Diagnostica 1980, Band XXVI, Heft 1, Seite 10–20

Statistischer und entscheidungsorientierter Vergleich der Testmodelle von Rasch und Birnbaum

Berthold Färber & Alf Zimmer

1. Einleitung

In der Diagnostik individuell variierender Merkmale ist für jede quantifizierende Aussage eine Theorie über den Zusammenhang zwischen dem Verhalten des Probanden als Konsequenz seiner Fähigkeit und den Eigenschaften einzelner Aufgaben – Schwierigkeit (d. h. Skalenwert auf einem ‚latent trait') und Diskriminationsfähigkeit – notwendig. Während explizite Annahmen hierüber in der ‚schwachen' klassischen Testtheorie nicht gegeben werden können, setzen die sogenannten ‚latent trait'-Modelle (Rasch, 1960, 1966a, und Birnbaum, 1968) in der Modellkonstruktion bei diesem Zusammenhang ein.

Die Beziehung zwischen individueller Fähigkeit von Probanden und der Schwierigkeit einer Aufgabe wird am besten in der sogenannten Item-Charakteristik-Kurve (ICC) sichtbar. Als Form dieser Kurve wird oft – ausgehend von psychophysischen und daraus entwickelten psychometrischen Überlegungen – die Ogive angenommen (Guilford, 1934).

So geht das Testmodell von Lord (1953) davon aus, daß die Lösungswahrscheinlichkeit p einer Person mit der Fähigkeit θ bei einer Aufgabe i bestimmt ist durch:

$$p_i(\theta) = \Phi(a_i[\theta - b_i]), \tag{1}$$

wobei

Φ die kumulative Verteilungsfunktion einer standardnormverteilten Zufallsvariablen ist,

a_i die Diskriminationsfähigkeit und

b_i die Schwierigkeit der Aufgabe charakterisiert.

Auf der anderen Seite nehmen logistische Testmodelle an, daß die Itemcharakteristiken einer logistischen Funktion

$$\Psi(x) = \frac{e^x}{1 + e^x}$$

folgen.

Modelle dieser Art sind einerseits mathematisch leichter handhabbar, andererseits konnte Haley (1952) zeigen, daß die Verteilungsfunktionen der Normalverteilung und der logistischen Funktion kaum unterscheidbar sind. Es gilt nämlich für alle x:

$$|\Phi(x) - \Psi(1.7\,x)| < 0.01 \tag{2}$$

Das logistische Modell des individuellen Antwortverhaltens nach Birnbaum (1968) setzt den Fähigkeitsparameter einer Person θ sowie den Schwierigkeits- und Diskriminationsparameter eines Items i (b_i bzw. a_i) folgendermaßen in Beziehung: bei

Vorgabe eines Items kommt ein Proband mit der Fähigkeit θ mit der folgenden Wahrscheinlichkeit $p_i(\theta)$ zu einer richtigen Lösung:

$$p_i(\theta) = \frac{e^{1.7(\theta - b_i)a_i}}{1 + e^{1.7(\theta - b_i)a_i}} \tag{3}$$

Der Wert 1.7 in Gleichung 3 stellt eine Skalierungskonstante dar, die eine Anpassung der logistischen Funktion an die Normalverteilungsfunktion ermöglicht (vgl. Gleichung 2, Haley, 1952). Diese Konstante ist für die weiteren Betrachtungen ohne Bedeutung und kann daher wegfallen. Das Postulat der spezifischen Objektivität führt beim Modell von Rasch (1960) zu der Annahme identischer Diskriminationsparameter (für alle i, i' gilt $a_i = a_{i'}$). Daraus resultiert ein Modell der Form:

$$p_i(\theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}} \tag{4}$$

Die Diskussion um Vor- und Nachteile dieser verschiedenen ‚latent trait'-Modelle in der Diagnostik wurde bislang von zwei Hauptargumentationssträngen bestimmt.

Zum einen wird von der Frage nach den meßtheoretischen Implikationen ausgegangen, so vor allem von Fischer (1974). In diesem Punkt und bei der damit zusammenhängenden Frage nach der effektiven Schätzung der Modellparameter erweist sich das 1-parametrige Rasch-Modell dem 2-parametrigen Birnbaum-Modell überlegen (siehe Fischer, 1974; Rasch, 1960, 1961, 1966a, 1966b).

Der alternative Argumentationsstrang, der die Frage der Übereinstimmung der Modelle mit gegebenen Datensätzen verfolgt, wurde u. E. zum ersten Mal von Andersen (1973) entwickelt. Ausgehend von der Entwicklung eines goodness-of-fit-Tests für das Rasch-Modell überprüft er die Frage, ob die Annahme gleicher Diskriminationsparameter bei diesem Modell mit den Daten übereinstimmt, und kommt zu dem Ergebnis, daß die von ihm verwendeten Daten auf unterschiedliche Diskriminationsparameter hindeuten – also auf das Birnbaum-Modell.

In der vorliegenden Arbeit wird nun zunächst ein vereinfachter Test zum Vergleich des goodness-of-fit beider Modelle vorgestellt, der es ermöglicht, die Abweichungen im goodness-of-fit eindeutig auf die Unterschiede der Annahmen über die Diskriminationsparameter zurückzuführen.

Beschränken sich die bisher dargestellten Argumentationslinien ausschließlich auf Meß- bzw. Skalierungsfragen, womit die Funktion von Tests, nämlich entscheidungsrelevante Information zu liefern, nicht explizit einbezogen wurde, soll hier versucht werden zu demonstrieren, welche Konsequenzen für Entscheidungen sich aus der Verwendung beider Testmodelle für die Parameterbestimmung in adaptiven Testvorgabesystemen (tailored testing) ergeben.

Die Daten, die hier als Grundlage dienen, entstammen einem Mathematiktest, der als Übertrittstest von der Grundschule zum Gymnasium konzipiert ist. Der Test besteht aus 38 Items und wurde von 219 Vpn bearbeitet.

2. Parameterschätzung

Die Parameter für das Rasch-Modell wurden nach der Methode TOTW geschätzt (Fischer, 1974). Alle Items erwiesen sich als modellverträglich ($\chi^2 = 13.66$; df = 37).

Bei der Schätzung der Parameter des Birnbaum-Modells mit dem Programm von P.O. White traten zunächst die von Fischer (1974) beschriebenen Schwierigkeiten auf – einige Trennschärfeparameter wuchsen über alle Maßen. Die Parameterschätzung nach dem Birnbaum-Modell ist – als iteratives Verfahren – stark von der Güte der Anfangsschätzung abhängig. Es bot sich deshalb an, die Parameterschätzungen des Rasch-Modells als beste Anfangsschätzungen für den Birnbaum-Algorithmus einzusetzen. Tatsächlich konnte dadurch das Ausufern einzelner Parameter verhindert werden, und das Verfahren erreichte nach 25 Iterationen das Abbruchkriterium. Das Schätzverfahren wird abgebrochen, wenn sich der Wert der Funktion nicht mehr wesentlich ändert, d. h. wenn $(F_{n-1} - F_n)/F_{n-1} < 0.05$ ist.

Die Ergebnisse der Parameterschätzungen für beide Modelle finden sich in Tabelle 1 im Anhang.

3. Analyse der Abweichungen der Parameterschätzungen

Bei den Itemparametern b (Itemschwierigkeit) und a (Trennschärfe) und bei den Personenparametern θ zeigten sich Abweichungen zwischen den beiden Modellen, die durch keine Skalierungskonstante erklärbar waren. Abbildung 1 verdeutlicht die Ab-

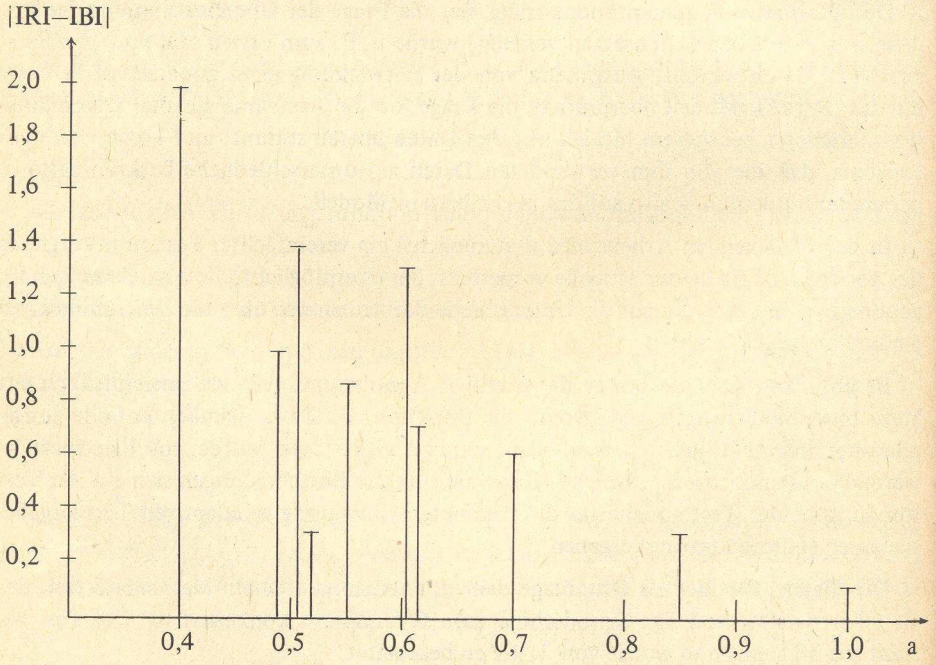

Abbildung 1: Absolute Abweichungen der Itemparameterschätzungen bei Rasch und Birnbaum in Abhängigkeit von a

weichungen der Itemschwierigkeitsparameter als Funktion des Trennschärfeparameters a, der aus der Schätzung nach dem Modell von Birnbaum resultierte.

Zur Klärung der Frage, ob diese Abweichungen durch das Birnbaum- oder Rasch-Modell bedingt sind, wurde die Anpassung der empirischen an die theoretisch erwarteten Häufigkeiten berechnet. Es war zu erwarten, daß das Modell von Birnbaum wegen des dritten Parameters a-priori besser mit den Daten übereinstimmt. Ist die Übereinstimmung aber nur um soviel besser, als aufgrund des zusätzlichen Parameters ohnehin zu erwarten ist, so ist im Sinne der Popper'schen „Sparsamkeit" dieser Parameter überflüssig und das Rasch-Modell vorzuziehen. Stimmen die empirischen und die theoretisch erwarteten Häufigkeiten bei dem Modell von Birnbaum jedoch wesentlich besser überein, so kann der zweite Parameter als nützlich und sinnvoll angesehen werden.

Für den χ^2-Test der Güte der Anpassung von theoretischen und empirischen Häufigkeiten war es nötig, die θ in Klassen zusammenzufassen, um genügend große Häufigkeiten für den Vergleich zu erhalten. Die 217 Vpn wurden in sechs ungefähr gleich große Klassen eingeteilt, deren neues θ_i der Mittelwert der r θ der entsprechenden Klasse ist:

$$\theta_i = \sum_{k=1}^{r} \theta_k / r.$$

3.1 Vergleich der empirischen und theoretischen Häufigkeiten

Für jedes Item konnte nun an sechs Stellen des Fähigkeitskontinuums θ die Lösungswahrscheinlichkeit auf der Item-Charakteristik-Kurve (ICC) errechnet werden. Multipliziert mit der Anzahl der Fälle in dieser Klasse ergaben sich die theoretischen Häufigkeiten. Über den χ^2-Test zur Güte der Anpassung wurden diese theoretischen Häufigkeiten mit den tatsächlich beobachteten Häufigkeiten in der jeweiligen Klasse verglichen.

Für jedes Item ergaben sich also zwei χ^2-Werte, einer für das Rasch-Modell und einer für das Birnbaum-Modell.

Die Anzahl der Freiheitsgrade errechnet sich wie folgt: Für die Festlegung der sechs θ und für jeden Modellparameter wird ein Freiheitsgrad abgezogen. Damit ergeben sich:
- für das Modell von Rasch ..... vier Freiheitsgrade,
- für das Modell von Birnbaum ..... drei Freiheitsgrade.

Die exakten χ^2-Wahrscheinlichkeiten p $(1-\alpha)$ wurden mit der SSP Subroutine CHISQ errechnet. Diese Wahrscheinlichkeiten sind umso kleiner, je besser die empirischen Häufigkeiten mit den theoretisch erwarteten Häufigkeiten übereinstimmen. Ist diese Übereinstimmung beim Birnbaum-Modell nur um soviel besser, als aufgrund des dritten Parameters erwartet werden kann, dann müßten die χ^2-Wahrscheinlichkeiten aufgrund der unterschiedlichen Freiheitsgrade annähernd gleich sein. Ob die Wahrscheinlichkeiten beim Birnbaum-Modell signifikant kleiner sind als beim Rasch-Modell, kann mit dem „Wilcoxon Matched-Pairs Signed-Ranks Test" überprüft werden.

Hypothesen:

$$H_1 : \chi^2_{p(1-\alpha)_{RASCH}} = \chi^2_{p(1-\alpha)_{BIRNBAUM}}$$

$$H_2 : \chi^2_{p(1-\alpha)_{RASCH}} > \chi^2_{p(1-\alpha)_{BIRNBAUM}}$$

Ergebnisse:

Mit einer Irrtumswahrscheinlichkeit von 2 % ist die Abweichung der empirischen von den theoretischen Werten beim Birnbaum-Modell geringer, als aufgrund des zusätzlichen Parameters ohnehin zu erwarten war.

Zur genaueren Analyse wurde der Itempool in Items mit niedriger, mittlerer und hoher Trennschärfe unterteilt.

- niedrige Trennschärfe $= 0.4015 \leqslant a \leqslant 0.8750$
- mittlere Trennschärfe $= 1.0105 \leqslant a \leqslant 1.2272$
- hohe Trennschärfe $= 1.2402 \leqslant a \leqslant 2.0648$

Für Items, die nach Birnbaum niedrige oder hohe Trennschärfe besitzen, ist die Anpassung der theoretischen und empirischen Werte beim Modell von Birnbaum besser als beim Modell von Rasch ($\alpha = .01$). Nur für Items mit mittlerer Trennschärfe konnten erwartungsgemäß keine signifikanten Unterschiede festgestellt werden.

Nun läßt sich einwenden, daß Items mit geringer Trennschärfe wenig Information zur Diskrimination zwischen verschiedenen Vpn beitragen und deshalb überflüssig sind. Es erschien deshalb sinnvoll, neben dem Vergleich über Anpassungstests, mittels der Informationsfunktion und einer Tailored-Testing-Vorgehensweise die beiden Testmodelle zu vergleichen.

4. Die Informationsfunktion

Die Informationsfunktion I ist eine Funktion von θ und ist geeignet, die Diskriminationsfähigkeit eines Items bzw. eines Tests für jedes Fähigkeitsniveau θ zu beschreiben. Allgemein hat die Informationsfunktion für logistische Modelle die Form:

$$I(\theta_v, x_i) = \frac{P'(\theta_v)^2}{P(\theta_v) \cdot Q(\theta_v)} \tag{5}$$

Gleichung (5) gibt das Informationsmaß eines Items x_i an, zur Diskrimination von Fähigkeiten um θ_v. Dabei ist

$$p'(\theta) = \frac{\delta}{\delta \theta} P(\theta) \text{ und } Q(\theta) = 1 - P(\theta).$$

Da

$$P'(\theta_v) = a_i P(\theta_v) \cdot Q(\theta_v),$$

läßt sich die Informationsfunktion für das Modell von Birnbaum in der Form

$$I(\theta_v, x_i) = a^2 \cdot P(\theta_v) \cdot Q(\theta_v) \tag{5.1}$$

schreiben.

Für das spezielle logistische Modell von Rasch vereinfacht sich die Informationsfunktion, so daß gilt:

$$I(\theta_v, x_i) = P(\theta_v) \cdot Q(\theta_v). \tag{6}$$

Da $$P'(\theta_v) = \left\{ \frac{\exp(\theta_v - b_i)}{1+\exp(\theta_v - b_i)} \right\}' = \frac{\exp(\theta_v - b_i)}{[1+\exp(\theta_v - b_i)]^2} = P(\theta_v) \cdot (1 - P[\theta_v]);$$

erhält man durch Einsetzen in (1)

$$I\ (\theta_v, x_i) = \frac{[P(\theta_v) \cdot Q(\theta_v)]^2}{P(\theta_v) \cdot Q(\theta_v)} = P(\theta_v) \cdot Q(\theta_v).$$

Wie aus den Gleichungen (5.1) und (6) leicht ersichtlich ist, hängt die Informationsfunktion von der ICC eines Items ab bzw. vom zugrundeliegenden Testmodell.

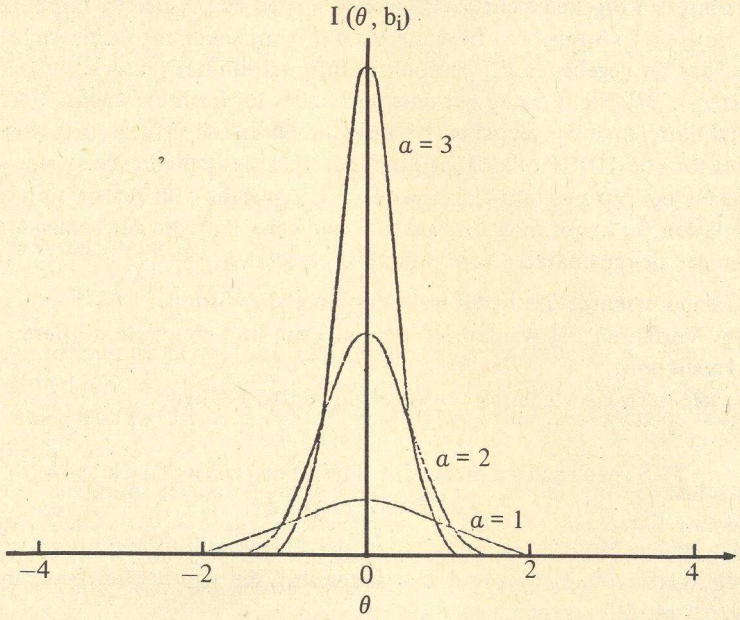

Abbildung 2: Informationsfunktion für logistische Items mit Schwierigkeit $b_i = 0$, in Abhängigkeit von drei verschiedenen a_i

Eine wichtige Forderung, die wir an ein Testmodell stellen, ist, daß es möglichst gut zwischen verschiedenen Fähigkeiten differenziert. Wenn also die beiden Modelle auf dieselbe Population angewendet werden, kann mit Hilfe der Informationsfunktion entschieden werden, welches Modell besser diskriminiert oder, mit anderen Worten, welches Modell die in den Daten enthaltene Information besser ausschöpft und deshalb zu bevorzugen ist.

Um unter Zuhilfenahme der Informationsfunktion die Diskriminationsfähigkeit eines jeden Modells zu untersuchen, erscheint eine adaptive Testvorgabe (tailored testing) besonders geeignet. Das Prinzip von tailored testing, zu dessen ausführlicher Beschreibung auf Lord (1970) verwiesen wird, ist dies: Eine Vp muß bei einem gegebenen Test nicht alle Aufgaben bearbeiten, sondern nur die, die ihrer Fähigkeit am besten entsprechen. Wenn man z. B. keine Information über die Fähigkeit einer Vp hat, wird man ihr zunächst eine mittelschwere Aufgabe vorlegen. Löst die Vp die Aufgabe, erhält sie eine schwierigere, löst sie die Aufgabe nicht, legt man ihr eine leichtere vor. Zahlreiche Untersuchungen (z. B. Lord, 1971; Zimmer & Lehmann, 1977) beschäftigen sich mit der optimalen „Schrittgröße" für das „auf und ab" in Abhängigkeit von der Lösung bzw. Nichtlösung einer Aufgabe und mit der Frage, bei welcher Schrittfolge die Abweichungen der wahren Fähigkeit der Vp von der durch tailored

testing geschätzten Fähigkeit am geringsten sind. Im Unterschied zu den bisherigen Untersuchungen, die von einer festen oder kontinuierlich abnehmenden Schrittgröße ausgingen, wird bei dieser Arbeit die Schrittgröße durch die Informationsfunktion bestimmt.

Das konkrete Vorgehen sieht also folgendermaßen aus: Einer Vp mit der Fähigkeit θ_v wird ein Item i vorgegeben. Löst die Vp das Item, bekommt sie als nächstschwierigeres das, das für gegebenes θ_v die höchste Information hat (ausgenommen natürlich das i-te Item). Bei Nichtlösung gilt entsprechendes für leichtere Items. Wechselt eine Vp viermal hintereinander zwischen Lösung und Nichtlösung (d. h. tritt eine Folge im Lösungsvektor von 10101 oder 01010 auf), so wird die Itemvorgabe abgebrochen und die Fähigkeit der Vp geschätzt. Verwendet man anstelle von realen Vpn eine Computersimulation, so kennt man das wahre θ und kann für jedes der beiden Modelle die Abweichungen des geschätzten vom wahren θ berechnen.

Eine Computersimulation bietet weiterhin noch die Vorteile:
- daß das Verfahren oft wiederholt werden kann und eine gute mittlere Schätzung für θ erzielt wird,
- daß dieselben Items wiederholt vorgegeben werden können.

5. Simulation mit der Informationsfunktion als Schrittgröße

Um die beiden Modelle an drei Stellen des Fähigkeitskontinuums zu vergleichen, wurden die Werte von θ_5, θ_{19} und θ_{33} ausgewählt, die sich bei der Berechnung nach dem Rasch-Modell ergeben hatten.

Bei der Schätzung nach dem Modell von Birnbaum ergeben sich durch die Gewichtung mit dem Trennschärfeparameter nicht nur andere Itemparameter, sondern auch vom Rasch-Modell abweichende Personenparameter. Die Verwendung der θ-Werte, die sich aus der Schätzung nach dem Rasch-Modell ergeben hatten, begünstigte daher eher das Rasch-Modell.

Für jedes der beiden Modelle werden die aus dem Mathematiktest resultierenden Parameter eingelesen und die ICC's und die Informationsfunktionen berechnet.

5.1. Ablauf der Simulation

Zu Beginn der Simulation wird die Fähigkeit θ_w (z. B. θ_5) eingelesen und die „Vp" bekommt Item 19 vorgelegt. Gleichzeitig wird ein Zufallsgenerator gestartet und sein Wert mit der Lösungswahrscheinlichkeit der „Vp" bei Item 19 verglichen. Ist der Wert des Zufallsgenerators kleiner oder gleich der Lösungswahrscheinlichkeit, so gilt das Item als gelöst. Als momentanes θ der „Vp" wird θ_{19} angenommen. Nun wird zwischen Item 20 und 38 dasjenige ausgewählt, das für θ_{19} die höchste Information hat. Hätte die „Vp" die Aufgabe nicht gelöst, so würde das Verfahren zwischen Aufgabe 1 und 18 dasjenige suchen, das für θ_{19} die höchste Information besitzt. Ab der Lösung bzw. Nichtlösung der zweiten Aufgabe kann das vorläufige θ der „Vp" genauer geschätzt werden. Dazu wird der Mittelwert von allen gelösten Aufgaben berechnet, die zwischen der leichtesten nicht gelösten und der schwierigsten gelösten liegen.

Angenommen, eine „Vp" hat folgenden, nach der Itemschwierigkeit geordneten Lösungsvektor:

13	14	15	17	19	20	22	23	24	Itemnummer
1	1	0	1	1	0	1	0	0	1 = gelöst 0 = nicht gelöst
		MIN				MAX			

so wird das vorläufige θ aus dem Mittelwert der Schwierigkeitsparameter von Item 17, 19 und 22 bestimmt.

Ausgehend von der Zufallsvariablen X_{vi}, mit den

Realisierungen x_{vi} = { 1 wenn Item i gelöst wird, / 0 wenn Item i nicht gelöst wird,

errechnet sich das vorläufige θ_v' einer Vp aus:

$$\theta_v' = \sum_{i=MIN}^{MAX} b_i^{\,x_{vi}}/N \tag{7}$$

Der Algorithmus vergleicht nun das momentan geschätzte θ_v' mit den möglichen, d. h. mit den empirisch gefundenen, wählt von den möglichen das aus, das θ_v' am nächsten liegt und sucht für dieses θ die Informationsfunktion nach dem am besten geeigneten Item ab.

Die „Vp" bekommt solange Items vorgelegt, bis sie viermal hintereinander zwischen Lösung und Nichtlösung hin und her oszilliert. Tritt also bei der Zufallsvariable x_{vi} die Folge 01010 oder 10101 auf, so bricht das Verfahren ab und berechnet das endgültige θ der „Vp" als Mittelwert der Schwierigkeit der letzten fünf Items.

$$\theta_v = \sum_{i=n-4}^{n} b_i/5 \tag{8}$$

Es ist klar, daß das Ergebnis der θ-Schätzung vom Verhalten des Zufallsgenerators mit beeinflußt ist. Um solche unerwünschten Effekte auszuschalten, wurde für jedes der beiden Modelle und jedes der drei θ die Simulation 100 mal durchgeführt.

5.2. Ergebnisse der Simulation

Um die Güte der θ-Schätzungen, die sich aus dem oben beschriebenen Verfahren ergaben, zu bestimmen, wurde die mittlere quadrierte Abweichung der geschätzten von den wahren Werten errechnet. Tabelle 2 zeigt die Ergebnisse.

Tabelle 2

Mittlere quadrierte Abweichungen (μ) der wahren von den geschätzten θ bei der tailored testing-Simulation für das Rasch- und das Birnbaum-Modell

		Rasch	Birnbaum
μ	θ_5	.8790	.7061
σ		.2942	.1262
μ	θ_{19}	.7010	.2176
σ		.2808	.0241
μ	θ_{33}	.3306	.3903
σ		.3060	.0854

5.3. Überprüfung der Signifikanz der Abweichungen

Um zu überprüfen, ob bei einem Modell die Abweichungen der geschätzten von den wahren Werten signifikant geringer sind, wurde der t-Test für unkorrelierte Stichproben durchgeführt.

Hypothesen: $H_0 : \mu_R = \mu_{BB}$

$H_1 : \mu_R > \mu_{BB}$

μ_R = mittlere quadrierte Abweichung der wahren von den durch tailored testing geschätzten θ bei Verwendung des Rasch-Modells,

μ_{BB} = mittlere quadrierte Abweichung der wahren von den durch tailored testing geschätzten θ unter Verwendung der Informationsfunktion des Birnbaum-Modells

Mit einer Irrtumswahrscheinlichkeit von 1 % wurde die H_0 für θ_5 und θ_{19} zurückgewiesen, d. h. die Abweichungen der wahren von den geschätzten θ sind bei Verwendung des Birnbaum-Modells geringer. Das heißt weiterhin, daß die Verwendung von trennschwachen Items durchaus sinnvoll ist und nicht zwangsläufig zu schlechten Schätzungen für θ führen muß. Insbesondere für tailored testing ist es sinnvoll, auf trennschwache Items zurückzugreifen, solange der grobe Bereich bestimmt wird, in dem die Vp liegt, und die trennscharfen Items erst für die exakte Bestimmung der Fähigkeit der Vp zu verwenden.

6. Zusammenfassung

Anhand eines neu konstruierten Mathematiktests werden das zwei- und das dreiparametrige logistische Testmodell miteinander verglichen. Der Test besteht aus 38 Aufgaben, die Aufgabenparameter für beide Modelle wurden an einer Stichprobe von 219 Schülern der 4. Klasse Grundschule geschätzt, die Modellkontrolle für das Rasch-Modell wies keine signifikanten Abweichungen auf. Durch Einsetzen der Parameter des Rasch-Modells als Anfangswerte in den Birnbaum-Algorithmus konnten die Schätzprobleme des Drei-Parameter-Modells von Birnbaum gelöst werden.

Über einen χ^2-Anpassungstest wird die Übereinstimmung der empirischen und der theoretisch erwarteten Lösungshäufigkeiten für beide Modelle verglichen. Trotz Abzugs eines Freiheitsgrades für den zweiten Aufgabenparameter zeigt das Modell von Birnbaum eine höhere Übereinstimmung der empirischen und theoretischen Werte.

Die Brauchbarkeit der Modelle für das adaptive Testen wird durch eine Computersimulation untersucht; sie soll Aufschluß geben, bei welchem Modell weniger Fehlklassifikationen auftreten. Die Verwendung der Informationsfunktion als Auswahlkriterium der jeweils vorzugebenden Aufgabe führte beim Modell von Birnbaum zu geringeren Abweichungen zwischen wahrem und durch adaptives Testen geschätztem Personenparameter.

Summary

The two- und three-parameter logistic models are compared using a (newly constructed) test on elementary mathematics. The parameters of the 38 item-test were

estimated for both models by studying a sample of 219 4th graders. No significant deviation between the Rasch-model and the data was found. Using the parameters of the Rasch-model as starting-values for the Birnbaum-algorithm, the estimation-problems of the 3-parameter model could be solved.

For both models the goodness of fit between the empirical and the theoretical frequencies was tested with a χ^2 test. The Birnbaum-model showed a better fit to the data even though one degree of freedom had been substracted for the additional model parameter.

The applicability of the two models to a tailored-testing-procedure was investigated with a computer-simulation designed to determine which model leads to less missclassifications. The information function was used as the criterion for item-selection. For this kind of tailored-testing procedure the Birnbaum-model showed significantly smaller deviations between the true parameters and the tailored-testing parameters.

Literatur

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

Birnbaum, A.: Some latent trait models and their use in inferring an examinee's ability, in: Lord, F. M. & Novick, M. R.: *Statistical Theories of mental test scores.* Addison-Wesley: Reading – London 1968.

Fischer, G. H.: *Einführung in die Theorie psychologischer Tests.* Huber: Bern 1974.

Guilford, J. P.: *Psychometric Methods.* McGraw-Hill: New York 1938.

Haley, D. C.: Estimation of the dosage mortality relationship when the dose is subject to error. *Techn. Rep. Nr. 15, Applied Mathematics and Statistics lab.* Stanford University: Stanford 1952.

Lord, F. M.: An application of confidence intervals and of maximum likelihood to the estimation of an examinees ability. *Psychometrika,* 1953, **18**, 57–76.

Lord, F. M.: An analysis of the verbal scholastic aptitude test using Birnbaum's threeparameter logistic model. *EPM,* 1968, **28**, 989–1020.

Lord, F. M.: Some test Theory for tailored testing, in: Holtzman, W. H.: *Computerassisted instruction, testing and guidance.* Harper & Row: New York 1970.

Lord, F. M.: Robbins-Monro procedures for tailored testing. *EPM,* 1971, **31**, 3–31.

Orth, B.: *Einführung in die Theorie des Messens.* Kohlhammer: Stuttgart 1974.

Rasch, G.: *Probabilistic models for some intelligence and attainment tests.* Nielson & Lydicke: Kopenhagen 1960.

Rasch, G.: An item analysis which takes individual differences into account. *Brit. J. Math. Stat. Ps.* 1966a, **19**, 49–57.

Rasch, G.: An individualistic approach to item analysis, in: Lazarsfeld, P. F. & Henry, N. W. (Eds.): *Readings in Mathematical Social Science.* Science Res. Ass.: Chicago 1966b, 89–108.

Zimmer, A., & Lehmann, B.: Individualisierte Vorgabe eines Persönlichkeitsinventars, in: Tack, W. H. (Hrsg.): *Bericht über den 30. Kongreß der DGfPs.* Hogrefe: Göttingen 1977.

Anmerkung:

Für die Schätzung der Parameter des Birnbaum-Modells wurde verwendet:

White, P. O.: *Program for Birnbaum 2-parameter model.* London 1974, adapted for Siemens 4004 by B. Färber & F. Sachse, Regensburg 1976.

Anschrift der Autoren:

Dipl.-Psych. Berthold Färber, Psychologisches Institut, Universität Tübingen, Friedrichstraße 21, 7400 Tübingen

Prof. Dr. Alf Zimmer, Fach Psychologie, Universität Oldenburg, Ammerländer Heerstraße 67–99, 2900 Oldenburg