägt ist.
Die Übersicht in Anhang A zeigt die Kriterien zur Lösung für alle Aufgaben, dabei flossen sämtliche Aufgaben der vier Termine in die Analyse ein, mit Ausnahme der jeweils letzten Aufgabe von Statistik I: Diese bestand aus jeweils fünf Wahr-Falsch-Aussagen, die jeweils unterschiedliche Themengebiete abdeckten. Die Lösung der einzelnen Teilaufgaben wurde nicht getrennt erfasst, eine Bewertung als „gelöst“ hätte somit willkürlich gewählt werden müssen, was die Validität der Modelle beeinträchtigt hätte. Anhang A ist auch als Referenz zum Thema der jeweiligen Aufgaben gedacht.

3 Analytische Grundlagen und Methoden

3.1 Probabilistische Testmodelle

Die probabilistische Testtheorie geht der Frage nach, wie man aus einer vorliegenden Datenmatrix aus Antworten bestimmter Testitems Rückschlüsse auf die interessierenden Fähigkeitsmerkmale oder Einstellungen der teilnehmenden Probanden schließen kann. Es wird also zwischen zwei Ebenen unterschieden: Zum einen die messbaren (oder manifesten) Antworten der Personen auf die gegebenen Items oder Aufgaben, zum anderen wird eine dahinterliegende latente, nicht direkt messbare Variable vermutet.
Nach Amelang und Schmidt-Atzert (2006) handle es sich bei den manifesten Variablen im Kontext der probabilistischen Testtheorie um das beobachtbare Antwortverhalten auf verschiedene Testitems, bei den latenten Variablen hingegen um nicht beobachtbare dahinterliegende Fähigkeiten oder Dispositionen, von welchen das manifeste Verhalten als abhängig angesehen werde.
Die probabilistische Testtheorie versucht nun anhand der Antworten der Probanden zu schätzen, welche Ausprägung sie auf der dahinterliegenden latenten Dimension annehmen. In allen späteren Analysen dieser Arbeit handelt es sich bei dieser latenten Variablen um die „Kompetenzen im Bereich Statistik im Grundstudium“, welche die Studierenden vor der Prüfung in den Kursen Statistik I und II erworben haben sollten.
Interessant daran ist die Tatsache, dass in den Modellen den Personen ein Fähigkeitsparameter zugewiesen wird. Abhängig von diesem Parameter lässt sich dann die Wahrscheinlichkeit bestimmen, mit der er eine Aufgabe korrekt lösen kann. Für die folgenden Ausführungen gilt:

- mit $i = 1, ..., m$ werden die Personen indiziert,
- mit $j = 1, ..., n$ werden die Aufgaben / Items indiziert,

- θ_i ist der Fähigkeitsparameter der Person i, der Vektor der Fähigkeitsparameter aller betrachteten Personen lautet dann $\Theta = (\theta_1, \theta_2, ..., \theta_m)$,
- $P(X_{ij} = 1|\theta_i)$ bezeichnet die Wahrscheinlichkeit mit der eine Person i mit dem Fähigkeitsparameter θ_i das Item j korrekt löst.
- X_{ij} kann nur die Werte 0 und 1 annehmen, in die Modelle fließen nur dichotome oder dichotomisierte Daten ein. Der Anwortvektor einer Person ist X_i, der einer Aufgabe X_j und die gesamte Datenmatrix wird mit X benannt.

Probabilistische Testmodelle unterstellen einen logistischen Zusammen zwischen Lösungswahrscheinlichkeit und Personenparametern der Form

$$P(X_{ij} = x_{ij}|\theta_i, A_j) = \frac{exp[f_j\theta_i]^{x_{ij}}}{1 - exp[f_j\theta_i]} \quad . \tag{3.1}$$

A_j sind die Parameter des Testitems j. Es werden zwei Modelle betrachet:

- Im Rasch-Modell ist $A_j = \sigma_j$. σ_j ist der Schwierigkeitsparameter für die Aufgabe j und im Rasch-Modell der einzige Parameter, der die Funktion $f_j(\theta_i)$ charakerisiert. Der Vektor aller Schwierigkeitsparameter eines Tests wird mit Σ bezeichnet.
- Im Birnbaum-Bodell ist $A_j = (\sigma_j, \lambda_j)$. Zusätzlich zum Schwierigkeitsparameter σ_j wie im Rasch-Modell wird die Funktion $f_j(\theta_j)$ durch den Trennschärfeparameter λ_j charakerisiert. Der Vektor aller Trennschärfeparameter eines Tests wird mit Λ bezeichnet.

Die folgenden Ausführungen orientieren sich im wesentlichen an dem mathematisch sehr sauberen Skript von Irtel (1995) und dem Lehrbuch von Rost (2004).

3.1.1 Lokale stochastische Unabhängigkeit

Die Annahme, dass alle im Modell enthaltenen Items die messbaren Manifestationen nur einer latenten Dimension Θ sind und die Forderung nach lokaler stochastischer Unabhängigkeit sind wesentliche Vorraussetzungen probabilistischer Testmodelle. *Lokale stochastische Unabhängigkeit* bedeutet in diesem Zusammenhang, dass die Lösungswahrscheinlichkeit eines Items durch eine Person stochastisch unabhängig von der Lösungswahrscheinlichkeit eines anderen Items durch dieselbe Person ist. Damit lässt sich die Wahrscheinlichkeit des Auftretens des Vektors X_i nun zerlegen in das Produkt der Lösungswahrscheinlichkeiten für alle einzelnen Aufgaben m:

$$P(X_i = x_i|\theta_i, \Sigma) = \prod_{j=1}^{n} P(X_i = x_i|\theta_i, \sigma_j) \tag{3.2}$$

Zusätzlich sollten natürlich, gesichert durch die Bedingungen der Testdurchführung, die Antwortvektoren aller Teilnehmer des Tests stochastisch unabhängig voneinander sein. Wie Irtel (1995) ausführlich zeigt, lässt sich damit die Wahrscheinlichkeit des Auftretens einer kompletten Datenmatrix auch als das Produkt der Wahrscheinlichkeit des Auftretens aller Zeilen schreiben:

$$P(X = x|\Theta, \Sigma) = \prod_{i=1}^{m} P(X_i = x_i|\theta_i, \Sigma) \tag{3.3}$$

$$= \prod_{i=1}^{m} \prod_{j=1}^{n} P(X_{ij} = x_{ij}|\theta_i, \sigma_j)$$

$$= \prod_{i=1}^{m} \prod_{j=1}^{n} \frac{\exp[x_{ij} f_j(\theta_i)]}{1 + \exp[f_j(\theta_i)]} \tag{3.4}$$

Die nächste Annahme des Rasch-Modells schließlich betrifft die Funktion $f(\theta_i)$: Ihre Spezifikation mittels den zwei subtraktiv verknüpften Parametern der latenten Variablen θ_i und σ_j führt zu einer logistischen Funktion der Form

$$P(X_{ij} = x_{ij}|\theta_i, \sigma_j) = \frac{\exp[x_{ij}(\theta_i - \sigma_j)]}{1 + \exp[(\theta_i - \sigma_j)]} \quad . \tag{3.5}$$

Abbildung 3.1 zeigt beispielhaft den logistischen Zusammenhang zwischen dem Schwierigkeitsparameter Θ und der Lösungswahrscheinlichkeit $P(X_{ij} = 1|\theta, \sigma)$ im Rasch-Modell. An der Stelle $P(X_{ij} = 1) = 0{,}5$ hat die Funktion ihren Wendepunkt. Für die Lösungswahrscheinlichkeit 0,5 ist für θ ein Wert von 1,0 nötig. Diese Werte, die Wendepunkte der logistischen Funktionen im Rasch-Modell, spiegeln den Schwierigkeitsparameter einer Aufgabe wider: In diesem Fall ist der Schwierigkeitsparameter σ der betrachteten Aufgabe gleich 1,0. Diese logistischen Funktionen werden im Rahmen der probabilistischen Testtheorie als itemcharakteristische Funktionen bezeichnet (im folgenden abgekürzt durch IC-Funktionen).

Birnbaum (1968) schlug eine allgemeinere Form des Zusammenhangs zwischen Lösungswahrscheinlichkeit und latenter Variablen θ vor. Er nahm die Möglichkeit, dass Items die latente Variable unterschiedlich genau messen können, in das Modell mit auf. Er fügte einen so genannten Trennschärfeparameter λ hinzu. Gleichung (3.5) ergibt sich somit zu

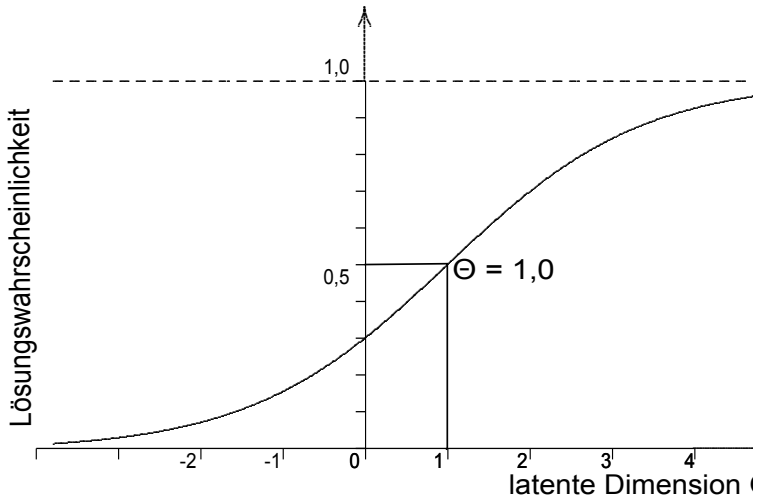

Abbildung 3.1: Beispielhafter Verlauf einer itemcharakteristischen Funktion des Rasch-Modells

$$P(X_{ij} = x_{ij}|\theta_i, \sigma_j, \lambda_j) = \frac{\exp[x_{ij}\lambda_j(\theta_i - \sigma_j)]}{1 + \exp[\lambda_j(\theta_i - \sigma_j)]} \quad . \tag{3.6}$$

Irtel (1995) zeigt, dass λ_j als Gewichtungsparameter der Aufgabe j fungiert. Während im Rasch-Modell die Distanz zwischen den Funktionsverläufen zweier IC-Funktionen der Items a und b ausreichend durch die Differenz der beiden zugehörigen Schwierigkeitsparameter $|\sigma_a - \sigma_b|$ bestimmt wird, die IC-Funktionen sich somit nicht überschneiden können, wird diese Tatsache durch die Einführung eines Trennschärfeparameters aufgeweicht. Abbildung 3.2 zeigt den möglichen Verlauf zweier IC-Funktionen in einem Birnbaum-Modell: Je größer der Trennschärfeparameter λ desto steiler Verlaufen die Kurven an ihrem Wendepunkt. Im Rasch-Modell sind genau genommen ebenso Trennschärfeparameter enthalten, die jedoch allesamt gleich 1 sein müssen. Durch einen Vergleich der Ausdrücke (3.5) und (3.6) kann man sich leicht von dieser Tatsache überzeugen.

Durch die Hinzunahme eines Trennschärfeparameters können sich die IC-Funktionen nun zwar überschneiden, dadurch kann aber auch eine bessere Anpassung an die Daten erreicht werden, da die Annahme unterschiedlicher Trennschärfen der Items in vielen Fällen realistischer ist als die Annahme identischer Funktionsverläufe (siehe Rost [2004]). Dennoch gilt wie immer, dass das einfachere Modell bei gleicher Anpassung bevorzugt werden sollte.

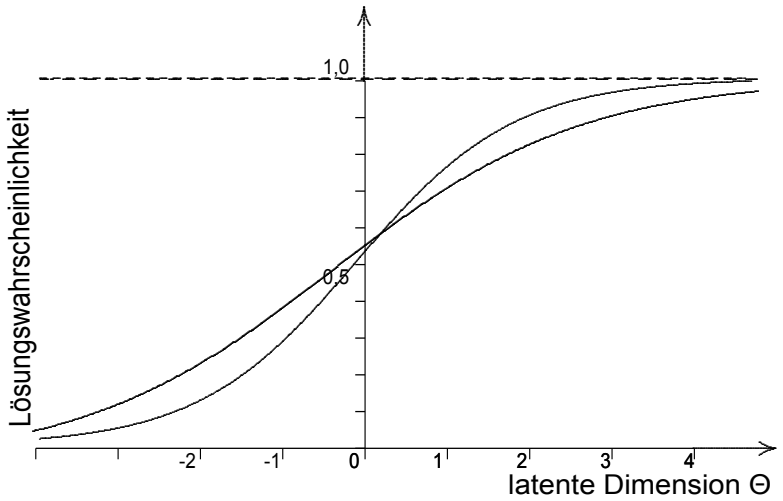

Abbildung 3.2: Beispielhafter Verlauf itemcharakteristischer Funktionen des Birnbaum-Modells

3.1.2 Ein Anpassungstest

Andersen (1973) schlug einen Anpassungstest für das Rasch-Modell vor. Dabei werden für unterschiedliche Personenteilmengen die Parameter des Rasch-Modells getrennt geschätzt und den Schätzungen der Gesamtmenge gegenübergestellt. Irtel (1995) erklärt die zugrunde liegende Teststatistik äußerst intuitiv: A sei die Gesamtheit aller Personen, eine Trennung von A in zwei disjunkte Teilmengen A_1 und A_2 mit anschließender Schätzung der Schwierigkeitsparameter, im Folgenden bezeichnet durch Σ_1 und Σ_2, führe zur Annahme, dass Σ_1 und Σ_2 jeweils optimal für die beiden Teilmengen A_1 und A_2 seien. H bezeichne die bedingte Likelihoodfunktion. Es lässt sich zeigen, dass für H bezogen auf die Teilmengen A_1 und A_2 folgende Beziehung gelten muss:

$$H(x|(r_1, r_2, ..., r_m), \Sigma) \leq H(x_{(1)}|(r_1, r_2, ..., r_m), \Sigma_1)H(x_{(2)}|(r_1, r_2, ..., r_m), \Sigma_2). \quad (3.7)$$

Wobei $x_{(1)}$ und $x_{(2)}$ die Antwortmatrizen der beiden Teilgruppen darstellen. Und r_j die Summe der korrekten Antworten einer Person i über alle Aufgaben j darstellt, die zur die bedingte Maximum-Likelihood-Schätzung verwendet werden (für eine ausführliche Darstellung der Herleitung des Maximum-Likelihood-Schätzers siehe Rost [2004]):

$$r_i = \sum_{j=1}^{n} x_{ij}$$

.

Ist das Rasch-Modell erfüllt, wird die Ungleichung (3.7) zur Gleichung, da alle Schätzungen der Schwierigkeitsparameter Σ, Σ_1 und Σ_2 folglich identisch sein müssen. Je stärker allerdings der linke Teil der Ungleichung vom rechten Teil überwogen wird, umso mehr unterscheiden sich Σ_1 und Σ_2, was gegen die Geltung eines Rasch-Modells spricht.

Die Anzahl der Items im Modell ist n, Andersen (1973) schlägt die Aufteilung der Personenmenge A in soviele Teilmengen, wie unterschiedliche Werte für r beobachtet werden. Unter Ausschluss von $r = 0$ und $r = n$ führt dies zu $n - 1$ disjunkten Personengruppen. Die Personengruppen werden nun mit k indiziert, die Anzahl ihrer korrekten Lösungen ist r_k. Gleichung (3.7) lässt sich somit verallgemeinern zu

$$H(x|(r_1, r_2, ..., r_m), \Sigma) \leq \prod_{k=1}^{n-1} H(x_{(k)}|r_k, \Sigma_k). \tag{3.8}$$

Im Falle der Gültigkeit des Rasch-Modells gilt $\Sigma = \Sigma_k$ und der Quotient:

$$\vartheta = \frac{H(x|(r_1, r_2, ..., r_m), \Sigma)}{\prod_{k=1}^{n-1} H(x_{(k)}|r_k, \Sigma_k)}$$

nimmt den Wert 1 an.
Mithilfe von ϑ lässt sich nun die Teststatistik Z bilden (siehe Irtel [1995] und Andersen [1973]):

$$Z = -2log\vartheta. \tag{3.9}$$

Bei Gültigkeit des zugrundeliegenden Rasch-Modells strebt Z gegen eine χ^2-Verteilung mit $(n-1)(n-2)$ Freiheitsgraden. Die Hypothesen lauten also:

- Nullhypothese H_0: Das Rasch-Modell ist gültig.
- Nullhypothese H_0: Das Rasch-Modell ist nicht gültig.

Im Falle der Annahme der Alternativhypothese sollte das statuierte Rasch-Modell also verworfen werden. Ist der Test nicht signifikant, kann weiterhin von der Gültigkeit des Modells ausgegangen werden.

Die empirischen Analysen der Klausurergebnisse im vierten Kapitel wurden mittels des R-Pakets 'eRm' durchgeführt (vgl. Mair und Hatzinger [2007] sowie Mair und Hatzinger [2009]), dabei wird der beschriebene Test durchgeführt, der einzige Unterschied besteht darin, dass die disjunkten Teilgruppen zufällig gewählt werden und die Anzahl der Freiheitsgrade der χ^2-verteilten Teststatistik Z sich dann aus der Anzahl der in den Teilgruppen geschätzten Parametern minus der Parameter im gesamten Datenset errechnet (siehe Mair und Hatzinger [2007]).

3.1.3 Die Testinformation

Verschiedene Items liefern auch verschiedene Informationsbeiträge über die Ausprägung des latenten Merkmals bei den betrachteten Personen. Die Steigung der IC-Funktionen ($P(X_{ij} = 1|\theta_j)$) erreicht ihr Maximum, wenn die Schwierigkeit des Items σ_j mit der Ausprägung der latenten Variablen θ_i übereinstimmt. Haben zwei Personen die Merkmalsdifferenz $|\theta_a - \theta_b|$, so sind nur im Bereich der Itemschwierigkeit deutliche Unterschiede der Wahrscheinlichkeit $P(X_{ij} = 1)$ zu erwarten, bei deutlichen Abweichungen von der Itemschwierigkeit werden die Unterschiede kleiner (Amelang und Schmidt-Atzert [2006]).

Die Untersuchung der Unterschiede der Lösungswahrscheinlichkeiten bei kleiner werdenden Differenzen des interessierenden Merkmals ergeben als Grenzfalls die Ableitung der IC-Funktion, ihre Steigungen; diese Steigung wird als Iteminformationsfunktion definiert (Amelang und Schmidt-Atzert [2006]), I_j bezeichne die Iteminformationsfunktion des Items j:

$$I_j = \frac{exp(\theta_i - \sigma_j)}{(1 - exp(\theta_i - \sigma_j))^2} \quad . \tag{3.10}$$

Die Testinformation lässt sich aufgrund der stochastischen Unabhängigkeit schließlich aus der Summe der Iteminformationsfunktionen berechnen:

$$I = \sum_{j=1}^{n} I_j \tag{3.11}$$

3.2 Mokken-Analyse

Eng verwandt mit den bereits besprochenen probabilistischen Modellen ist die so genannte Mokken-Analyse. Van der Ark (2007) hat sie in R umgesetzt, wobei er zwischen zwei Teilen einer Mokken-Analyse unterscheidet:

- Methoden um die Vorraussetzungen nicht-parametrischer Testmodelle zu überprüfen, sowie
- einen automatisierten Algorithmus, um aus einem Pool von Items die passenden für eine Mokkenskala auszuwählen.

Die Basiskonzepte einer Mokken-Analyse werden im folgenden kurz dargelegt:

3.2.1 Grundlagen der Mokken-Analyse

Das hauptsächliche Ziel einer Mokkenanalyse besteht in der Überprüfung der Annahme der Überschneidungsfreiheit der Itemfunktionen. Ihr Typ wird dabei nicht genauer spezifiziert. Rost (2004) ordnet das Mokken-Testmodell deshalb den nichtparametrischen Modellen zu: Die Itemfunktionen werden nicht als metrische Funktionen von Modellparametern spezifiziert.

Das Prinzip der doppelten Monotonie

Die wesentliche Annahme des Modells ist unter dem Begriff „doppelte Monotonie“ bekannt. Demnach sollten einerseits alle Probanden bezüglich der Lösungswahrscheinlichkeit eines Items dieselbe Ordnung aufweisen, was sich in die Annahme monoton steigender Itemfunktionen übersetzen lässt:

$$P(x_j = 1|\theta_a) \leq P(x_j = 1|\theta_b) \text{ für alle } \theta_a < \theta_b \text{ und alle Items j.}$$

Dies bedeutet, dass sobald die Wahrscheinlichkeit, dass Item j zu lösen für Proband a kleiner oder gleich der Lösungswahrscheinlichkeit des Probanden b ist, muss dies auch für alle anderen Items gelten. In einfachen Worten: Sobald Proband b eine höhere Ausprägung der latenten Variablen θ_b aufweist, muss sich dies auch in einer höheren Lösungswahrscheinlichkeit für die anderen Items einer Mokken-Skala niederschlagen, was nur erfüllt ist, wenn alle Itemfunktionen monoton ansteigen.

Andererseits beinhaltet die Annahme der doppelten Monotonie, dass alle Items bezüglich der Wahrscheinlichkeit ihrer Lösung für alle Probanden dieselbe Ordnung aufweisen:

Überschneidungsfreiheit. Wenn für einen fixen Wert θ_0 gilt $P(x_i = 1|\theta_0) \geq P(x_j = 1|\theta_0)$ dann muss auch für alle anderen Werte von θ_0 gelten: $P(x_i = 1|\theta) \geq P(x_j = 1|\theta)$ für alle Paare von Items $i \neq j$.

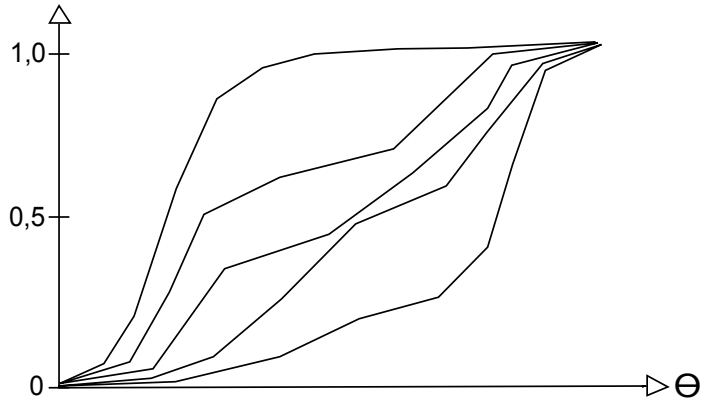

Abbildung 3.3: Beispielhafter Verlauf der Item-charakteristischen Funktionen für 5 Items einer Mokkenskala

Die Anwendung der Mokken-Analyse besteht nun in der Überprüfung dieser Grundannahmen. Abbildung 3.3 verdeutlicht eine korrekte Mokken-Skala beispielhaft für fünf Items: Die geschätzen Itemfunktionen überschneiden sich nicht und steigen monoton.

3.2.2 Berechnung von Homogenitätskoeffizienten

Die Beurteilung, ob die Annahme bezüglich der Skalen erfüllt sind erfolgt bei einer Mokken-Analyse mit Hilfe der sogenannten Homogenitätskoeffizienten. Diese geben, wie der Name bereits vermuten lässt, Auskunft über die Homogenität zweier Items oder eine kompletten Itembatterie. Nach Molenaar (1991) berechnet er sich aus der Kovarianz zwischen den Items X_i und X_j und der maximalen Kovarianz zwischen denselben Items, gegeben die Randverteilungen von X_i und X_j: $COV(X_i, X_j)^{max}$.

$$H_{ij} = \frac{COV(X_i, X_j)}{COV(X_i, X_j)^{max}} \tag{3.12}$$

Bei einer Mokken-Skala sollten alle Itempaare einen strikt positiven Homogenitätskoeffizienten aufweisen. Möchte man allerdings die Homogenität eines Items in Bezug zur restlichen Skala betrachten, muss obige Formel (3.12) verallgemeinert werden:

$$H_j = \frac{COV(X_j, R_{-j})}{COV(X_j, R_{-j})^{max}} \tag{3.13}$$

wobei $H_j (= X_+ - X_j)$ als „rest score“ bezeichnet wird. H_j ist dann die normierte Kovarianz zwischen Item j und den übrigen Items einer Skala. Sijtsma and Molenaar (2002) führen an, dass der Itemhomogenitätskoeffizient größer als eine positive un-

tere Grenze c sein sollte, wobei $c = 0,3$ als Daumenregel gilt.

Für eine ganze Itembatterie wiederum, lässt sich der Skalenhomogenitätskoeffizient berechnen:

$$H = \frac{\sum_{j=1}^{J} COV(X_j, R_{-j})}{\sum_{j=1}^{J} COV(X_j, R_{-j})^{max}} \tag{3.14}$$

Dabei sind die Koeffizienten nach Mokken (1971) wie folgt zu bewerten:

- starke Skala: $H \geq 0,5$
- angemessene Skala: $0,4 \geq H > 0,5$
- schwache Skala: $0,3 \geq H > 0.4$

Eine schwächere Skala ($H < 0,3$) sei nicht zu empfehlen.

3.2.3 Der automatisierte Itemauswahlalgorithmus

Van der Ark (2007) implementierte in sein R-Paket „mokken“ für den Endnutzer einen Algorithmus zur Itemauswahl. Dieser wählt aus einem gegebenen Set Items die passenden für eine Skala aus, welche die Kriterien einer Mokken-Skala erfüllt. Mokken (1971) definierte eine Mokken-Skala als ein Set dichotomer Items, deren Inter-Itemkovarianzen allesamt positiv sind $H_j \geq c > 0$. Bei dem implementierten Algorithmus ist der Wert von c standardmäßig auf den Wert 0,3. Dies entspricht der oben genannten Daumenregel. Für eine detailierte Beschreibung des Itemauswahlmechanismus sei auf Sijtsma und Molenaar (2002) verwiesen, dies kann im Rahmen dieser Arbeit nicht geleistet werden.

3.2.4 Überprüfung der Überschneidungsfreiheit

Zur Prüfung der Annahme der Überschneidungsfreiheit schlagen Molenaar und Sijtsma (2000) drei verschiedene Methoden vor. Intuitiv am verständlichsten ist die „rest score“ Methode, welche auch für die Analyse im vierten Kapitel Anwendung fand.

Wenn die Bedingung der lokalen Unabhängigkeit erfüllt ist, impliziert Überschneidungsfreiheit für alle Werte der latenten Variablen θ bei dichotomen Items X_i und X_j

$$P(X_i = 1|\theta) \geq P(X_j = 1|\theta), \tag{3.15}$$

was sich manifest durch

$$P(X_i = 1|W = w) \geq P(X_j = 1|W = w), \tag{3.16}$$

beobachten lassen müsste, wobei W unabhängig von X_i und X_j ist. Bei der Methode „rest score“ wird für W, wie der Name schon sagt, der „rest score“ verwendet:

$$R_{-i,-j} = X_+ - X_i - X_j. \tag{3.17}$$

Ausdruck 3.16 wird dann zu

$$P(X_i = 1|R_{-i,-j} = r) \geq P(X_j = 1|R_{-i,-j} = r) \text{ für alle } r. \tag{3.18}$$

Van der Ark (2007) kritisiert daran, dass X_+ eine inadequate Wahl ist, da es von X_i und X_j abhängt. Für alle Paare der Items wird nun die Überschneidungsfreiheit mittels der berechneten Werte für R überprüft, wobei das Paket „mokken“ nur Verletzungen der Annahme berichtet, die größer als 0,03 sind.

4 Empirische Ergebnisse

4.1 Rasch-Modell

Die Analysen des Rasch-Modells wurden ausnahmslos mittels des für die Open-Source-Software R programmierten Funktionspakets „eRm“ geschätzt. Mair und Hatzinger (2009) sind für das Paket verantwortlich und veröffentlichten auch eine detaillierte Beschreibung der implementierten Funktionen (vgl. Mair und Hatzinger [2007]).

4.1.1 Ergebnisse für Statistik I - 1. Termin

Zentral im Rasch-Modell ist die Schätzung der Schwierigkeitsparameter σ für die im Modell enthaltenen Aufgaben. Tabelle 4.1 zeigt für den ersten Termin des Kurses Statistik I (2004) die Ergebnisse der Parameterschätzungen. Zusätzlich sind der Standardfehler und ein 95%-Konfidenzintervall für alle Parameterschätzungen aufgelistet. Es sticht ins Auge, dass der Schwierigkeitsgrad der Aufgaben sich recht gleichmäßig über die Dimension der latenten Variablen verteilt. Zusätzlich sind der Standardfehler und die Grenzen des 95%-Konfidenzintervalles in der Tabelle enthalten.

Aufgabe	Schwierigkeits-paramter σ	Standardfehler	obere Grenze Konfidenzintervall	untere Grenze Konfidenzintervall
1	-1,185	0,189	-0,815	-1,555
2	-0,207	0,169	0,124	-0,537
3	1,312	0,178	1,661	0,962
4	-0,435	0,172	-0,098	-0,771
5	-1,921	0,222	-1,485	-2,357
6	0,560	0,167	0,887	0,234
7	1,875	0,199	2,265	1,486

Tabelle 4.1: Geschätzte Parameter für Rasch-Modell Statistik I - Termin 1 mit 95% Konfidenzintervall

Interessant sind an dieser Stelle vor allem die Blicke auf die als extrem einfach

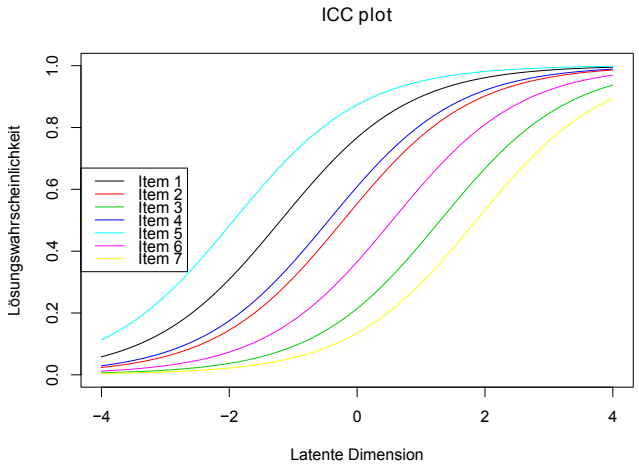

Abbildung 4.1: Item-charakteristische Kurven, Rasch-Modell, Statistik I, Termin 1

(Item 5) und extrem schwierig (Item 7) eingestuften Aufgaben. Diese Ergebnisse überraschen: Aufgabe 7 verlangte von den Studierenden eine Zuordnung von korrekten Maßzahlen für jeweils eine nominal- und metrisch-skalierte Variable. Damit handelt es sich um absolutes statistisches Grundwissen. Dies impliziert auch, dass in der zugehörigen Lehrveranstaltung mehr Zeit für die Differenzierung der Skalenniveaus verwendet werden sollte. Die 5. Aufgabe, es handelte sich um die Berechnung eines Medians von klassierten Daten weist hingegen weniger intellektuellen Tiefgang auf, da nur die richtige Formel angewendet werden musste, überrascht die Schätzung des Schwierigkeitsgrades wenig.

Modellanpassungstest:

- Wert der Teststatistik Z: 9,314
- χ^2-Freiheitsgrade: 6
- P-Wert: 0,157

Der Modellanpassungtest wie in Kapitel 3.1.2 vorgestellt, ergibt einen P-Wert von 0,157. Das heißt, die Wahrscheinlichkeit dass diese oder eine schlechtere Anpassung bei Gültigkeit des Modells erzielt wird liegt bei 15,7%. Bei einem Signifikanznivau von $\alpha = 5\%$, wie es die Konvention im allgemeinen erfordert, kann die Nullhypothese (Das geschätzte Rasch-Modell besitzt Gültigkeit) nicht zurückgewiesen werden. Mit Blick auf die hohe Teststatistik erscheint die Anpassung an die Daten aber allenfalls gerade noch befriedigend.

Die IC-Funktionen in Abbildung 4.1, wie in Kapitel 3.1.3 eingeführt, verdeutlichen noch einmal die Verteilung der geschätzten Schwierigkeitsgrade über die latente

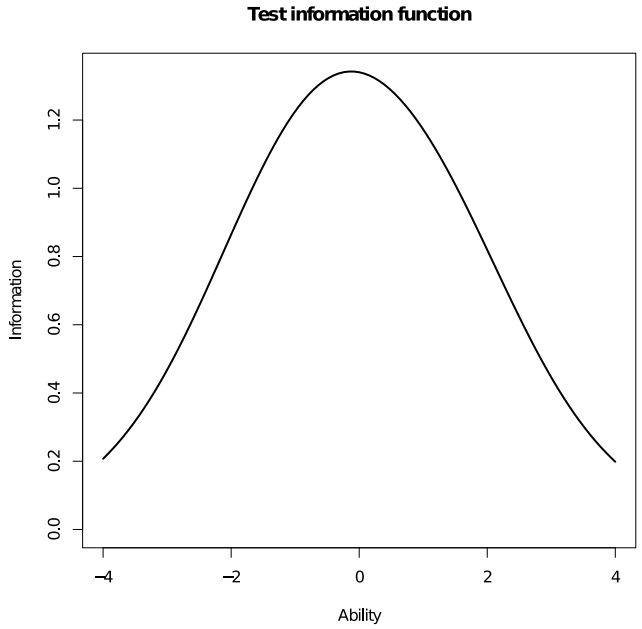

Abbildung 4.2: Testinformationsfunktion, Rasch-Modell, Statistik I, Termin 1

Dimension „Kompetenzen der statistischen Grundausbildung“. Verteilen sich die Parameter derart gleichmäßig über die Dimension Θ, ist auch für extremere Werte für die Personenparameter θ noch eine relativ genaue Schätzung möglich (siehe Amelang und Schmidt-Atzert [2006]). Dies spricht also im allgemeinen für eine gute Auswahl der Items.

Verdeutlichen kann man sich diesen Sachverhalt auch an der Testinformationsfunktion (TIF) in Abbildung 4.2: Sie zeigt ihr Maximum für mittlere Werte der latenten Dimension Θ und weist im Intervall zwischen -2 und +2 recht hohe Genauigkeit nach. Damit nimmt diese TIF die Form einer TIF eines sehr ausgewogen konstruierten Tests an.

4.1.2 Ergebnisse für Statistik I - 2. Termin

Die statistisch vorteilhafte Verteilung der Itemschwierigkeiten über die gesamte Dimension von Θ zeigt sich für den 2. Termin des Kurses Statistik I leider nicht mehr. Alle Itemschwierigkeiten werden kleiner als 0,3 geschätzt mit einer Ausnahme: Aufgabe 7 wird als extrem schwierig eingestuft. Dabei handelt es sich im die Berechnung einer Varianz eines gepoolten Datensatzes; ein Konzept das offenbar äußerst schwierig mit dem vermittelten Grundwissen zu verstehen ist. Die restlichen Items verteilen sich hingegen recht gleichmäßig über den mittleren und negativen Bereich

der latenten Dimension.

Aufgabe	Schwierigkeits-paramter σ	Standardfehler	obere Grenze Konfidenzintervall	untere Grenze Konfidenzintervall
1	-0,085	0,162	0,232	-0,402
2	-0,741	0,168	-0,412	-1,069
3	0,153	0,162	0,470	-0,165
4	-1,170	0,177	-0,822	-1,517
5	0,259	0,163	0,578	-0,061
6	-1,641	0,194	-1,260	-2,022
7 3,224	0,363	3,936	2,512	

Tabelle 4.2: Geschätzte Parameter für Rasch-Modell Statistik I - Termin 2 mit 95% Konfidenzintervall

Item 6, die Berechnung der Grenzen einer stetig gleichverteilten Variablen, wird als sehr einfach eingestuft. Diese Aufgabe lässt sich, hat man das Prinzip der Wahrscheinlichkeitsdichte erst einmal erfasst, sehr leicht lösen. Bei der stetigen Gleichverteilung handelt es sich um das einfachste der im Kurs besprochenen Verteilungsmodelle.

Modellanpassungstest:

- Wert der Teststatistik Z: 1,324
- χ^2-Freiheitsgrade: 6
- P-Wert: 0,97

Der Modellanpassungstest verspricht eine äußerst gute Anpassung an die Daten. Der äußerst kleine Wert der Teststatistik von 1,325 und der P-Wert von 0,97 deuten nicht im geringsten auf die Ungültigkeit des geschätzten Rasch-Modells hin.

Abbildung 4.3 zeigt die Verläufe der geschätzten IC-Funktionen und verdeutlicht nochmals die ungleiche Verteilung der Items.

Die Testinformationsfunktion in Abbildung 4.4 spiegelt die schiefe Messgenauigkeit des geschätzten Modells wider: Die genauesten Schätzungen der Personenparameter θ erhält man demnach für Werte die knapp unter 0 liegen. Trotzdem weist die TIF für Werte bis etwa $\theta = 2$ noch hinreichend genaue Schätzgenauigkeit nach.

4.1.3 Ergebnisse für Statistik II - 1. Termin

Die Schätzung der Schwierigkeitsparameter σ ergibt, dass sich die tendentiell eher den mittleren bis leichten Bereich der latenten Dimension erfassen. Wobei die Aufgabe 3, die Schätzung der Parameter einer Regression, als besonders leicht zu meistern

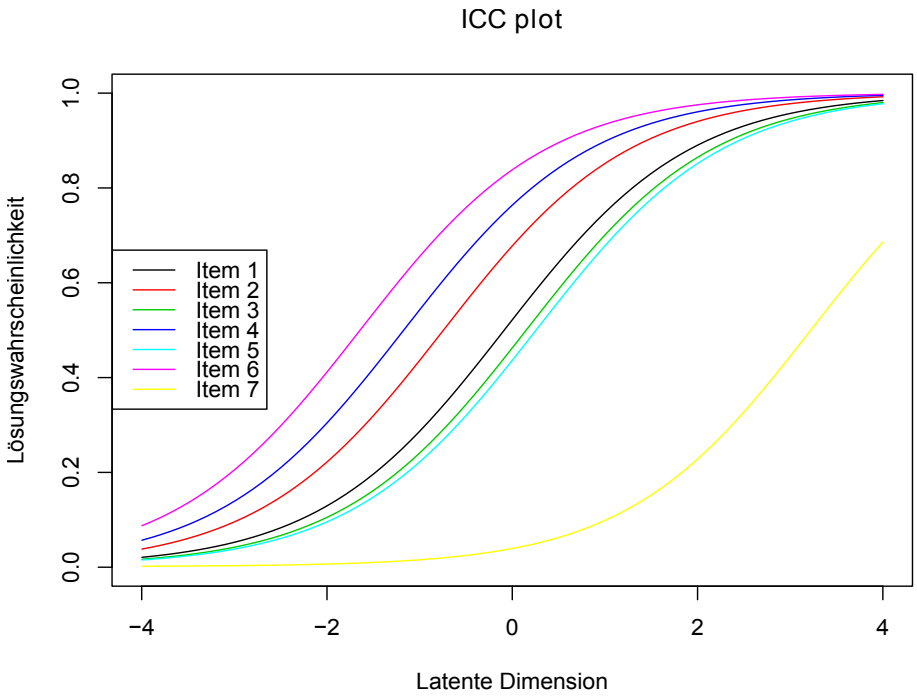

Abbildung 4.3: Item-charakteristische Kurven, Rasch-Modell, Statistik I, Termin 2

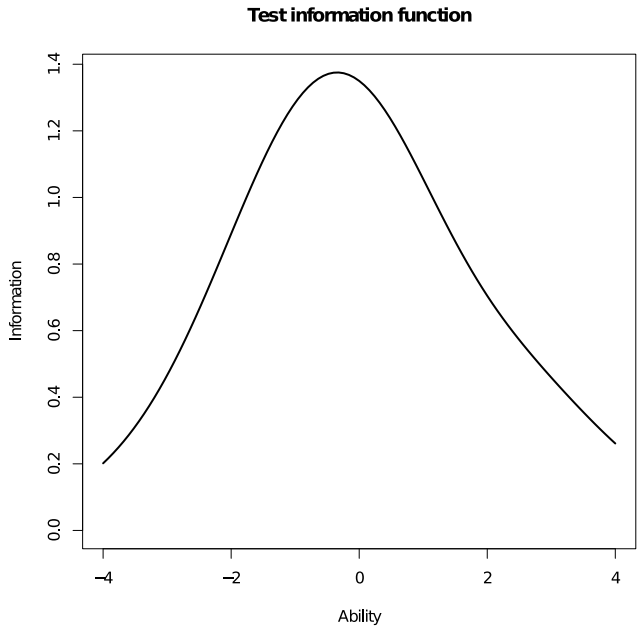

Abbildung 4.4: Testinformationsfunktion, Rasch-Modell, Statistik I, Termin 2

hervorsticht. Das schwierigste Item der Batterie ist Aufgabe 6, wobei es sich um die Durchführung eines Tests auf den Anteilswert handelte. Dies ist wenig überraschend, diese Aufgaben verlangen gute Kenntnisse in vielen Bereichen wie Teststatistiken, Approximation von Zufallsvariablen und den Umgang mit binomialverteilten Zufallsvariablen. Dabei fällt auf, dass die korrekte Formulierung der Hypothesen des Tests (Item 5) noch das geringere Problem darstellt: Dieses Konzept wird durch die Lehrveranstaltungen,Übungen und Materialien offensichtlich gut vermittelt.

Aufgabe	Schwierigkeits-paramter σ	Standardfehler	obere Grenze Konfidenzintervall	untere Grenze Konfidenzintervall
1	-0,623	0,179	-0,272	-0,975
2	0,809	0,150	1,103	0,514
3	-1,365	0,220	-0,935	-1,796
4	-0,586	0,178	-0,237	-0,935
5	0,252	0,155	0,556	-0,052
6	1,001	0,150	1,296	0,706
7	0,513	0,152	0,810	0,215

Tabelle 4.3: Geschätzte Parameter für Rasch-Modell Statistik II - Termin 1 mit 95% Konfidenzintervall

Modellanpassungstest:

- Wert der Teststatistik Z: 9,546
- χ^2-Freiheitsgrade: 6
- P-Wert: 0,145

Ähnlich, wie für das geschätzte Rasch-Modell für den ersten Termin des Kurses Statistk I kann für dieses Modell die Nullhypothese der Gültigkeit des Modells bei einem P-Wert von 0,145 und einem Signifikanzniveau von $\alpha = 5\%$ nicht zurückgewiesen werden. Die Anpassung an die Daten kann aber wiederum allenfalls als befriedigend bezeichnet werden.

Wie im Kapitel zu den deskriptiven Statistiken bereits angesprochen, handelte es sich bei dieser Klausur um die mit den besten Ergebnissen, was die erreichten Punkte betrifft. Die IC-Funktionen (Abb. 4.5) lassen darüber noch keinen genauere Einschätzung zu, die TIT allerdings (Abb. 4.6) zeigt ganz eindeutig, wie die Aufgaben nur den einfachen Bereich der latenten Dimension der statistischen Kompetenzen erfassen. Für Werte von θ, die kleiner als 0 sind, fällt die Testinformationskurve rapide ab. Die Schätzungen der Personenparameter werden also sehr ungenau.

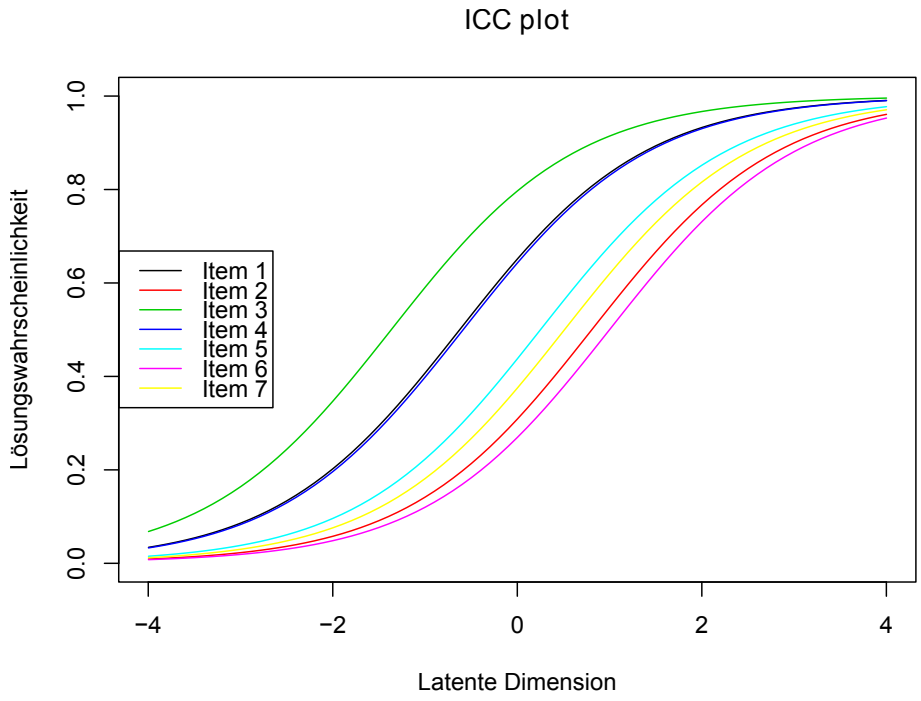

Abbildung 4.5: Item-charakteristische Kurven, Rasch-Modell, Statistik II, Termin 1

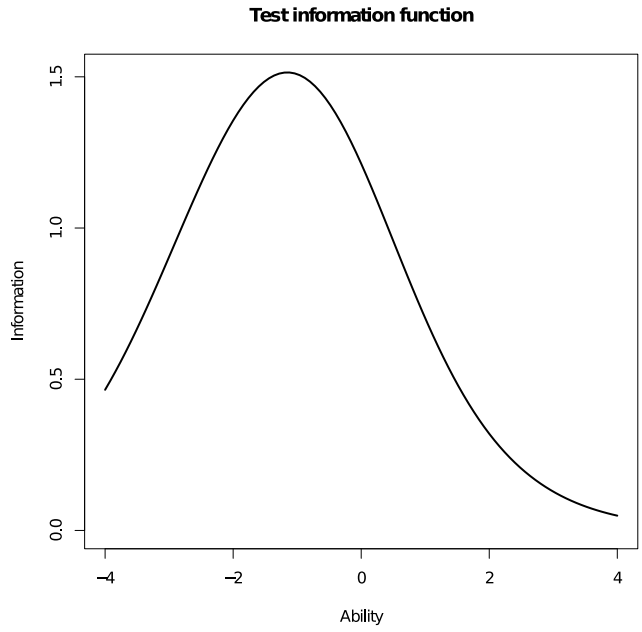

Abbildung 4.6: Testinformationsfunktion, Rasch-Modell, Statistik II, Termin 1

Dies wirft das Argument auf, auch in Hinblick der vergleichsweise guten Ergebnisse, ob für die Klausur zu leichte Items ausgewählt wurden und die Ergebnisse damit nicht ausreichend die wahren Kompetenzen der Studierenden widerspiegeln, eine ausgewogenere Itemauswahl also zu Ergebnissen vergleichbar mit den restlichen Klausuren geführt hätte.

4.1.4 Ergebnisse für Statistik II - 2. Termin

Ein Blick auf die Schätzungen für σ zeigt, dass für die Aufgaben 3 und 6 dieselben Parameter geschätzt wurden, beide Aufgaben werden als leicht zu lösen eingestuft. Bei Aufgabe 3 handelt es sich um die Berechnung einer mittleren Entwicklungsrate, eine Standardprozedur der Zeitreihenanalyse, bei Aufgabe 6 um die Berechnung eines Konfidenzintervalls für einen Mittelwert, ebenfalls ein Standardvorgang der Schätztheorie und ein ausführlich behandeltes Thema des Kurses.

Aufgabe	Schwierigkeits-paramter σ	Standardfehler	obere Grenze Konfidenzintervall	untere Grenze Konfidenzintervall
1	0,563	0,199	0,953	0,173
2	0,381	0,197	0,766	-0,005
3	-1,110	0,216	-0,687	-1,533
4	1,207	0,217	1,633	0,781
5	0,070	0,195	0,451	-0,312
6	-1,110	0,216	-0,687	-1,533

Tabelle 4.4: Geschätzte Parameter für Rasch-Modell Statistik II - Termin 2 mit 95% Konfidenzintervall

Die restlichen Aufgaben decken den mittleren Schwierigkeitsgrad der latenten Dimension sehr gut ab. Mit einem Wert von 1,207 wird Aufgabe 4 am schwierigsten eingestuft. Dabei handelte es sich um einen Zweistichprobentest auf Gleichheit der Mittelwerte, eines der komplizierteren Themen des Kurses Statistik II.

Modellanpassungstest:

- Wert der Teststatistik Z: 10,223
- χ^2-Freiheitsgrade: 5
- P-Wert: 0,069

Der Wert der Teststatistk ist für den zweiten Termin noch höher als für den ersten, der P-Wert liegt knapp über dem Signifikanzniveau von $\alpha = 5\%$, die Nullhypothese kann also nicht zurückgewiesen werden. Allerdings sollte in diesem Fall, trotz der

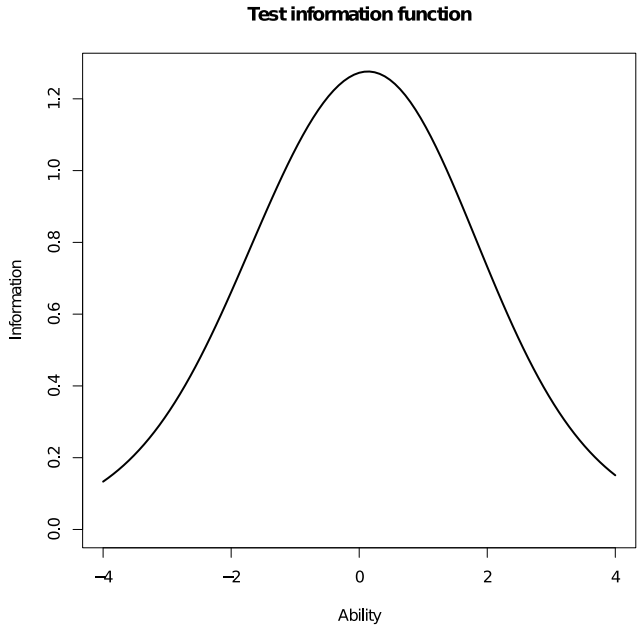

Abbildung 4.7: Testinformationsfunktion, Rasch-Modell, Statistik II, Termin 2

Tatsache, dass einfacheren Modellen bevorzugt werden sollten, geprüft werden, ob andere Modelle eine bessere Anpassung erzielen.

In Abbildung 4.8 überlagern sich die IC-Funktionen des 3. und des 6. Items, sie haben den gleichen Verlauf.

Abbildung 4.9 stellt die geschätzten Schwierigkeitsparameter des Raschmodells für die vier Klausurtermine abschließend übersichtlich dar. Die Übersicht ist selbsterklärend und sollte einen schnellen Überblick über den Schwierigkeitsgrad der behandelten Themengebiete erlauben, für die genaueren Werte sollte allerdings auf obige Tabellen zurückgegriffen werden.

4.2 Birnbaum-Modell

Für mehrparametrische Modelle hat sich das R-Paket „irtoys“, programmiert von Partchev (2009), zur Schätzung besonders bewährt. Das Paket selbst nutzt dabei das Paket „ltm“, programmiert von Rizopoulos (2009). Aus Platzgründen werden im Folgenden nur kurz die Parameterschätzungen für die vier betrachteten Klausurtermine aufgeführt sowie die zugehörigen IC-Funktionen. Besondere Berücksichtigung erfahren dabei die im Rasch-Modell als für alle Items gleich angenommenen Schätzungen der Trennschärfeparameter λ. Auch erzielte das Rasch-Modell größtenteils befriedigende Anpassungen an die Daten und die Konvention, dem einfacheren Mo-

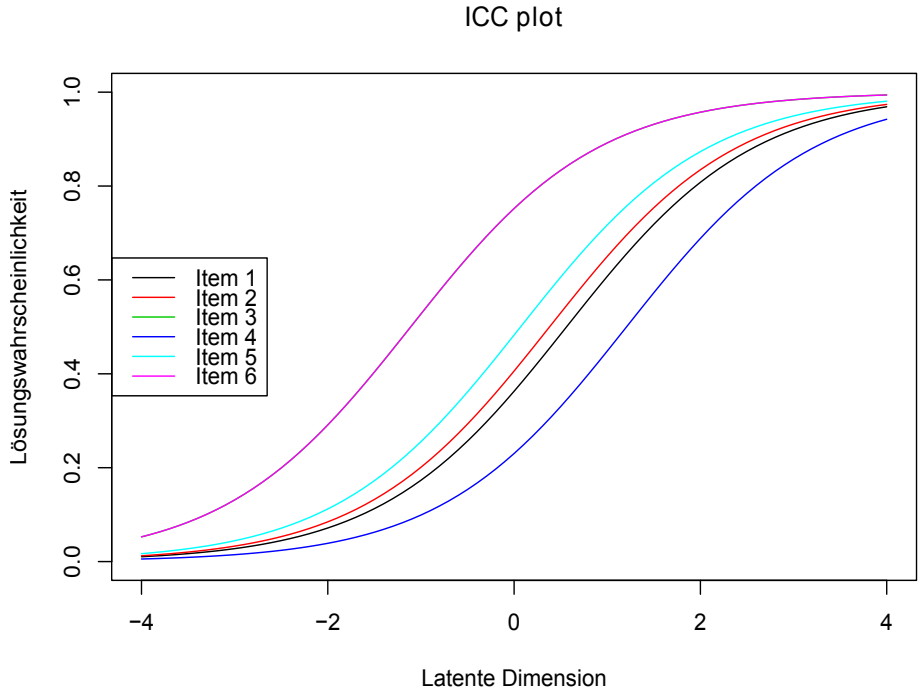

Abbildung 4.8: Item-charakteristische Kurven, Rasch-Modell, Statistik II, Termin 2

dell den Vorzug zu gewähren, erfährt ihre Würdigung.

Aufgabe	1	2	3	4	5	6	7
Schwierigkeitsparameter σ	-1,024	-0,173	0,828	-0,340	-1,058	0,531	2,044
Trennschärfeparameter λ	1,121	2,151	1,737	1,686	2,450	0,727	0,784

Tabelle 4.5: Geschätzte Parameter für Birnbaum-Modell Statistik I - Termin 1

Ein Vergleich der Schätzwerte für das einparametrische Rasch-Modell zeigt, dass sich die Schätzungen für σ kaum zwischen den beiden Modellen unterscheiden. Die Analyse der Trennschärfen λ zeigt allerdings, dass jeweils zwei Items besonders gut bzw. schlecht zwischen den Fähigkeiten diskriminieren können. Die Items 2 und 5, Berechnung von Wahrscheinlichkeiten einer exponential-verteilten Zufallsvariablen und Berechnung des Medians für klassierte Daten, trennen offenbar sehr gut zwischen guten und schlechten Studenten. In Abbildung 4.10 lässt sich diese Tatsache an der Steigung der IC-Funktion an der Stelle $\theta = \sigma$ ablesen, die Funktionen sind dann besonders steil. Für die Items 6 und 7, Aufstellen einer Likelihoodfunktion und Wahl der korrekten Maßzahlen, bescheinigt das Modell äußert geringe Trennschärfe.

Für den zweiten Termin des Kurses Statistik I zeigt sich, dass die Aufgabe 7 besonders hohe Trennschärfe besitzt: Dabei handelt es sich auch um die schwierigste Aufgabe der Klausur, die Berechnung der Varianz eines gepoolten Datensatzes. Die

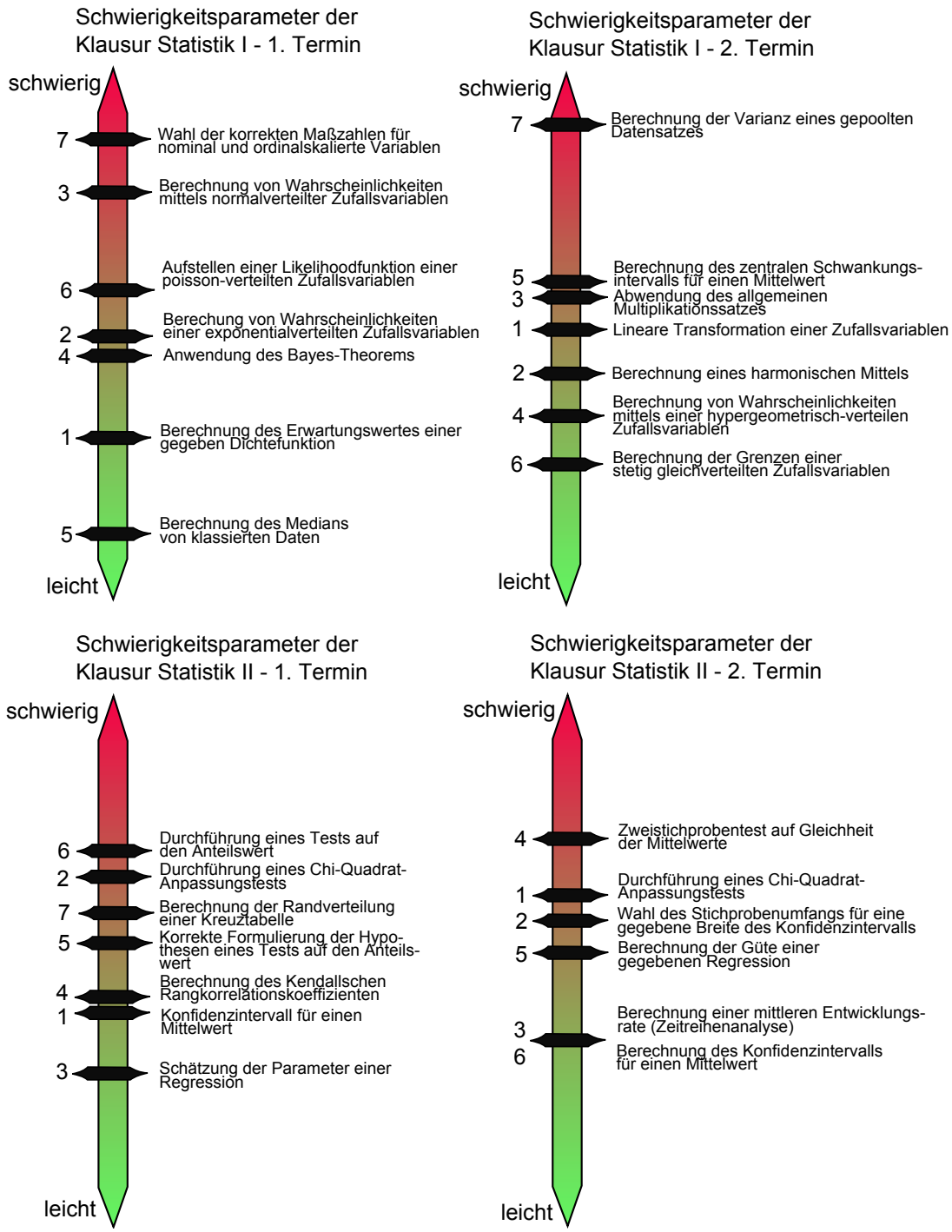

Abbildung 4.9: Übersicht: Einordnung der Schwierigkeitsparameter aller betrachteten Klausuren

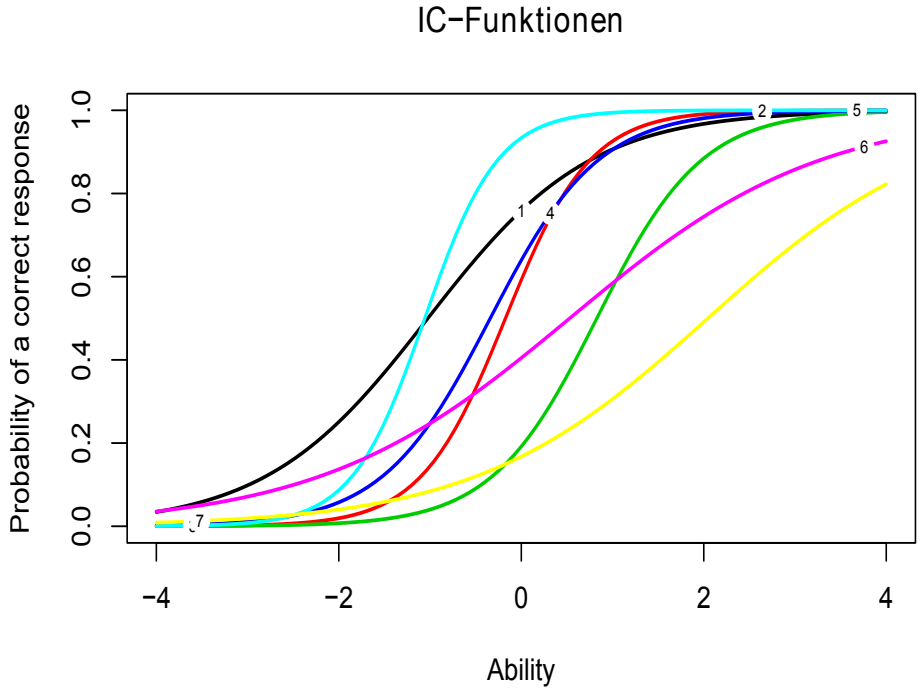

Abbildung 4.10: Item-charakteristische Kurven, Birnbaum-Modell, Statistik I, Termin 1

restlichen Aufgaben besitzen mittlere Trennkraft, wobei die Items 2 und 3 besonders negativ auffallen: Die Berechnung eines harmonischen Mittels und die Anwendung des allgemeinen Multiplikationssatzes erfassen die latente Dimension offenbar nur ungenau.

Aufgabe	1	2	3	4	5	6	7
Schwierigkeitsparameter σ	-0,011	-0,803	0,272	-0,931	0,324	-1,343	1,943
Trennschärfeparameter λ	1,356	0,783	0,785	1,267	1,047	1,223	1,794

Tabelle 4.6: Geschätzte Parameter für Birnbaum-Modell Statistik I - Termin 2

Anhand von Abbildung 4.11 kann man sich diese Schätzungen des Modells mittels der IC-Funktionen noch einmal verdeutlichen. Wie im Rasch-Modell erfassen die Items den mittleren bis negativen Bereich gut, während in den Bereich der höheren Werte von θ nur das Item 7 fällt.

Für den ersten Termin von Statistik II zeigen die geschätzten Schwierigkeitsparameter ein interessantes Bild, dies sollte erwähnt werden, da es sich in einem Punkt vom Rasch-Modell deutlich unterscheidet: Alle Paramter σ sind negativ, es wird nur der negative Bereich der latenten Dimension durch den Test erfasst. Dies passt zur Einschätzung der TIF im Rasch-Modell und damit der Einschätzung der guten

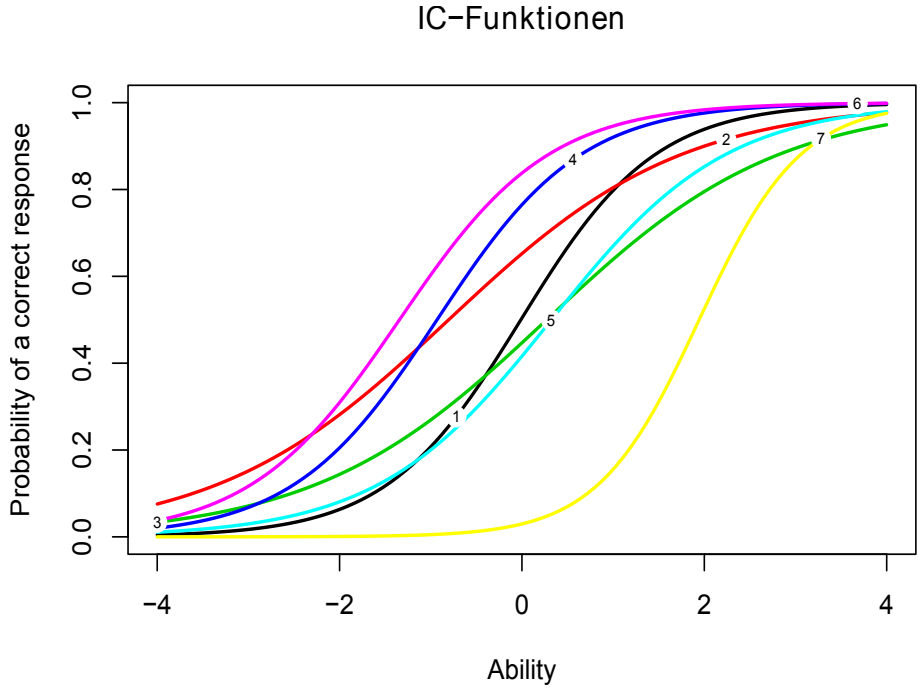

Abbildung 4.11: Item-charakteristische Kurven, Birnbaum-Modell, Statistik I, Termin 2

Klausurergebnisse.

Aufgabe	1	2	3	4	5	6	7
Schwierigkeitsparameter σ	-2,056	-0,403	-2.940	-1,926	-0,830	-0,151	-1,012
Trennschärfeparameter λ	0,855	1,055	0,837	0,906	1,258	2,248	0,633

Tabelle 4.7: Geschätzte Parameter für Birnbaum-Modell Statistik II - Termin 1

Die Berechnung eines Anteilswerts (Item 6) hat eine äußerst hohe Trennkraft. Die restlichen Aufgaben trennen, wie man auch in Abbildung 4.12 sehen, kann größtenteils allenfalls mittelmäßig. Wobei die Trennkraft von Item 7, dabei handelte es sich um die Berechnung der Randverteilung einer Kreuztabelle, besonders unterdurchschnittlich ist.

Für den zweiten Termin von Statistik II zeigt sich wieder dasselbe Muster der Schwierigkeitsparameter wie bei der Schätzung des Rasch-Modells. Besonders trennscharf ist die Aufgabe 1, die Durchführung eines Chi-Quadrat-Anpassungstests. Besonders negativ hinsichtlich der Trennschärfe fällt Aufgabe 5 auf: Die Berechnung der Güte einer Regression kann als zu ungenau zur Messung der interessierenden latenten Variablen eingestuft werden.

Die restlichen Items weisen mittlere Trennschärfen zwischen 0,8 und 1,3 auf. Ab-

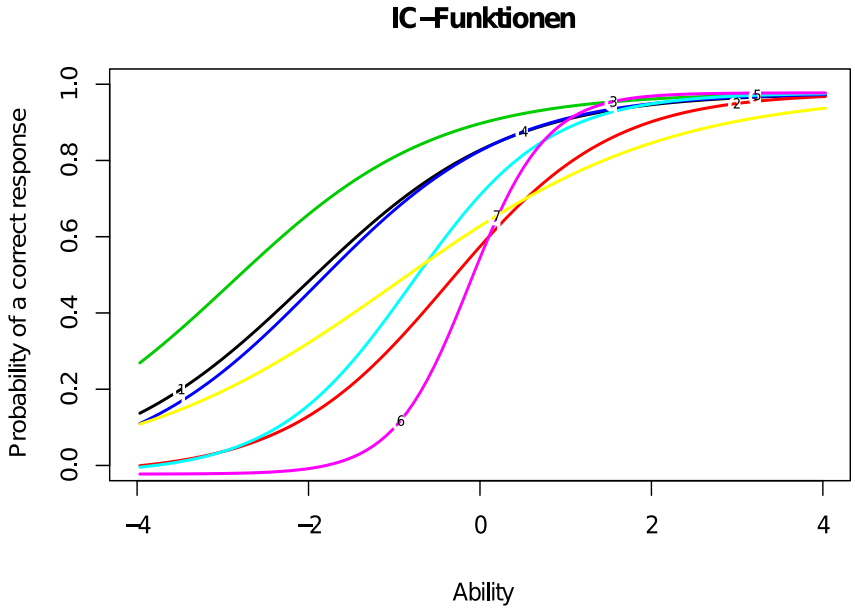

Abbildung 4.12: Item-charakteristische Kurven, Birnbaum-Modell, Statistik II, Termin 1

Aufgabe	1	2	3	4	5	6
Schwierigkeitsparameter σ	0,417	0,413	-1,135	1,114	0,216	-1,193
Trennschärfeparameter λ	2,214	1,126	0,861	1,246	0,544	0,807

Tabelle 4.8: Geschätzte Parameter für Birnbaum-Modell Statistik II - Termin 2

bildung 4.13 zeigt nochmals den Plot der IC-Funktionen des geschätzten Birnbaum-Modells für diese Klausur.

Die Testinformationsfunktionen für die Birnbaum-Modelle zeigen in etwa dasselbe Bild, wie für die geschätzten Rasch-Modelle im vorangehenden Kapitel. In Anhang C kann man sich davon genauer überzeugen.

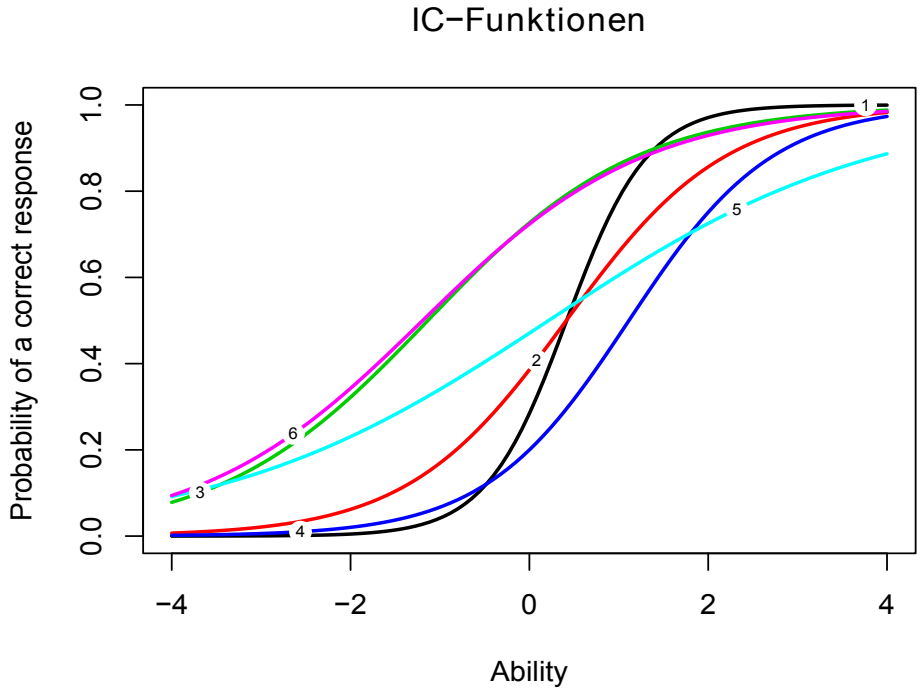

Abbildung 4.13: Item-charakteristische Kurven, Birnbaum-Modell, Statistik II, Termin 2

4.3 Mokken-Analyse

Wie im dritten Kapitel erwähnt, wurden die Mokken-Analysen allesamt mittels des von van der Ark (2007) in R implementierten Pakets „mokken“ durchgeführt.

4.3.1 Ergebnisse für Statistik I - Termin 1

Der erste Schritt der Mokken-Analyse besteht in der Überprüfung der Homogenität der Items, für die Klausur Statistik I - Termin 1 zeigt Tabelle 4.9 beispielhaft die Interitemhomogenitätskoeffizienten H_{ij}, berechnet wie in Formel (3.12) dargestellt. Es zeigt sich, dass alle Koeffizienten positiv und nur die Itemhomogenität der 6. mit der 7. Aufgabe ($H_{67} = 0,10$) und der 6. mit der 1. Aufgabe ($H_{16}) = 0,13$) mit Koeffizienten von jeweils deutlich geringer als 0,2 als schlecht zu bezeichnen sind.

Interessanter ist bei der Mokkenanalyse allerdings die Homogenität eines Items mit den restlichen Items (H_j wie in Formel (3.13) dargestellt.) Tabelle 4.10 gibt Auskunft, inwiefern die Items zur restlichen Mokken-Skala passen: Auch hier fällt auf, dass Item 6 die geringste Homogenität aufweist. Nach Mokken sollte dieser Koeffizient über 0,3 liegen, damit man das Item zur Skala hinzufügen kann. Die 6. Aufgabe verlangte von den Studierenden das Aufstellen einer Likelihoodfunktion einer poisson-verteilten Zufallsvariablen. Ein Grund für dieses Ergebnis könnte sein,

Aufgabe	1	2	3	4	5	6	7
1	1,00						
2	0,43	1,00					
3	0,43	0,59	1,00				
4	0,43	0,39	0,60	1,00			
5	0,30	0,74	0,79	0,51	1,00		
6	0,13	0,26	0,20	0,25	0,65	1,00	
7	0,38	0,20	0,23	0,42	0,85	0,10	1,00

Tabelle 4.9: Interitemhomogenitätskoeffizienten H_{ij} für Statistik I - Termin 1

dass die Anforderungen bei Aufgabe 6 sich deutlich von denen der anderen Aufgaben unterschied, etwa im Bereich mathematischer Methoden anzusiedeln sind. Das Aufstellen einer Funktion ist ein Konzept, dass die Studierenden bereits in den Mathematikkursen lernen, kein in Statistik I komplett neu eingeführtes Konzept. Erwähnt sei auch, dass derartige Schätzungen inzwischen im Kurs Statistik II behandelt werden.

Aufgabe	1	2	3	4	5	6	7
	0,351	0,421	0,416	0,417	0,573	0,243	0,284

Tabelle 4.10: Itemhomogenitätskoeffizienten H_j für Statistik I - Termin 1

- Homogenitätskoeffizient H für die Klausur Statistik 1 - Termin 1: 0,382

Der Homogenitätskoeffizient der betrachteten sieben Aufgaben, liegt bei 0,382, was als schwache Mokken-Skala zu bezeichnen ist.

Die Überprüfung der Annahme der Überschneidungsfreiheit ergibt, dass zwei Items sich überschneiden: Aufgaben 3 und 7; mit einem Wert von 0,03 ist die Überschneidung allerdings äußerst gering und man kann allenfalls von einer leichten Verletzung der Annahmen sprechen.

Tabelle 4.11 zeigt das Ergebnis des automatisierten Algorithmuses zur Itempartionierung in Mokken-Skalen: Wenig überraschend zeigt sich, dass die Aufgaben 1 - 5 und 7 eine äußerst homogene Mokken-Skala darstellen. Nach Ausschluss der 6. Aufgabe steigt der H auf einen Wert von 0,45 und die Skala ist als angemessen zu bezeichnen.

	Aufgabe	Thema
Skala 1	1	Berechnung des Erwartungswertes einer gegebenen Dichtefunktion
	2	Berechnung der Eintrittswahrscheinlichkeit einer exponentialverteilten Zufallsvariablen
	3	Berechnung von Wahrscheinlichkeiten mittels einer normalverteilten Zufallsvariablen
	4	Anwendung des Bayes-Theorems
	5	Berechnung des Medians von klassierten Daten
	7	Wahl der korrekten Maßzahlen für metrisch- und ordinal-klassierte Daten
zu keiner Skala zugehörig	6	Aufstellen der Likelihoodfunktion einer poissonverteilten Zufallsvariablen

Tabelle 4.11: Ergebnis der Itempartitionierung für Statistik I - Termin 1

4.3.2 Ergebnisse für Statistik I - Termin 2

Für den zweiten Termin der Klausur Statistik I verändert sich bei der Überprüfung der Itemhomogenität das Bild deutlich: Nur für ein Item, die 7. Aufgabe (Berechnung der Varianz eines gepoolten Datensatzes), zeigt sich eine hoher Itemhomogenitätskoeffizient ($H_7 = 0{,}543$). Ein Blick in Tabelle 4.12 zeigt, dass die restlichen Items zwischen 0,2 und 0,3 liegen, also als nicht-zugehörig zu dieser Mokken-Skala bezeichnet werden können.

Aufgabe	1	2	3	4	5	6	7
	0,271	0,218	0,220	0,279	0,245	0,290	0,543

Tabelle 4.12: Itemhomogenitätskoeffizienten H_j für Statistik I - Termin 2

- Homogenitätskoeffizient H für die Klausur Statistik I - Termin 2: 0,263

Ebenso verhält es sich mit dem Homogenitätskoeffizienten der gesamten Itembatterie, mit 0,263 liegt er deutlich unter den von Mokken als Untergrenze vorgeschlagenen 0,3. Die Skala kann demnach nicht einmal als „schwach“ bezeichnet werden.

Dagegen verläuft die Überprüfung der Annahme der Überschneidungsfreiheit mittels „rest score“ problemlos. Die Prozedur berichtet keine Verletzungen der Annahme. Die Partitionierung der Itembatterie in verschiedene Mokken-Skalen mittels des automatisierten Algorithmuses zeigt, dass sich aus den Aufgaben 1 und 4 - 7 eine Mokken-Skala bilden lässt, deren Homogenitätskoeffizient anschließend auf 0,36

($H_{Skala1} = 0,36$) geschätzt wird. Die verbleibenden zwei Aufgaben lassen sich zu einer Skala zusammenfassen, mit einem Koeffizienten von ebenfals 0,36 ($H_{Skala2} = 0,36$).

	Aufgabe	Thema
Skala 1	1	Lineare Transformation einer Zufallsvariablen
	4	Berechnung der Wahrscheinlichkeiten einer hypergeometrisch-verteilten Zufallsvariablen
	5	Berechnung eines zentralen Schwankungsintervalls (Mittelwert)
	6	Berechnung der Grenzen einer stetig gleichverteilten Zufallsvariablen
	7	Berechnung der Varianz eines gepoolten Datensatzes
Skala 2	2	Berechnung des harmonischen Mittels
	3	Anwendung des allgemeinen Multiplikationssatzes

Tabelle 4.13: Ergebnis der Itempartitionierung für Statistik 1 - Termin 2

4.3.3 Ergebnisse für Statistik II - Termin 1

Die Überprüfung der Homogenität der Items der Klausur zu Statistik II (1. Termin) zeigt ein nochmals schlechteres Ergebnis. Nur ein Koeffizient ($H_6 = 0,316$) liegt über dem geforderten Wert von 0,3. Aufgabe 7, gefordert war die Berechnung der Randverteilung einer Kreuztabelle, weist nur einen Wert von $H_7 = 0,162$ auf. Ein Grund dafür könnte sein, dass zur Lösung mehr logisches Denken als statistische Kompetenzen erforderlich sind, die Aufgabe zum Testen der latenten Variablen also eher ungeeignet ist.

Aufgabe	1	2	3	4	5	6	7
	0,215	0,231	0,258	0,207	0,224	0,316	0,162

Tabelle 4.14: Itemhomogenitätskoeffizienten H_j für Statistik II - Termin 1

- Homogenitätskoeffizient H für die Klausur Statistik II - Termin 1: 0,229

Für die gesamte Klausur zeigt sich ein äußerst niedriger Homogenitätskoeffizient von 0,229. Die Überprüfung der Überschneidungsfreiheit zeigt eine deutliche Verletzung durch die Items 2 und 6. Tabelle 4.15 zeigt das Ergebnis des Versuchs homogene Mokken-Skalen aus den vorliegenden Items zu isolieren: Aufgaben 2 - 4 und 6 bilden Skala 1 mit einem Koeffizienten $H_{Skala1} = 0,34$. Die Aufgaben 1 und 5 werden durch

den Algorithmus zu einer Skala zusammengefasst ($H_{Skala2} = 0{,}37$). Das siebte Item, das schon die geringste Itemhomogenität aufweist, lässt sich keiner Skala zuordnen.

	Aufgabe	Thema
Skala 1	2	Chi-Quadrat-Anpassungstest
	3	Schätzung der Parameter einer Regression
	4	Berechnung des Kendallschen Rangkorrelationskoeffizienten
	6	Durchführung eines Tests auf den Anteilswert
Skala 2	1	Konfidenzintervall für einen Mittelwert
	5	Hypothesenformulierung (Test auf Anteilswert)
zu keiner Skala zugehörig	7	Berechung der Randverteilung einer Kreuztabelle

Tabelle 4.15: Ergebnis der Itempartitionierung für Statistik II - Termin 1

4.3.4 Ergebnisse für Statistik II - Termin 2

Die größte Itemhomogenität beim 2. Termin weist wiederum eine Aufgabe aus dem Bereich der Testtheorie auf. Bei der 1. Aufgabe handelte es sich um einen Chi-Quadrat-Anpassungstest (H_1) = 0,318. Der Bereich Testverfahren ist im Kurs Statistik II zentral, er baut in vielerlei Hinsicht auf die die vorhergehenden Kapitel desselben Kurses sowie auf den Kurs Statistik I auf. Dass die homogensten Items aus diesem Bereich stammen, ist aus dieser Perspektive weniger verwunderlich: Testverfahren scheinen repräsentativ für die angeeigneten Kompetenzen im Kurs zu sein. Auffällig ist ebenso die geringe Homogenität des 5. Items, die Berechnung der Güte einer gegebenen Regression. Dabei handelte es sich um das simple Einsetzen der gegebenen Werte in die richtige Formel: aus didaktischer Perspektive kann man bezweifeln, dass damit die latente Dimension „stastitische Kompetenzen“ genau genug erfasst wird.

Aufgabe	1	2	3	4	5	6
	0,318	0,246	0,273	0,239	0,168	0,232

Tabelle 4.16: Itemhomogenitätskoeffizienten H_j für Statistik II - Termin 2

- Homogenitätskoeffizient H für die Klausur Statistik II - Termin 1: 0,245

Für diese Mokken-Skala zeigt sich wiederum ein äußerst niedriger Homogenitätskoeffizient von 0,245. Die Partitionierung zeigt in diesem Fall ein interessantes Bild:

Zwar werden die beiden Aufgaben aus den Bereichen Regression (Item 5) und Zeitreihen (Item 3), zwei Themengebiete, die sich von den restlichen des Kurses unterscheiden, in sich aber viele parallelen haben, zu einer Skala zusammengefasst ($H_{Skala2} = 0{,}42$) und auch die Items 1, 2 und 4, allesamt zum Thema Schätzen und Testen und zentral im Kurs Statistik II, ergeben eine schwache Mokken-Skala ($H_{Skala1} = 0{,}37$). Nicht in dieses Bild passt allerdings Aufgabe 6, Berechnung eines Konfidenzintervalles für einen Mittelwert, die thematisch in Skala 1 anzusiedeln wäre.

	Aufgabe	Thema
Skala 1	1	Chi-Quadrat-Anpassungstest
	2	Wahl des Stichprobenumfangs für ein geg. Konfidenzintervall
	4	Zweistichprobentest auf Gleichheit der Mittelwerte
Skala 2	3	Berechnung der mittleren Entwicklungsrate (Zeitreihenanalyse)
	5	Berechnung der Güte einer Regression
zu keiner Skala zugehörig	6	Konfidenzintervall für einen Mittelwert

Tabelle 4.17: Ergebnis der Itempartitionierung für Statistik II - Termin 2

Zusammenfassend stellt man fest, dass der Versuch mittels den vorliegenden Daten Mokken-Skalen für die vier Klausurtermine zu bilden von wenig Erfolg gekrönt ist. Anzumerken bleibt auch die Tatsache, dass die Skalenhomogenität des Kurses Statistik I deutlich über der Homogenität des Kurses Statistik II liegt, man also der Hypothese, dass Statistik I thematisch stringenter ist, in weiteren Untersuchungen nachgehen könnte. Dagegen spricht die Tatsache, dass die Kompetenzen, welche die Studierenden im Kurs Statistik I erlernen, völlig neuartig sind und Themen wie Wahrscheinlichkeitsrechnung und Verteilungsmodelle thematisch zwar verwandt aber nicht derart verknüpft wie die Themen des Kurses Statistik II sind. Zu erwarten wäre also das gegenteilige Bild gewesen, abgesehen von der Unterscheidung der Bereiche Testen und Schätzen sowie Regression und Zeitreihen, die sich bei der Itempartitionierung des zweiten Termins von Statistik II andeutete.
Darüber hinaus bleibt zu bezweifeln, dass die Annahme der Überschneidungsfreiheit bei Hinzunahme von weiteren Items so zu halten wäre, wie es in der vorliegenden Untersuchung größtenteils der Fall war.

5 Fazit

Die Anpassung der Daten der vier betrachteten Klausuren gelang leider nur unbefriedigend. Die Mokken-Analyse eignet sich auch eher für einen großen Itempool aus dem Items mittels des bereitgestellten Algorithmuses zu homogenen Mokken-Skalen partitioniert werden, welche die Annahmen des Modells erfüllen.
Bei Rasch- und Birnbaum-Modellen ist dies im Prinzip nicht anders, auch hier werden Verfahren angeboten, um aus einem Itempool möglichst gute Skalen (im Sinne der Anpassung) zu schätzen. Dennoch gelang die Anpassung mittels des Rasch-Models an die Daten größtenteils zufriedenstellend.
Sicherlich sollte den Birnbaum-Analysen noch Modelanpassungstests folgen, um zwischen dem Modell mit der besseren Anpassung wählen zu können, außerdem hätten die von den Modellen geschätzten Parameter der Personenfähigkeiten noch genauer analysiert werden können. Dies alles hätte aber den Rahmen dieser Arbeit bei weitem gesprengt.
Das grundsätzliche Ziel, eine Einordnung der Modelle und ihre Überprüfung an den Daten mit der Fragestellung, ob sie im universitären Umfeld Anwendung finden sollten, konnte geleistet werden. Letztere Frage kann eindeutig positiv beantworet werden. Die vielen guten statistischen Eigenschaften, die Rasch- und Birnbaum-Modelle mit sich bringen, finden nicht umsonst auch in großen Studien wie der eingangs erwähnten PISA-Studie Anwendung.
Problematisch daran ist allerdings, dass die Erstellung der Items, sollte man sich auf diesen Versuch einlassen, bei weitem mehr Aufwand abverlangen als das einfache Erstellen einer Klausur durch den Professor oder Dozenten. Für eine genaue Schätzung der Personenparameter ist vielmehr eine große Itembatterie, die den gesamten Bereich des Schwierigkeitsgrades umfasst vonnöten. Aus Zeitgründen müssten also viele kurze Aufgaben in den Klausuren gestellt werden. So interessant und statistisch gesichert die Ergebnisse der Überprüfung der Kenntnisse mittels probabilistischen Testmodellen wären, aus personeller Sicht ist dies von den Universitäten kaum zu leisten.

A Kriterien zur Lösung der Klausuraufgaben

Aufgabe	1	2
Thema	Berechnung des Erwartungswerts einer gegebenen Dichtefunktion	Berechnung der Wahrscheinlichkeit einer exponentialverteilten Zufallsvariable
nötiger Anteil der Punkte	50 %	60 %
Kommentar	Anwendung der korrekten Methode: 50 % numerisches Ergebnis: 50 %	Anwendung der korrekten Methode: 60 % numerisches Ergebnis: 40 %
Aufgabe	3	4
Thema	Berechnung der Wahrscheinlichkeit einer normalverteilten Zufallsvariable	Anwendung des Bayes-Theorems
nötiger Anteil der Punkte	50 %	100 %
Kommentar	Anwendung der korrekten Methode: 50 % numerisches Ergebnis: 50 %	Diese Aufgabe wurde mit 100 % bei Lösung bewertet, ansonsten gab es 0 Punkte
Aufgabe	5	6
Thema	Berechnung des Medians von klassierten Daten	Aufstellen einer Likelihoodfunktion einer poissonverteilten Zufallsvariablen
nötiger Anteil der Punkte	60 %	50 %
Kommentar	Anwendung der korrekten Methode: 60 % numerisches Ergebnis: 40 %	korrektes Aufstellen der Likelihoodfunktion: 50 % numerisches Ergebnis für λ: 50 %
Aufgabe	7	
Thema	Wahl der korrekten Maßzahlen für metrisch- und nominal-skalierte Variablen	
nötiger Anteil der Punkte	75 %	
Kommentar	Pro falscher oder nichtgetätigter Zuordnung wurde die Aufgabe einen Punkt schlechter bewertet (insg. 4 Punkte)	

Tabelle A.1: Kriterien zur Lösung der Klausuraufgaben für Statistik I - Termin 1 (20.07.2004)

Aufgabe	1	2
Thema	Lineare Transformation einer Zufallsvariablen	Berechnung eines harmonischen Mittels
nötiger Anteil der Punkte	50 %	50 %
Kommentar	Anwendung der korrekten Methode: 50 % numerisches Ergebnis: 50 %	Anwendung der korrekten Methode: 50 % numerisches Ergebnis: 50 %
Aufgabe	3	4
Thema	Anwendung des allgemeinen Multiplikationssatzes	Berechnung der Wahrscheinlichkeiten einer hypergeometrisch-verteilten Zufallsvariablen
nötiger Anteil der Punkte	50 %	50 %
Kommentar	Anwendung der korrekten Methode: 50 % numerisches Ergebnis: 50 %	Anwendung der korrekten Methode: 50 % numerisches Ergebnis: 50 %
Aufgabe	5	6
Thema	Berechnung eines zentralen Schwankungsintervalls für einen Mittelwert	Berechnung der Grenzen einer stetig gleichverteilten Zufallsvariablen
nötiger Anteil der Punkte	50 %	50 %
Kommentar	Anwendung der korrekten Methode: 50 % numerisches Ergebnis: 50 %	Anwendung der korrekten Methode: 50 % numerisches Ergebnis: 50 %
Aufgabe	7	
Thema	Berechnung der Varianz eines gepoolten Datensatzes	
nötiger Anteil der Punkte	50 %	
Kommentar	Anwendung der korrekten Methode: 50 % numerisches Ergebnis: 50 %	

Tabelle A.2: Kriterien zur Lösung der Klausuraufgaben für Statistik I - Termin 2 (15.10.2004)

Aufgabe	1	2
Thema	Schätzintervall für den Mittelwert	Chi-Quadrat-Anpassungstest
nötiger Anteil der Punkte	55,6 %	50 %
Kommentar	Anwendung der korrekten Methode: 55,6 % numerisches Ergebnis: 44,4 %	Anwendung der korrekten Methode: 50 % numerisches Ergebnis: 50 %
Aufgabe	3	4
Thema	Schätzung der Parameter einer Regression	Berechnung des Kendallschen Rangkorrelationskoeffizienten
nötiger Anteil der Punkte	57 %	50 %
Kommentar	Anwendung der korrekten Methode: 57 % numerisches Ergebnis: 43 %	Anwendung der korrekten Methode: 50 % numerisches Ergebnis: 50 %
Aufgabe	5	6
Thema	korrekte Formulierung der Hypothesen eines Tests auf den Anteilswert	Durchführung des Tests auf den Anteilswert
nötiger Anteil der Punkte	83,3 %	66,7 %
Kommentar	für kleinere formale Fehler wurde ein Punkt abgezogen (entspricht 16,7 %)	Anwendung der korrekten Methode: 66,7 % numerisches Ergebnis: 33,3 %
Aufgabe	7	
Thema	Berechnung der Randverteilung einer Kreuztabelle	
nötiger Anteil der Punkte	50 %	
Kommentar	Anwendung der korrekten Methode: 50 % numerisches Ergebnis: 50 %	

Tabelle A.3: Kriterien zur Lösung der Klausuraufgaben für Statistik II - Termin 1 (22.02.2005)

Aufgabe	1	2
Thema	Durchführung eines Chi-Quadrat-Anpassungstests	Wahl des Stichprobenumfangs für ein gegebenes Konfidenzintervall
nötiger Anteil der Punkte	55,6 %	57 %
Kommentar	Anwendung der korrekten Methode: 55,6 % numerisches Ergebnis: 44,4 %	Anwendung der korrekten Methode: 57 % numerisches Ergebnis: 43 %
Aufgabe	3	4
Thema	Berechnung einer mittleren Entwicklungsrate (Zeitreihenanalyse)	Zweistichprobentest auf Gleichheit der Mittelwerte
nötiger Anteil der Punkte	55,6 %	66,7 %
Kommentar	Anwendung der korrekten Methode: 55,6 % numerisches Ergebnis: 44,4 %	korrekte Teststatistik und kritischer Wert: 66,7 % Ergebnis der Teststatistik: 33,3 %
Aufgabe	5	6
Thema	Berechnung der Güte einer gegebenen Regression	Konfidenzintervall für einen Mittelwert
nötiger Anteil der Punkte	57 %	44,4 %
Kommentar	Anwendung der korrekten Methode: 57 % numerisches Ergebnis: 43 %	Anwendung der korrekten Methode: 44,4 % numerisches Ergebnis: 55,6 %

Tabelle A.4: Kriterien zur Lösung der Klausuraufgaben für Statistik II - Termin 2 (08.04.2005)

B Tabellen der Interitemkoeffizienten H_{ij}

Aufgabe	1	2	3	4	5	6	7
1	1,000						
2	0,247	1,000					
3	0,166	0,365	1,000				
4	0,388	0,183	0,236	1,000			
5	0,226	0,156	0,133	0,241	1,000		
6	0,286	0,072	0,208	0,316	0,583	1,000	
7	0,732	0,452	0,391	0,536	0,531	0,691	1,000

Tabelle B.1: Interitemhomogenitätskoeffizienten H_{ij} für Statistik I - Termin 2

Aufgabe	1	2	3	4	5	6	7
1	1,000						
2	0,262	1,000					
3	0,191	0,512	1,000				
4	0,076	0,371	0,302	1,000			
5	0,374	0,107	0,103	0,085	1,000		
6	0,210	0,328	0,130	0,375	0,494	1,000	
7	0,201	0,067	0,333	0,140	0,138	0,221	1,000

Tabelle B.2: Interitemhomogenitätskoeffizienten H_{ij} für Statistik II - Termin 1

Aufgabe	1	2	3	4	5	6
1	1,000					
2	0,257	1,000				
3	0,350	0,259	1,000			
4	0,431	0,456	0,286	1,000		
5	0,214	0,031	0,423	-0,084	1,000	
6	0,431	0,333	0,134	-0,071	0,295	1,000

Tabelle B.3: Interitemhomogenitätskoeffizienten H_{ij} für Statistik II - Termin 2

C Graphen der Testinformationsfunktionen für die geschätzten Birnbaummodelle

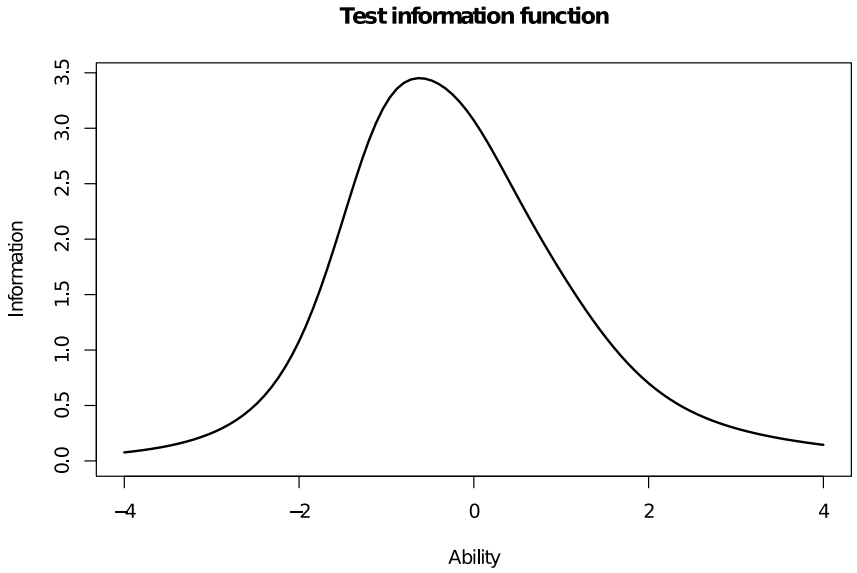

Abbildung C.1: TIF für Birnbaum-Modell Statistik I, Termin 1

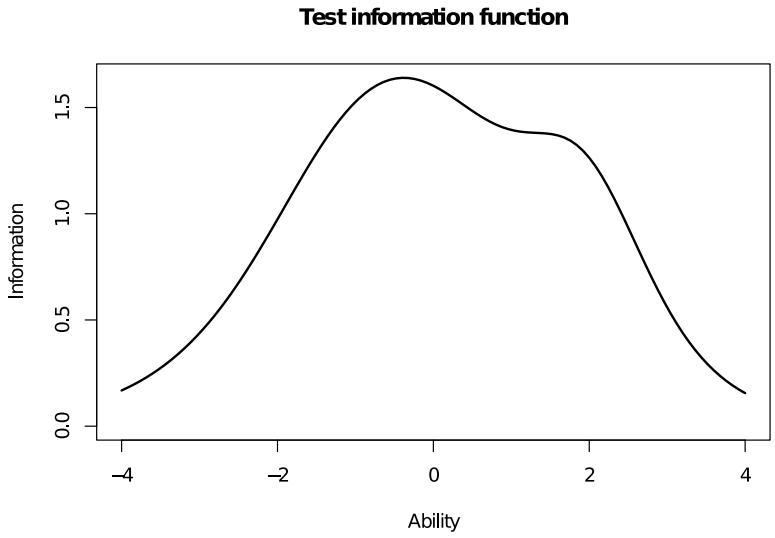

Abbildung C.2: TIF für Birnbaum-Modell Statistik I, Termin 2

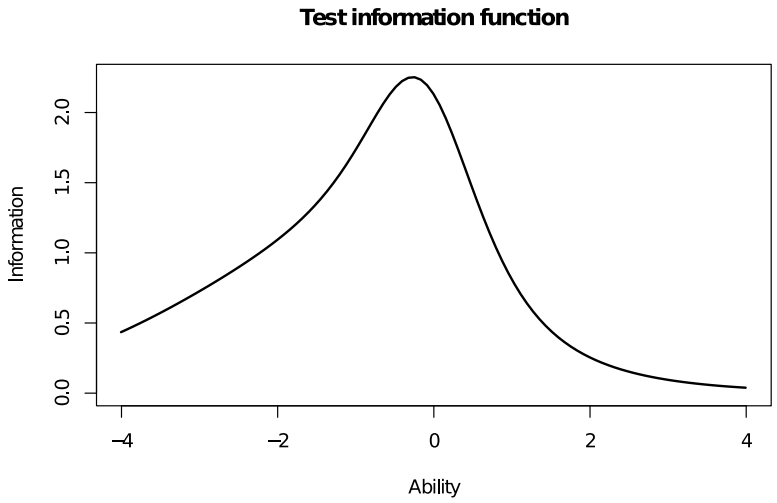

Abbildung C.3: TIF für Birnbaum-Modell Statistik II, Termin 1

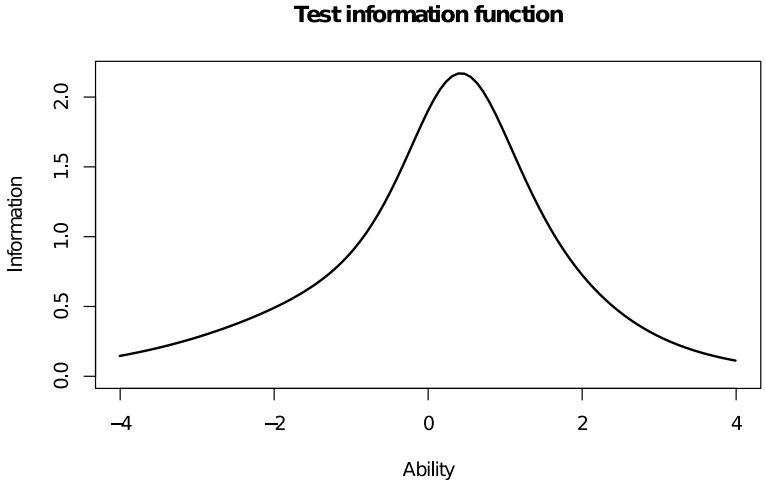

Abbildung C.4: TIF für Birnbaum-Modell Statistik II, Termin 2

Literaturverzeichnis

[1] Amelang, M. und Schmidt-Atzert, L.(2006): *Psychologische Diagnostik und Intervention*. Springer, Berlin.

[2] van der Ark, L. A.(2007): Mokken scala analysis in R. In: *Journal of Statistical Software* Vol.20, 11: S. 1 - 19.

[3] Andersen, E. B.(1973): A goodness of fit test for the Rasch model. In: *Psychometrika* 38: S. 123 - 140.

[4] Bartolucci, F. und Forcina, A.(2005): Likelihood inference on the underlying structure of IRT models. In: *Psychometrika* 70: 31 - 43.

[5] Birnbaum, A.(1968): Some latent trait models. In: F. M. Lord und M. R. Novick (Hrsg.) *Statistical theories of mental test scores*: S.395-479. Addison-Wesley.

[6] de Boeck, P.(2008): Random item IRT models. In: *Psychometrika* 73: 533 - 559.

[7] Irtel, H.(1995): *Entscheidungs- und testtheoretische Grundlagen der psychologischen Diagnostik*. Skript, Universität Mannheim.

[8] Mair, P. und Hatzinger, R.(2009): Extended Rasch Modeling: The eRm Package fhe Application of IRT Models in R. In: *Journal of Statistical Software* 20(9): S. 1 - 20.

[9] Mair, P. und Hatzinger, R.(2009): Referenz zu Package 'eRm'. Download: http://cran.r-project.org/web/packages/eRm/eRm.pdf.

[10] Mokken, R.J.(1971): *A theory and procedure of scale analysis*. Mouton, Den Haag.

[11] Molenaar, I. W.(1991): A weighted Loevinger H-Coefficient Extrending Mokken Scaling to Multicategory Items. In: *Quantitative Methoden* Vol.12, 37: S. 97 - 117.

[12] Molenaar I. W. und Sijtsma K. (2000): *User's Manual MSP5 for Windows*. IEC ProGAMMA, Groningen.

[13] Moosbrugger, H.(2007): Item-Response-Theory (IRT). In: *Testtheorie und Fragebogenkonstruktion*: S. 215 - 259. Springer, Berlin.

[14] Ortner, T.; Proyer, R. und Kubinger, K.(2006): Theorie und Praxis Objektiver Persönlichkeitstests. Verlag Hans Huber, Bern.

[15] Partchev, I.(2009): Referenz zu Package 'irtoys'. Download: http://cran.r-project.org/web/packages/irtoys/irtoys.pdf.

[16] Raiche, G.(2009): Referenz zu Package 'irtProb'. Download: http://cran.r-project.org/web/packages/irtProb/irtProb.pdf.

[17] Rasch G.(1960): *Probabilistic models for some intelligence and attainment tests*. The Danish Institute for Educational Research, Kopenhagen.

[18] Rizopoulos, D.(2009): Referenz zu Package 'ltm'. Download: http://cran.r-project.org/web/packages/ltm/ltm.pdf.

[19] Rost, J.(2004): *Testtheorie - Testkonstruktion*. Verlag Hans Huber, Bern.

[20] Sijtsma K. und Molenaar I. W.(2002): Introduction to Nonparametric Item Response Theory. Sage, Thousand Oaks (California).

[21] Tent, L. und Stelzl, I.(1993): *Pädagogisch-psychologische Diagnostik*. Hogrefe, Göttingen.

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegenden Arbeit zum Thema *Anwendung probabilistisch-testtheoretischer Modelle auf Statistikklausuren des Grundstudiums* selbstständig verfasst, keine anderen als die angegebenen Quellen verwendet, keine unzulässigen Hilfsmittel benutzt sowie alle Zitate als solche kenntlich gemacht habe. Die Arbeit hat außerdem keiner anderen Prüfungsbehörde vorgelegen.

Christian Westermeier

Berlin, 18. Januar 2010