Der schiefe Turm von PISA: die logistischen Parameter des Rasch-Modells sollten revidiert werden

Harney, Klaus; Fuhrmann, Christoph; Harney, Hanns L.

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Der schiefe Turm von PISA – die logistischen Parameter des Rasch-Modells sollten revidiert werden

von Klaus Harney, Christoph Fuhrmann und Hanns L. Harney[1]

*Die Kompetenz- und Anforderungsmessungen in den PISA-Studien beruhen auf dem logistischen **Rasch**-Modell, welches der probabilistischen Testtheorie zu Grunde liegt. Dieses Modell weist Schwächen auf. Wegen der logistischen item response Funktion lässt es Sätze von Antworten zu, die zwar legitim aber mit den vorgesehenen Parametern nicht auswertbar sind. Es handelt sich um die uniform beantworteten Fragebögen. Mit deren Sonderstellung hängt zusammen, dass die Schätzer für besonders hohe wie auch besonders niedrige Kompetenzen systematisch vom wahren Wert des Parameters abweichen und dass die Fehlerintervalle beliebig groß werden. Dies erschwert die Interpretation der Schätzer sowie die sozialwissenschaftliche Verwendung der Resultate – z.B. in Regressionsanalysen. Es sind aber gerade die oberen und unteren Kompetenzniveaus und Schwierigkeitsstufen, denen das besondere Interesse der Bildungsforschung und Bildungspolitik gilt. Dieses durchaus bekannte Problem wurde bislang formal nicht gelöst. Im vorliegenden Aufsatz wird gezeigt, dass man es lösen kann, indem man zu einer anderen – der trigonometrischen – item response Funktion übergeht.*

*The attainment of persons and the difficulties of items are measured via the logistic **Rasch** model which is the basis of probabilistic test theory. This model has a weakness. Due to the logistic item response function it admits sets of answers that – although they are legitimate – cannot be analyzed in terms of the parameters of the model. We refer to persons that uniformly answer all items either correctly or incorrectly. As a consequence, the estimators of especially high and especially low attainment are*

1 Dr. ***Klaus Harney*** ist Professor, ***Christoph Fuhrmann*** wissenschaftlicher Mitarbeiter am Institut für Pädagogik, Lehrstuhl für Berufs- und Wirtschaftspädagogik, Ruhr-Universität Bochum, D-44780 Bochum; email: Klaus.Harney@ruhr-uni-bochum.de. Dr. ***Hanns L. Harney*** ist Professor am ***Max-Planck***-Institut für Kernphysik, 69117 Heidelberg.

Danksagung: In Gesprächen mit Herrn Professor ***Andreas Müller*** zum Thema des ***Rasch***-Modells haben wir wertvolle Hinweise bekommen.

systematically biased and the error intervals can become indefinitely large. This obscures the interpretation of the estimators, and makes the results difficult to use in the social sciences. However, the special interest of research and politics is focussed on the upper and lower levels of attainment. The problem is known but has remained unsolved so far. We show that one can solve it within the binomial framework of the Rasch model by introducing a new – the so-called trigonometrical – item response function.

1 Einleitung

Mit den PISA-Studien ist die Methodik von Kompetenz- und Anforderungsmessungen aus dem engeren Bereich der psychologischen Testtheorie herausgetreten. Sie ist heute auch für die Bildungsforschung und Bildungssoziologie von grundlegender Bedeutung. Die Modellierung von Kompetenzen und Kompetenzniveaus wird mittlerweile gestützt auf Erkenntnisse der Sozialwissenschaften, der Psychologie und der Testtheorie im engeren Sinne. Dies geht auf den Fortschritt der Teststatistik in der sogenannten item response theory (IRT) zurück. Mit ihr steht ein Instrumentarium statistischer Schlussfolgerungen zur Verfügung, das zwar in Messungen von Kompetenzen – wie in den PISA-Studien – seinen angestammten Ort hat, das sich jedoch auf die Messung von politischen und geschmacklichen Einstellungen verallgemeinern lässt. Trotz dieses beachtlichen theoretischen Unterbaus weist die IRT Schwächen auf. Diese Schwächen werden beim Verarbeiten uniformer Antwortmuster und beim Abschätzen von Fehlerintervallen deutlich. Je zutreffender oder unzutreffender Tests beantwortet werden umso unzuverlässiger werden die Schätzer und umso größer werden Fehlerintervalle. Die Information über mittelmäßig kompetente Personen fällt systematisch genauer aus als die über stark oder schwach kompetente Personen. Daher weist die Information, die man den Daten entnehmen kann, eine Tendenz zur Mitte auf. Sie ist am größten für Ausprägungen im Bereich durchschnittlicher Werte. Komplementär dazu scheinen die Anforderungen an den Umfang des Tests zu wachsen, wenn stark oder schwach kompetente Personen gemessen werden sollen. Der Test muss dann um Aufgaben erweitert werden, die die Kompetenz von starken oder schwachen Personen diagnostizieren können. Daraus ergeben sich erneut stark oder schwach kompetente Personen, und der Test muss wiederum erweitert werden. Der Umfang eines Tests lässt sich also nicht planen bevor man die Ergebnisse gesehen hat. Die bisher eingeführten Korrekturverfahren haben dieses Problem nicht strukturell lösen können. Im Folgenden legen wir einen Vorschlag vor, der das Auswerten uniformer Antwortmuster, eine einfache Angabe des Fehlerintervalles sowie eine planbare, vom Wert des Parameters unabhängige Information möglich macht.

In Abschnitt 2 wird beschrieben wie das Konzept der Kompetenz im ***Rasch***-Modell formuliert wird.[2] Das geläufige Modell der IRT wird präzisiert durch die Schilderung des logistischen ***Rasch***-Modells in Abschnitt 3. Das ***Rasch***-Modell (***Rasch*** 1980) ist die mathematische Gestalt der IRT. Es geht von binären Antwortformaten und damit vom Binomialmodell aus. Man kann den Gehalt des Binomialmodelles von der Form der in ihm enthaltenen item response Funktion unterscheiden, siehe Abschnitt 4. Die logistische Funktion wird üblicherweise durch Analogien zur Physik begründet. In Abschnitt 5 kritisieren wir diese Auffassung. Wir argumentieren dort, dass die logistische Funktion dem Modell keineswegs die spezifische Objektivität verleiht, die dem Modell von ***Fischer*** zugeschrieben wurde. Wir zeigen in den Abschnitten 5 und 6, dass die trigonometrische Form dem ***Rasch***-Modell Vollständigkeit und spezifische Objektivität gibt. In Abschnitt 7 werden die Ergebnisse zusammengefasst, und es werden Schlussfolgerungen gezogen. Daran schließen sich die Anhänge A und B mit mathematischen Details an.

2 Die IRT als Form der Kompetenzmessung

2.1 Kontext und Spezifikation der Fragestellung

Kompetenzen werden in der Bildungsforschung als das Vermögen von Menschen aufgefasst, Anforderungen ihrer sozialen und natürlichen Umwelt zu erkennen und erfolgreich zu bearbeiten. Wie die traditionsreiche Diskussion des Kompetenzbegriffs in den Geisteswissenschaften zeigt, besitzen Kompetenzzuschreibungen eine prognostische Bedeutung: Kompetenten Personen wird die erfolgreiche Bewältigung künftig eintretender Anforderungen zugerechnet. Die PISA-Studien gehen laut ***Baumert*** und ***Artelt*** (2003, S.12), von Basiskompetenzen aus, die in modernen Gesellschaften für eine befriedigende Lebensführung in persönlicher und wirtschaftlicher Hinsicht sowie für eine aktive Teilhabe am gesellschaftlichen Leben notwendig sind. Eine besondere Variante des prognostischen Charakters, der dem Kompetenzbegriff zukommt, ist in den PISA-Studien zu erkennen: Dort findet sich die Unterscheidung von Kompetenzstufen wie auch die Konstruktion von Kompetenzstufenmodellen für zentrale Bereiche des schulischen Kanons.

2 Bezeichnet wird das ***Rasch***-Modell auch als die 1-pl-Variante (one-parameter logistic version) der IRT. Die Berücksichtigung der sog. Trennschärfe, ***Rost*** (2004, S. 98), und des sog. Zufalls, ***Rost*** (2004, S. 135) in den 2-pl- und 3-pl-Modellen der IRT diskutieren wir nicht, weil es uns um Messungen unter dem Kriterium der spezifischen Objektivität geht, siehe Abschnitt 5 und ***Rost*** (2004, S. 133). Die Auseinandersetzung mit den anderen Modellen würde den Rahmen des vorliegenden Artikels sprengen.

Die empirischen Messungen, die den Modellen unterlegt sind, stützen sich auf die sogenannte probabilistische Testtheorie – auch bezeichnet als IRT – deren bekannteste Form auf die Weiterentwicklung der Teststatistik durch den dänischen Mathematiker ***Georg Rasch*** zurückgeht. In der Grundstruktur unterscheidet ***Rasch*** (1980) "two types of parameters, a 'difficulty' for each test (or item) and an 'ability' for each person". Beide Parameter werden aus einem durch Fragebögen o.ä. zustande gekommenen Datensatz geschätzt. Die Itemschwierigkeit ergibt sich im Prinzip aus dem Anteil von Personen, die ein gegebenes Item gelöst haben. Im Fall eines einzigen Items entspricht dies der relativen Häufigkeit zutreffender Reaktionen. Die Personenfähigkeit ergibt sich im Prinzip aus dem Anteil gelöster Items an der Gesamtheit lösbarer Items. Im ersten Fall entspricht dies der Frage, wie oft ein Item gelöst wurde, im zweiten Fall der Frage, wie viele Items eine Person gelöst hat. Beide Parameterschätzungen unterscheiden sich lediglich durch die Richtung der Datenaggregation: Für die Schätzung der Itemschwierigkeit benötigt man im Prinzip die jeweiligen Häufigkeiten richtig antwortender Personen (im Folgenden Erträge genannt). Für die Schätzung der Personenfähigkeit benötigt man die Häufigkeiten richtig beantworteter Items (im Folgenden scores genannt). Die möglichen Antworten auf eine Zahl M_D von Items lassen viele verschiedene Antwortmuster zu. Träger der vollständigen Information über die Personenfähigkeit sind dem ***Rasch***-Modell zufolge nicht die individuellen Antwortmuster, die sich hinter dem score verbergen, sondern ausschließlich der score selbst. Das ist für die Parameterschätzung zentral (***Davier, Molenaar*** und ***Person*** 2003).

Die Kompetenzstufen der PISA-Studie[3,4] basieren auf der aus den Antworthäufigkeiten geschätzten durchschnittlichen Schwierigkeit der Items, die in Form von Schwierigkeitsniveaus klassifiziert werden.

Entscheidend für die Logik der IRT ist die Zurückführung der Lösung einer Aufgabe auf zwei Komponenten: auf die Personenfähigkeit und auf die Aufgabenschwierigkeit. Je fähiger eine Person desto leichter fällt ihr die Lösung der Aufgabe, dieser Logik zufolge. Die Aufgabenschwierigkeit fällt also mit wachsender Ausprägung der Personenfähigkeit immer weniger ins Gewicht (allerdings: die Schwierigkeit an sich bleibt dabei immer gleich) – et vice versa. In der mathematischen Modellierung setzt sich deshalb die Lösungswahrscheinlichkeit einer Aufgabe für eine Person aus der Differenz zwischen der Ausprägung des Personenparameters

3 ***Deutsches PISA-Konsortium PISA 2000*** (2003): Ein differenzierter Blick auf die Länder der Bundesrepublik, Verlag Leske und Budrich, Opladen 2003.

4 ***PISA-Konsortium Deutschland*** (2004): Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs, Waxmann Verlag, Münster.

und derjenigen des Itemparameters zusammen. Beide Parameter werden auf derselben Skala gemessen.[5]

Das Vorstehende beschreibt die Struktur des einfachen ***Rasch***-Modells, welches von einer eindimensionalen Anordnung der gemessenen Kompetenz ausgeht. Komplexere mehrdimensionale Erweiterungen des einfachen ***Rasch***-Modells, die das Zustandekommen der Aufgabenlösungen in eine mehrdimensionale Auffächerung von Parametern mit jeweils spezifischer Bedeutung für die korrekte Antwort auflösen (***Adams***, ***Wilson*** and ***Wang*** 1997), können außer Betracht bleiben, da sie von denselben formalen Prinzipien ausgehen, auf die sich auch unsere Überlegungen richten.

In dieser formalen Eigenschaft der IRT ist der Anschluss an den besonderen prognostischen Charakter des Kompetenzbegriffs enthalten: Kompetenzen werden durch Personenparameter indiziert, siehe ***Baumert*** und ***Artelt*** (2003, S. 43). Die Parameter selber sind letztlich Zahlenwerte, die den durchschnittlichen Anteil einer gemessenen Ausprägung der jeweiligen Personenfähigkeit an der Lösbarkeit von Items repräsentieren. In Verbindung mit dem Itemparameter, dessen Wert den durchschnittlichen Anteil der Schwierigkeit eines Items an dessen Lösbarkeit vertritt, ergibt sich die auf Personen beziehbare Lösungswahrscheinlichkeit. Man kann nun einem bestimmten Personenparameter verschiedene Aufgaben gleicher Schwierigkeit zuordnen: die durch den Personenparameter charakterisierte Person wird die Aufgabe mit gleich bleibender Wahrscheinlichkeit lösen. Folglich bildet der Personenparameter den prognostischen Gehalt des Kompetenzbegriffs dadurch ab, dass er – einem Messgerät ähnlich – die personenspezifische Lösbarkeit von Aufgaben auf einer zwischen leicht und schwierig normierten Skala vergleichbar macht. D.h. wenn die Werte von Personenparametern bekannt sind, ist die Prognose der Lösungswahrscheinlichkeit mit Hilfe von Aufgabenstellungen möglich, die sich der Logik der IRT folgend auf einer Anordnung von Schwierigkeitsstufen haben skalieren lassen. Das Andere trifft auch zu: Wenn die Werte von Itemparametern bekannt sind, ist die Prognose der Lösungswahrscheinlichkeit mit Hilfe von Personenfähigkeiten möglich, deren Ausprägung sich – ebenfalls der Logik der IRT folgend – auf einer

5 Dies wird in der anwendungsorientierten Literatur allgemein als die mit dem ***Rasch***-Modell verbundene Eigenschaft der spezifischen Objektivität beschrieben. Zur Problematik dieser Eigenschaft siehe ***Harney*** und ***Harney*** (2006, S. 105-126). Es kann der Fall eintreten, dass die Differenzierung leistungsstarker und -schwacher Personen auf einer Skala nicht vollständig gelingt, z.B. dann wenn im unteren Bereich der leistungsstarken und im oberen Bereich der leistungsschwachen Personen unterschiedlich Antwortprofile für die Zugehörigkeit zur einen oder anderen Gruppe charakteristisch sind. ***Rost*** (2004, S. 662-678) schlägt dafür die Entmischung von Teilnehmerhäufigkeiten im Rahmen des sogenannten mixed-***Rasch***-Modells vor.

Anordnung von Fähigkeitsstufen haben skalieren lassen. Man kann also festhalten, dass der prognostische Gehalt des Kompetenzbegriffs in der „Proportionalität“[6] von Fähigkeiten und Schwierigkeiten abgebildet wird: In der Prognose der Lösungswahrscheinlichkeit wird die „Proportionalität“ der beiden Komponenten zum Ausdruck gebracht. So gesehen sind die Parameter nichts anderes als „Proportionalitätsmaße“.

Grundsätzlich besteht das formale Dilemma der Messung von Kompetenzen darin, dass von beobachteten Aufgabenlösungen ausgegangen werden muss, deren Realisierung bereits stattgefunden hat, die also der Vergangenheit angehören, wohingegen der Sinn des Kompetenzbegriffs nicht auf vergangene, sondern auf künftige Ereignisse gerichtet ist. Die Auflösung dieses Dilemmas besteht in der Festlegung von Proportionalitätsbedingungen und damit in der Festlegung eines formalen Prinzips der Erfüllung von Restriktionen, deren Geltung fortgeschrieben wird. Die Form, in der die Fortschreibbarkeit der Proportionalität von Fähigkeiten und Schwierigkeiten als allgemein abgesichert gilt, ist die der Schätzung: Personen werden zur Schätzung von Itemschwierigkeiten gebraucht. Entsprechendes gilt für die Schätzung der Personenfähigkeiten durch Items.

Dieses conjoint measurement (***Andrich*** 1988, ***Fischer*** 1974) im gemeinsamen Datenpool bindet die Reichweite der gemessenen Fähigkeiten an die durch die Items begrenzte Domäne. Items, die einer Domäne angehören, kann man grundsätzlich durch andere Items ersetzen, die das Kriterium der Zugehörigkeit ebenfalls erfüllen. Sobald man sich vorstellt, man würde ein Item aus der Domäne herauslösen und einer anderen zuführen, auf deren Struktur es nicht passt, der es also im Sinne der ***Rasch***-Modell-Bildung nicht angehört, würde es seinen Informationswert für die Festlegung der Personenfähigkeit verlieren. Es würde m. a. W. keine Rückschlüsse auf Kompetenz mehr zulassen. Schätzungen im Rahmen der IRT sind also auf den Geltungsbereich der gestellten Aufgaben (Items) beschränkt. Es sind Schätzungen im Rahmen des Geltungsbereichs einer Domäne: Von der Domäne der beobachteten Itemlösungen ausgehend werden Personenfähigkeiten und Itemschwierigkeiten geschätzt und in der anwendenden Bildungsforschung als Kompetenzen bzw. Niveaustufen beschrieben, vgl. ***Neubrand*** et al. (2002).

Wenn wir im Folgenden die beobachtete Liste x der Itemlösungen als die Observable bezeichnen, dann stellen wir darauf ab, dass weder die Itemschwierigkeit

6 Im gegebenen Zusammenhang setzen wir Anführungszeichen, weil nicht die Proportionalität eines linearen Zusammenhangs gemeint ist, sondern die allgemeinere Vorstellung, dass das Eine monoton vom Anderen abhängt.

noch die Personenfähigkeit direkt beobachtbar sind. Es handelt sich in beiden Fällen um Abstraktionen: Kompetenzen im Sinne der formalen Logik der IRT stellen eine Form der abstrahierenden Klassifikation der durch sie eingeschlossenen Aufgabenmenge (Domäne) dar. Umgekehrt stellen Aufgabenschwierigkeiten lediglich eine Form der abstrahierenden Klassifikation der durch sie eingeschlossenen Kompetenzen dar.

Diese Feststellung führt an sich noch nicht zu einer kritischen Beurteilung der IRT, sondern dient lediglich der begrifflichen Präzisierung: Die Reichweite der gemessenen Kompetenzen bzw. Schwierigkeiten ist auf die in der Domäne vertretenen Aufgabenmerkmale beschränkt. Gerade deshalb ist die besondere Leistung der IRT hervorzuheben: Sie ermöglicht es, durch großflächige Erhebungen geschätzte Itemschwierigkeiten zu verwenden, um die Domäne in niveaugeordnete Aufgabenstellungen zu untergliedern. Auf diese Weise werden Anschlussuntersuchungen in die Lage versetzt, mit materiell an Probanden ausgegebenen unterschiedlichen Aufgabenstellungen das Gleiche zu messen. Die Werte der Parameter der IRT stellen einen Tauschwert dar, der geldähnlich funktioniert: Wie kann man Menschen, denen ein Reihenhaus gehört, mit solchen vergleichen, die sieben Rolex Uhren besitzen? Auf der Ebene der Observablen kann man sie eigentlich gar nicht vergleichen. Das ist nur durch die Transformation in Geldwerte möglich, die man aber am Objekt selbst nicht beobachten kann (***Luhmann*** 1989, S. 14ff.[7]) Wie kann man Menschen, die den cosinus ausrechnen, mit solchen vergleichen, die quadratische Gleichungen lösen können? Auf der Ebene der Observablen muss man die gleiche Antwort geben: Eigentlich nicht. Die Äquivalenz zum Geld wird durch die Item- und Personenparameter hergestellt. Sie sind der Tauschwert, der von der Materialität – man könnte auch sagen: von der Wirklichkeitsverhaftung – der Aufgaben wie auch der Personen abstrahiert und dadurch die Vergleichbarkeit begründet.

Mit dem Ausschluss uniformer Antworten aus dem statistischen Verfahren wird die in einem Datensatz vorhandene Information beschnitten. Nicht der Karikatur, sondern der Zuspitzung wegen wählen wir hierzu eine Veranschaulichung. Nehmen wir einen Test an, der die Fortbewegung von Personen durch zwei Items prüft, nämlich ob sie laufen können und ob sie Rad fahren können. Wenn die uniform antwortenden Personen ausgeschieden werden, dann verbleiben nur Personen, die laufen aber

7 Vgl. auch ***Simmel, G.*** (1958): Philosophie des Geldes, 6. Aufl., unveränderter Nachdruck der 5. Aufl.1930, Duncker & Humblot, Berlin. Siehe darin die Hinweise auf den spezifischen Informationsverlust von Zahlungen gegenüber dem Substrat, für das gezahlt wird. Die skalierende Bedeutung von Zahlungen wird auf diese Weise „erkauft". Diese Stelle liefert die Parallele.

nicht Rad fahren und umgekehrt Personen, die Rad fahren aber nicht laufen können. Dann ist also die Herstellung der Proportionalität, mit der die IRT Kompetenzen identifiziert (s. o.) nur durch den Vergleich dieser beiden Gruppen möglich. Personen, die beides können, tragen innerhalb der IRT nicht zum Informationsgewinn bei. Mit diesem Personenkreis kann die IRT in ihrer Grundstruktur nichts anfangen. Dass man mit diesen Personen im Alltag hingegen sehr viel anfangen kann, dürfte evident sein. Allgemein ist die Grundstruktur der IRT nicht anwendbar auf Personen, die uniform antworten, d.h. alle Fragen mit Ja oder alle Fragen mit Nein beantworten. Dies wurde als methodisches Problem angesehen, was man u. a. daran erkennen kann, dass uniforme Antwortmuster zum Ausgangspunkt eines Korrekturverfahrens gemacht worden sind, siehe Abschnitt 4 und ***Rost*** (2004, S. 313f.)

Die angesprochene Problematik hat im Rahmen der IRT Weiterentwicklungen der Parameterschätzung ausgelöst. Gleichwohl haben sich die Probleme nicht rückstandslos beseitigen lassen. Vor allem in den Bereichen der sehr leichten und der sehr schwierigen Aufgaben bzw. der sehr leistungsschwachen und der sehr leistungsstarken Personen gibt es eine nach wie vor strukturelle Genauigkeitsschwäche der Schätzungen. Auf diese in der Literatur bekannte Thematik beziehen wir uns im Folgenden. Wir zeigen, dass diese Schwäche nicht mit der IRT selbst, nicht mit dem ***Rasch***-Modell selbst, zusammenhängt sondern aus der item response Funktion resultiert, mit der die IRT umgesetzt wird. Dies wird in der vorhandenen Literatur nicht fokussiert. Vielmehr wird die logistische Funktion, die auch ***Rasch*** selbst gewählt hat, allgemein akzeptiert. Wir zeigen, dass die logistische Funktion zu mehr (korrigierender) Theorie führt als man eigentlich braucht.

3 Das logistische Rasch-Modell

Einer gegebenen Person werden M_D Fragen gestellt. Die Fragen seien einfache Alternativen. D.h. die Antwort auf die l-te Frage kann nur lauten $x_l = 1$ oder $x_l = 0$; das entspricht Ja oder Nein. Diese Antworten kann man nach Durchführen des Tests in richtig und falsch transformieren. Es sei λ_l die Wahrscheinlichkeit, dass die Antwort $x_l = 1$ gegeben wird. Dann ist $1 - \lambda_l$ die Wahrscheinlichkeit, dass die Antwort $x_l = 0$ lautet. Diese Verteilung lautet also

$$q(x_l) = \lambda_l^{x_l}(1-\lambda_l)^{1-x_l}\,; \tag{1}$$

sie ist als Binomialverteilung bekannt.

Die Wahrscheinlichkeit, dass die Person die Liste

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_{M_D} \end{pmatrix} \tag{2}$$

von Antworten auf die M_D Fragen gibt, ist durch das Produkt $p(x) = \prod_{l=1}^{M_D} q(x_l)$ der Binomialverteilungen gegeben.

Die Wahrscheinlichkeit λ_l hänge von einem Parameter $\tilde{\xi}$ ab, der für die Person charakteristisch ist. Wenn der Kontext es erlaubt, wird $\tilde{\xi}$ als Fähigkeit der Person interpretiert. Wir schreiben die Gleichung (1) daher genauer in der Form

$$q(x_l|\tilde{\xi}) = \left[\lambda_l(\tilde{\xi})\right]^{x_l} \left[1 - \lambda_l(\tilde{\xi})\right]^{1-x_l} . \tag{3}$$

dann hängt auch die Verbundverteilung über alle Fragen

$$p(x|\tilde{\xi}) = \prod_{l=1}^{M_D} q(x_l|\tilde{\xi}) . \tag{4}$$

parametrisch von $\tilde{\xi}$ ab. Der Parameter $\tilde{\xi}$ ist zu erschließen aus den Antworten x, den „Observablen“. Ein mathematischer Ausdruck für die Verteilung der Observablen – gegeben ein Parameter – wird ein statistisches Modell genannt. Um das von ***Rasch*** (1980) geprägte Modell der IRT zu definieren, muss nur noch die item response Funktion $\lambda = \lambda(\tilde{\xi})$ festgelegt werden. Gebräuchliche Einführungen wurden z. B. von ***Andrich*** (1988) und ***Rost*** (2004) geschrieben. Die für das ***Rasch***-Modell grundlegende Binomialverteilung (1) wird z. B. in den Kapiteln 5.1 und 11.3 von ***Harney*** (2003) diskutiert.

Bei der ursprünglichen Formulierung des ***Rasch***-Modells wurde die item response Funktion

$$\lambda(\tilde{\xi}) = \frac{\exp(\xi)}{1 + \exp(\tilde{\xi})} , \quad -\infty < \xi < \infty , \tag{5}$$

gewählt. Sie erhielt den Namen logistische Funktion; das motiviert hier die Wahl des Buchstabens λ. Es wurde versucht, diese Funktion als die Natürliche oder zwingend Nötige zu erweisen, siehe ***Andrich*** (1988, Kapitel 2 und 3) und ***Rost*** (2004, S. 116ff.). Mit wachsendem $\tilde{\xi}$ steigt die logistische Funktion monoton von $\lambda = 0$

nach $\lambda = 1$, was intuitiv einleuchtet: Je ausgeprägter die Kompetenz einer Person ist desto wahrscheinlicher wird die erfolgreiche Lösung einer Aufgabe. Der Definitionsbereich von $\tilde{\xi}$ umfasst die ganze reelle Achse. Er scheint also eine unendliche Ausdehnung zu haben.

Hier setzt die Kritik der vorliegenden Arbeit an. Man erwartet, dass die durch Fragebögen gemessenen Kompetenzen weder unbegrenzt steigerungs- noch unbegrenzt abbaufähig sind. Entweder gibt es sie dann irgendwann gar nicht mehr – dann kann man sie auch nicht weiter abbauen – oder sie erreichen ihr Wachstumslimit in der Endlichkeit der Aufgaben, an denen sie sich bewähren können. Diese Endlichkeit besteht darin, dass die Observable aus den Antworten auf eine endliche Liste einfacher Alternativen besteht – dass sie also durch eine endliche Liste der Einträge Null oder Eins darstellbar ist. Der Informationsgehalt einer solchen Observablen ist endlich. Dieser Sachverhalt wird durch die logistischen Parameter $\tilde{\xi}$ jedoch nicht abgebildet. In der logistischen Form kann die Kompetenz ins Unendliche wachsen; so wie es bei den Ergebnissen von Längenmessungen im Prinzip der Fall ist. Die Literatur hält dies für eine vernachlässigbare, eher pedantische Thematik. Wir zeigen in den Abschnitten 4 und 6, dass das nicht der Fall ist.

Die Fragen, die man der getesteten Person stellt, haben charakteristische Schwierigkeiten. Auch die Schwierigkeiten $\tilde{\delta}_l$ der Fragen müssen als Parameter in das Modell eingeführt werden. Dazu ersetzt man $\tilde{\xi}$ durch die Differenz $\tilde{\beta} - \tilde{\delta}_l$ des Personenparameters $\tilde{\beta}$ und des Fragenparameters $\tilde{\delta}_l$. Indem man diese Differenz benutzt, sucht man das Prinzip der spezifischen Objektivität (***Fischer*** 1974 und 1988) zu beachten. Wir definieren dies Prinzip im Abschnitt 5 und zeigen dort, dass die logistische Form des ***Rasch***-Modells nicht spezifisch objektiv ist.

Es gibt kein anderes Messgerät für die Fragenparameter als die Art von Befragung, die auch zum Abschätzen der Personenparameter verwendet wird. Tatsächlich gelingt es, die Fragenparameter zusammen mit den Personenparametern zu ermitteln, indem man nicht nur *eine* Person sondern eine ausreichende Zahl M_B von Personen befragt. Die k-te Person, $k = 1,..., M_B$, habe den Personenparameter $\tilde{\beta}_k$. Die Antworten müssen nun doppelt indiziert werden: Es sei x_{kl} die Antwort der k-ten Person auf die l-te Frage, $k = 1,..., M_B$, $l = 1,..., M_D$,. Das Modell hängt dann von der Liste

$$\tilde{\delta} = \begin{pmatrix} \tilde{\delta}_1 \\ \vdots \\ \tilde{\delta}_{M_D} \end{pmatrix} \tag{6}$$

der Fragenparameter sowie von der Liste

$$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_1 \\ \vdots \\ \tilde{\beta}_{M_B} \end{pmatrix} \tag{7}$$

der Personenparameter ab. Die Liste der Observablen ist

$$x = (x_{kl})_{k=1\ldots M_B, l=1\ldots M_D} \,. \tag{8}$$

Auch λ ist nun doppelt zu indizieren, und die Gleichung (5) wird verallgemeinert zu

$$\lambda_{kl} = \frac{\exp(\tilde{\beta}_k - \tilde{\delta}_l)}{1 + \exp(\tilde{\beta}_k - \tilde{\delta}_l)} \,. \tag{9}$$

Das Binomialmodell (1) wird verallgemeinert zu

$$q(x_{kl}|\tilde{\beta}_k, \tilde{\delta}_l) = \lambda_{kl}^{x_{kl}} (1 - \lambda_{kl})^{1 - x_{kl}} \,; \tag{10}$$

und damit lautet das vollständige Modell

$$p(x|\tilde{\beta}, \tilde{\delta}) = \prod_{k=1}^{M_B} \prod_{l=1}^{M_D} q(x|\tilde{\beta}_k, \tilde{\delta}_l) \,. \tag{11}$$

Tabelle 1 Die Matrix der x_{kl} des einfachen Beispiels im Text

	Frage 1 Laufen	Frage 2 Rad fahren
Person 1	1	0
Person 2	0	1

Wenn die Zahl $M_B \cdot M_D$ größer ist als $M_B + M_D - 1$ dann erlaubt das ***Rasch***-Modell (11) grundsätzlich, alle Parameter zu schätzen bis auf einen, der als Referenzgröße willkürlich gewählt wird. Zur sinngemäßen Veranschaulichung: Gegeben sei der schon erwähnte Test, der $M_B = 2$ Personen $M_D = 2$ Fragen stellt: Ob sie laufen und ob sie Rad fahren können. Mögliche Antworten x_{kl} sind in der Tabelle 1 dargestellt. Die eine Person kann laufen aber nicht Rad fahren; für die andere ist es umgekehrt. Wegen der Symmetrie der Matrix der Antworten wird man vermuten, dass die Personenparameter einander gleich sind und dass die Fragenparameter einander gleich

sind. (Das ergibt sich tatsächlich aus den Gleichungen (26, 27)). Die numerischen Werte der Parameter liegen damit aber noch nicht fest. Denn wenn man alle Parameter des Modells (11) um denselben Wert verschiebt, dann ändert sich das Modell nicht; es hängt nur von Differenzen seiner Parameter ab. Man muss also einen der Parameter willkürlich festlegen. Durch die Konvention

$$\tilde{\delta}_{M_D} = 0 \tag{12}$$

machen wir das ***Rasch***-Modell benutzbar.[8] Im obigen Beispiel ergeben sich dann die Schätzer[9] $\widetilde{\beta}_1^{\text{opt}} = \widetilde{\beta}_2^{\text{opt}} = \widetilde{\delta}_1^{\text{opt}} = \widetilde{\delta}_2^{\text{opt}} = 0$. Die Parallele zu den Maßeinheiten der Naturwissenschaften – z. B. der Festlegung der Celsius- oder der Fahrenheit-Skala der Temperatur – ist offensichtlich. Auch hier wird ein Standard gewählt, der zur Vergleichbarkeit der gemessenen Differenzen benötigt wird.

Das logistische ***Rasch***-Modell hängt nicht von dem Muster der Antworten x_{kl} einer jeden Person k ab; vielmehr gehen die Observablen nur über die „scores“

$$s_k = \sum_{l=1}^{M_D} x_{kl} \tag{13}$$

und „Erträge“

$$t_l = \sum_{k=1}^{M_B} x_{kl} \tag{14}$$

in das Modell ein. Der score s_k ist die Zahl aller von der k-ten Person mit Ja oder richtig beantworteten Fragen. Der Ertrag t_l ist die Zahl der Personen, die die l-te Frage mit Ja oder richtig beantworten. Dass das Modell nur von den scores und den Erträgen abhängt, erkennt man, indem man es auf die Form

$$p(x|\tilde{\beta},\tilde{\delta}) = \exp\left(\sum_{k=1}^{M_B} s_k\tilde{\beta}_k - \sum_{l=1}^{M_D} t_l\tilde{\delta}_l\right)\prod_{k,l}\left(1+\exp(\tilde{\beta}_k - \tilde{\delta}_l)\right)^{-1} \tag{15}$$

bringt. Die Rechnung, die von den Gleichungen (9, 10, 11) dahin führt, ist nicht schwierig. Sie ist bei ***Rost*** (2004) auf S. 123f. angegeben. Im Rahmen des im nächsten Abschnitt eingeführten trigonometrischen ***Rasch***-Modells existiert eine vergleich-

8 Oft wird auch festgesetzt, dass die Summe aller Fragenparameter gleich Null ist.

9 Es handelt sich um die in Abschnitt 6 definierten maximum likelihood Schätzer.

bare Darstellung nicht. Dort werden die Parameter nicht allein aus den scores und den Erträgen bestimmt. Die Parameter hängen dort nicht nur davon ab wie viele Fragen, sondern auch davon, *welche* Fragen eine Person richtig beantwortet hat.

Trotz der im vorliegenden Aufsatz formulierten Kritik sind wir uns ***Rasch***s wissenschaftstheoretischer Leistung bewusst. ***Rasch*** (1980) hat ein statistisches Modell des Messens schlechthin geschaffen: An seinem Modell kann man studieren wie es möglich ist, in ein- und demselben Rahmen sowohl zu messen als auch die Messgeräte zu eichen.

4 Das trigonometrische Rasch-Modell

Das im letzten Abschnitt beschriebene Modell hat auf Grund seiner logistischen Form die Eigenart, dass die Werte $\lambda_{kl} = 1$ und $\lambda_{kl} = 0$ nur asymptotisch erreicht werden, nämlich für $\widetilde{\beta}_k - \widetilde{\delta}_l \to \pm\infty$. Die Werte 0 und 1 markieren das Zustandekommen von ceiling- und floor-Effekten. Beide resultieren aus der Annäherung der Parameter an die Grenzen, zwischen denen beide, die Item- und die Personenparameter, liegen, siehe ***Andrich*** (1988, S. 58). Die Situation, in der die Person *k* die Frage *l* mit deterministischer Sicherheit richtig (ceiling) bzw. falsch (floor) beantwortet, kommt im Definitionsbereich der Parameter nicht vor. Dies Problem hat zwar Korrekturvorschläge für die Schätzer ausgelöst; zur systematischen Beseitigung des Problems ist es jedoch bisher nicht gekommen. Der Definitionsbereich der Parameter bildet nicht alle Daten ab, die möglich sind. Anschaulich kann man von einem durch die Ränder begrenzten Tunnelblick sprechen, den die Parameter auf die Kompetenz- und Schwierigkeitsniveaus werfen. Das logistische ***Rasch***-Modell vermittelt nämlich nicht zwischen deterministischer und probabilistischer Auffassung des Verhaltens von Personen. Es ist rein probabilistisch. Es kennt keine Personen, die auf irgendeine Frage mit Sicherheit Ja oder mit Sicherheit Nein antworten.

Auf den ersten Blick erscheint diese Feststellung als Pedanterie. Denn der Definitionsbereich der Parameter lässt eine beliebige Annäherung an $\lambda_{kl} = 1$ und $\lambda_{kl} = 0$ zu. Es stellt sich aber heraus, dass die Ausklammerung dieser beiden Werte eine keineswegs triviale sondern – im Gegenteil – substanziell bedeutsame methodische Restriktion für die Deutung der empirischen Sachverhalte darstellt. Es kann vorkommen, dass eine Person alle M_D Fragen des Tests mit Ja beantwortet. In diesem Fall existiert der – in Abschnitt 6 definierte – bestmögliche Wert der logistischen

Rasch-Parameter nicht. Es war und ist üblich, solche Antwortbögen zu verwerfen[10], siehe Kapitel 4.2 bei ***Rost*** (2004). Man kann der Meinung sein, dass keine Person unbegrenzte Fähigkeiten haben kann. Aber Fragebögen sind endlich; und deswegen kommt es vor, dass ein Fragebogen uniform, d.h. vollständig richtig oder vollständig falsch, beantwortet wird – ehrlich. Ihn auszuscheiden heißt, Beobachtungen zu verwerfen, die vom Modell zugelassen sind. Das Modell der wechselseitigen Messbarkeit von Kompetenzen und Aufgabenschwierigkeiten schließt die Möglichkeit der uniformen Beantwortung der Fragen weder logisch noch inhaltlich aus.

Wir stoßen an dieser Stelle auf ein Problem des Beobachtungszusammenhangs: Beschränkt man die Beobachtung auf die Endlichkeit eines Tests, wie in den PISA-Untersuchungen, dann erweisen sich Probanden, die alles oder nichts lösen, als solche, denen der deterministische Grenzfall zugerechnet werden kann, also eine Wahrscheinlichkeit von $\lambda_{kl} = 0$ und $\lambda_{kl} = 1$. Dass das beobachtete Ergebnis auch durch eine zufällige Abweichung von der wahren Fähigkeit zustande kommen kann, ist kein Gegenargument an dieser Stelle, sondern unterstreicht vielmehr die Notwendigkeit, Personen- und Aufgabenparameter auch im Fall eines uniformen Antwortmusters jeweils zu schätzen und damit die Zugehörigkeit uniformer Antworten zum ***Rasch***-Modell statistisch abzubilden. Lösungswahrscheinlichkeiten von $\lambda_{kl} = 0$ und $\lambda_{kl} = 1$ – den deterministischen Grenzen der Wahrscheinlichkeitslogik – werden deshalb erreicht, weil die Menge der formulierten und formulierbaren Aufgaben – sogar bei PISA – endlich ist.

Daher hat auch ein uniformes Antwortmuster eine geltende Erwartbarkeit. Der Determinismus des uniformen Antwortmusters ist ja selber ein Ereignis, das auf nichtdeterministische Weise zustande kommen kann. Die verschiedenen Varianten der IRT können die uniformen Antworten nicht zwanglos einbeziehen.

Wenn die zu den uniformen Beobachtungen am Besten passenden Werte der Parameter nicht definiert sind, ist es außerordentlich schwierig, Fehlerintervalle für die

10 Das Programm STATA (siehe unter http://www.stata.com/support/faqs/stat/rasch.html) meldet das dem Benutzer durch die Warnung „... has dropped xx subjects from the analysis. These subjects responded either to all items correctly or to all items incorrectly; in a conditional likelihood, these subjects carry no information about the difficulty of the items“. Das Programm WINMIRA gibt den Hinweis: Die Fragenparameter „are generated based on CML estimates (Conditional-Maximum-Likelihood)“. Die Methode der CML wird in ***Rost*** (2004, S. 126) definiert und besteht darin, die Likelihood Funktion der dortigen Gleichung (13) zu maximieren. Wenn der score verschwindet oder seinen maximalen Wert annimmt, dann hängt diese Funktion von den Fragenparametern aber nicht ab. Unter diesen Umständen kommt man auch mit einem Korrekturverfahren – wie z.B. ***Warm***'s WLE – strukturell nicht wirklich weiter (was von den WINMIRA-Autoren ***Rost*** und ***Davier*** auch nicht behauptet wird); siehe auch ***Bühner*** (2006, S. 384f.).

Parameter anzugeben. Ferner: Wenn man sich entschließt, uniform beantwortete Fragebögen zu verwerfen, wo hört man auf? Verwirft man auch Fragebögen, die mit der Ausnahme *einer* Frage uniform beantwortet wurden? Das Problem ist bekannt. Es wird in der Literatur breit erörtert. Allerdings ist diese Erörterung weniger von der Suche nach einer alternativen Form gekennzeichnet als vielmehr davon, die erkennbare Schwäche des logistischen Modells korrigierend zu behandeln.

Neben dem Verwerfen von Antwortbögen wurde die Lösung darin gesucht, die von ***Fisher*** (1925, S. 700-725) gegebene und gut begründete Definition des Schätzers – also das Prinzip der maximum likelihood – zu modifizieren, siehe Anhang B sowie Kapitel 4 bei ***Rost*** (2004). Mit Hilfe des Prinzips der maximum likelihood werden die Parameter geschätzt. Die Daten werden als von den Parametern bedingt aufgefasst; und die Parameter werden dadurch geschätzt, dass man die Erwartbarkeit der gegebenen Daten maximal macht. Das seit den achtziger Jahren weithin benutzte Verfahren besteht darin, die likelihood Funktion mit einer geeigneten Gewichtung zu maximieren. Diese weighted likelihood estimation (WLE) ist als Standardsoftware verfügbar, siehe Anhang B und ***Rost*** (2004, S. 313). Mit der WLE ist die Uniformität von Antwortmustern allein kein Grund mehr, diese außer Betracht zu lassen. Unsere Auffassung ist dennoch, dass die Beibehaltung der logistischen item response Funktion zwar technische Fortschritte wie die der WLE von großer Bedeutung – z.B. für die Anwendung in den PISA-Studien – ermöglicht, aber keinen strukturellen Fortschritt für das Schätzen der Parameter gebracht hat. Das Problem wird zwar verringert aber nicht gelöst.

Man kann die Schwierigkeiten jedoch durch eine Form der item response Funktion beheben, welche die Werte Null und Eins wirklich annimmt. Das heißt, die strukturelle Antwort auf das Problem sehen wir in einer Parameterbildung, die die logistische Funktion ersetzt, so dass die uniformen Antwortmuster durch endliche Schätzer interpretiert werden. Wir schlagen vor, an die Stelle der logistischen Funktion (5, 9) den Ansatz

$$\begin{aligned} \tau(\xi) &= \sin^2 \xi\,, \\ \tau_{kl} &= \sin^2(\beta_k - \delta_l) \end{aligned} \tag{16}$$

treten zu lassen. Dies nennen wir die trigonometrische Form, und das motiviert die Wahl des Buchstabens τ. Um die logistischen von den trigonometrischen Parametern zu unterscheiden, wird an den Letzteren die Tilde weggelassen. In Abschnitt 5 wird gezeigt, dass die Form (16) auf ein spezifisch objektives Modell führt.

Die Funktion (16) ist im Ansatz schon vor längerer Zeit von ***Samejima*** (1983)[11] vorgeschlagen worden. Der Autor versuchte, die item response Funktion aus der Forderung herzuleiten, dass die ***Fisher***-Information (40) unabhängig von den Parametern sein solle. Dies wird durch die trigonometrische Parametrisierung geleistet, siehe Anhang B. ***Samejima***s Vorschlag hat in der weiteren Entwicklung der IRT keine Rolle gespielt. Zwei Gründe sehen wir dafür. Erstens macht ***Samejima*** die Informationsfunktion nur stückweise konstant, denn er lässt sie identisch verschwinden auf einem Teil des Definitionsbereichs der Parameter. Dort hängt das Modell von den Parametern nicht ab und ist daher kein sinnvolles Modell, siehe Kapitel 9.3 von ***Harney*** (2003). Zweitens wurde durch die Arbeiten von ***Fischer*** (1988) nahe gelegt, dass die logistische Parametrisierung des ***Rasch***-Modells die einzige sei, die wissenschaftstheoretisch befriedige, weil nur sie das Postulat der spezifischen Objektivität erfülle. Dies trifft jedoch nicht zu. Darauf gehen wir im Abschnitt 5 ein.

Da das Quadrat der sinus-Funktion die Periode π hat, werden die Parameter β, δ modulo π definiert. Dies liefert eine eindeutige Rechenvorschrift beim Bilden von Linearkombinationen der Parameter. Die τ_{kl} hängen dann stetig und differenzierbar von den Parametern ab. Man kann sich nämlich den Anfang und das Ende des Parameterbereichs zu einer ringförmigen Topologie zusammengefügt denken; dann besteht Differenzierbarkeit überall. Allerdings durchläuft man den Wertebereich von τ_{kl} (das ist das Intervall [0,1]) zweimal, wenn einer der trigonometrischen Parameter seinen Definitionsbereich einmal durchläuft. Dennoch können die Parameter alle eindeutig aus den Observablen x erschlossen werden (bis auf die sogleich besprochene Spiegelungssymmetrie). Die Parameter werden eindeutig bestimmt, weil die Observablen von allen Differenzen der Personen- und Fragenparameter abhängen. Daher ist – sagen wir – β_1 nicht äquivalent zu $\pi-\beta_1$. Nimmt man diese Transformation allerdings an allen Parametern β, δ vor, dann bleibt die trigonometrische Form unverändert; d. h. dann erhält man äquivalente Parameter. Da die Parameter modulo π definiert sind, kann man diese Transformation einfacher als den Übergang aller β_k , δ_l zu $\beta'_k = -\beta_k$, $\delta'_l = -\delta_l$ beschreiben. Wir nennen diese Transformation die Spiegelung. Unter der Spiegelung bleiben alle τ_{kl} unverändert. Daher gelten alle Schlüsse für β, δ auch für die gespiegelten Parameter β', δ'. Diese beschreiben also dieselbe Situation. Um die Parameter eindeutig zu machen, wird in Abschnitt 6.2 eine Konvention getroffen.

11 Siehe auch die Literaturhinweise darin.

Die Binomialverteilung q lautet nun

$$\begin{aligned} q(x_{kl}|\beta_k,\delta_l) &= \tau_{kl}^{x_{kl}}(1-\tau_{kl})^{1-x_{kl}} \\ &= \left[\sin^2(\beta_k-\delta_l)\right]^{x_{kl}}\left[\cos^2(\beta_k-\delta_l)\right]^{1-x_{kl}}, \end{aligned} \tag{17}$$

wobei wir 1 - τ_{kl} durch die cosinus-Funktion ausgedrückt haben. In die Formel

$$p(x|\beta,\delta)=\prod_{k=1}^{M_B}\prod_{l=1}^{M_D} q(x_{kl}|\beta_k,\delta_l)\,. \tag{18}$$

für das ***Rasch***-Modell ist nun der Ausdruck (17) einzusetzen. Wiederum muss man einen Parameter willkürlich festsetzen. Ähnlich wie in Gleichung (12) verfügen wir

$$\delta_{M_D}=0\,. \tag{19}$$

Damit ist die trigonometrische Form des ***Rasch***-Modells definiert. Sie ist keine Reparametrisierung der logistischen Variante des ***Rasch***-Modells. Unter einer Reparametrisierung versteht man eine umkehrbar eindeutige Abbildung der alten Parameter auf neue; eine solche lässt die Eigenschaften des statistischen Modells unverändert. Eine Reparametrisierung führt (außer wenn sie linear ist) nicht dazu, dass eine Differenz der alten Parameter in die entsprechende Differenz der neuen Parameter übergeht.

Die logistischen Parameter hängen keineswegs linear mit den trigonometrischen zusammen. Daher ist der Übergang von der logistischen zur trigonometrischen item response Funktion keine Reparametrisierung des ursprünglichen ***Rasch***-Modells. Das trigonometrische Modell hat andere Eigenschaften als das logistische Modell – wie man in Abschnitt 6 sehen wird.

In Abschnitt 5 wird gezeigt, dass das trigonometrische ***Rasch***-Modell eine Vollständigkeit besitzt, die dem logistischen Modell ermangelt. Dazu geben wir in Abschnitt 6.2 numerische Beispiele. Zuvor greifen wir im folgenden Abschnitt die bei der Grundlegung des ***Rasch***-Modells diskutierten Fragen auf nach dem „fundamental measurement“ (***Andrich*** 1988) und der spezifischen Objektivität (***Fischer*** 1988).

5 Forminvarianz und Spezifische Objektivität

Die in den Humanwissenschaften verwendeten Messmethoden werden gelegentlich durch Analogien aus den Naturwissenschaften gerechtfertigt. Für das ***Rasch***-Modell sind Analogien zur Mechanik verbreitet (***Rasch*** 1972 und ***Andrich*** 1988). Nach

Newton[12] ist die Beschleunigung eines Körpers proportional zur wirkenden Kraft und umgekehrt proportional zur trägen Masse des Körpers. Dies kann nahe legen, die exponentielle Fähigkeit $\exp\widetilde{\beta}_k$ einer Person in Analogie zur Kraft und die Schwierigkeit $\exp\widetilde{\delta}_l$ einer Frage in Analogie zur trägen Masse zu interpretieren. Daraus zieht ***Andrich*** (1988) den Schluss, dass das Verhältnis von Fähigkeit und Schwierigkeit die Wahrscheinlichkeit λ_{kl} der Antwort bestimmen müsse. Implizit bedeutet das: Der Definitionsbereich eines jeden der Parameter $\widetilde{\beta},\widetilde{\delta}$ ist die reelle Achse, ist also unendlich ausgedehnt. Der unendlich ausgedehnte Definitionsbereich der Parameter $\widetilde{\beta},\widetilde{\delta}$ ist in dem Bild des von ***Galilei***[13] untersuchten freien Falls enthalten. Denn dieser findet im Prinzip auf einer Geraden im euklidischen Raum statt.

Gibt die Physik also einen allgemein gültigen Hinweis darauf, welches die richtige Wahl ist? Das ist nicht der Fall. Beide Typen von Parametern – zyklische und euklidische – kommen vor. Welcher Typ vorliegt, ergibt sich aus den Eigenschaften des statistischen Modells $p(x|\xi)$, das die Verknüpfung zwischen den beobachteten Daten x und den daraus zu erschließenden Parametern ξ festlegt. Genauer gesagt: Welcher Typ von Parameter vorliegt, ergibt sich aus der Symmetrie, die ein statistisches Modell besitzen muss, um (i) vollständig zu sein und (ii) unvoreingenommene Schlüsse zu ermöglichen. In ***Harney*** (2003) wurde diese Symmetrie als Forminvarianz bezeichnet. Bei dem hier vorliegenden Modell, wo die Daten nach einem gemeinsamen Maß die Personenparameter und die Fragenparameter festlegen, führt die Eigenschaft der Forminvarianz gerade auf die Eigenschaft der spezifischen Objektivität, die in ***Fischer*** (1988) gefordert und diskutiert wurde. Wir geben eine kurze Schilderung von Forminvarianz und von spezifischer Objektivität. Außerdem geben wir zwei Beispiele von spezifischer Objektivität. Das eine Beispiel ist die Messung von Längen bei gleichzeitiger Eichung der Maßstäbe. Längen und Maßstäbe sind euklidische Parameter. Das andere Beispiel ist der Gegenstand des vorliegenden Aufsatzes – das trigonometrische ***Rasch***-Modell. Dessen Parameter sind zyklisch.

12 Sir ***Isaac Newton*** (1643–1727) begründete durch seine Philosophiae naturalis principia mathematica (1686) die klassische Mechanik.

13 ***Galileo Galilei*** (1564–1642) war von 1589 bis 1592 Professor an der Universität Pisa. Eine Legende der Wissenschaft erzählt, dass er die Bestätigung des von ihm theoretisch aufgestellten Fallgesetzes in Experimenten am schiefen Turm von Pisa gesucht habe. Die in Padua durchgeführten Laborversuche an der von ihm erfundenen Fallrinne waren nicht so augenfällig – aber erfolgreich.

5.1 Zur Forminvarianz

Ein statistisches Modell $p(x|\xi)$ ist eine Verteilung der Observablen x, welche parametrisch von ξ abhängt. Sie ist für jedes ξ auf 1 normiert. Der Übergang von einem Wert ξ des Parameters zu einem anderen Wert ξ' entspricht einer Transformation der Wahrscheinlichkeiten. Man kann einem willkürlich herausgegriffenen Parameter den Wert 0 zuweisen. Dann entstehen die $p(x|\xi)$ dadurch, dass man einen Satz von Transformationen $\mathbf{G}_\xi$ auf die herausgegriffene Verteilung $p(x|\xi = 0)$ anwendet. Wenn diese Transformationen im mathematischen Sinn eine Gruppe bilden, dann ist das Modell p forminvariant. Die Existenz einer solchen Gruppe bedeutet, dass das Modell eine Symmetrie besitzt.

Es ist nicht möglich, im vorliegenden Aufsatz eine wirklich präzise, alle Feinheiten erfassende Definition dieser Symmetrie zu geben. Sie wurde in der Monographie von ***Harney*** (2003) gegeben. Wir hoffen aber, mit den folgenden Beispielen zu illustrieren, was gemeint ist.

Das ***Gauß***sche Modell

$$w(x|\xi) = (2\pi)^{-1/2} \exp\left(-(x-\xi)^2/2\right) . \tag{20}$$

hat eine Observable x, die jeden Wert auf der reellen Achse annehmen kann. Man denke an die Messung einer Länge. Der Parameter ξ ist die tatsächliche Länge des vermessenen Objekts. Auch ξ ist auf der ganzen reellen Achse definiert. Denn ξ muss der Observablen folgen können. Die sämtlichen Verteilungen dieses Modells entstehen dadurch, dass man x um $-\xi$ verschiebt. Eine solche Translation ist eine Transformation $\mathbf{G}_\xi$. Die Menge aller Translationen bildet eine mathematische Gruppe. Durch die Anwendung aller Transformationen der Gruppe auf die Verteilung mit $\xi = 0$ entstehen genau die Verteilungen dieses Modells. Die Menge der Transformationen muss eine Gruppe bilden, damit jede mögliche Observable x interpretierbar ist. Wenn man irgendeinen Wert – sagen wir $\xi = 1$ – als Parameter ausschließt, dann schließt man die Verteilung aus, die ihren wahrscheinlichsten Wert bei $x = 1$ hat, welche man also als gegeben vermuten wird, wenn die Observable den Wert $x = 1$ angenommen hat. Wenn $\xi = 1$ ausgeschlossen wird, dann bilden die verbleibenden Translationen keine Gruppe mehr. Eine Gruppe liegt nur dann vor, wenn sämtliche reellen Verschiebungen vorgesehen sind. So sorgt Forminvarianz dafür, dass alle möglichen Beobachtungen interpretierbar sind, dass also das Modell p vollständig ist. – Forminvarianz sorgt aber auch dafür, dass die Information, die man erhält, für jede Beobachtung dieselbe ist. Denn alle Verteilungen des Modells (20) haben dieselbe Breite. Die Streuung ist stets gleich 1. Hinge die

Streuung von ξ ab, so gingen die Verteilungen nicht durch Translationen auseinander hervor, und die Transformationen bildeten wiederum keine mathematische Gruppe. Da die Information, die man erhält, unabhängig vom Schätzer des Parameters ist, wird der Wert von ξ unvoreingenommen erschlossen.

Wir wenden diese Gedanken auf das im vorliegenden Aufsatz diskutierte Binomialmodell

$$q(x|\xi) = \left[\sin^2(\xi)\right]^x \left[\cos^2(\xi)\right]^{1-x} \tag{21}$$

an. Die Symmetrie der trigonometrischen Form erkennt man, indem man den Vektor *a(ξ)* mit den Komponenten

$$\begin{pmatrix} a_1(\xi) \\ a_2(\xi) \end{pmatrix} = \begin{pmatrix} \sqrt{q(0|\xi)} \\ \sqrt{q(1|\xi)} \end{pmatrix} \tag{22}$$

bildet. Das liefert

$$\begin{pmatrix} a_1(\xi) \\ a_2(\xi) \end{pmatrix} = \begin{pmatrix} \cos(\xi) \\ \sin(\xi) \end{pmatrix} . \tag{23}$$

Man sieht, dass der Vektor *a(ξ)* aus dem Vektor

$$a(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{24}$$

hervorgeht, indem man ihn um den Winkel ξ dreht. Die Drehungen $\mathbf{G}_\xi$ eines zweidimensionalen Vektors bilden hier die Gruppe von Transformationen, welche alle Verteilungen des binomischen Modells erzeugt (vgl. ***Harney*** 2003, Kap. 11.3). Der Definitionsbereich eines Winkels ist endlich; er kann als kreisförmig aufgefasst werden. Indem man den Drehwinkel als Parameter wählt, macht man die Gruppe der Drehungen äquivalent zur Gruppe der Translationen des Winkels.

Bezogen auf das ***Rasch***-Modell bzw. die Identifikation von Aufgabenschwierigkeiten heißt das: Die Aufgaben sind je nach Ausprägung der Personenfähigkeit auf der Skala der Personenfähigkeiten verschiebbar, ohne dass sich deren jeweilige

Schwierigkeit ändert bzw. ändern darf. Das Einzige, was sich ändert, ist die fähigkeitsbedingte Wahrscheinlichkeit der Aufgabenlösung.

Auch für das Binomialmodell garantiert die Existenz der Symmetriegruppe, dass jeder Wert der Observablen interpretierbar ist. Wenn man den Wert $\xi = \pi / 2$ ausschließt, dann schließt man aus, dass die Observable x mit Sicherheit gleich 1 ist. Es ist aber durchaus zugelassen, dass eine Serie von Beobachtungen stets $x = 1$ erbringt. Dann ist die Annahme $\xi = \pi / 2$ die optimale Interpretation der Observablen. Die Forderung, dass die Menge der $\mathbf{G}_{\xi}$ eine Gruppe bildet, sorgt dafür dass man $\xi = \pi / 2$ nicht ausschließen kann. Denn nur sämtliche Drehungen bilden eine Gruppe. – Setzt man die logistische Funktion (5) an die Stelle von $\sin^2$ in Gleichung (21), dann schließt man u. a. aus, dass die Observable mit deterministischer Sicherheit $x = 1$ ist. Es gibt daher keine Reparametrisierung – also keine eineindeutige Abbildung – des logistischen Parameters auf den trigonometrischen Parameter. Aus diesem Grunde ist das logistische ***Rasch***-Modell nicht forminvariant. Es ist daher nicht vollständig.

Es gibt ein Rezept dafür, die Vollständigkeit eines Modells zu prüfen. Dazu reparametrisiert man es so, dass die in Gleichung (40) angegebene a priori Verteilung – auch Informationsfunktion genannt – konstant wird. Das nennen wir die natürliche Parametrisierung. Wenn das Modell nur einen einzigen Parameter hat – so wie das Modell der Gleichung (21) – dann ist die natürliche Parametrisierung immer zu erreichen (siehe ***Harney*** 2003, Kap. 2). Wenn das Modell forminvariant ist oder forminvariant gemacht werden kann, dann ist es möglich, den Definitionsbereich des natürlichen Parameters so auszudehnen, dass er entweder die gesamte reelle Achse umfasst oder zyklisch wird. Denn in der natürlichen Parametrisierung sind die oben erwähnten Transformationen $\mathbf{G}_{\xi}$ des Modells äquivalent zu den Translationen des Parameters. Durch Wählen der natürlichen Parametrisierung findet man automatisch die trigonometrische item response Funktion. Wenn man die logistische Funktion reparametrisiert, findet man allerdings nur den Bereich $0 < \xi < \pi / 2$ des natürlichen Parameters ξ. Auf diese Weise erkennt man, dass die logistische item response Funktion keine vollständige Darstellung des binomischen Modells liefert; d.h. mit der logistischen item response Funktion ist das Modell nicht forminvariant. Diese Argumente zeigen: Es ist die Forderung nach Forminvarianz, die die item response Funktion festlegt, nicht die oben zitierte Analogie zur Mechanik.

Wir erwähnen, dass bei der Faktorenanalyse (***Wirtz*** und ***Nachtigall*** 1998, S. 209ff. sowie ***Bortz*** 1999) ein (teilweiser) Übergang zu trigonometrischen Parametern vorkommt – zu Parametern also, deren Definitionsbereich endlich ist. In der Faktorenanalyse wird eine Korrelationsmatrix durch ihre Eigenwerte und die Richtungen ihrer

Eigenvektoren parametrisiert. Die Einführung der Richtungen bedeutet einen teilweisen Übergang zu trigonometrischen Parametern. Zu methodischen Problemen der Faktorenmodellbildung vgl. ***Rohwer*** und ***Pötter*** (2001, S. 239ff.).

5.2 Zur spezifischen Objektivität

Die im letzten Abschnitt an Beispielen dargestellte Forminvarianz statistischer Modelle hängt mit dem Begriff der spezifischen Objektivität zusammen, welcher von ***Rasch*** (1967, 1977) und ***Fischer*** (1988) zur Begründung des ***Rasch***-Modells herangezogen wurde. Danach sollte das statistische Modell für die Personen- und Fragenparameter β, δ den Vergleich von Personen ermöglichen ohne dass die Schwierigkeiten der Fragen eingehen. Die Bedeutung dieses Postulats wird vielleicht deutlicher, wenn man es auf die Messung von Längen überträgt. Die Längen zweier Gegenstände sollen verglichen werden können, ohne dass die Verfahren der Längenmessungen eine Rolle spielen. Diese Eigenschaft eines Messverfahrens oder eines statistischen Modells heißt spezifische Objektivität.

Die Forderung nach spezifischer Objektivität ist nach Gleichung (3) in ***Fischer*** (1988) äquivalent zu einem Satz von Gleichungen für die Verknüpfung der Parameter. Diese können nach ***Fischer*** (1988) Abschnitt 4 gelöst werden, wenn es möglich ist, so zu parametrisieren, dass für die k-te Person die Wahrscheinlichkeit der richtigen Antwort auf die l-te Frage eine Funktion der Differenz $\xi = \beta_k - \delta_l$ ist. – Beide, das logistische ***Rasch***-Modell wie auch die trigonometrische Form des Modells, haben diese Eigenschaft. Welches Modell ist nun spezifisch objektiv?

In ***Fischer*** (1988) wird nicht explizit gesagt, dass das spezifisch objektive Modell als Funktion der Differenz ξ vollständig und unvoreingenommen – also forminvariant – sein muss. Dies wird insofern als selbstverständlich angenommen als ξ ja alle denkbaren Werte zu durchlaufen scheint, wenn es einer beliebigen Translation unterworfen wird. Unvollständigkeit bemerkt man jedoch nur, wenn man vom gegebenen Parameter zum natürlichen Parameter transformiert. Dies ist das fehlende Stück der Argumentation in ***Fischer*** (1988). In der natürlichen Parametrisierung des binomischen Modells (21) ergibt sich von selbst die trigonometrische item response Funktion, welche in Gleichung (21) bereits eingetragen wurde. Ihre Herleitung bedarf keiner zusätzlichen Annahme wie sie in Abschnitt 5 bei ***Fischer*** (1988) unter Punkt (v) – wenig plausibel – eingeführt wurde.

Zum Vergleich mit der IRT schreiben wir noch ein spezifisch objektives Modell der Längenmessungen auf. Es seien β_k, k = 1,..., M_B, die Längen von M_B Objekten. Jedes Objekt wird M_D-mal vermessen. Die Länge, die das l-te Messgerät, l = 1,..., M_D,

dem *k*-ten Objekt zuweist, sei x_{ik}. Die x_{ik} seien gemäß dem Modell (20) statistisch verteilt. Auf diese Weise kommt der statistische Messfehler hinein. Die Verbundverteilung aller $M_B \cdot M_D$ Messungen ist dann

$$p(x|\beta,\delta) = \prod_{k,l} w(x_{kl}|\beta_k - \delta_l)\,. \tag{25}$$

Jeder Faktor *w* in diesem Produkt ist forminvariant. Die ***Gauß***schen Faktoren *w* haben die Eigenschaft, dass die zugehörige a priori Verteilung (40) uniform ist. Bildet man in dieser Parametrisierung der Faktoren die Differenzen β_k - δ_l der Parameter, so ist das entstehende Modell *p* spezifisch objektiv. Das heißt, man kann Objektparameter β_k , $\beta_{k'}$ ermitteln und vergleichen, ohne die Eichung δ_l der Messgeräte zu kennen. Das folgt daraus, dass das Gesamtmodell (25) forminvariant ist – und zwar auf eine besonders angenehme Weise: Jeden der Parameter β, δ kann man ermitteln, ohne die anderen zu kennen. Die anderen Parameter müssen nur implizit durch die Observablen festgelegt sein. Diese Eigenschaft eines Modells wird in Kapitel 12 bei ***Harney*** (2003) diskutiert. Sie wird dort damit benannt, dass die Parameter kommutieren.

6 Schätzer für die Parameter des Rasch-Modells

Im vorliegenden Abschnitt werden die Schätzer der Parameter des ***Rasch***-Modells ausgerechnet. Leider verdunkelt dies fest eingebürgerte Wort, dass zum Schätzen eines Parameters die Angabe eines einzigen Wertes unzureichend ist. Zum Schätzen gehört in den messenden Wissenschaften notwendig die Angabe des Fehlerintervalls (siehe ***Harney*** et al. 2006).

Seit ***Fisher*** (1925) wird als Schätzer des (multidimensionalen) Parameters ξ der Wert ξ^{opt} angesehen, der die likelihood Funktion maximiert. Diese Funktion ist das Modell *p(x| ξ)*, betrachtet als Funktion von ξ bei gegebenem Satz von Daten *x*. Das Besondere an diesem Maximum lässt sich am Besten im Rahmen der ***Bayes***schen Statistik erkennen. Dort kann man den Fehlerbereich von ξ definieren als den kleinsten Bereich, in dem ξ mit einer vorgegebenen Wahrscheinlichkeit *K* liegt. (Je größer *K* gewählt wird desto größer ist die Sicherheit, dass der wahre Wert des Parameters im Fehlerintervall liegt, desto länger ist das Intervall aber auch.) Der Schätzer ξ^{opt} liegt für jedes *K* im Fehlerbereich (siehe ***Harney*** 2003, Kapitel 3.3). In diesem Sinne ist er der optimale Wert von ξ. Er wird intuitiv oft als Intervallmitte angesehen; das trifft aber wegen der bias der Schätzer im Allgemeinen nicht zu, siehe Anhang B.

Um die Likelihood Funktion des ***Rasch***-Modells zu maximieren, werden die Nullstellen der nach den Parametern erfolgenden logarithmischen Ableitungen von p gesucht. Das führt auf das Gleichungssystem (36, 37) des Anhangs A; es gilt sowohl für das trigonometrische als auch – mit $\tilde{\beta},\tilde{\delta}$ an Stelle von β, δ – für das logistische ***Rasch***-Modell. Das System ist nicht-linear. Im Allgemeinen ist es nicht in geschlossener Form lösbar; d. h. man kann die Lösung nicht durch die bekannten Funktionen der Analysis ausdrücken. Zu geschlossenen Lösungen kommt man nur unter Vereinfachungen. Wir diskutieren vereinfachte Situationen im folgenden Abschnitt 6.1 für das logistische ***Rasch***-Modell.

6.1 Schätzer der logistischen Parameter

Das Gleichungssystem (36, 37) für die Schätzer geht unter der logistischen item response Funktion (9) über in

$$s_k = \sum_{l=1}^{M_D} \lambda_{kl}\,, \quad k = 1, \ldots, M_B, \tag{26}$$

$$t_l = \sum_{k=1}^{M_B} \lambda_{kl}\,, \quad l = 1, \ldots, (M_D - 1)\,. \tag{27}$$

Diese Gleichungen besagen, dass für $\tilde{\beta} = \tilde{\beta}^{\,\mathrm{opt}}, \tilde{\delta} = \tilde{\delta}^{\,\mathrm{opt}}$ die Erwartungswerte der scores und Erträge – das sind die rechten Seiten der Gleichungen (26, 27) – mit den Ergebnissen s_k, t_l des Tests übereinstimmen. Das scheint plausibel, ist aber nicht zwingend. Diese Gleichungen ergeben sich nur im logistischen ***Rasch***-Modell.

Die Lösungen der Gleichungen (26, 27) werden im Folgenden für zwei einfache Situationen diskutiert.

(i) *Daten gleicher Erträge.*
Wenn alle Fragen denselben Ertrag $t = t_l$, $l = 1,..., M_D$ erbringen, dann wird das Gleichungssystem (26, 27) mit dem Ansatz gleicher Fragenparameter

$$\begin{aligned} \tilde{\delta}_l^{\mathrm{opt}} &= \tilde{\delta}_0\,, \\ &= 0\,, \quad l = 1, \ldots, M_D, \end{aligned} \tag{28}$$

gelöst. Denn die Gleichungen (27) werden durch diesen Ansatz befriedigt, und die Gleichungen (26) gehen über in

$$s_k = M_D \frac{\exp \tilde{\beta}_k^{\text{opt}}}{1 + \exp \tilde{\beta}_k^{\text{opt}}}, \quad k = 1, \ldots, M_B, \tag{29}$$

liefern also die Schätzer $\widetilde{\beta}_k^{\text{opt}}$ auf Grund des jeweiligen scores s_k. Die Personenparameter sind scheinbar unabhängig von der Schwierigkeit der Fragen. Das liegt jedoch an der Konvention (12).

(ii) *Daten gleicher scores.*
Wenn alle getesteten Personen denselben score $s = s_k$, $k = 1,...,$ M_B haben, dann kann man die (Differenzen der) Fragenparameter allein aus den Erträgen ermitteln. Ganz analog zum vorherigen Fall wird das Gleichungssystem (26, 27) dann durch den Ansatz gleicher Personenparameter $\widetilde{\beta}_k^{\text{opt}} = \widetilde{\beta}_0$; $k = 1,...,$ M_B, gelöst. Denn damit geht es über in

$$s = \sum_{l=1}^{M_D} \frac{\exp(\tilde{\beta}_0 - \tilde{\delta}_l^{\text{opt}})}{1 + \exp(\tilde{\beta}_0 - \tilde{\delta}_l^{\text{opt}})}, \tag{30}$$

$$t_l = M_B \frac{\exp(\tilde{\beta}_0 - \tilde{\delta}_l^{\text{opt}})}{1 + \exp(\tilde{\beta}_0 - \tilde{\delta}_l^{\text{opt}})}, \quad l = 1, \ldots, (M_D - 1). \tag{31}$$

Die Summe über alle s_k ist stets gleich der Summe über alle t_l. Daraus ergibt sich die Möglichkeit, die Gleichung (30) so umzuformen, dass das ganze System übergeht in

$$t_l = M_B \frac{\exp(\tilde{\beta}_0 - \tilde{\delta}_l^{\text{opt}})}{1 + \exp(\tilde{\beta}_0 - \tilde{\delta}_l^{\text{opt}})}, \quad l = 1, \ldots, M_D. \tag{32}$$

Die Fähigkeit der Personen schlägt sich in der Verschiebung der Fragenparameter um $\widetilde{\beta}_0$ nieder.

Erstaunlich an diesem Resultat – und mutatis mutandis auch an dem Resultat (28) – ist, dass der Personenparameter für alle Personen derselbe sein soll unabhängig von den Schwierigkeiten der Aufgaben. Dieser Umstand kann durchaus Zweifel an der Konstruktion des Modells wecken.

Im Allgemeinen werden die Daten verschiedene scores aufweisen. Sollte man die Daten nach den Personen gleicher scores sortieren[14] und für jeden score die Bestwerte der Fragenparameter bestimmen? Da wir annehmen, dass die Fragenparameter wohldefiniert sind, sollte dann das Ergebnis nicht jedes Mal dasselbe sein? Nein, da die Daten statistisch schwanken, würde zu jedem score ein etwas anderer Satz von Fragenparametern erschlossen werden. Wie die verschiedenen Werte für denselben Parameter zu mitteln wären, bliebe offen. Diese Prozedur ist nicht richtig.

Wenn ein score vom Wert 0 oder M_D vorkommt (das sind die uniform beantworteten Fragebögen) oder wenn ein Ertrag gleich 0 oder M_B ist, dann gibt es keine Schätzer für $\widetilde{\beta},\widetilde{\delta}$. Man erkennt das aus der Form (15) des Modells. Es sei z. B. $s_1 = 0$. Dann hängt der erste Faktor in Gleichung (15) von $\widetilde{\beta}_1$ nicht ab. Die Likelihood Funktion hängt also von $\widetilde{\beta}_1$ nur über das doppelte Produkt ab. Sie wächst daher monoton mit fallendem $\widetilde{\beta}_1$. Da der Definitionsbereich von $\widetilde{\beta}_1$ unbeschränkt ist, gibt es kein Maximum als Funktion von $\widetilde{\beta}_1$, also auch kein Maximum als Funktion von $\widetilde{\beta},\widetilde{\delta}$. Das Entsprechende ergibt sich, falls einer der Erträge verschwindet. Für $\widetilde{\beta}_1 \to \infty$ wird p unabhängig von der Person 1 – man kann dann $k = 1$ aus dem Modell p einfach weglassen. Entsprechende Ergebnisse findet man auch, wenn ein score gleich M_D oder ein Ertrag gleich M_B ist. Denn wenn man jede Frage durch ihre Verneinung ersetzt, gehen $s_1 = M_D$ in $s_1 = 0$ und $t_1 = M_B$ in $t_1 = 0$ über. Damit die Konvention (12) sinnvoll ist, darf t_{M_D} also nicht gleich 0 oder gleich M_B sein.

In der Literatur findet man einen Vorschlag, die uniform beantworteten Fragebögen auszuwerten. Nach ***Warm*** (1989) sollte man die ***Bayes***sche a posteriori Verteilung von $\widetilde{\beta}$ (gegeben $x,\widetilde{\delta}$) maximieren, um $\widetilde{\beta}^{\text{opt}}$ zu erhalten. Das ist eine Modifikation des ***Fisher***schen Prinzips der maximum likelihood. Sie wird als gewichtete likelihood Schätzung bezeichnet (WLE für weighted likelihood estimation). Die WLE wird derzeit als Standardverfahren betrachtet. Einzelheiten sind im Anhang B zu finden.

Der Vorschlag von ***Warm*** verringert die bias der Schätzer. Unter bias versteht man den Mangel, dass der Schätzer im Mittel über die Observablen nicht mit dem wahren Wert des Parameters übereinstimmt. Um die bias zu verringern, wurde die WLE erarbeitet. Es stellte sich heraus, dass die WLE Schätzer stets endlich sind – auch für

14 Damit behandelt man die Daten unter der Bedingung eines festen scores. Das dementsprechende statistische Modell weicht von dem Modell des Abschnitts 3 ab. Es hängt von den Personenparametern überhaupt nicht ab (siehe ***Rost*** 2004, S. 126). Die Schätzer der Fragenparameter daraus zu gewinnen, wird als Methode der CML (conditional maximum likelihood) bezeichnet. Diese Schätzer hängen mit den Daten nicht vermöge Gleichung (32) zusammen. Die hier formulierte Kritik betrifft auch die CML.

uniform antwortende Personen (siehe ***Rost*** 2004, S. 314). Deshalb wurde dieser Schätzer in der statistischen Testtheorie wichtig. Im Anhang B wird genauer ausgeführt, dass wir diesen Vorschlag dennoch nicht für überzeugend halten. Erstens ist die ***Warm***sche Modifikation keine kleine Korrektur – ein Anschein, den die Formulierungen von ***Rost*** (2004) auf S. 314 nahe legen. Parameterwerte, die eigentlich im Unendlichen liegen sollten, werden durch ***Warm***s Modifikation endlich gemacht. Was immer der endliche Wert ist, der Parameter wird um ein unendlich langes Intervall verschoben. Zweitens führt ***Warm***s Vorschlag dennoch nicht zum Erfolg. Gäbe man nämlich das Fehlerintervall des Parameters an – was in den messenden Wissenschaften als zwingend angesehen wird – dann reichte es bei uniformem Antwortbogen auch nach der ***Warm***schen Modifikation bis ins Unendliche, denn es muss den unendlich weit entfernten Punkt als von den Daten implizierte Möglichkeit umfassen. Der Vorschlag von ***Warm*** wird von ***Rost***, (2004, S. 123) auch nicht argumentativ eingeführt sondern als das gängige Standardverfahren berichtet.

Der an der ***Bayes***schen a posteriori Verteilung orientierte Vorschlag von ***Warm*** nutzt die Informationsfunktion zum Auswerten uniformer Antworten, siehe Anhang B. Die Informationsfunktion wird in der Literatur auch zum Vergleich unterschiedlicher Verfahren des Schätzens eingesetzt. Der von ***Eggen*** (2000)[15] für die Schätzverfahren beschriebene jeweilige verfahrensspezifische Informationsverlust ist in unserer Sicht eine Folge der logistischen item response Funktion. Er ist daher auch durch weitere Schätzer wie die in den PISA-Studien benutzten EAP Schätzer oder plausible values nicht aufgelöst worden.[16]

All die Probleme des Schätzens – inexistenter Parameterwert, unendlich langes Fehlerintervall, ***Warm***sche bias der Bestwerte – verschwinden durch den Übergang zur trigonometrischen Form.

6.2 Schätzer der trigonometrischen Parameter

Man kann beweisen, dass für das trigonometrische ***Rasch***-Modell der Gleichung (18) die Schätzer für jeden vom Modell zugelassenen Satz von Observablen existieren. Ein gut konstruiertes statistisches Modell sollte diese Eigenschaft haben, wenn die

15 Der Autor vergleicht unterschiedliche Schätzverfahren unter dem Aspekt des Informationsverlusts. Ein solcher Vergleich erübrigt sich im Rahmen der trigonometrischen Parametrisierung.

16 In EAP Schätzungen (expected a posteriori) werden im Allgemeinen nicht einzelne Personenparameter geschätzt – wie wir es in Anhang B der Einfachheit halber beschrieben haben. Vielmehr werden latente Variablen geschatzt, die die Verteilung der Personenparameter beschreiben. Näheres dazu findet sich auf S. 314 bei ***Rost*** (2004) sowie auf S. 135 in ***Mislevy***, ***R.J. et al.*** (1992).

Zahl der Observablen groß genug ist. Der Beweis liegt vor, wird aber nicht im Rahmen der vorliegenden sondern einer späteren Publikation ausgeführt (***Harney*** et al. 2006).

Das Gleichungssystem (36, 37) für die Schätzer geht unter der trigonometrischen item response Funktion (16) über in

$$0 = \sum_{l=1}^{M_D} [x_{kl}\cot(\beta_k - \delta_l) - (1 - x_{kl})\tan(\beta_k - \delta_l)] , \quad k = 1 \ldots, M_B, \tag{33}$$

$$0 = \sum_{k=1}^{M_B} [x_{kl}\cot(\beta_k - \delta_l) - (1 - x_{kl})\tan(\beta_k - \delta_l)] , \quad l = 1, \ldots, (M_D - 1) . \tag{34}$$

Die Lösungen dieser Gleichungen hängen von den Daten *x* keineswegs nur über die scores und die Erträge ab; im Allgemeinen spielen die individuellen Muster der Antworten eine Rolle. Dies wird weiter unten an Beispielen gezeigt. Wir haben keine einfachen Datensätze gefunden, die zu einfachen Lösungen führten. Insbesondere lassen sich die Fälle (i) und (ii) des Abschnitts 6.1 nicht übertragen. Daher diskutieren wir die Lösungen im Folgenden an Hand von numerischen Beispielen.

So einfache Beispiele wie das der Tabelle 1 vermögen die Eigenschaften des trigonometrischen Modells nicht zu repräsentieren. Man braucht mindestens $M_D = 3$ Fragen, um zu zeigen, dass bei identischen scores die Schätzer der Personenparameter nicht identisch sein müssen. Wir zeigen das mit dem Beispiel in Tabelle 2.

Tabelle 2 Die Matrix x_{kl} des Beispiels, welches identische scores $2 = s_k$, $k = 1,...,4$ enthält. Die Schätzer der Personenparameter stehen in der letzten Spalte; die Schätzer der Fragenparameter stehen in der letzten Zeile der Tabelle.

	Frage 1	Frage 2	Frage 3	β_k^{opt}
Person 1	1	1	0	0.81
Person 2	1	0	1	0.76
Person 3	0	1	1	0.76
Person 4	1	1	0	0.81
δ_l^{opt}	-0,27	-0,27	0	

korrigierte Werte gegenüber Druckfassung

Trotz identischer scores sind hier die Personenparameter nicht identisch; denn die Fragenparameter sind verschieden. Ein solches Resultat ist unseres Erachtens eher zu

erwarten als das Resultat unter Punkt (ii) in Abschnitt 6.1, wo identische scores stets identische Personenparameter zur Folge haben. Die Frage 3 mit dem Ertrag $t_3 = 2$ hat den höchsten Fragenparameter. Die beiden anderen Fragen mit den Erträgen $t_1 = t_2 = 3$ haben kleinere Parameter. Das erwartet man, wenn der Datensatz im Sinne eines Leistungsvergleichs interpretierbar ist. Er ist es möglicherweise nicht; denn die Personen 2 und 3 die die schwierigere Frage lösen, haben den geringeren Personenparameter. Ohne Diskussion der Fehler – die erst in ***Harney*** et al. (2006) ausgerechnet werden – kann man weder sagen, ob dieser Unterschied signifikant ist noch ob der Datensatz mit dem Modell im Sinne eines goodness of fit Tests vereinbar ist. Wenn der Datensatz mit dem Modell vereinbar und der Unterschied signifikant ist, dann ist der Datensatz nicht im Sinne des Leistungsvergleichs interpretierbar. So liefert die trigonometrische Form des ***Rasch***-Modells möglicherweise Information darüber, ob das meistens zu Grunde gelegte Leistungsmodell von den Daten erfüllt wird.

In allen Tabellen wird derjenige Satz von Schätzern angegeben, für den

$$\beta_1^{\text{opt}} > 0 \tag{35}$$

ist. Wegen der in Abschnitt 4 besprochenen Spiegelungssymmetrie des trigonometrischen Modells gibt es zwei Sätze von Schätzern. Die Konvention (35) macht die Schätzer eindeutig. Der folgende Datensatz mit identischen Erträgen ist mit dem Leistungsmodell vereinbar.

Tabelle 3 Die Matrix x_{kl} des Beispiels, welches identische Erträge $4 = t_l$, $l = 1,2,3$ enthält. Die Schätzer der Personenparameter stehen in der letzten Spalte; die Schätzer der Fragenparameter stehen in der letzten Zeile der Tabelle.

	Frage 1	Frage 2	Frage 3	β_k^{opt}
Person 1	1	1	0	0.96
Person 2	1	0	1	0.96
Person 3	0	1	1	0.96
Person 4	1	0	0	0.62
Person 5	0	0	1	0.62
Person 6	0	1	0	0.62
Person 7	1	1	1	1.57
δ_l^{opt}	0	0	0	

Die Fragenparameter erweisen sich hier als identisch. Die Personenparameter steigen monoton mit dem score an. Die Person, welche monoton geantwortet hat, bekommt den höchsten Fragenparameter vom Wert $\pi/2$ zugewiesen. Diese Parameter sind mit dem Leistungsmodell vereinbar.

Man braucht mindestens $M_D = 4$ Fragen, um trotz identischer Erträge unterschiedliche Fragenparameter zu erhalten. Der folgende Datensatz ist ein Beispiel.

Tabelle 4 Die Matrix x_{kl} eines Beispiels, welches identische Erträge $1 = t_l$, $l = 1,...,4$ enthält. Die Schätzer der Personenparameter stehen in der letzten Spalte; die Schätzer der Fragenparameter stehen in der letzten Zeile der Tabelle.

	Frage 1	Frage 2	Frage 3	Frage 4	β_k^{opt}
Person 1	1	0	0	0	0.54
Person 2	0	0	0	1	0.54
Person 3	0	1	1	0	0.81
δ_l^{opt}	0	0.056	0.056	0	

Hier sind die Fragenparameter nicht identisch, obgleich die Erträge identisch sind. Dieser Datensatz ist mit dem Leistungsmodell verträglich; denn die als schwieriger eingestuften Fragen werden von der leistungsfähigeren Person beantwortet.

Bei den Datensätzen der Tabellen 2, 3 und 4 kann man ohne einen goodness-of-fit-Test nicht wissen, ob sie mit dem Modell verträglich sind. Ein solcher Test liegt vor (***Harney*** et al. 2006). Er ist mit dem üblichen Chi-Quadrat-Test nicht identisch. Der Datensatz der Tabelle 5 ist ganz sicher mit dem ***Rasch***-Modell verträglich. Denn jedes mögliche Muster von Antworten kommt genau einmal vor. Indem man die Parameter so wählt, dass alle Muster dieselbe Wahrscheinlichkeit ihres Auftretens haben, erhält man einen Parametersatz, der mit diesem Datensatz bestimmt vereinbar ist. Dieser Parametersatz, der alle Personen gleich behandelt, existiert und lautet $\beta_k = \pi/4$ für alle k sowie $\delta_l = 0$ für alle l. Dann ist die Wahrscheinlichkeit $\tau(\beta_k - \delta_l)$ gleich 1/2 für jede Differenz eines Personen- und eines Fragenparameters. Das bedeutet aber, dass die Wahrscheinlichkeit für $x_{kl} = 0$ ebenso groß ist wie die Wahrscheinlichkeit für $x_{kl} = 1$. Kurz, die Wahrscheinlichkeiten für das Auftreten von 0 oder 1 in den Antwortmustern sind gleich groß. Alle Antwortmuster haben also dieselbe Wahrscheinlichkeit – in Übereinstimmung mit dem Datensatz. Als Schätzer für eine Person mit dem score 4 ergibt sich nicht der Personenparameter $\pi/4$ sondern etwas deutlich Größeres – nämlich $\pi/2$ – denn als reiner Zufall ist das Auftauchen eines uniform richtigen Antwortmusters recht unwahrscheinlich. Wir

weisen darauf hin, dass die uniformen Antwortmuster der Personen 1 und 16 problemlos analysiert wurden. Der Datensatz in Tabelle 5 kann als Leistungstest interpretiert werden.

Tabelle 5 Die Matrix x_{kl} eines Beispiels gleicher Wahrscheinlichkeit für jedes Antwortmuster. Die Schätzer der Personenparameter stehen in der letzten Spalte; die Schätzer der Fragenparameter stehen in der letzten Zeile der Tabelle

	Frage 1	Frage 2	Frage 3	Frage 4	β_k^{opt}
Person 1	1	1	1	1	1.57
Person 2	1	1	1	0	1.05
Person 3	1	1	0	1	1.05
Person 4	1	0	1	1	1.05
Person 5	0	1	1	1	1.05
Person 6	1	1	0	0	0.79
Person 7	1	0	1	0	0.79
Person 8	0	1	1	0	0.79
Person 9	1	0	0	1	0.79
Person 10	0	1	0	1	0.79
Person 11	0	0	1	1	0.79
Person 12	0	0	0	1	0.52
Person 13	0	0	1	0	0.52
Person 14	0	1	0	0	0.52
Person 15	1	0	0	0	0.52
Person 16	0	0	0	0	0
δ_l^{opt}	0	0	0	0	

Dieses Beispiel zeigt, dass es Datensätze gibt, welche im Rahmen des trigonometrischen Modells als Ergebnisse eines Leistungstests interpretiert werden können. Mit $M_D = 5$ Fragen anstatt $M_D = 4$ wurden ganz entsprechende Ergebnisse erzielt. Das Beispiel der Tabelle 5 ist insofern unrealistisch als die Fragenparameter untereinander gleich sind. Man kann jedoch auf ähnliche Weise einen Satz weiterer Fragen einführen, die einen anderen Fragenparameter haben. In diesem Fall wird die Konstruktion des Datensatzes, der unbedingt mit dem Modell übereinstimmt, etwas schwieriger, weil

die Antwortmuster nicht mehr identische Wahrscheinlichkeiten haben. Derartige Untersuchungen könnten in Zukunft zu einer Regel führen, die zu erkennen erlaubt, welche Datensätze im Sinne des trigonometrischen ***Rasch***-Modells skaliert, d. h. mit dem Modell vereinbar und als Leistungstest interpretierbar sind.

7 Schlussfolgerungen

Es wurde gezeigt, dass die Parameter des ***Rasch***-Modells für jeden Datensatz stabil geschätzt werden können, wenn man die logistische item response Funktion durch eine trigonometrische Form ersetzt. Die trigonometrische Funktion ergibt sich aus der inneren Symmetrie – der Forminvarianz – des Binomialmodells. Jedes gut konstruierte statistische Modell sollte diese Symmetrie besitzen. Denn sie garantiert, dass das Modell vollständig ist und dass seine Parameter unvoreingenommen ermittelt werden. Es bedarf keiner weiteren Annahmen, um die item response Funktion festzulegen.

Für die Anwendung ergibt sich folgende Bedeutung. Mit der Einbeziehung der uniformen Antwortmuster wird die Häufigkeit der deterministisch antwortenden Personen formal einwandfrei in das Modell integriert. Damit wird es möglich, eine strukturelle Formschwäche des ***Rasch***-Modells zu beheben. Die Schwäche ist: Wenn Personen uniform antworten, also perfekt kompetent oder inkompetent auf einen Satz von Fragen reagieren, so liefert das Modell keinen Schätzer. Das Fehlerintervall wächst ins Unendliche. Einerseits führt in oberen und unteren Kompetenzbereichen die geringer werdende Anzahl von Aufgaben, mit denen man diese Personen charakterisieren kann, zur Vergrößerung des Fehlerintervalls; andererseits gehen in die Vergrößerung des Fehlerintervalls aber auch nicht-empirische Folgen der logistischen Funktion ein. Das ist problematisch und deutet auf eine strukturelle bias hin, die durch die item response Funktion erzeugt wird.

In der Anwendung gehört die Verringerung von Fehlerintervallen zu den zentralen methodischen Fragen. Die PISA-Studie arbeitet sowohl mit gewichteten Maximum Likelihood Schätzern (***Carstensen*** et al. 2004, S. 381) – WLE genannt – als auch mit Hintergrundmodellen, in deren Rahmen weitere Variable regressionsanalytisch zur Erhöhung der Messgenauigkeit herangezogen werden (ebenda). Messgenauigkeit im statistischen Sinne liegt auch für die PISA-Kompetenzstufenmessungen erst dann vor, wenn *sämtlichen* Antwortmustern, also auch den uniformen Antwortmustern, Schätzer zugewiesen werden können.

Generell bleibt die Schätzung der uniformen Antwortmuster im logistischen Modell unbefriedigend. Formal betrachtet ist nämlich die Schätzung für jeden Wert der

Parameter von der bias betroffen, welche ***Warm*** (1989) beschreibt. Mit den Schätzern der WLE ist es unmöglich, das uniforme Antwortmuster genau zu treffen, also die Übereinstimmung zwischen Schätzwerten und Daten herzustellen. Denn im Fall eines uniformen Antwortmusters würde der Parameter gegen Unendlich laufen. Dieser Defekt geht auf den Definitionsbereich der Parameter zurück; und er wirkt sich bei allen anderen Antwortmustern als systematische Fehleinstellung der Schätzer aus. Man kann den Defekt auf zweierlei Weise bearbeiten, nämlich (i) algorithmisch also von der Form der Informationsverarbeitung her (wie in der vorliegenden Arbeit) oder (ii) inkrementalistisch also durch pragmatisches, auf Versuch und Irrtum beruhendem Vorgehen. Letzteres ist der Fall, wenn zum Verringern des Schätzfehlers neben den angesprochenen Kompensationen (Regressionsmodelle, WLE) die Menge der Aufgaben erweitert und mit Hilfe der IRT skaliert wird – was zu den Notwendigkeiten jeder Testentwicklung gehört. Die Logik der WLE ist eine Vermengung der Ebenen (i) und (ii). Dabei greift die WLE das Problem sozusagen vom Ende her auf. Sie verhindert, dass für uniforme Antwortmuster die Schätzer gegen Unendlich gehen um den Preis dass die Schätzer bedeutungslos werden; denn die uniformen Antwortmuster sind immer noch mit einem bis ins Unendliche ausgedehnten Bereich der Parameter verträglich.

Unter der hier vorgeschlagenen trigonometrischen item response Funktion tritt das nicht auf. Die für die Bildungsforschung wichtigen oberen und unteren Kompetenz- und Schwierigkeitsbereiche werden von vorn herein in den Definitionsbereich der Parameter einbezogen. Die Kompetenzniveaus sind kontingent und endlich. Auf der Zeitachse der Lebensläufe von Personen, Gruppen und Institutionen schließen sie aneinander an, verflechten sich und unterliegen dabei Prozessen der Auf- und Abwertung. Was innerhalb des einen zeitlich-sozialen Kontextes eine normalisierte, von jedermann erwartete Kompetenz ist, das kann in einem anderen Kontext ausgesprochen exklusiv sein. Das logistische Modell bildet diesen Zusammenhang der Kontingenz, d. h. Überlappung, des Übergangs von einer Domäne zur nächsten und der damit einhergehenden Veränderung, nicht angemessen ab. Streng genommen kennt es ihn gar nicht. Denn es läuft in den Grenzbereichen ins Unendliche aus und wird leer von Information. So ist es bezeichnend, dass die WLE der nachträglich eingeschobenen Informationsfunktion bedarf, die diese Formschwäche kompensieren soll (Anhang B). Unter der Bedingung der trigonometrischen Form gibt es keine Daten, die leer von Information wären. Eine kompensatorische Ergänzung des ***Rasch***-Modells wird überflüssig. In diesem Rahmen sind die maximum likelihood Schätzer der Parameter nicht von der bias betroffen, die durch das Ausweichen auf andere Schätzer – wie z.B. WLE oder EAP – verringert werden soll.

Dennoch bleibt es eine unzulässige Vereinfachung, das Schätzen der Parameter auf die Bestimmung eines ihrer Werte zu beschränken. Zum Schätzen der Parameter gehört die Angabe ihrer Fehlerintervalle.

Wir haben nachgewiesen, dass die trigonometrische Form des ***Rasch***-Modells als Leistungstest interpretiert werden kann – also dem Ziel entspricht, das die meisten Anwendungen der IRT im Auge haben. Die trigonometrische Form verfügt jedoch im Vergleich zum logistischen Modell über einen erweiterten Definitionsbereich der Parameter. Daher ist es möglich, dass die trigonometrische Form noch andere als Leistungsvergleiche zulässt. Dies muss durch zukünftige numerische Studien geklärt werden.

Literatur

Adams, R.J., Wilson, M. and ***Wang, W.*** (1997): The multidimensional random coefficients of the multinomial logit model, Applied Psychological Measurement 21, S. 1-23.

Andrich D. (1988): ***Rasch-***Models for Measurement, Quantitative Applications in the Social Sciences Vol. 68, Sage Publications, Newbury Park.

Baumert, J. und ***Artelt, C.*** (2003): Konzeption und technische Grundlagen der Studie. In: ***Deutsches PISA-Konsortium PISA 2000***: Ein differenzierter Blick auf die Länder der Bundesrepublik, Verlag Leske und Budrich, Opladen, S. 11-50.

Bock, R.D. (1983): The discrete Bayesian. In: ***Wainer, H.*** and ***Messick, S.*** (Eds.) Principles of modern psychological measurement: A festschrift for ***Frederick M Lord,*** Lawrence Erlbaum, New Jersey, S. 103-115.

Bortz, J. (1999): Statistik für Sozialwissenschaftler, Springer Verlag, Berlin, S.517 f.

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

Carstensen, Ch., Knoll, S., Rost, J. und ***Prenzel, M.*** (2004): Technische Grundlagen in: ***PISA-Konsortium Deutschland***: Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs, Waxmann Verlag, S. 371-387.

Cox, D.R. and ***Hinkley, D.V.*** (1974): Theoretical Statistics, Chapman and Hall, New York.

Davier, M.; Molenaar, I.W. and ***Person, A.*** (2003): FIT Index for Polytomous ***Rasch***-Models, Latent Class Models, and their Mixture Generalizations, Psychometrika 68, S. 213-228.

Deutsches PISA-Konsortium PISA 2000 (2003): Ein differenzierter Blick auf die Länder der Bundesrepublik, Verlag Leske und Budrich, Opladen.

Eggen, Th.J. (2000): On the Loss of Information in Conditional Maximum Likelihood of Item Parameters, Psychometrika 65, S. 337-362.

Fischer, G.H. (1974): Einführung in die Theorie psychologischer Tests, Verlag Hans Huber, Bern.

Fischer, G.H. (1988): Spezifische Objektivität: Eine wissenschaftstheoretische Grundlage des ***Rasch***-Modells. In: ***Kubinger, K.*** (Hrsg.): Moderne Testtheorie, Psychologie Verlagsunion, Weinheim und München, S. 87-111.

Fisher, R.A. (1925): Theory of Statistical Information, Proc. Cambridge Phil. Soc. 22, S. 700-725. Dieser Aufsatz ist auch enthalten in: ***Fisher, R.A.*** (1950): Contributions to mathematical statistics, Wiley, New York.

Harney, H.L. (2003): Bayesian Statistics – Parameter Estimation and Decisions, Springer Verlag, Heidelberg. Korrekturen und Kommentare dazu bei http://www.mpi-hd.mpg.de/harney/home.

Harney, K. und ***Harney, H.L.*** (2006): Zwischen PISA, Rasch und Godot: Die statistische Form und ihre Bedeutung in der Bildungsforschung. In: ***Rustemeyer, D.*** (Hrsg.): Formfelder – Genealogien von Ordnung, Wittener Kulturwissenschaftliche Studien Bd. 5, Königshausen & Neumann, Würzburg, S. 105-126.

Harney, H.L., Fuhrmann, Ch. und ***Harney, K.*** (2006): Unveröffentlichtes Manuskript.

Luhmann, N. (1989): Die Wirtschaft der Gesellschaft, Suhrkamp Verlag, Frankfurt/M., S. 14ff.

Mislevy, ***R.J.; Beaton, A.E.; Kaplan, B.*** and ***Shehan, K.M.*** (1992): Estimating Population Characteristics from Sparse Matrix Samples of Item Responses, Educational Measurement 29, S. 133-161.

Neubrand, M., Klieme, E., Lüdtke, ***O***. und ***Neubrand, J.*** (2002): Kompetenzstufen und Schwierigkeitsmodelle für den PISA-Test zur mathematischen Grundbildung, Unterrichtswissenschaft 30, S. 100-119.

PISA-Konsortium Deutschland (2004): Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs, Waxmann Verlag, Münster.

Rasch, G. (1967): An informal report on a theory of objectivity in comparisons. In: ***van der Kamp, L.J.Th.*** and ***Vlek, C.A.J.*** (Hrsg.): Measurement theory. Proceedings of the NUFFIC international summer session in science in "Het Oude Hof", The Hague, July 14-28, 1966. University of Leiden, Leiden, S. 1-19.

Rasch, G. (1972): Objectivitet i samfundsvidenskaberne et metodeproblem, Nationalokonomisk Tidsskrift 110, S. 161-196.

Rasch, G. (1977): On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In: ***Blegvad, M.*** (Hrsg.) The Danish yearbook of Philosophy, Munksgaard, Kopenhagen, S. 58-94.

Rasch, G. (1980): Probabilistic Models for some Intelligence and Attainment Tests, The University of Chicago Press, Chicago.

Rohwer, G. und ***Pötter, U.*** (2001): Methoden sozialwissenschaftlicher Datenkonstruktion, Juventa Verlag, Weinheim.

Rohwer, G. und ***Pötter, U.*** (2002): Wahrscheinlichkeit. Begriff und Rhetorik in der Sozialforschung, Juventa Verlag, Weinheim.

Rost, J. (2004): Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen, Zeitschrift für Pädagogik 50, S. 662-678.

Rost, J. (2004): Lehrbuch Testtheorie – Testkonstruktion, Verlag Hans Huber, Bern.

Samejima, F. (1983): Constant Information Model on Dichotomous Response Level. In: ***Weiss, D.J.*** (Hrsg.): New Horizons in Testing, Academic Press, New York, S. 63-79.

Simmel, G. (1958): Philosophie des Geldes, 6. Aufl., unveränderter Nachdruck der 5. Aufl.1930, Duncker & Humblot, Berlin.

Überla, K. (1977): Faktorenanalyse. Eine systematische Einführung für Psychologen, Mediziner, Wirtschafts- und Sozialwissenschaftler, Springer Verlag, Berlin.

Warm, T.A. (1989): Weighted Likelihood Estimation of Ability in Item Response Theory, Psychometrika 54, S. 427-450.

Wirtz, M. und ***Nachtigall, Ch.*** (1998): Deskriptive Statistik Teil 1, Juventa Verlag, Weinheim, 3. Auflage, S. 209 ff.

Wolter, K.M. (1985): Introduction to Variance Estimation, Springer Verlag, New York.

Anhang A: Einzelheiten zu den Schätzern der Parameter

Um die likelihood Funktion des ***Rasch***-Modells zu maximieren, werden die Nullstellen der nach den Parametern erfolgenden logarithmischen Ableitungen von *p* gesucht. Das führt auf das Gleichungssystem

$$0 = \sum_{l=1}^{M_D} \frac{\partial \phi_{kl}}{\partial \beta_k} \cdot \left[\frac{x_{kl}}{\phi_{kl}} - \frac{1 - x_{kl}}{1 - \phi_{kl}} \right], \quad k = 1, \ldots, M_B; \tag{36}$$

$$0 = \sum_{k=1}^{M_B} \frac{\partial \phi_{kl}}{\partial \delta_l} \cdot \left[\frac{x_{kl}}{\phi_{kl}} - \frac{1 - x_{kl}}{1 - \phi_{kl}} \right], \quad l = 1, \ldots, (M_D - 1) \tag{37}$$

für die Schätzer. Dieses Gleichungssystem gilt sowohl für das trigonometrische – mit $\phi = \tau$ – als auch für das logistische Modell – mit $\phi = \lambda$ und $\widetilde{\beta}, \widetilde{\delta}$ an Stelle von β, δ.

Für das logistische Modell (9) geht das Gleichungssystem (36, 37) über in das System (26, 27). Das erkennt man aus den Beziehungen

$$\begin{aligned} \frac{\partial \lambda_{kl}}{\partial \tilde{\beta}_k} &= \lambda_{kl}(1 - \lambda_{kl}), \\ \frac{\partial \lambda_{kl}}{\partial \tilde{\delta}_l} &= -\lambda_{kl}(1 - \lambda_{kl}). \end{aligned} \tag{38}$$

Die Gleichungen (26, 27) sind keine Definitionen sondern ein gekoppeltes Gleichungssystem für die Bestwerte der Parameter $\widetilde{\beta}, \widetilde{\delta}$. Das System ist gekoppelt, weil in beiden Gleichungen die λ_{kl} Funktionen sowohl der $\widetilde{\beta}$ wie auch der $\widetilde{\delta}$ sind. Es wird in Abschnitt 6.1 besprochen.

Für das trigonometrische Modell geht das Gleichungssystem (36, 37) über in das System (33, 34) und wird in Abschnitt 6.2 diskutiert.

Anhang B: Die Schätzer der Methoden WLE und EAP

Warm (1989) hat vorgeschlagen, zur Ermittlung der Schätzer der logistischen Parameter die Methode der maximum likelihood (***Fisher*** 1925) durch ein Vorgehen zu ersetzen, das er weighted likelihood estimation (WLE) nennt. Durch analytische Betrachtungen (***Cox*** and ***Hinkley*** 1974) und durch Monte Carlo Rechnungen hatte man festgestellt, dass die durch Maximieren der likelihood definierten Schätzer $\widetilde{\beta}^{\text{opt}}$ der logistischen Parameter von einer bias betroffen sind. Das bedeutet: Der wahre Wert von $\widetilde{\beta}$ weicht von dem über die Ereignisse gemittelten $\widetilde{\beta}^{\text{opt}}$ ab. Man kann das auch so ausdrücken: Die Wahrscheinlichkeit, dass der wahre Wert kleiner ist als $\widetilde{\beta}^{\text{opt}}$, ist verschieden von der Wahrscheinlichkeit, dass er größer ist. ***Warm*** definiert einen Schätzer $\widetilde{\beta}^{W}$, der eine geringere bias hat. Es erwies sich, dass $\widetilde{\beta}^{W}$ immer, also auch für uniform beantwortete Tests, endlich ist (siehe ***Rost*** 2004, S. 314).

Wir besprechen die WLE am Beispiel eines ***Rasch***-Modells $p(x \mid \widetilde{\beta}_1)$, welches von einem einzigen Personenparameter abhängt; die Fragenparameter seien gegeben und nicht aus den Daten *x* zu erschließen. Der Maximum Likelihood Schätzer $\widetilde{\beta}_1^{\text{opt}}$ ergibt sich durch Maximieren von $p(x \mid \widetilde{\beta}_1)$ bei festem *x*. Der WLE Schätzer $\widetilde{\beta}^{W}$ ist der Ort des Maximums der ***Bayes***schen a posteriori Verteilung

$$P(\tilde{\beta}_1|x) = \frac{p(x|\tilde{\beta}_1) f(\tilde{\beta}_1)}{\int \mathrm{d}\tilde{\beta}_1' \, p(x|\tilde{\beta}_1') f(\tilde{\beta}_1')} \,. \tag{39}$$

Dabei ist $f(\widetilde{\beta}_1)$ die a priori Verteilung des Parameters. Sie wird in Kapitel 9 von ***Harney*** (2003) definiert als

$$f(\tilde{\beta}_1) \propto \overline{\left(-\frac{\partial^2}{\partial \tilde{\beta}_1^2} \ln p(x|\tilde{\beta}_1) \right)}^{1/2} \,. \tag{40}$$

Der Überstrich bedeutet die Mittelung in Bezug auf *x*. Wir zeigen, dass der Ausdruck (40) mit der von ***Warm*** benutzten Funktion *f* übereinstimmt. Dazu berechnet man die zweite Ableitung von ln *p*. Man findet eine lineare Funktion von x_{1l}.

Indem man den Mittelwert

$$\overline{x_{1l}} = \phi_{1l} \tag{41}$$

bildet, folgt

$$\overline{\left(\frac{\partial^2}{\partial\tilde{\beta}_1^2}\ln p(x|\tilde{\beta}_1)\right)} = -\sum_l \frac{(\phi'_{1l})^2}{\phi_{1l}(1-\phi_{1l})}\,. \tag{42}$$

Hier steht ϕ für eine der item response Funktionen λ in Gleichung (5) oder τ in Gleichung (16). Es bedeutet ϕ' die Ableitung nach dem Parameter. Dies gilt sowohl für das logistische wie auch das trigonometrische Modell.

Zunächst betrachten wir das logistische Modell, das von ***Warm*** vorausgesetzt wird. In diesem gilt

$$\lambda'_{1l} = \lambda_{1l}(1-\lambda_{1l})\,. \tag{43}$$

Die a priori Verteilung ist also

$$f(\tilde{\beta}_1) \propto \sqrt{\sum_l \lambda'_{1l}}\,, \tag{44}$$

woraus sich ergibt

$$\frac{\partial}{\partial\tilde{\beta}_1}\ln f(\tilde{\beta}_1) = 2^{-1}\frac{\sum_l \lambda''_{1l}}{\sum_l \lambda'_{1l}}\,. \tag{45}$$

Unter Benutzen der Größen $I = \sum_i(\lambda'_{1l})^2/\lambda_{1l}(1-\lambda_{1l})$ und $J = \sum_i \lambda'_{1l}\lambda''_{1l}/\lambda_{1l}(1-\lambda_{1l})$, welche ***Warm*** einführt, und vermöge der Gleichung (43) geht die Gleichung (45) in

$$\frac{\partial}{\partial\tilde{\beta}_1}\ln f(\tilde{\beta}_1) = \frac{J}{2I} \tag{46}$$

über. Dies ist identisch mit der Gleichung (9) bei ***Warm*** (1989). Die dortige a priori Verteilung stimmt also mit der Definition (40) überein.[17]

Es gibt eine Parametrisierung, in der die a priori Verteilung *f* eine Konstante ist, (siehe ***Harney*** 2003, Kap. 2) Unter dieser natürlichen Parametrisierung verschwindet die Ableitung von *f*. Dann geht der WLE Schätzer in den Maximum Likelihood Schätzer über. Genau das wird aber durch die trigonometrische item response Funktion

17 Man beachte, dass in den Gleichungen (3, 5, 10) von ***Warm*** (1989) ein Druckfehler ist. Der Faktor $(PQ)^{-1}$ sollte innerhalb und nicht außerhalb der Summe $\sum_i$ stehen.

$$\tau_{1l}(\beta_1) = \sin^2(\beta_1 - \delta_l) \tag{47}$$

bewirkt. Denn mit Hilfe von Gleichung (42) sieht man, dass

$$f(\beta_1) \equiv \text{const} \tag{48}$$

liegt.

Das heißt, für den Bestwert β^{opt} verschwindet die bias-Korrektur.[18]

Die Funktion f^2 wird von ***Warm*** als „test information", von ***Rost*** (2004) auf S. 313f. als Informationsfunktion bezeichnet. Das geht auf ***Fisher*** (1925) zurück. Daran anschließend wurde für die Sozialforschung der Begriff der Information in ***Rohwer*** und ***Pötter*** (2002, S. 136 ff.) neuerlich aufgegriffen. ***Fisher*** erkannte, dass das Inverse von f – genommen an der Stelle des Bestwertes – als der Fehler angesehen werden kann, der dem Maximum Likelihood Schätzer zuzuweisen ist (siehe auch ***Rost*** 2004, S. 358 ff.) In diesem Sinne ist die Information zu verstehen, die man über den Parameter erhält. Für das logistische Modell ist die a priori Verteilung

$$f(\tilde{\beta}_1) = \left(\sum_l \frac{\exp(\tilde{\beta}_1 - \tilde{\delta}_l)}{\left(1 + \exp(\tilde{\beta}_1 - \tilde{\delta}_l)\right)^2} \right)^{1/2} \tag{49}$$

Asymptotisch verhält sich diese Funktion wie

$$f(\tilde{\beta}_1) \lim_{|\tilde{\beta}_1| \to \infty} \text{const} \, \exp(-|\tilde{\beta}_1|/2) \, ; \tag{50}$$

sie wird also exponentiell klein. Für großes $\left|\tilde{\beta}_1\right|$ wächst daher der Fehler – genauer die Länge des Fehlerintervalls – exponentiell an.

Im Rahmen der trigonometrischen Form ist der Fehler für jedes β^{opt} derselbe; denn $f(\beta_1)$ ist eine Konstante. Im vorliegenden Anhang haben wir das für einen einzigen Personenparameter bei gegebenen Fragenparametern ausgerechnet. Die Fehlerbetrachtung für alle Parameter des trigonometrischen ***Rasch***-Modells zugleich ist wesentlich aufwändiger; sie führt aber zu demselben Ergebnis, siehe ***Harney*** et al.

18 Bekanntlich ist die Schätzung der Streuung $\tilde{\sigma}$ ***Gauß***verteilter Daten von einer bias betroffen (***Wolter*** 1985). Auch hier verschwindet die Korrektur von ***Warm***, wenn man zu derjenigen Parametrisierung übergeht, die die a priori Verteilung konstant macht. Der erforderliche Parameter ist $\sigma = \ln \tilde{\sigma}$.

(2006). Das trigonometrische Modell ist forminvariant im Sinne der Kapitel 2.3, 6 und 11 von ***Harney*** (2003). Die Werbung des vorliegenden Artikels ist zugleich eine Beschreibung der in ***Harney*** (2003) definierten Forminvarianz: Ein forminvariantes statistisches Modell vermag jeden legitimen Datensatz zu interpretieren und zieht aus jedem Datensatz gleich viel Information.

Im Rahmen der PISA-Studien (***Carstensen*** et al. 2004) wurde neben der WLE von ***Warm*** ein weiterer Schätzer wichtig, der Erwartungswert der a posteriori Verteilung (EAP) oder plausible value, siehe Fußnote 15, ***Bock*** (1983) und ***Rost*** (2004, S. 315f.)

Dieser Schätzer ist

$$\tilde{\beta}_1^{\mathrm{EAP}} = \int \mathrm{d}\tilde{\beta}_1 \, \tilde{\beta}_1 P(\tilde{\beta}_1|x) \,. \tag{51}$$

Auch er ist stets endlich wie der WLE Schätzer von ***Warm***. Auch er hat eine geringere bias als der maximum likelihood Schätzer der logistischen Parameter. Beide Schätzer haben aber auch die ungelösten Probleme gemeinsam. Die bias ist zwar verringert, ist aber noch immer von Bedeutung. Die Fehlerintervalle wachsen ins Unendliche, wenn das Antwortmuster dem uniformen nahe kommt; denn die logistische item response Funktion ist geblieben.

Wir halten die Diskussion der Schätzer vor dem Hintergrund unserer Ergebnisse nicht für ergiebig. Es ist besser, das Problem der bias zu minimieren, indem man das Modell so konstruiert, dass die a priori Verteilung aller Parameter eine Konstante ist. Auf jeden Fall sollte man das korrekte Verfahren anwenden, indem man das Fehlerintervall angibt. Das ist das kleinste Intervall, in welchem der Parameter mit einer vorgegebenen Wahrscheinlichkeit liegt (siehe ***Harney*** 2003, Kap. 3).