GOODNESS-OF-FIT TESTS FÜR DAS RASCH-MODELL

Reinhold HATZINGER*)

Forschungsbericht/
Research Memorandum No. 173

Juli 1982

*) Scholar am Institut für höhere Studien

Die in diesem Forschungsbericht getroffenen Aussagen liegen im Verantwortungsbereich des Autors und sollen daher nicht als Aussagen des Instituts für Höhere Studien wiedergegeben werden.

INHALT

GOODNESS-OF-FIT TESTS FÜR DAS RASCH-MODELL

Zusammenfassung:

Die Erfüllung von vier Voraussetzungen bildet die Grundlage für die Gültigkeit des Rasch-Modells und seiner Folgerungen wie spezifische Objektivität oder Stichprobenunabhängigkeit der Ergebnisse. Während die meisten Goodness-of-fit Statistiken die Annahmen über Suffizienz und Monotonizität zu prüfen erlauben, wurden erst unlängst einige Tests vorgeschlagen mit denen Unidimensionalität und lokale stochastische Unabhängigkeit einschätzbar wird. In dieser Arbeit werden die Grundlagen des Rasch-Modells dargestellt sowie die entsprechenden Goodness-of-fit Tests diskutiert. Zusätzlich sind einige Vorschläge zur graphischen Analyse enthalten. Es wurde der Versuch unternommen, einen Überblick über diagnostische Hilfsmittel zur Erkennung von Verletzungen der Annahmen des Rasch-Modells zu geben.

Abstract:

Four requirements to be met are basic to the validity of the Rasch-model and its consequences such as specific objectivity or results independent of the sample. Whereas most goodness-of-fit statistics concentrate on the assumptions of sufficiency and monotonicity more recently some tests have been suggested to prove unidimensionality and stochastic independence. This paper presents the axioms of the Rasch-model and gives a discussion of the corresponding goodness-of-fit tests. Additionally suggestions for graphical analyses are included. The attempt has been made to provide a survey on diagnostical aids for detecting failures of the Rasch-model.

1. EINLEITUNG

Zwei wesentliche Arbeitsrichtungen lassen sich im Bereich psychometrischer Forschung unterscheiden: die klassische Testtheorie und die Analyse latenter Strukturen. (Diese Methodenunterscheidung gilt natürlich nicht nur für die Psychometrie sondern wird auch im Rahmen der Soziometrie u.a. getroffen. Die größere Verbreitung der entsprechenden Ideen in der Psychologie dürfte in erster Linie auf die längere Tradition und spezielle Problematik rückführbar sein.)

Bei der Gegenüberstellung der beiden Ansätze stößt man notwendigerweise auf das Problem der Messung in den Sozialwissenschaften, seine wissenschaftstheoretischen Implikationen und die Behandlung von Meßfehlern. Das Hauptanliegen, von beiden Richtungen in verschiedener Weise formuliert, besteht darin, auf Grund gewisser beobachtbarer Merkmale oder bestimmter Tatbestände (z.B. angekreuzte Antwort einer Fragestellung in einem Intelligenztest[§]) entweder auf strukturelle Beziehungen zwischen diesen Tatbeständen (z.B. Zusammenhänge in der Beantwortung verschiedener Fragen des Intelligenztests) oder auf zugrundeliegende Variablen im Rahmen hypothetischer oder theoretischer Konstruktionen (z.B. was ist Intelligenz, wie ist sie strukturiert und wie läßt sie sich messen) zu schließen und solches zu formalisieren.

(§) In der vorliegenden Arbeit werden die Beispiele vor allem aus dem Bereich der psychologischen Intelligenzforschung gewählt, da sich diese Beispiele recht gut zur Illustration der entsprechenden Probleme eignen. Dies bedeutet jedoch nicht, daß die hier dargestellten Ergebnisse nur auf dieses spezielle Gebiet oder den Bereich der Fragebogenmessung im allgemeinen beschränkt sind.

Die Grundgleichung ("basic equation") der klassischen Testtheorie lautet:

(1) $$y = \tau + \varepsilon$$

und beschreibt eine Beobachtung y als zusammengesetzt aus einem "wahren" Wert oder "true-score" τ und einem Fehlerterm. Das methodische Gerüst besteht im wesentlichen aus dem allgemeinen linearen Modell unter der Annahme normalverteilter Fehler. (Verschiedene Arten der Regressionstechnik, multivariate Methoden und die Faktorenanalyse im speziellen stehen in enger Beziehung zum klassischen Ansatz.) Da der Erwartungswert des Fehlers definitionsgemäß gleich Null ist und somit der Erwartungswert der Beobachtungen gleich dem wahren Wert ist, kann die klassische Testtheorie eigentlich als Fehlertheorie bezeichnet werden. Das Beobachtbare ist deterministisch an das "zu messende" geknüpft; Variation, die nicht auf die Struktur des true-scores zurückführbar ist, wird als Fehler der Messung aufgefaßt.

Der zweite Ansatz, die Analyse latenter Strukturen, geht in erster Linie auf Lazarsfeld (1950a,1950b) zurück, der erstmals die Trennung zwischen zugrundeliegenden und beobachtbaren Variablen vollzog, wobei diese nur als Indikatoren oder Symptome jener aufgefaßt werden. Die Berücksichtigung des stochastischen Charakters der Beobachtungen führt zur Aufweichung der deterministischen Beziehung, wie sie im Rahmen der klassischen Theorie postuliert wird. Die Einführung latenter Strukturen erlaubt es, Kovariationen zwischen Beobachtungen als deren gemeinsame Abhängigkeit von Parametern, die die latente Struktur beschreiben, zu interpretieren.

Lazarsfelds Überlegungen wurden von G.Rasch (1960,1961,1966) erweitert und führten zu einer allgemeinen Meßtheorie, deren wesentlichste Forderungen die Stichprobenunabhängigkeit und

spezifische Objektivität der Ergebnisse betreffen. Zur Erfüllung dieser Eigenschaften, auf die noch näher einzugehen sein wird, ist allerdings Voraussetzung, daß das Modell mit all seinen Annahmen gilt, das G.Rasch (1960) als Formalisierung seiner Theorie zugrundelegte.

Eine Vielzahl von Arbeiten, die in den letzten Jahren zum Rasch-Modell publiziert wurden, befassen sich mit Goodness-of-fit Tests als diagnostischem Mittel zur Prüfung der Modellgeltung. Eine kritische Gegenüberstellung und abgerundete Diskussion steht allerdings noch aus. Die hier vorgelegte Arbeit versucht einen Beitrag zu solch einer Integration zu leisten und einige offen scheinende Bereiche zu ergänzen.

Als erstes sollen nun die Voraussetzungen des Modells und ihre Implikationen genauer behandelt werden.

2. DAS RASCH-MODELL UND SEINE VORAUSSETZUNGEN

Es wurde bereits erwähnt, daß im Rahmen der klassischen Testtheorie eine Beobachtung als fehlerbehaftete Realisation einer zugrundeliegenden Zufallsvariable aufgefaßt wird. Die Zuordnung ist durch eine deterministische Äquivalenzrelation festgelegt, der Einfluß des Meßinstruments auf die Messung wird nicht direkt berücksichtigt, sondern ist Teil der Zufallsschwankung, also des Fehlerterms.

Der wesentliche Unterschied zwischen dem klassischen Ansatz und der neueren Theorie zur Analyse latenter Strukturen betrifft nun die Annahme einer latenten Dimension (wie sich schon aus der Bezeichnung erkennen läßt). Diese wird hier als nicht direkt beobachtbar formuliert und ist somit nicht, wie vorher schon dargestellt, deterministisch mit der beobachtbaren oder manifesten Variable verbunden. In der probabilistischen Meßtheorie im allgemeinen und im Ansatz von Rasch im speziellen wird die Wahrscheinlichkeit eine bestimmte Realisation zu beobachten und nicht die Beobachtung selbst modelliert. Das Prinzip der Suffizienz, auf das noch einzugehen sein wird, dient zur Bestimmung von Parametern, die die Quantifikation der latenten Dimension leisten sollen. Kennt man die Anzahl richtiger Lösungen, die eine Versuchsperson in einem bestimmten Test erzielt hat, so soll dieser Wert als Indikator für den Grad der Leistungsfähigkeit ebendieser Person dienen. Bezeichnete man die Häufigkeit korrekter Lösungen schon als deren Fähigkeit, wäre dies ebenso unsinnig wie etwa die Verwechslung des Begriffs Temperatur mit der Verbiegung eines Bimetallstreifens oder dem elektrischen Widerstand eines Drahtstückes (s.G.Fischer,1974,S.183). Es wird hier also nicht die suffiziente Statistik als zu messende Eigenschaft aufgefaßt, sondern diese dient zur Schätzung der Parameter. Erst diese werden als das zu messende aufgefaßt.

Im Modell von Rasch wird überdies noch der Einfluß des Meßinstruments auf die Messung berücksichtigt, hängt doch eine Messung, um es im psychologischen Kontext zu formulieren, nicht nur vom Verhalten oder der Eigenschaft des Meßobjekts sondern auch von den Charakteristika der Reizsituation oder des Meßinstruments ab. Im Rasch-Modell wird also zusätzlich diese Wechselbeziehung mitbeachtet. Kehrt man zum Beispiel des Intelligenztests zurück, so wird die Wahrscheinlichkeit der richtigen Beantwortung einer Aufgabe nicht nur von der Fähigkeit ξ_i der antwortenden Person i sondern auch von der Schwierigkeit ε_j des j-ten Testitems (Aufgabe) abhängen.

Betrachtet man nun die Ergebnisse der Untersuchung von n Personen i (i=1,..,n), die k Testitems j (j=1,..,k) zu lösen hatten, so lautet Raschs Modellierung einer Einzelbeobachtung

$$(2) \qquad p(Y_{ij}=y_{ij} \mid \xi_i, \varepsilon_j) = \frac{\exp(\xi_i - \varepsilon_j)}{1 + \exp(\xi_i - \varepsilon_j)}$$

wobei die Realisierung y_{ij} der Bernoulli-Variable Y_{ij} mit 1 bei einer richtigen Lösung, bzw. mit 0 sonst kodiert wird.

Folgende Bedingungen, die auch als notwendige Forderungen an eine sozialwissenschaftliche Messung im weitesten Sinn aufgefaßt werden können, führen zur Definition von (2), bzw. sind für die Gültigkeit von (2) notwendig und hinreichend:

i) AXIOM DER UNIDIMENSIONALITÄT

Die Meßinstrumente (in unserem Beispiel die Aufgaben des Intelligenztests) sind eine endliche Stichprobe aus einem Universum denkbarer Instrumente, die im einfachsten Fall

(auf den diese Arbeit beschränkt bleibt) dichotome Messungen zulassen (also Vorhandensein oder Nichtvorhandensein eines Symptoms - richtige oder falsche Lösung). Da eine bestimmte Eigenschaft gemessen werden soll, müssen die Meßinstrumente dasselbe messen, oder in anderen Worten, sie sollen der latenten Dimension homogen sein. (Eine Motivation im sozialwissenschaftlichen Kontext hiefür läßt sich anhand folgender Überlegung aufzeigen: Die einmalige Beobachtung eines Ereignisses reicht zur Bestimmung seiner Auftretenswahrscheinlichkeit nicht aus. Mehrmalige Beobachtungen erfordern aber oft verschiedene Beobachtungsinstrumente. Wird im Beispiel des Intelligenztests einmal eine Aufgabe bearbeitet oder vielleicht sogar richtig gelöst, so scheint eine wiederholte Vorgabe derselben "Prüfungsfrage" sinnlos. Die Vorgabe mehrerer verschiedener Aufgaben macht es aber notwendig, daß diese Testitems dasselbe, also in unserem Beispiel eine bestimmte Intelligenzdimension wie etwa räumliche Vorstellungskraft messen.)

ii) AXIOM DER MONOTONIZITÄT

Die Wahrscheinlichkeitsfunktion für ein bestimmtes j entlang der latenten Dimension ξ (oder Itemcharakteristikkurve (ICC) für das Item j), in Zeichen $f_j(\xi)$ soll für alle j monoton steigend sein und es soll gelten:

$0 \leq f_j(\xi) \leq 1$, wobei $\xi \in R$,

$f_j(\xi) \to 0$ wenn $\xi \to -\infty$, und $f_j(\xi) \to 1$ wenn $\xi \to +\infty$.

Im Beispiel bedeutet dies: mit zunehmender Leistungsfähigkeit einer Person, die sich in einem hohen Wert auf der latenten Dimension ξ repräsentiert, soll die Wahrscheinlichkeit Aufgabe j richtig zu lösen immer größer werden und umgekehrt. Zusätzlich soll gewährleistet sein, daß keine

Person mit absoluter Sicherheit richtige oder falsche Antworten gibt. In diesem Fall wäre wieder eine deterministische Beziehung postuliert, die aber offensichtlich empirisch nicht gerechtfertigt werden kann.

iii) AXIOM DER (LOKALEN) STOCHASTISCHEN UNABHÄNGIGKEIT

Lokale stochastische Unabhängigkeit der Beobachtungen, als dritte Voraussetzung, bedeutet, daß die Kovariation der Einzelbeobachtungen nur von der Parameterstruktur auf der latenten Dimension abhängt. Es gilt

$$p(Y_{ij}=y_{ij}, Y_{ij+1}=y_{ij+1} \mid \xi) = f_j(\xi_i) \cdot f_{j+1}(\xi_i)$$

und

$$p(Y_{ij}=y_{ij}, Y_{i+1j}=y_{i+1j} \mid \xi) = f_j(\xi_i) \cdot f_j(\xi_{i+1})$$

für alle i und alle j. Die Beantwortung einer Frage soll demnach nicht davon abhängen, ob und welche Frage vorher schon richtig beantwortet wurde, bzw. ob und wie eine andere Person diese Frage beantwortet.

iv) AXIOM DER SUFFIZIENZ

Die Summe der Einzelbeobachtungen sollen erschöpfende Statistiken für die zu schätzenden Parameter sein. Weiß man also wieviele Testitems von einer Person richtig gelöst wurden, so soll dies die gesamte Information ausschöpfen, die zur Schätzung der Leistungsfähigkeit dieser Person notwendig ist.

Sei

$$t_i = \sum_j y_{ij} \ ,$$

wobei y_{ij} wie oben eine einzelne Beobachtung bezeichnet, die entsprechend mit 1 (richtig) oder 0 (falsch) kodiert wird, so soll die Größe t_i ausreichen den Parameter ξ_i zu schätzen. Es soll demnach die Kenntnis welche einzelnen Items gelöst wurden keine zusätzliche Information zur Bestimmung des Grades der Leistungsfähigkeit der Person i liefern. Damit diese Bedingung erfüllt ist soll nach dem Faktorisierungstheorem von Fisher und Neyman die gemeinsame Wahrscheinlichkeitsverteilungsfunktion der Y_{ij} wie folgt zerlegt werden können:

$$(3) \qquad f(y_{i1},\ldots,y_{ik} \mid \xi_i) = g(t_i \mid \xi_i) \cdot h(y_{i1},\ldots,y_{ik})$$

wobei g von y_{ij} nur durch t_i abhängt und h von ξ_i völlig unabhängig ist. (Vgl.G.G.Roussas,1973).

In der Praxis wird nun folgenderweise vorgegangen. Sind einmal Daten erhoben, so werden die Parameter des Modells (2) unter der Annahme geschätzt, daß die eben dargestellten Axiome i) - iv) gelten. Daran anschließend versucht man, die Gültigkeit dieser Annahmen zu überprüfen und gegebenenfalls zu bestätigen, wozu verschiedene inferenzstatistische und graphisch explorative Verfahren entwickelt wurden.

Die Methoden zur Parameterschätzung und ihre wissenschaftstheoretischen Implikationen werden im nächsten Kapitel behandelt. In den darauf folgenden Abschnitten wird auf die einzelnen Tests zur Modellkontrolle eingegangen.

3. BEDINGTE UND UNBEDINGTE MAXIMUM LIKELIHOOD SCHÄTZUNG IM RASCH-MODELL

Gegeben seien k Realisationen einer Bernoulli-Variable, wobei t_i Erfolge beobachtet werden konnten. Unter der Annahme stochastischer Unabhängigkeit und suffizienter Statistiken (Axiome iii) und iv);) läßt sich die Wahrscheinlichkeit, daß eine Person i eine bestimmte Sequenz oder ein Muster A_i von Antworten gibt, als das Produkt der Einzelwahrscheinlichkeiten (2) anschreiben:

$$(4) \qquad p(Y_{i1}=y_{i1},..,Y_{ik}=y_{ik} \mid \xi_i, \varepsilon_1,..,\varepsilon_k) =$$

$$p(A_i=a_i \mid \xi_i, \varepsilon_1,..,\varepsilon_k) = \prod_j p(Y_{ij}=y_{ij} \mid \xi_i, \varepsilon_j) =$$

$$\prod_j \frac{\exp(y_{ij}[\xi_i-\varepsilon_j])}{1+\exp(\xi_i-\varepsilon_j)} =$$

$$\frac{\theta_i^{t_i} \cdot \exp(-\sum_j \varepsilon_j y_{ij})}{\prod_j [1+\exp(\xi_i-\varepsilon_j)]} \quad ; \quad \text{mit} \quad \theta_i = \ln \xi_i \, .$$

Nimmt man zusätzlich Unabhängigkeit zwischen den Personen an (die z.B. dann nicht gegeben wäre, wenn bei gleichzeitiger Testung mehrerer Personen diese voneinander abschreiben könnten), so erhält man die "Gesamtwahrscheinlichkeit" oder Likelihood der beobachteten Daten

$$(5) \qquad L_u = \frac{\exp(-\sum \varepsilon_j s_j) \cdot \exp(\sum \xi_i t_i)}{\prod_i \prod_j [1+\exp(\xi_i-\varepsilon_j)]} \, , \text{ wobei } s_j = \sum_i y_{ij} \, .$$

Maximieren der logarithmierten Likelihood $L_u(5)$, u steht für unbedingt oder "unconditional", ergibt die Parameterschätzer für ξ_i bzw. ε_j . Die ersten partiellen Ableitungen von log L_u nach ξ und ε liefern die Schätzer, die Inverse der negativen Matrix der zweiten partiellen Ableitungen ergibt die asymptotische Varianz-Kovarianz-Matrix der Parameterschätzer.

Die eben beschriebene Vorgangsweise entspricht den üblichen Gepflogenheiten zur Bestimmung von Parameterwerten und beruht auf gut fundierten Kenntnissen der Exponentialfamilie von Verteilungen. Im hier behandelten Fall lassen sich allerdings drei Argumente anführen, die eine Verwendung der Methode der unbedingten Maximum Likelihood-Schätzung (UML) nicht empfehlenswert erscheinen lassen. Das betrifft erstens die Konsistenz der Parameterschätzer, weiters die Unterscheidung von inzidentellen und strukturellen Parametern nach Neyman und Scott (1948) und schließlich den Begriff der Stichprobenunabhängigkeit der Ergebnisse oder spezifische Objektivität, der als wesentlichste Schlußfolgerung der Theorie von Rasch aufgefaßt wird. Diese drei Argumente führen zur bedingten oder "conditional" Maximum Likelihood-Schätzung (CML), wie im folgenden zu zeigen sein wird.

1) zur Konsistenz

E.B.Andersen (1973a) konnte zeigen, daß die Maximierung der unbedingten Likelihood (5) inkonsistente Parameterschätzer liefert, die einen asymptotischen Bias von $k/(k-1)$ aufweisen wenn $n \to \infty$. Der Beweis ist für $k = 2$ Items relativ einfach zu geben; allerdings ist die folgende Überlegung voranzustellen, die auch im Fall $k > 2$ gilt: die Parameter ξ_i und ε_j sind nur bis auf eine additive Konstante bestimmbar, da aus $\xi_i^* = \xi_i + c$ und $\varepsilon_j^* = \varepsilon_j + c$ folgt, daß $\xi_i^* - \varepsilon_j^* = \xi_i - \varepsilon_j$. Diese Mehrdeutigkeit läßt sich

aber durch Einführen einer Normierungsbedingung, wie z.B. $\varepsilon_1 = 0$ oder der in weiterer Folge verwendeten $\sum \varepsilon_j = 0$, leicht beheben. Weiters benötigt man noch die ML-Gleichungen

$$(6) \qquad t_i = E(T_i) = \sum_j p(Y_{ij}=y_{ij} \mid \xi_i, \varepsilon_j)$$

und

$$(7) \qquad s_j = E(S_j) = \sum_i p(Y_{ij}=y_{ij} \mid \xi_i, \varepsilon_j) \quad ,$$

die sich nach partiellem Ableiten von (5) ergeben.

Sei also die Anzahl der Items $k = 2$. Aufgrund der Normierung ist dann $\varepsilon_1 = -\varepsilon_2$. Einsetzen in (6) ergibt die Schätzgleichung für $\hat{\xi}_i$

$$(8) \qquad t_i = \frac{\exp(\xi_i - \varepsilon_1)}{1 + \exp(\xi_i - \varepsilon_1)} + \frac{\exp(\xi_i + \varepsilon_1)}{1 + \exp(\xi_i + \varepsilon_1)} \quad .$$

Da im Fall $k = 2$ t_i die Werte 0, 1 oder 2 annehmen kann, ergeben sich folgende Lösungen für $\hat{\xi}_i$:

$$\hat{\xi}_i = \begin{cases} -\infty & \text{für } t_i = 0 \\ 0 & \text{für } t_i = 1 \\ +\infty & \text{für } t_i = 2 \end{cases}$$

Daraus erhält man nun unter Verwendung von (7) die Lösungs-

gleichung für $\hat{\varepsilon}_1$ durch Einsetzen der entsprechenden Werte von $\hat{\xi}_i$

$$(9) \qquad s_1 = n_0 \cdot 0 + n_1 \cdot e^{-\varepsilon_1}/(1+e^{-\varepsilon_1}) + n_2 \cdot 1 \ ,$$

wobei n_r für die Anzahl der Individuen mit Rohscore $t_i = r$, $(r = 0, 1$ oder $2)$, steht. Da

$$n_1 - s_1 + n_2 = \text{Anzahl der Personen mit } (0,1) \ ,$$

d.h. mit genau dem Antwortmuster: erstes Item nicht gelöst, zweites gelöst, und

$$s_1 - n_2 = \text{Anzahl der Personen mit } (1,0) \ ,$$

gilt auf Grund des Gesetzes der großen Zahlen

$$\frac{n_1-s_1+n_2}{n} \xrightarrow{P} \lim \left\{ \frac{1}{n} \sum_i \frac{\exp(\xi_i+\varepsilon_1)}{[1+\exp(\xi_i-\varepsilon_1)]\,[1+\exp(\xi_i+\varepsilon_1)]} \right\}$$

sowie

$$\frac{s_1-n_2}{n} \xrightarrow{P} \lim \left\{ \frac{1}{n} \sum_i \frac{\exp(\xi_i-\varepsilon_1)}{[1+\exp(\xi_i-\varepsilon_1)]\,[1+\exp(\xi_i+\varepsilon_1)]} \right\} .$$

Daraus folgt, daß das Verhältnis $(n_1-s_2+n_2)/(s_1-n_2)$ in Wahrscheinlichkeit gegen das Verhältnis der beiden Limiten konvergiert und daher gilt

$$\hat{\varepsilon}_1 = \ln(e^{\varepsilon_1}/e^{-\varepsilon_1}) = 2\hat{\varepsilon}_1 \ .$$

Konsistente Schätzungen für die $\hat{\varepsilon}_j$ erhält man unter Verwendung der Korrektur

$$\hat{\varepsilon}^*_j = k/(k-1)\ \hat{\varepsilon}_j \quad .$$

Simulationsstudien ergaben, daß eine solche Vorgangsweise dann gerechtfertigt ist, wenn die bedingte ML-Schätzung aus welchen Gründen auch immer versagt (pers. Mitteilung von E.B.Andersen;s.a. G.Fischer,1974,S.260).

2) Unterscheidung zwischen strukturellen und inzidentellen Parametern

Solche eben beschriebenen Situationen, wo ML-Schätzer inkonsistent sind, wurden von Neyman und Scott (1948) erkannt, die darauf aufmerksam machten, daß eine Identifizierung von Modellparametern dann in Frage gestellt ist, wenn jede Person bei Vergrößerung der Stichprobe mindestens einen neuen, unbestimmten Parameter mit sich bringt, da in diesem Fall die Präzision der Parameterschätzung nicht ohne weiteres erhöht werden kann, wenn die Zahl der Personen erhöht wird. Mehr Personen bedeuten dann nicht präzisere Schätzung sondern immer neue, unbekannte Parameter.

Eine Lösung dieser Problematik ergibt sich aus der Definition sogenannter Strukturparameter und inzidenteller Parameter, wie sie von Neyman und Scott gegeben wurde. Der Unterschied zwischen beiden besteht im wesentlichen nur bezüglich ihrer Anzahl. Strukturparameter eines Modells sind dadurch definiert, daß ihre Anzahl endlich ist und daß sie in beliebig vielen Zufallsvariablen als Verteilungsparameter vorkommen. Je größer die Stichprobe der Beobachtungen, umso größer wird der auszuschöpfende Betrag an statistischer Information in Bezug auf diese Strukturparameter, d.h. umso genauer können diese geschätzt werden. Strebt die Anzahl der Beobachtungen

gegen unendlich, so konvergiert die Schätzfehlervarianz der Strukturparameter gegen Null, die Schätzungen sind demnach konsistent.

Inzidentelle Parameter hingegen sind dadurch definiert, daß jeder von ihnen in höchstens endlich vielen beobachtbaren Zuvallsvariablen als Verteilungsparameter enthalten sind, ihre Anzahl aber mit wachsender Stichprobengröße gegen unendlich strebt. Über inzidentelle Parameter können also keine genaueren Angaben gemacht werden. Damit aber diese nicht näher bestimmbaren Parameter die Strukturparameterschätzung nicht beeinträchtigen, erhebt sich die Forderung, sie im Zuge des Schätzvorgangs zu eliminieren. Genau diese Forderung wird durch das Modell von Rasch erfüllt. Die Verwendung der bedingten Likelihood ermöglicht einerseits eine Schätzung der Strukturparameter unabhängig von den inzidentellen Parametern und erlaubt andererseits aus Symetriegründen auch Aussagen über die inzidentellen Parameter. Auf die Herleitung der bedingten Maximum Likelihood soll im folgenden näher eingegangen werden.

3) bedingte ML-Schätzung und spezifische Objektivität

Betrachtet man nicht nur eine einzelne Realisation eines Antwortvektors A_i wie in (4) sondern die Menge aller möglichen Muster A, die einen bestimmten Randscore $T_i = t_i$ erfüllen, so erhält man die Wahrscheinlichkeit eben Rohscore t_i zu beobachten aus

$$(8) \qquad p(T_i=t_i) = \sum_{A_i|t_i} p(A_i=a_i) = \frac{\exp(\xi_i t_i) \sum_{A_i|t_i} \exp(-\sum_j \varepsilon_j y_{ij})}{\prod_j [1 + \exp(\xi_i - \varepsilon_j)]} \quad .$$

Die Summe im Zähler von (8) läßt sich auch als

$$(9) \qquad \gamma(t_i; \varepsilon_1, .., \varepsilon_k) = \sum_{A_i|t_i} \exp(- \sum_j \varepsilon_j y_{ij})$$

schreiben und wird als elementarsymetrische Funktion der Ordnung t_i bezeichnet. Der Quotient

$$p(A_i=a_i | T_i=t_i) = \frac{p(A_i=a_i)}{p(T_i=t_i)}$$

gibt die bedingte Wahrscheinlichkeit eines Antwortmusters a_i bei gegebenem Rohscore t_i an und ist unabhängig von ξ_i.

Produktbilden über i liefert die bedingte Likelihood L_c (c für conditional)

$$(10) \qquad L_c = \exp(-\sum_j \varepsilon_j s_j) / \prod_r \gamma(r; \varepsilon_1, .., \varepsilon_k)^{n_r} \cdot K$$

wobei n_r die Anzahl der Personen mit $t_i=r$, $(r=0,..,k)$, bezeichnet. Aus Symetriegründen gilt ebenso

$$(11) \qquad L_c' = \exp(\sum_i \xi_i t_i) / \prod_q \gamma(q; \xi_1, .., \xi_n)^{n_q} \cdot K$$

Die Konstante K bezeichnet eine kombinatorische Größe, die die Anzahl der Datenmatrizen (y_{ij}) beschreibt, die mit den Randvektoren (t_i) bzw. (s_j) vereinbar sind.

Die inhaltliche Bedeutung für (10) und (11) liegt in der Unabhängigkeit der Personen- von den Aufgabenparametern während des Schätzvorgangs. Darin begründet sich auch die Kernaussage der Theorie der spezifischen Objektivität: (10) erlaubt einen Vergleich der Meßinstrumente unabhängig

von den Meßobjekten. Dies führt auch zum Begriff der Stichprobenunabhängigkeit. Im Beispiel des Intelligenztests bedeutet das die Möglichkeit, Testaufgaben und deren Struktur unabhängig von den getesteten Personen, also der Stichprobe, zu evaluieren. (Ebenso gilt der umgekehrte Fall, wo Personen unabhängig von der spezifischen Testsituation verglichen werden können.) Sind nämlich die ξ_i's unabhängig von den ε_j's, so müssen selbst in extrem verzerrten Stichproben gleiche Schätzungen für die ε_j's resultieren. Die strenge Forderung des klassischen Ansatzes nach Repräsentativität der Stichprobe, die in experimentellen oder praktischen Anwendungen sozialwissenschaftlicher Fragestellungen oft nur schwer oder gar nicht erfüllt werden kann, ist hier völlig vernachlässigbar. Dazu ist allerdings notwendig, daß die Axiome i) - iv) gelten, was offensichtlich nicht ohne weiteres vorausgesetzt werden kann.

Um die Gültigkeit dieser Forderungen überprüfbar zu machen, wurden verschiedene Tests entwickelt, die im folgenden genau beschrieben werden sollen. Die Möglichkeit einer Prüfung der Modellannahmen ist ein weiterer Vorteil probabilistischer Meßmodelle. Im klassischen Ansatz geht man im wesentlichen nicht über eine Analyse der Residuen hinaus. Das probabilistische Modell hingegen kann schon als Formalisierung der Messung aufgefaßt werden, die zusätzlich die Möglichkeit bietet, ihre Angemessenheit in spezifischen Situationen überprüfen zu können.

4. GOODNESS-OF-FIT TESTS FÜR DAS RASCH-MODELL

Spezifische Objektivität wird allgemein als die wichtigste Eigenschaft des dichotomen logistischen Modells von Rasch aufgefaßt. Wie erwähnt impliziert diese Eigenschaft, daß Itemparameterschätzungen aus verschiedensten Stichproben bis auf Zufallsschwankungen gleiche Ergebnisse liefern müssen. Dies bedeutet weiters, daß bei einer Unterteilung einer Stichprobe in Subsamples, die Itemparameterschätzer in allen Teilstichproben sowie in der Gesamtstichprobe gleich sein sollen. Die meisten Goodness-of-fit Tests für das Rasch-Modell beruhen auf dieser Überlegung. Versucht man ein Kriterium zu einer Systematisierung dieser Modelltests finden, so bietet sich hiefür die Axiomatik, wie in Kapitel 2. dargestellt, nahezu von selbst an. Da alle Tests mehr oder minder sensibel auf Verletzung der verschiedenen Voraussetzungen reagieren, erscheint eine Gegenüberstellung der Prüfverfahren anhand dieses Klassifikationskriteriums sinnvoll. Die meisten Teststatistiken entdecken Verletzungen der Suffizienz und der Monotonizität recht gut, Multidimensionalität und stochastische Abhängigkeiten werden aber kaum erfaßt. Erst in jüngster Zeit wurde dieser Mangel an diagnostischen Hilfsmitteln behoben, so daß davon gesprochen werden kann, ein geeignetes Instrumentarium zur Prüfung der Gültigkeit des Rasch-Modells zur Verfügung zu haben. In den folgenden Abschnitten sollen nun diese beiden Klassen von Tests dargestellt werden.

4.1 TESTS ZUR ÜBERPRÜFUNG DER SUFFIZIENZ UND DER MONOTONIZITÄT

4.1.1 ANDERSENS χ^2 - DER BEDINGTE LIKELIHOOD RATIO TEST

Der wohl am besten bekannte und am meisten verwendete Modelltest, der bedingte Likelihood Ratio Test, wurde von Andersen (1973b) vorgestellt und beruht auf einem Vergleich der bedingten Likelihood (10) aus verschiedenen Teilstichproben. Die Argumentation ist hierbei die folgende: unterteilt man die gesamte Stichprobe in k-1 Gruppen entsprechend dem Rohscore r, (r=1,..k-1), - Personen mit den trivialen Scores 0 bzw. k werden nicht berücksichtigt, da ihr Beitrag zur Likelihood 0 bzw. 1 ist - so kann man für jede dieser Gruppen die Likelihood

$$(12) \qquad L_c^{(r)} = \exp(-\sum_j \varepsilon_j s_j^{(r)}) / \gamma(r;\varepsilon_1,..,\varepsilon_k)^{n_r}$$

erhalten. $s_j^{(r)}$ bedeutet hier, wie oft Item j in der Scoregruppe r richtig gelöst wurde. Offensichtlich kann die bedingte Likelihood (10) in gleichwertiger Form auch als das Produkt von (12) über alle r angeschrieben werden

$$(13) \qquad L_c = \prod_r L_c^{(r)}$$

Natürlich gilt (13) nur dann, wenn die Likelihoods (12) Funktionen der gleichen Parameterschätzungen in allen Teilstichproben sind.

Diesen Überlegungen entsprechend wird in Andersens Testkonstruktion die Stichprobe in Rohscoregruppen unterteilt

und es werden getrennt sowohl in allen Subsamples als auch in der Gesamtstichprobe die Parameter geschätzt. Allerdings gilt die Gleichung (13) nur mehr approximativ, da die Likelihoods jetzt auf verschiedenen Parameterschätzern beruhen. Unter der Voraussetzung der Gültigkeit des Modells müßten aber gleiche Parameterschätzer in allen Scoregruppen resultieren und daher (13) bis auf Zufallsschwankungen erfüllt sein. Dies läßt sich auch als

$$(14) \qquad \lambda = \frac{L_c}{\prod_r L_c^{(r)}} \approx 1$$

anschreiben, λ ist der bedingte Likelihoodquotient. Andersen zeigte (1973b), daß die Prüfgröße

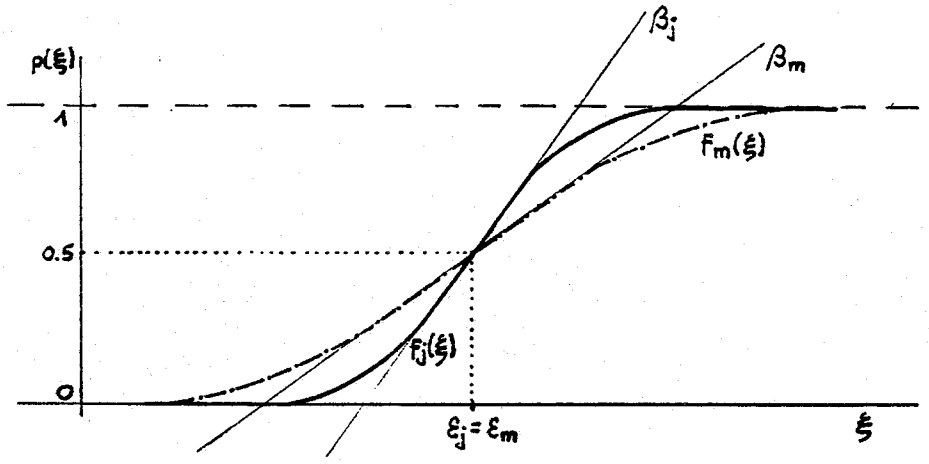
$$(15) \qquad Z = -2 \ln \lambda = 2 \sum_r \ln L_c^{(r)} - 2 \ln L_c$$

asymptotisch χ^2-verteilt ist mit df=(k-2)(k-1) Freiheitsgraden, wenn $n_r \to \infty$ für alle r.

Eine mögliche Alternativhypothese zum Rasch-Modell ist das sogenannte zweiparametrige Modell von Birnbaum (1968)

$$(16) \qquad p(Y_{ij}=y_{ij} \mid \xi_i, \varepsilon_j, \beta_j) = \frac{\exp([\xi_i - \varepsilon_j]\,\beta_j)}{1 + \exp([\xi_i - \varepsilon_j]\,\beta_j)} \quad .$$

β_j beschreibt den Anstieg der logistischen Funktion (16) an der Stelle p=.5 . Den Unterschied zum Rasch-Modell, in dem ja alle β's als gleich und mit dem Wert 1 angenommen werden, mag Abbildung 1. veranschaulichen.



(Abbildung 1.)

Der Begriff Trennschärfe ("item discriminating power") eines Items stammt von der inhaltlichen Überlegung, daß ein solches Item, das einen steileren Anstieg der logistischen Funktion oder ICC aufweist, besser zwischen zwei Personen diskriminiert. Da deren unterschiedliche Leistungsfähigkeit durch einen konstanten Abstand auf der Dimension ξ beschrieben wird, nähern sich die entsprechenden Lösungswahrscheinlichkeiten schneller den Werten 0 oder 1.

Die suffizienten Statistiken für die $\hat{\xi}_i$'s aus (16) sind

$$t_i^* = \sum_j \beta_j y_{ij} \quad . \tag{17}$$

Allerdings sind die β_j's in (17) nicht bekannt und müssen erst geschätzt werden. Aus diesem Grund ist eine unabhängige Schätzung der Parameter, wie sie durch die Verwendung der bedingten Likelihood im Rasch-Modell ermöglicht wird, hier nicht gegeben. Daher muß für (16) auch der Begriff der Stichprobenunabhängigkeit aufgegeben werden.

Wird nun der Likelihood Ratio Test (15) signifikant, dann mag dies darauf zurückzuführen sein, daß unterschiedlich steile ICC's auftreten und somit das Axiom der suffizienten Statistiken verletzt ist. Die folgenden Überlegungen sollen

nun zeigen, daß die Prüfgröße (15) sensibel gegenüber solchen Verletzungen ist.

Nimmt man an, daß eine Anordnung der t_i ihrer Größe nach in gewisser Übereinstimmung mit einer entsprechenden Anordnung der ξ_i steht, so müßte eine solche Übereinstimmung auch zwischen den t_i^* und den t_i gegeben sein, wie man sich leicht für kleine k und verschiedene Werte für die β_j's vor Augen führen kann. Das folgende Beispiel stammt von Andersen (1973b).

$(\beta_1,..,\beta_4)=$	(4.0,2.0,.5,.25)	(2.0,1.0,1.0,.5)
$T_i=$ 1	$T_i^*=$ .25,.5,2.0,4.0	$T_i^*=$ .5,1.0,2.0
2	.75,2.25,2.5,4.25,4.5,6.0	1.5,2.0,2.5,3.0
3	2.75,4.75,6.25,6.5	2.5,3.5,4.0

Dies ermöglicht nun auch den Schluß, daß die Variation der ξ_i innerhalb einer Scoregruppe substantiell kleiner sein müßte als zwischen Rohscoregruppen. Daraus folgt aber nun, daß die restringierten Itemparameterschätzer $\hat{\varepsilon}_j^{(r)}$ in den einzelnen Teilstichproben voneinander abweichen, da ja die Variation der ξ_i's in die Likelihoods (12) eingeht, wenn diese fälschlicherweise zur Schätzung der ε_j's aus (16) herangezogen wird. Approximation (14) wird dann auch nicht erfüllt sein.

Ähnliche Argumente lassen sich auch zur Beurteilung der diagnostischen Zuverlässigkeit des Likelihood Ratio Tests gegenüber Verletzungen des Axioms der Monotonizität anführen. Weist die ICC an irgendeiner Stelle ein Maximum auf und fällt dann wieder, so wird die Lösungswahrscheinlichkeit des entsprechenden Items bei größeren t_i wieder

sinken. Entsprechende Variation der Itemparameterschätzer in verschiedenen Subsamples sind die Folge und es gelten die obigen Überlegungen.

Gewisse Schwierigkeiten in der Anwendung des Likelihood Ratio Tests von Andersen können dann auftreten, wenn die Anzahl der Items und die Größe der Stichprobe in einem Mißverhältnis stehen, d.h. wenn auf Grund weniger Versuchspersonen und einer relativ großen Anzahl der Items die Besetzung der Teilstichproben klein wird. Da in jeder Teilstichprobe getrennt Itemparameter zu schätzen sind, kann in solchen Fällen leicht eintreten, daß der Algorithmus zum Berechnen der elementarsymetrischen Funktionen versagt.

Allerdings kann diesen Problemen insoferne ausgewichen werden, als man weniger Teilstichproben bildet, indem man Rohscoregruppen mit kleinen Besetzungen zusammenfaßt. (14) bzw. (15) gilt für jede beliebige disjunkte und erschöpfende Teilung des erhobenen Datenmaterials in $2 \leq r < k$ Gruppen. Häufig verwendet man den Median der ranggereihten t_i als Trennungskriterium zur Bildung von zwei Teilstichproben. Man erhält dann also eine Gruppe mit niedrigem und eine mit hohem Rohscore.

Allerdings ist der LR-Test nicht auf eine Teilung nach dem im Test erzielten Score beschränkt. Alter, Geschlecht oder andere externe Variablen können ebenso benützt werden. Diese Vorgangsweise bietet zusätzlich die Möglichkeit Variablen zu finden, die einen Einfluß auf die zu überprüfende Fragestellung haben. Insoferne kann man den LR-Test auch als Verfahren zur Generierung von Hypothesen bezeichnen.

4.1.2 DIE STATISTIK S VON FISCHER UND SCHEIBLECHNER

G.H.Fischer und H.Scheiblechner (1970) entwickelten einen Test, der dann zur Anwendung gebracht werden kann, wenn eine Teilung der Stichprobe in zwei Untergruppen sinnvoll erscheint. Das Verfahren beruht auf der wohlbekannten Tatsache, daß ML-Schätzer asymptotisch normalverteilt sind mit einer asymptotischen Varianz-Kovarianzmatrix C, die sich aus der Matrix der zweiten partiellen Ableitungen der log Likelihood ergibt:

$$(18) \qquad C^{-1} = -E\left\{\frac{\partial \ln L^2}{\partial \varepsilon_j \partial \varepsilon_m}\right\}$$

Die Diagonalelemente von C sind die Varianzen der Parameterschätzer, ihre Wurzeln die Standardabweichungen.

Schätzt man nun die Itemparameter in den beiden Subsamples getrennt, so erhält man die beiden Schätzer $\hat{\varepsilon}_j^{(1)}$ bzw. $\hat{\varepsilon}_j^{(2)}$, sowie die zugehörigen Standardabweichungen. Die Statistik

$$(19) \qquad S_j = (\hat{\varepsilon}_j^{(1)} - \hat{\varepsilon}_j^{(2)})/(s_j^{(1)} + s_j^{(2)})^{1/2}$$

ist asymptotisch standardnormal verteilt und kann zur Beurteilung einzelner Items herangezogen werden.

Quadrieren von (19) und Summieren über alle j ergibt die Prüfgröße für den Gesamttest

$$(20) \qquad S = \sum_j S_j^2 \quad .$$

Unter der Annahme der Unabhängigkeit der Parameterschätzer wäre (20) χ^2-verteilt mit df=k Freiheitsgraden. Fischer und Scheiblechner meinen aber, daß auf Grund der Normierungsbedingung $\sum_i \varepsilon_j = 0$ ein Freiheitsgrad verloren ginge und daher die richtige Zahl df=k-1 sei.

Dem läßt sich aber die folgende Argumentation von A.L.van den Wollenberg (1979,S.32) entgegenhalten: Das Quadrat der Statistik (19) ist χ^2-verteilt mit df=1. Es ist bekannt, daß der Erwartungswert einer χ^2-verteilten Größe gleich ist der Anzahl der Freiheitsgrade, und ferner, daß der Erwartungswert einer Summe gleich ist der Summe der Erwartungswerte. (20) ist die Summe von k Variablen, jede mit Erwartungswert 1. Der Erwartungswert von (20) ist daher k, folglich kann (20) nicht χ^2-verteilt sein mit df=k-1. (20) kann aber auch nicht χ^2-verteilt sein mit k Freiheitsgraden, da nur k-1 Parameter frei variieren. Die Approximation an eine χ^2-Verteilung ist also nicht sehr gut, Fischer bezeichnet (1974,S.297) den Test auch als konservativ.

Eine mögliche Ursache für diesen Widerspruch mag darin begründet sein, daß nur die Diagonalelemente der Varianz-Kovarianzmatrix C verwendet werden. Eine Lösungsmöglichkeit bietet sich daher in der Anwendung des Hottellings T^2 an, wo auch Kovariationen berücksichtigt werden. Seien $C^{(1)}$ und $C^{(2)}$ die beiden Varianz-Kovarianzmatrizen aus den Teilstichproben, dann erhält man gemeinsame Matrix $C^{(o)}$ aus

$$C^{(o)} = (C^{(1)} + C^{(2)}) \qquad \text{mit } df = n^{(1)} + n^{(2)} - 2 \ .$$

Will man die Hypothese testen, daß die Differenzen zwischen allen Parameterschätzern Null sind, dann ist der Test von Hottelling gleich der Prüfgröße D^2 von Mahalanobis. Sei

$$(21) \qquad D^2 = \sum_i \sum_j s_{ij}^{-1} \, d_i \, d_j$$

mit $s_{ij} = c_{ij}^{(o)}/(n^{(1)}+n^{(2)}-2)$ und $d_i = \hat{\varepsilon}_i^{(1)} - \hat{\varepsilon}_i^{(2)}$ dann ist

das auf D^2 basierende Varianzverhältnis

$$(22) \qquad D^2 \cdot \frac{n^{(1)}+n^{(2)}-k-1}{k} \cdot \frac{n^{(1)}n^{(2)}}{(n^{(1)}+n^{(2)})(n^{(1)}+n^{(2)}-2)}$$

zentral F-verteilt mit k und $n^{(1)}+n^{(2)}-k-1$ Freiheitsgraden (vgl.C.R.Rao,1973,S.565f.). $n^{(1)}$ und $n^{(2)}$ bezeichnen die Stichprobengrößen der beiden Subsamples.

Ein Vorteil der Prüfgröße S_j von Fischer und Scheiblechner gegenüber dem LR-Test (15) liegt aber trotz der diskutierten Probleme darin, daß ein diagnostisches Hilfsmittel zur Evaluierung einzelner Items zur Verfügung gestellt wurde. Dies kann sich dann als besonders nützlich erweisen, wenn Ursachen für Modellverletzungen gesuchtwerden. In der praktischen Anwendung erweist sich (19) als wertvolle Ergänzung zum Likelihoodquotientenverfahren.

Allerdings wies van den Wollenberg (1979) auf ein mögliches Artefakt hin, das bei der Anwendung von (19) besonders dann auftreten kann, wenn die Parameter unter der Normierung $\sum \varepsilon_j = 0$ geschätzt wurden. Offensichtlich beeinflussen Unterschiede zwischen Itempaaren $\hat{\varepsilon}_j^{(1)}$ und $\hat{\varepsilon}_j^{(2)}$ die Differenzen zwischen den anderen geschätzten Parametern. Er gab dazu folgenden Illustration:

Gegeben seien 5 Itemparameter, die in den beiden Stichproben als gleich geschätzt wurden mit folgenden Werten:

ITEM	ITEMPARAMETERSCHÄTZER IN TEILSTICHPROBE	
	(1)	(2)
1	-2	-2
2	-1	-1
3	0	0
4	1	1
5	2	2

Fügt man nun ein Item hinzu, das extrem verschiedene Parameterschätzer in den beiden Subsamples aufweist, so ändern sich die Differenzen der anderen Itemparameterwerte auf Grund der Normierungsbedingung.

ITEM	ITEMPARAMETERSCHÄTZER IN TEILSTICHPROBE	
	(1)	(2)
1	0	-2
2	1	-1
3	2	0
4	3	1
5	4	2
6	10	0

Anhand dieser Überlegung ließe sich argumentieren, daß die Prüfgröße (20) solange keine Aussagekraft besitzt, als sie nicht sehr groß wird. Dennoch ist sie aber, wie eben dargestellt, sehr sensibel auf Abweichungen einzelner Items und somit gegenüber Verletzungen der Suffizienz und Monotonizität.

4.1.3 MARTIN-LÖFS STATISTIK T

Auch die von Martin-Löf (1973) vorgeschlagene Prüfgröße beruht auf einer Unterteilung der gesamten Stichprobe in k-1 Gruppen wie im Likelihoodquotiententest von Andersen. Ausgangspunkt ist hier ein Vergleich beobachteter Häufigkeiten a_{rj}, der Anzahl von Personen in Scoregruppe r, die Item j positiv beantwortet haben, mit den entsprechenden Erwartungswerten $E(n_{rj})$. Für jedes Item erhält man die bedingte Lösungswahrscheinlichkeit für Scoregruppe r aus

$$(23) \qquad \pi_{rj} = \frac{\varepsilon_j \, \gamma(r-1;\varepsilon_1,..,\varepsilon_{j-1},\varepsilon_{j+1},..,\varepsilon_k)}{\gamma(r;\varepsilon_1,..,\varepsilon_k)} \quad .$$

Die bedingte Wahrscheinlichkeit einer gemeinsamen Beobachtung richtiger Lösungen bei Items j und m für eine gegebene Rohscoregruppe r ist

$$(24) \qquad \pi_{rjm} = \frac{\varepsilon_j \, \varepsilon_m \, \gamma_{r-2}^{(j,m)}}{\gamma_r}$$

mit $\gamma_{r-2}^{(j,m)} = \gamma(r-2;\varepsilon_1,..,\varepsilon_{j-1},\varepsilon_{j+1},...,\varepsilon_{m-1},\varepsilon_{m+1},..,\varepsilon_k)$, der elementarsymetrischen Grundfunktion der Ordnung r-2 unter Nichtberücksichtigung von ε_j und ε_m. Die γ-Funktionen in (23), bzw. (24) erhält man durch ein- bzw. zweimaliges Ableiten von γ_r nach ε_j und ε_m . (Vgl.G.Fischer,1974,S.236f.)

Verwendet man nun die geschätzten Parameter $\hat{\varepsilon}$ zur Bestimmung von $\hat{\pi}_{rj}$ und $\hat{\pi}_{rjm}$ so sind Varianz und Kovarianz gegeben durch

(25) $$Var(\pi_{rj}) = \pi_{rj}\,(1-\pi_{rj})$$

bzw.

(26) $$Cov(\pi_{rjm}) = \pi_{rjm} - \pi_{rj}\,\pi_{rm} \quad .$$

Im folgenden werden noch der Erwartungswert $E(n_{rj})$ sowie die entsprechenden Varianzen und Kovarianzen benötigt. Diese sind (Vgl.van den Wollenberg,1979,p.35):

(27) $$E(n_{rj}) = E(n_r.\pi_{rj}) = \pi_{rj}\cdot E(n_r)$$

(28) $$Var(n_{rj}) = \pi_{rj}^2\ Var(n_r) + \pi_{rj}\,(1-\pi_{rj}).E(n_r)$$

(29) $$Cov(n_{rj},n_{rm}) = \pi_{rj}\,\pi_{rm}\,Var(n_r) + (\pi_{rjm} - \pi_{rj}\pi_{rm}).E(n_r)$$

Martin-Löf nimmt nun an, daß die n_r einer Poisson Verteilung mit dem Parameter $\lambda\xi_r$ folgen, n_r ist dann der zugehörige ML-Schätzer, d.h. $\hat{\lambda\xi}_r = n_r$. Die Argumentation hiefür ist die folgende: Die Versuchspersonen bilden eine repräsentative Stichprobe, die aus einer unendlich großen Population gezogen wurde. Wenn die Auswahl einer Person zufällig erfolgte, dann hat die Wahrscheinlichkeit einen bestimmten Rohscore r zu beobachten für jede Person einen konstanten Wert. (Dies entspricht der Annahme einer randomisierten Version des Rasch-Modells. Eine Darstellung dieser Modelltypen und ihre Beziehung zu log-linearen Modellen findet sich in R.Dittrich u. R.Hatzinger,1981.) Die n_r folgen dann einer Binomialverteilung, die durch die Poissonverteilung angenähert werden kann. Unter der Poissonverteilung sind Erwartungswert und Varianz gleich $\lambda\xi_r$, was die Formeln (27) - (29) wesentlich vereinfacht.

Dementsprechend werden Erwartungswert, Varianz und Kovarianz zu

$$(30) \qquad E(n_{rj}) = \lambda \xi_r \pi_{rj}$$

$$(31) \qquad Var(n_{rj}) = \lambda \xi_r \pi_{rj}$$

$$(32) \qquad Cov(n_{rj}) = \lambda \xi_r \pi_{rjm}$$

Durch Aufsummieren der quadratischen Terme

$$(33) \qquad T_r = d_r' \cdot V_r^{-1} \cdot d_r$$

erhält man die kombinierte Statistik für den Gesamttest

$$(34) \qquad T = \sum T_r \; .$$

d_r' ist hierbei der (1xk)-Vektor mit Elementen $d_{rj}=(n_{rj}-n_r\pi_{rj})$, die (kxk)-Varianz-Kovarianzmatrix V_r besteht aus

$$(35) \qquad v_{jj} = n_r \varepsilon_j \gamma_{r-1}^{(j)} / \gamma_r$$

und

$$(36) \qquad v_{jm} = n_r \varepsilon_j \varepsilon_m \gamma_{r-2}^{(j,m)} / \gamma_r \; .$$

Statistik (34) ist χ^2-verteilt mit df=(k-1)(k-2), für $n_r \to \infty$ für alle r.

Auch für Martin-Löfs Teststatistik lassen sich die schon dargestellten Argumente anführen. Sensibilität gegenüber Verletzungen der Suffizienz und der Monotonizität kann erwartet werden. Voraussetzung ist wieder eine relativ große Stichprobe und/oder eine kleine Anzahl von Items.

4.1.4 DIE WRIGHT UND PANCHAPAKESAN STATISTIK Y

Wie bei Martin-Löfs Testgröße beruht der von Wright und Panchapakesan (1969) vorgestellte Modelltest auf einem Vergleich beobachteter und erwarteter Häufigkeiten n_{rj}. Dieser Goodness-of-fit Test soll hier aber nur kurz und der Vollständigkeit halber dargestellt werden, da er auf falschen Voraussetzungen beruht.

Für jedes Item in jeder Scoregruppe wird eine Approximation an eine normalverteilte Größe aus

$$(37) \qquad y_{rj} = n_{rj} - E(n_{rj}) \; / \; Var(n_{rj})^{\frac{1}{2}}$$

bestimmt, wobei man den Erwartungswert aus

$$(38) \qquad E(n_{rj}) = n_r \cdot \pi_{rj}$$

und die Varianz aus

$$(39) \qquad Var(n_{rj}) = n_r \cdot \pi_{rj} \cdot (1 - \pi_{rj})$$

erhält. Die Wahrscheinlichkeit $\hat{\pi}_{rj}$ ergibt sich unter Einsetzen der Schätzer $\hat{\xi}_i$ und $\hat{\varepsilon}_j$ in (2) als

$$(40) \qquad \hat{\pi}_{rj} = \exp(\hat{\xi}_r - \hat{\varepsilon}_j)/[1 + \exp(\hat{\xi}_r - \varepsilon_j)] \; .$$

Wright und Panchapakesan meinen nun die y_{rj} seien approximativ

standardnormalverteilt und daher folge ihr Quadrat einer χ^2-Verteilung mit df = 1. Weiters sei die Summe

$$(41) \qquad Y = \sum_{r}^{g} \sum_{j} y_{rj}^2$$

asymptotisch χ^2-verteilt mit (k-1)(g-1) Freiheitsgraden (g ist hierbei die Anzahl der verschiedenen, beobachteten Rohscores).

Vergleicht man (40) mit (23) so wird ersichtlich, daß zur Verwendung von (40) auch Schätzer für die Personenparameter vorliegen müssen. Dies setzt voraus, daß die Parameter mittels der UML (5) geschätzt wurden, da die elementarsymetrischen Grundfunktionen in (23) nur bei Anwendung der CML-Methode (10) berechnet werden. Scheint (40) eher praktischen Wert zu besitzen, so ist (23) aus theoretischen Gründen vorzuziehen.

Stärkere Bedenken sind hingegen bezüglich der Verteilungsannahmen über (41) angebracht. Die y_{rj} können nicht als standardnormal aufgefaßt werden, da die Häufigkeiten n_{rj} schon zur Schätzung der Itemparameter herangezogen werden. Diese werden aber wieder für die Berechnung der Erwartungswerte verwendet. Zusätzlich kann (41) nicht χ^2-verteilt sein - selbst wenn die y_{rj} standardnormale Werte wären - , da hiezu die Summe unabhängiger y_{rj}^2 erforderlich ist. In jeder Scoregruppe ist aber die Anzahl positiver Antworten gleich $n_r.r$ (gegeben sind n_r Personen in Scoregruppe r, jede mit r richtigen Antworten), daher gilt für jede Teilstichprobe die Restriktion $n_r.r = \sum_j n_{rj}$. Die Statistiken y_{rj} können aus diesem Grund nicht als unabhängig aufgefaßt werden, nur k-1 anstatt k Beobachtungen variieren frei.

4.1.5 VAN DEN WOLLENBERGS Q_1

Als Neuformulierung der Wright und Panchapakesan Statistik schlug van den Wollenberg (1979) folgendes Testverfahren vor:

Unterteilt man die beobachteten Daten nach Items und Scoregruppen, so erhält man für jedes Item die folgende Kontingenztafel:

Roh-score	Antwort 1	Antwort 0	
1	n_{1j}	n_1-n_{1j}	n_1
⋮			
r	n_{rj}	n_r-n_{rj}	n_r
⋮			
k-1	n_{k-1j}	$n_{k-1}-n_{k-1j}$	n_{k-1}
	s_j	$n-s_j$	n

In dieser Tafel bedeuten die Zeilensummen die Anzahl der Personen in Scoregruppe r. Die Spaltensumme s_j ist die Anzahl der Personen über alle r, die das Item j positiv gelöst haben. s_j bedeutet also wie oft Item j insgesamt richtig beantwortet wurde und ist daher die minimal suffiziente Statistik für ε_j. Sind einmal die Randsummen gegeben, so bleiben k-2 Häufigkeiten frei zu variieren.

Insgesamt lassen sich k solche Tafeln anschreiben, die auf beobachteten Häufigkeiten beruhen. Ebenso können k Kontingenztafeln für die erwarteten Häufigkeiten gebildet werden. Hiezu verwendet man die aus der Gesamtstichprobe mittels

CML (10) geschätzten Itemparameter und erhält die Erwartungswerte

$$(42) \qquad E(n_{rj}|n_r) = n_r \, \pi_{rj}$$

mit π_{rj} aus (23). Van den Wollenbergs Statistik beruht nun auf dem Vergleich der Kontingenztafeln beobachteter und erwarteter Häufigkeiten:

$$(43) \qquad q_j = \sum_{r=1}^{k-1} \frac{(n_{rj}-E(nr_j))^2}{E(n_{rj})} + \frac{((n_r-n_{rj})-E(n_r-n_{rj}))^2}{E(n_r-n_{rj})} =$$

$$= \sum_{r=1}^{k-1} \frac{(n_{rj}-E(n_{rj}))^2}{E(n_{rj})} + \frac{(n_{rj}-E(n_{rj}))^2}{E(n_r-n_{rj})}$$

Sind einmal die Itemparameter gegeben so ist q_j eine Summe von k-1 unabhängigen Ausdrücke, jeder χ^2-verteilt mit df=1. Allerdings sind die Itemparameter nicht a priori bekannt und müssen erst aus den Daten geschätzt werden. Wird also zur Schätzung des Parameters für Item j die minimal suffiziente Statistik s_j herangezogen, so verliert man einen Freiheitsgrad und die Anzahl der Freiheitsgrade für q_j verringert sich zu k-2. Dies entspricht auch der Anzahl frei variierender Häufigkeiten in der Kontingenztafel.

Aufsummieren der q_j's über alle Items ergibt die globale Testgröße

$$(44) \qquad Q_1^* = \sum q_j \, .$$

(44) ist wieder asymptotisch χ^2-verteilt mit df=(k-1)(k-2).

Wären alle q_j voneinander unabhängig, so müßte die Anzahl der Freiheitsgrade k(k-2) sein. Dies ist aber wie schon bei der Prüfgröße von Wright und Panchapakesan nicht der Fall, da auch hier die Restriktion $n_r.r = \sum_j n_{rj}$ gilt. Es bleiben also nur (k-1)(k-2) Freiheitsgrade über.

Allerdings sollte der Erwartungswert einer χ^2-verteilten Größe gleich sein der Anzahl der Freiheitsgrade. Der Erwartungswert von (40) sollte demnach (k-1)(k-2) und nicht k(k-2) sein. Van den Wollenberg schlägt aus diesem Grund auch die korrigierte Testgröße

$$(45) \qquad Q_1 = (k-1)(k-2)/k(k-1)\ Q_1^* = (k-1)/k\ Q_1^*$$

vor. Ausgedehnte Simulationsstudien (van den Wollenberg,1979, 1980) scheinen die Korrektur (45) zu rechtfertigen. Im Rahmen dieser Simulationen wurde auch gefunden, daß die Statistik Q_1 und der Likelihoodquotiententest von Andersen (s.Kap.4.1.1) gleiche Fehlerwahrscheinlichkeiten für die Nullhypothese des Modells liefern. Der Grund hiefür liegt einerseits in der Verwendung der bedingten Wahrscheinlichkeiten (23) in beiden Testverfahren, andererseits in der Behandlung der Parameterschätzer. Im Q_1-Ansatz werden aus den global geschätzten Itemparametern erwartete Häufigkeiten bestimmt und diese dann mit den beobachteten Häufigkeiten verglichen, die ihrerseits als Grundlage zur Schätzung der restringierten Itemparametern in den einzelnen Scoregruppen im LR-Test herangezogen werden. Diese werden dann mit den aus der gesamten Stichprobe geschätzten Parametern verglichen. Beide Tests verwenden also das gleiche Verfahren. Die Unterschiedlichkeit besteht nur in einer verschiedenen Sichtweise der gleichen Aspekte in den Daten.

Es lassen sich nun einige Vorteile der Statistik Q_1 gegenüber

dem Likelihoodquotiententest anführen. Dies betrifft vor allem die Möglichkeit einzelne Items zu isolieren, die die Modellkonformität der gesamten Itemmenge global betrachtet in Frage stellen. Alle Argumente, die für das Verfahren von Fischer und Scheiblechner (Kap.4.1.2) vorgebracht wurden, treffen auch hier zu. Überdies kann eine Teilung der Stichprobe in beliebige, disjunkte und erschöpfende Subsamples vorgenommen werden (dies wird besonders dann günstig sein, wenn die Güte des Tests durch kleine n so verringert wird, daß nicht alle Rohscoregruppen getrennt sondern durch Zusammenschluß mehrerer Rohscoregruppen insgesamt weniger Subsamples betrachtet werden), wohingegen in jenem Verfahren nur eine Zweiteilung vorgesehen ist.

Ein weiterer Vorteil liegt in der relativen Einfachheit der Berechnung. Braucht man für den LR-Test iterative Parameterschätzungen zusätzlich für jede Scoregruppe, so genügt zur Bestimmung von Q_1 die Parameterschätzung aus der Gesamtstichprobe. Wurden nur kleine n und somit kleine n_r beobachtet, wird sich dies als besonders vorteilhaft erweisen, da in diesem Fall bei getrennter Schätzung gravierende Rechenungenauigkeiten nicht auszuschließen sind.

Da keine geringere Sensibilität gegenüber Verletzungen des Suffizienz- und des Monotonizitätsaxioms als beim LR-Test anzunehmen sind, wird im allgemeinen die Verwendung der Q_1-Statistik dem LR-Test vorzuziehen sein.

Abschließend sei noch auf die Koinzidenz von Q_1 und der Martin-Löf Prüfgröße T im Falle gleicher Werte für alle ε_j, d.h. $\varepsilon_1=\ldots=\varepsilon_k$, verwiesen. Dies gilt dann, wenn in der Stichprobe

$$(46) \qquad s_j = \sum n_{rj}$$

den gleichen Wert für alle j hat. (vgl. Molenaar,1981)

4.1.6 MOLENAARS STATISTIK U_j

In einer neueren Arbeit diskutiert I.Molenaar (1981) verschiedene Teststatistiken, wobei er im wesentlichen zwei Gruppen unterscheidet: Binomialtests für einzelne Items über Scoregruppen und erweiterte ("extended") hypergeometrische Tests für Itempaare je Scoregruppe. Diese seine Unterscheidung entspricht der hier getroffenen Einteilung von Statistiken in solche, die zu einer Überprüfung der Axiome der Monotonizität und der Suffizienz dienen und solche, womit die Gültigkeit der Axiome der Unidimensionalität und der stochastischen Unabhängigkeit getestet werden soll. Beide Vorgehensweisen unterstützt er durch verschiedene Methoden graphischer Analysen, auf die aber erst in Kap. 5 eingegangen werden soll.

Die wesentlichste Aussage der Arbeit Molenaars besteht in der Argumentation für eine spezielle Untersuchung solcher Items, die eine Verletzung der Axiome für die gesamte Itemmenge verursachen, wenn diese global betrachtet werden. Insbesonders spricht er sich gegen das "mechanische" Ausscheiden solcher nichthomogenen Items aus und plädiert für eine Beachtung des jeweiligen speziellen Inhalts und seiner Beziehung zu den Inhalten anderer, homogener Items. Aus diesem Grund wendet er sich auch gegen eine Verwendung globaler Statistiken, die meist die Unterschiede zwischen Items innerhalb einer Scoregruppe beleuchten. Seiner Meinung nach ist eine Untersuchung eines Items über verschiedene Scoregruppen informativer. Wichtiger als die Kenntnis, daß eine bestimmte Scoregruppe einen schlechten fit verursacht, sei es, einzelne Items zu isolieren, um ihre Beziehung zu den anderen weiter studieren zu können.

Der hier getroffenen Einteilung entsprechend soll zunächst ein von Molenaar vorgeschlagenes diagnostisches Mittel zur Entdeckung etwaiger Verletzungen des Suffizienz- und des Mono-

tonizitätsaxioms besprochen werden. Gegeben sei ein Item j mit einem flacheren Anstieg der ICC als bei den anderen Items. In diesem Fall wird die Lösungswahrscheinlichkeit nur langsam mit zunehmendem Parameterwert ξ_i bzw. zunehmendem Rohscore t_i steigen oder anders ausgedrückt die beobachteten Häufigkeiten n_{rj} (Anzahl der Personen in Scoregruppe r die das Item j richtig gelöst haben) werden sich von den erwarteten Häufigkeiten $n_r \pi_{rj}$ (mit π_{rj} aus (23)) systematisch so unterscheiden daß gilt

$$n_{rj} > n_r \pi_{rj} \qquad \text{für kleine r} \tag{47}$$

$$n_{rj} < n_r \pi_{rj} \qquad \text{für große r}$$

Für ein Item mit steilerer ICC gelten umgekehrte Vorzeichen in (47). Diese Beziehung läßt sich allerdings nur bei Betrachtung einzelner Scoregruppen feststellen. Summiert man nämlich über alle r erhält man die Äquivalenz

$$\sum_r n_{rj} = \sum_r n_r \pi_{rj} \quad . \tag{48}$$

Ein Signifikantwerden der Ungleichungen (47) läßt sich durch folgende einseitige Binomialtests prüfen

$$P_{rj} = \sum_{i=n_{rj}}^{n_r} \binom{n_r}{i} \pi_{rj}^{\,i} (1 - \pi_{rj})^{n_r - i} \quad , \quad \text{für } n_{rj} > n_r \pi_{rj} \tag{49}$$

und

$$P_{rj} = \sum_{i=0}^{n_{rj}} \binom{n_r}{i} \pi_{rj}^{\,i} (1 - \pi_{rj})^{n_r - i} \quad , \quad \text{für } n_{rj} < n_r \pi_{rj}. \tag{50}$$

Dasselbe Verfahren eignet sich auch zur Entdeckung von Items mit nicht monotoner ICC. Hier werden die Differenzen

zwischen beobachteten und erwarteten Häufigkeiten nicht einem Muster wie in (47) folgen sondern eine ungeordnete Struktur aufweisen.

Natürlich lassen sich die Testgrößen in (49) und (50) zu einer globalen Statistik für Item j kombinieren, und der obigen Argumentation entsprechend scheint es sinnvoller einzelne Items über Scoregruppen als eine Scoregruppe über Items zu betrachten. Allerdings ergibt sich bei der Erstellung einer solchen kombinierten Statistik das Problem, daß in manchen Scoregruppen große n_r, in anderen relativ kleine n_r auftreten können und daher die Güte der Einzeltests variiert. Eine Approximation an eine Standardnormalverteilung könnte daher in vielen Fällen fehlschlagen.

Molenaar umgeht dieses Problem durch die Einführung einer Unterteilung der Gesamtstichprobe in drei, einander ausschließende und erschöpfende Teilmengen L, M und R der Indexmenge der Scoregruppen, wobei

$$\begin{aligned} L &= \{1,2,\ldots,r_1\} \\ M &= \{r_1+1,\ldots,r_2-1\} \\ R &= \{r_2,\ldots,k-1\} \end{aligned} \tag{51}$$

Sei $n_m = n - n_0 - n_k$ so sind die Grenzen r_1 und r_2 definiert durch

$$\begin{aligned} \sum_{r=1}^{r_1-1} n_r < n_m/4 \leqslant \sum_{r=1}^{r_1} n_r \\ \sum_{r=r_2}^{k-1} n_r \geqslant n_m/4 > \sum_{r=r_2+1}^{k-1} n_r \end{aligned} \tag{52}$$

Unter Verwendung von (37) (wobei hier z_{rj} anstatt y_{rj} ge-

schrieben werden soll) als standardisierte Form der binomialverteilten Größe n_{rj} schlägt Molenaar die Testgröße

$$(53) \qquad U_j = (\sum_{L} z_{rj} - \sum_{R} z_{rj})/(r_1 + k - r_2)^{\frac{1}{2}}$$

vor. Während große positive U_j's einen zu flachen Anstieg der ICC andeuten, lassen große negative Werte für U_j auf einen steileren Anstieg der ICC schließen. Dieses Item mißt die Skala "zu gut".

Unter H_o wäre (53) asymptotisch standardnormal (wenn alle $n_r \to \infty$). Dies trifft aber auf Grund der Beschränkung (48) und der Verwendung der geschätzten Itemparameter für $E(n_{rj})$ in (37) nicht zu, da unabhängige z_{rj} hiezu Bedingung wären. Allerdings werden stärkere Abweichungen von der Standardnormalverteilung nur dann auftreten, wenn die π_{rj} nahe 0 oder 1 sind oder nur sehr kleine n_r beobachtet wurden. In solchen Fällen wird man eine bessere Abschätzung U_j erzielen, wenn die Terme in (53) für die entsprechenden n_r weggelassen werden.

Zusammenfassend scheint die Verwendung von U_j günstig, wenn auf Grund globaler Testgrößen und graphischer Analysen der Eindruck entsteht, daß einzelne Items der gesamten Skala nicht homogen sind. Die Statistik (53) erlaubt dann eine genauere Untersuchung dieser Items und möglicher Ursachen ihrer schlechten Anpassung an das Modell.

4.2 TESTS ZUR ÜBERPRÜFUNG DER UNIDIMENSIONALITÄT UND DER LOKALEN STOCHASTISCHEN UNABHÄNGIGKEIT

Wurden bis jetzt vor allem solche Verfahren besprochen, die speziell geeignet sind die Axiome der Suffizienz und der Monotonizität zu überprüfen, so soll im folgenden eine Besprechung jener Tests im Mittelpunkt stehen, die eine Überprüfung der wichtigen Eigenschaften der Unidimensionalität und der stochastischen Unabhängigkeit einzelner Antworten auf die Items j ermöglichen. Der Zusammenhang zwischen diesen beiden Forderungen läßt sich leicht durch folgende Überlegung illustrieren: Erzielen verschiedene Personen in einem Test, der aus zwei oder mehreren zugrundeliegenden Dimensionen zusammengesetzt ist, den gleichen Rohscore t_i, so werden sie trotzdem verschiedene Positionen $\xi_i^{(1)}$, $\xi_i^{(2)}$,... auf den zugrundeliegenden Skalen einnehmen. Das aber führt wieder dazu, daß solche Items, die einer bestimmten latenten Dimension entsprechen, positiv miteinander korreliert sind. Somit wird auch die Forderung nach stochastischer Unabhängigkeit verletzt sein.

Die Wichtigkeit dieser beiden Axiome und ihrer Überprüfungsmöglichkeit wird aber besonders dann deutlich, wenn solche Fragestellungen bearbeitet werden sollen, in denen die Struktur bestimmter Variablen, die nicht ohne weiteres beobachtbar sind, von Interesse ist. Im Beispiel der Intelligenzforschung kann etwa die Frage auftauchen, ob zwei Dimensionen wie verbale Flüssigkeit und logisches Denken, die auf Grund einer Faktorenanalyse isoliert wurden, tatsächlich strukturell verschiedene Eigenschaften besitzen. Die bisher besprochenen Modelltests sind allerdings kaum dazu geeignet die Axiome der Unidimensionalität und der stochastischen Unabhängigkeit zu überprüfen. Dies wird aus einem Theorem van den Wollenbergs (1979,S.100) einerseits und ausgedehnten Simulationsstudien desselben Autors deutlich. So konnte van den Wollenberg zeigen, daß unter der

Voraussetzung zweier Rasch-homogener Tests mit gleichen Itemparametervektoren und gleicher Verteilung der Eigenschaftsparameter der Personen, diese beiden Skalen als eindimensional in dem Sinn aufgefaßt werden müssen, als die Itemparameterschätzer über alle Scoregruppen konstant bleiben. Die entsprechenden bisher referierten Teststatistiken, die auf einer Teilung der Stichprobe nach Rohscoregruppen basieren, versagen bei der Entdeckung dieser Inkonsistenz des Modells.

Aber auch dann, wenn sowohl die Itemparameter der beiden Dimensionen und die Verteilungen der Personenparameter variiert werden, steigen die χ^2-Werte solcher Tests, die auf Rohscore-Teilung beruhen, nur geringfügig an, wie in Simulationsstudien festgestellt wurde. Da die Veränderungen der entsprechenden Prüfgrößen sich nur in einem sehr bescheidenen Rahmen bewegten, muß der Schluß gezogen werden, daß Zufallsschwankungen die erwähnten Effekte leicht überlagern können und somit diese Tests als ungeeignet zur Prüfung der Unidimensionalität und der stochastischen Unabhängigkeit erscheinen.

Einige neuere Arbeiten beschäftigen sich mit der Entwicklung adäquater Verfahren, um diesen Mangel an diagnostischen Instrumenten zur Feststellung der Geltung des Rasch-Modells zu beheben. Diese sollen im folgenden genauer dargestellt werden.

4.2.1 VAN DEN WOLLENBERGS Q_2

Unterliegt bestimmten Fragebögen mehr als eine latente Dimension, so wird eine Assoziation zwischen einzelnen Items auch dann nicht aufgehoben werden können, wenn auf den globalen Testwert t_i bedingt wird. (s.Kap.2) Lokale stochastische Unabhängigkeit wird dann auch nicht erfüllt sein. Wurde in den bisher besprochenen Verfahren vor allem Häufigkeiten erster Ordnung berücksichtigt und blieben somit Interaktionen zwischen Items außer Acht, so beruht der Q_2-Test von A. van den Wollenberg(1979) auf der Untersuchung von Effekten zweiter Ordnung.

Information über Interaktionen zwischen zwei Items kann man anhand einer Analyse folgender 2x2-Kontingenztafel erhalten. Gegeben sei Scoregruppe r und die Häufigkeiten der richtigen Antworten zu Item j und Item m bei gemeinsamer Betrachtung. Die entsprechende 2x2-Tafel ist dann

		Item m: 1	Item m: 0	
Item j	1	n_{rjm}	$n_{rj\bar{m}}$	n_{rj}
	0	$n_{r\bar{j}m}$	$n_{r\bar{j}\bar{m}}$	$n_{r\bar{j}}$
		n_{rm}	$n_{r\bar{m}}$	n_r

Unter der Voraussetzung, daß die Itemparameter getrennt für jede Scoregruppe r mittels CML (10) geschätzt wurden, erhält man die bedingten Erwartungswerte erster und zweiter Ordnung aus (23) und (24).

Sei $d^2 = (n_{rjm} - E(n_{rjm}))^2$ dann können unter Verwendung von (23) und (24) alle möglichen 2x2-Tafeln erwarteter und beobachteter Häufigkeiten in Scoregruppe r durch die Statistik

$$(54) \qquad q_{rjm} = \frac{d^2}{E(n_{rjm})} + \frac{d^2}{E(n_{r\bar{j}m})} + \frac{d^2}{E(n_{rj\bar{m}})} + \frac{d^2}{E(n_{r\bar{j}\bar{m}})}$$

miteinander verglichen werden. Die Randsummen n_{rj} und n_{rm} sind suffiziente Statistiken für $\hat{\varepsilon}_j$ und $\hat{\varepsilon}_m$ und die Erwartungswerte in (54) erhält man durch die geschätzten Itemparameter. Dies impliziert die Äquivalenz der beobachteten und der erwarteten Häufigkeiten erster Ordnung. Die Werte für q_{rjm} ergeben sich nur auf Grund der Unterschiede zweiter Ordnung. Daraus folgt weiter, daß (54) χ^2-verteilt ist mit df=1.

In jeder Scoregruppe können k(k-1)/2 Itempaare und somit ebensoviele χ^2-Statistiken beobachtet werden. Summation dieser Prüfgrößen führt zu

$$(55) \qquad Q^{*}_{2(r)} = \sum_j \sum_m q_{rjm} \; ; \; (j=1,..,k-1; \quad m=j+1,..,k) \; .$$

Wären die einzelnen q_{rjm} unabhängig, so wäre $Q_{2(r)}$ χ^2-verteilt mit df=k(k-1)/2 Freiheitsgraden. Dies ist allerdings wie schon bei der Statistik Q_1 (44) nicht der Fall. Bei gegebenen Randsummen n_{rj} wird Item j n_{rj}-mal mit r-1 anderen Items gemeinsam auftreten, da das Antwortmuster insgesamt r positive Antworten enthalten muß. Item j wird daher $n_{rj}.(r-1)$ - mal Element eines Paares sein. Daraus folgt die Restriktion, die für jedes Item gilt

$$(56) \qquad n_{rj}.(r-1) = \sum_m n_{rjm} \; .$$

Es bleiben also nur $k(k-1)/2 - k = (k/2)(k-3)$ Häufigkeiten frei zu variieren. Analog zur Testgröße Q_1^* ist eine korrigierte Version von Q_2^* nötig, um eine χ^2-Verteilung mit entsprechenden Freiheitsgraden zu approximieren. Dies führt zur Q_2-Statistik van den Wollenbergs

$$(57) \qquad Q_2 = \frac{k-3}{k-1} Q_2^*$$

mit $df = (k/2)(k-3)$ Freiheitsgraden.

Simulationsstudien (van den Wollenberg, 1979) ergaben, daß der Q_2-Test sehr sensibel auf Verletzungen der Unidimensionalität und der lokalen stochastischen Unabhängigkeit reagiert. Die praktische Anwendung wirft allerdings einige Probleme auf. Dies betrifft vor allem die Zahl zu analysierender Itempaare, die bei größeren k rasch ansteigt. Sind zusätzlich die Scoregruppen auf Grund kleiner Stichproben nur gering besetzt, so kann die Parameterschätzung leicht versagen. Selbst wenn eine solche aber möglich ist, können die erwarteten Häufigkeiten so klein werden, daß die Stabilität der Statistik beeinträchtigt wird. I.Molenaar(1981) gab einige Beispiele für 2x2-Tafeln, in denen eine Besetzung gewisser Zellen aufgrund zu kleiner, gegebener Randsummen unmöglich wird.

Auf diese Probleme geht van den Wollenberg in einer neueren Arbeit (1981) näher ein. Wird die Gesamtstichprobe nicht nach einzelnen Scoregruppen sondern nach einem anderen Kriterium, das disjunkte und erschöpfende Teilstichproben zu bilden erlaubt (also etwa niedriger und hoher Score oder niedriger, mittlerer und hoher Score), und sei der Index dieser Teilstichproben g (g=1,..,G), so erhält man die erwarteten Häufigkeiten erster und zweiter Ordnung analog zu (23) und (24) aus

$$(58) \qquad E(n_{gj}) = n_r \, \varepsilon_j^{(g)} \, \gamma_{r-1}^{(j)}(\underset{\sim}{\varepsilon}^{(g)}) / \gamma_r(\underset{\sim}{\varepsilon}^{(g)})$$

und

$$(59) \qquad E(n_{gjm}) = n_r \, \varepsilon_j^{(g)} \varepsilon_m^{(g)} \, \gamma_{r-2}^{(j,m)}(\underset{\sim}{\varepsilon}^{(g)}) / \gamma_r(\underset{\sim}{\varepsilon}^{(g)}) \; .$$

Alle weiteren oben dargestellten Argumente, die zu Q_2 führen, treffen auch jetzt zu.

Abschließend sei erwähnt, daß sich die Unterscheidung in drei Teilstichproben (niedriger, mittlerer und hoher Rohscore) aus folgenden Gründen als recht günstig erweist. Sei der gesamte Tests aus zwei homogenen Subtests zusammengesetzt. Personen mit hohem Rohscore werden für beide Skalen hohe Werte erzielen und somit relativ homogen sein. Das Gleiche trifft auch für Scoregruppen mit kleinen Parameterwerten zu. Hat man die Stichprobe in drei anstatt zwei Gruppen unterteilt, dann wird die mittlere insoferne am heterogensten sein, als manche Personen für die eine Skala höhere und für die andere niedrigere Werte erzielen werden. In dieser Teilstichprobe wird also die Verletzung der stochastischen Unabhängigkeit am ehesten zu beobachten sein. Ein weiteres Argument betrifft die Stichprobengröße der Subsamples. Für drei Teilgruppen ist die Zellenbesetzung in den 2x2-Tafeln größer als bei stärkerer Unterteilung der gesamten Stichprobe. Die Anpassung an die χ^2-Verteilung wird also besser sein, die Teststatistik somit genauer.

Allgemein läßt sich auf Grund graphischer Analysen meist eine Abschätzung der Dimensionalität des Fragebogens erzielen (s.Kap.5). Die jeweilige Gruppenbildung wird dann dieser Abschätzung entsprechend erfolgen, wobei man einen Kompromiß zwischen den beiden erwähnten, entgegenlaufenden Forderungen eingehen wird.

4.2.2 MOLENAARS Δ

Wie van den Wollenberg (s.Kap.4.2.1) schlägt Molenaar (1981) eine Analyse von 2x2-Kontingenztafeln unter gegebenen Randsummen vor. Seine Überlegungen beruhen vor allem auf dem schon erwähnten Einhergehen von Verletzungen der Unidimensionalität und der stochastischen Unabhängigkeit. Unter der Annahme zweier Subskalen, aus denen ein bestimmter Fragebogen zusammengesetzt ist, werden Items, die derselben Teilskala angehören, stärker miteinander korreliert sein als solche, die von verschiedenen Subtests stammen. Wenn die Wahrscheinlichkeiten richtiger Lösungen der Items j und m voneinander unabhängig sind, dann müßte das logarithmische Verhältnis der Chancen einer korrekten Antwort (log odds ratio)

$$(60) \qquad \ln \Delta_{ijm} = \ln \left\{ \frac{\pi_{ij} \, (1 - \pi_{im})}{(1 - \pi_{ij}) \, \pi_{im}} \right\} = 0$$

sein, wobei die π_{ij} wie in (23) definiert sind. Einen Schätzer für Δ_{ijm} erhält man aus den beobachteten Häufigkeiten der 2x2-Tafel(wie in Kap.4.2.1)

$$(61) \qquad \hat{\Delta}_{rjm} = \frac{n_{rjm} \, (n_r - n_{rj} - n_{rm} + n_{rjm})}{(n_{rj} - n_{rjm})(n_{rm} - n_{rjm})} \; .$$

Items von verschiedenen Subskalen müßten den obigen Überlegungen entsprechend nahezu unkorreliert sein, es sollte also (60) erfüllt sein, während für Items derselben Teilskala ein $\Delta \neq 1$ zu beobachten sein wird. Bei einem Vergleich der aus den Daten geschätzten $\hat{\Delta}_{rjm}$ und den aus dem Rasch-Modell vorhergesagten Δ_{rjm} (60) sollte im Falle zweier

homogener Subskalen folgende Ungleichungen gelten:

$$(62)\qquad \begin{aligned} \hat{\Delta}_{rjm} &> \Delta_{rjm} \qquad \text{für Items derselben Teilskala} \\ \hat{\Delta}_{rjm} &< \Delta_{rjm} \qquad \text{für Items verschiedener Skalen,} \end{aligned}$$

wobei dieses Muster über alle Scoregruppen r sowie für alle Itempaare zu beobachten sein sollte.

Einzelne Unterschiede wie in (62) lassen sich auch mittels der hypergeometrischen Verteilung testen. Die Wahrscheinlichkeit einer Zellenbesetzung der 2x2-Tafel n_{rjm} unter gegebenen Randsummen ist

$$(63)\qquad p(N_{rjm}=n_{rjm} \mid N_{rj}=n_{rj}, N_{rm}=n_{rm}, N_r=n_r, \xi) =$$

$$K \binom{n_{rj}}{n_{rjm}} \binom{n_r-n_{rj}}{n_{rm}-n_{rjm}} \Delta_{rjm}^{\,n_{rjm}} \quad ,$$

mit

$$K = 1 \Big/ \sum_{n'_{rjm}} \binom{n_{rj}}{n'_{rjm}} \binom{n_r-n_{rj}}{n_{rm}-n'_{rjm}} \Delta_{rjm}^{\,n'_{rjm}} \, ,$$

wobei im Nenner über alle möglichen n'_{rjm} summiert wird, die unter den gegebenen Randsummen möglich sind, also $\max(0, n_{rj}+n_{rm}-n_r) \leqslant n'_{rjm} \leqslant \min(n_{rj}, n_{rm})$. Einseitige Tests erhält man durch Aufsummieren von (63) über n_{rjm} zwischen $\max(0, n_{rj}+n_{rm}-n_r)$ und N_{rjm} für den unteren Wahrscheinlichkeitsbereich und zwischen N_{rjm} und $\min(n_{rj}, n_{rm})$ für den oberen. Im Falle $\Delta_{rjm}=1$ reduziert sich der Test zum sogenannten "Fisher's exact test" oder auch "Fisher-Irwin-test".(Die Eigenschaften dieser Statistiken werden

von Lehmann,1959,S.140ff. diskutiert. Bei D.R.Cox findet man eine Darstellung mit Angaben über Konfidenzintervalle im Rahmen bedingter Analysen von Parametern des linearen logistischen Modells.)

Sind alle beobachteten Häufigkeiten groß, läßt sich auch eine Approximation an die Standardnormalverteilung angeben, die von Harkness (1965) vorgeschlagen und von Molenaar auf die hier referierte Situation angewandt wurde. Sei n_{rjm} hypergeometrisch verteilt, wie in (63), so schlägt Molenaar (1981) folgende einseitige Tests unter Verwendung einer Kontinuitätskorrektur vor:

$$p(N_{rjm} = n_{rjm}) = \Phi((n_{rjm}+1/2-n_r\hat{\pi}_{rjm}).H)$$

(64)

$$p(N_{rjm} \quad n_{rjm}) = 1 - \Phi((n_{rjm}-1/2-\hat{\pi}_{rjm}).H)$$

mit

$$H^2 = \frac{1}{\hat{\pi}_{rjm}} + \frac{1}{\hat{\pi}_{rj}-\hat{\pi}_{rjm}} + \frac{1}{\hat{\pi}_{rm}-\hat{\pi}_{rjm}} + \frac{1}{1-\hat{\pi}_{rj}-\hat{\pi}_{rm}+\hat{\pi}_{rjm}} ,$$

und $\Phi(.)$ als der kumulativen Standardnormalverteilungsfunktion mit entsprechenden Integrationsgrenzen.

Abschließend sei noch auf die Möglichkeit verwiesen, die berechneten Werte und Differenzen in (62) zur explorativen Analyse heranzuziehen. Ein Hilfsmittel zur graphischen Repräsentation der entsprechenden Muster wird in Kap.5 vorgeschlagen.

4.2.3 T.TJURS VORSCHLÄGE

Als letztes sollen noch kurz zwei Ansätze zur Kontrolle der Unidimensionalität von T.Tjur (1980) Erwähnung finden. Im Rahmen einer Arbeit über den Zusammenhang des Rasch-Modells mit dem log-linearen Modell empfiehlt Tjur, ebenso wie van den Wollenberg (Kap.4.2.1) und Molenaar (Kap.4.2.2) eine Analyse von Kontingenztafeln. Ein erster Vorschlag basiert auf der Formulierung folgender 2x(k-1)-Tafel:

		Item j gelöst Item m nicht gelöst	Item j nicht gel. Item m gelöst
Rohscore	1		
	2		
	⋮	⋮	⋮
	r		
	⋮	⋮	⋮
	k-1		

Hier wird also analog zu den Testansätzen der vorigen Kapiteln ein Itempaar (j,m) zur Analyse herangezogen. Die zugrundeliegende Idee entspricht Molenaars Argumentation besser Unterschiede zwischen Items über Rohscoregruppen zu beachten als umgekehrt. Überdies ist die Erstellung einer kombinierten Statistik wie bei van den Wollenberg nicht notwendig, da die Information aller Rohscoregruppen schon in der Kontingenztafel enthalten ist. Ein größerer Vorteil aber scheint darin zu bestehen, daß alle relevanten Rohscoregruppen r (r=1,..,k-1) Berücksichtigung finden. Da sowohl bei van den Wollenbergs Q_2 als auch bei Molenaars Δ die 2x2-Tafel so formuliert ist, daß richtige Beantwortung beider Items j und m in die Analyse miteinbezogen werden existieren dort nur k-3 nichttriviale Scores r (r=2,..,k-2). Dies kann sich aber beson-

ders bei kleinen k negativ auswirken, da dann eventuell relativ viele Personen in diese Gruppen fallen und somit Information verlorengeht.

Konkrete Statistiken werden von Tjur nicht präsentiert, es lassen sich aber die gleichen Verfahren wie bei van den Wollenberg oder Molenaar verwenden. Da in erster Linie das Vorhandensein von Interaktionseffekten interessiert, wird Fisher's exact test (s.a.Kap.4.2.2) ein geeignetes Instrumentarium hiezu sein. Alles bisher erwähnte gilt auch für Tjurs zweiten Vorschlag. Dabei wird der durch die Rohscoregruppen definierte Faktor durch ein weiteres Item ersetzt. Die folgende 2x2-Tafel erlaubt also die Untersuchung von Item-Triples:

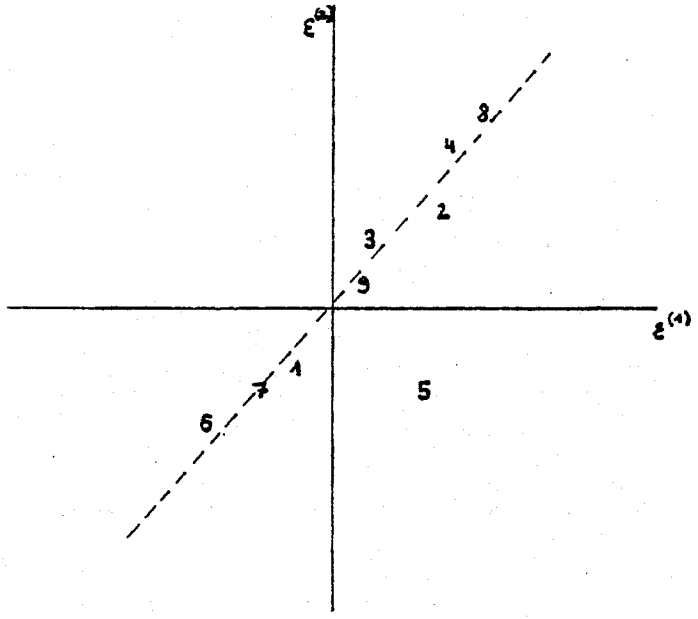
	Item j gelöst Item m nicht gel.	Item j nicht gel. Item m gelöst
Item q gelöst		
Item q nicht gel.		

Unter der Annahme einer homogenen Skala müßten die beiden Dimensionen unabhängig sein. Findet man allerdings substantielle Interaktionen, dann können entsprechende Items einer gemeinsamen Subskala angehören. Die 2x2-Kontingenztafeln lassen sich für alle möglichen Item-Triples und beliebige, disjunkte und erschöpfende Teilstichproben analysieren. Auf Grund der dabei beobachteten Interaktionsmuster können dann Rückschlüsse auf die Dimensionalität des Fragebogens gezogen werden.

5. GRAPHISCHE ANALYSEMETHODEN

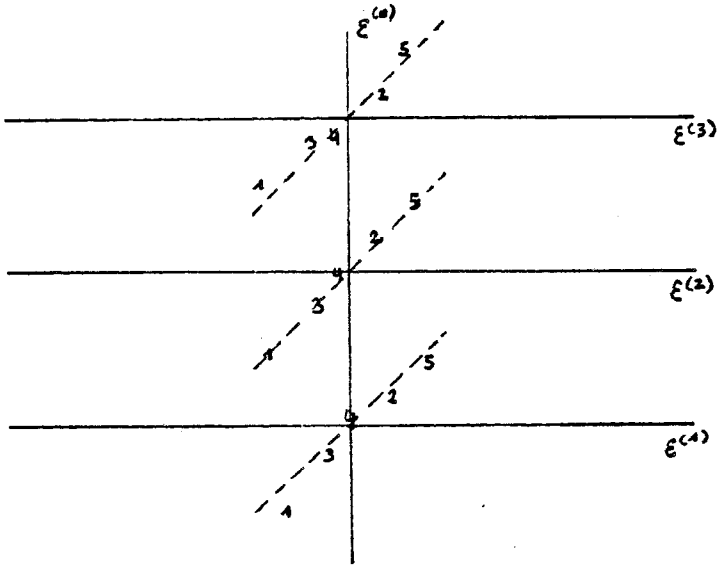
Neben den im vorigen Abschnitt dargestellten, numerisch teils recht aufwendigen, inferenzstatistischen Verfahren existieren auch einige Vorschläge zur graphischen Exploration eventueller Verletzungen der Modellannahmen. Im Gegensatz zu jenen sind diese mit relativ wenig Rechenaufwand verbunden, erlauben aber trotzdem eine Evaluierung erhobener Daten unter der Voraussetzung des Rasch-Modells. Darüber hinaus stellen sie geeignete Hilfsmittel dazu dar, die Art eventueller Modellverletzungen zu erkennen und somit die Auswahl einer geeigneten inferenzstatistischen Methode zu erleichtern.

Die einfachste Analyse besteht darin, die Itemparameterschätzer, die aus zwei oder mehreren Teilstichproben gewonnen wurden, gegeneinander zu plotten. Im Falle perfekter Modellkonformität sollten die Punkte auf einer 45^{o}-Geraden durch den Ursprung liegen bzw. auf Grund von Zufallsschwankungen nur gering von dieser abweichen.Abbildung 2. zeigt eine homogene Skala mit Ausnahme des Items 5, das in Teilstichprobe 2 wesentlich leichter als in Teilstichprobe 1 ist.



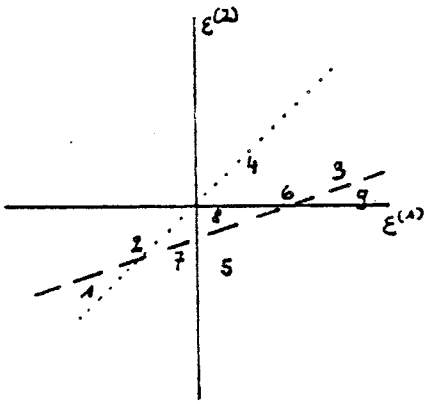
(Abbildung 2.)

Unterteilt man die Stichprobe in mehrere Subsamples, so scheint es günstig, die Parameterschätzer der Teilstichproben gegen jene der Gesamtstichprobe zu plotten (vgl.Fischer, 1974,S.282ff.), und hierbei den Ursprung um einen konstanten Betrag an der Achse für die Subsamples zu verschieben.Abbildung 3. gibt ein Beispiel für den Fall einer Teilung in drei Samples.



(Abbildung 3.)

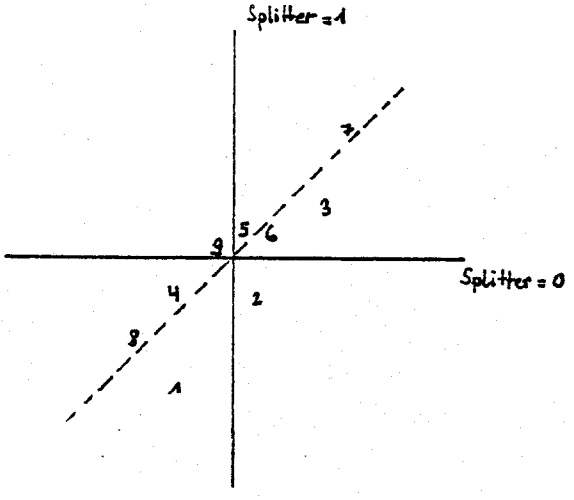
Das zusätzliche Einzeichnen der Regressionsgerade erlaubt eine Abschätzung der globalen Tendenz der Testitems. Abbildung 4. zeigt eine Skala, die in Teilstichprobe 2 leichter ist (Intercept negativ) und in der die Variation der Itemschwierigkeiten für Sample 1 größer ist (Anstieg < 1).



(Abbildung 4.)

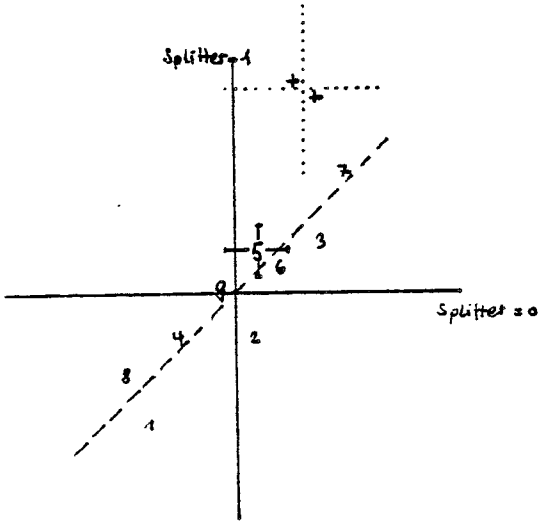
Das häufigste Teilungskriterium ist die Summe der richtigen Lösungen im untersuchten Fragebogen. Der Vorteil dieses Kriteriums, also etwa die Teilung nach dem Median in Gruppen mit hohem bzw. niedrigem Score, besteht darin, daß alle Einflüsse auf die experimentelle Situation zumindest indirekt erfaßt werden. Unterliegen dem Test allerdings zwei oder mehrere homogene Skalen, so wird diese Verletzung der Modellannahmen bei Verwendung des totalen Scores nicht entdeckt werden (s.Kap.4.2).

Eine Alternative hiezu ist die Einführung sogenannter Splitter-Items.(Dieses Verfahren wurde von van den Wollenberg (1979) vorgeschlagen und von Molenaar (1981) ausführlich diskutiert.) Man wählt dabei ein bestimmtes Item - das Splitter-Item -, unterteilt die Gesamtstichprobe je nach Lösung oder Nichtlösung dieses Items und schätzt w.o. getrennt die Parameter. Das Splitter-Item darf allerdings nicht mitgeschätzt werden, da sonst Artefakte auftreten (s.van den Wollenberg,1979,S.111ff.). Aus den resultierenden Plots lassen sich folgende Schlüsse ziehen: Items die zu derselben Skala gehören wie das Splitter-Item werden für Personen,die dieses nicht gelöst haben, schwieriger sein, während Items die nicht zur Skala gehören gleich schwierig sein sollten. Ein Beispiel gibt Abbildung 5.



(Abbildung 5.)

Da aber auf Grund der Normierung $\Sigma\varepsilon_j = 0$ in beiden Teilstichproben die Mittelwerte der Parameterschätzer zum Ursprung verschoben werden ist es günstig solche Items einzuführen, die nichts mit der Skala zu tun haben. Molenaar (1981) schlägt beispielsweise vor, die Versuchspersonen zu numerieren und jeder Person mit gerader Nummer oder einer Nummer größer als n/2 eine "richtige" Lösung dieser artifiziellen Items zuzuordnen. Die Darstellung des Plots einer homogenen Skala könnte dann beispielsweise wie in Abb. 6 aussehen. Die beiden artifiziellen Items, gekennzeichnet durch (+), deuten den "wahren Ursprung" an.



(Abbildung 6.)

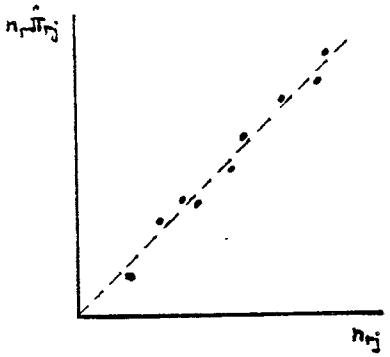
Obwohl natürlich alle Items als Splitter verwendet werden können, erweist es sich als geeigneter solche auszuwählen, die etwa gleichoft gelöst und nicht gelöst wurden. Für Items, die nur von wenigen Personen gelöst oder nicht gelöst wurden, sind die Parameterschätzer relativ ungenau, die Varianzen der Schätzer werden relativ groß. Eine Berücksichtigung von Konfidenzintervallen in den Plots (wie in Abb. 6. für Item 5) kann in solchen Fällen günstig sein.

Das Verfahren der graphischen Analyse mittels Splitter-Items dient neben der Entdeckung einzelner nicht-homogener Items auch zur Auffindung mehrdimensionaler Parameterstrukturen. Besteht der Gesamttest zum Beispiel aus zwei homogenen Skalen, so werden die Items des einen Subtests immer dann entlang einer Geraden liegen, wenn irgend ein Item dieser Skala als Splitter verwendet wurde.

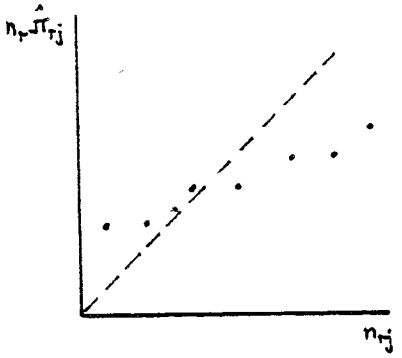
Generell läßt sich sagen, daß Items, die wesentlich unter der 45°-Geraden durch den Ursprung liegen, positiv mit dem Splitter-Item korreliert sind und einen steileren Anstieg der ICC aufweisen. Items oberhalb der Geraden haben einen flacheren Anstieg; sind sie nahe dem durch die sinnlosen Items festgelegten Ursprung zu finden, deutet dies auf Verletzung der Monotonizität. Das entsprechende Item wird der zu analysierenden Skala nicht homogen sein.

Zusammenfassend scheint die Anwendung graphischer Methoden, bei denen Parameterschätzer unter Verwendung von Splitter-Items gegeneinander geplottet werden, besonders dann vorteilhaft, wenn der gesamte Fragebogen nicht a priori als eindimensional im Sinne einer latenten Struktur, aufgefaßt werden kann. Sowohl einzelne nicht homogene Items als auch Subskalen, die zwar intern homogen zusammen mit anderen aber nicht konsistent sind, können relativ einfach mit Hilfe dieses Verfahrens erkannt und isoliert werden.

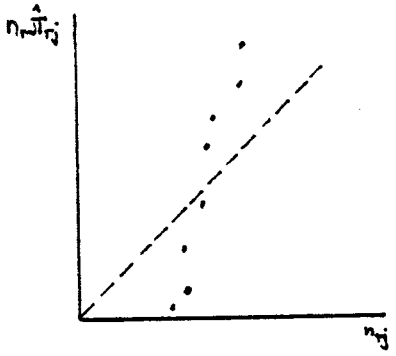
Zum Abschluß seien noch kurz zwei weitere, etwas andere Möglichkeiten graphischer Modellkontrollen dargestellt. Dies betrifft einerseits die Untersuchung der "Parallelität" der ICC's, die wie schon in Kap. 4.1.6 auf einem Vergleich beobachteter und erwarteter Häufigkeiten der Lösung eines bestimmten Items über Rohscoregruppen beruhen. Plottet man die n_{rj} gegen die $n_r \hat{\pi}_{rj}$ so lassen sich folgende vier Möglichkeiten auffinden:



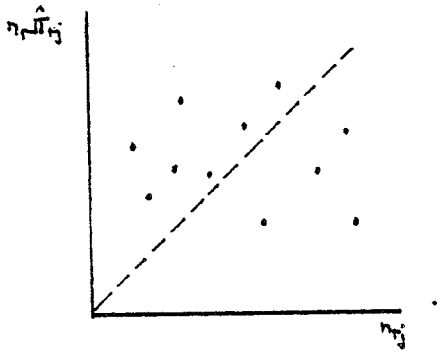
(Abb.7.1)



(Abb.7.2)



(Abb.7.3)



(Abb.7.4)

Während in Abb.7.1 und 7.4 relativ eindeutig Konformität bzw. Unabhängigkeit des Items von der zu analysierenden Skala festgestellt werden kann, deuten Muster wie in Abb. 7.2 und 7.3 auf zu flache oder zu steile ICC. Solche Bilder können wohl auf Zufallsschwankungen beruhen (entsprechende Tests (49),(5o) geben genauere Aufschlüsse), finden sich aber mehrere Items mit gleichen Abweichungen, so weist dies auf die mögliche Existenz einer durch diese Items repräsentierten Subskala hin.

Eine genauere Untersuchung der Frage, ob solch eine Subskala isoliert werden kann, könnte ein zweites Verfahren ermöglichen, das auf einer graphischen Darstellung der Unterschiede zwischen den Δ und $\hat{\Delta}$ basiert. Wie in Kap.4.2.2 dargestellt, sollten solche Items, die eine bestimmte Subskala

besser messen als die anderen, über alle Scoregruppen ein höheres $\hat{\Delta}$ aufweisen, während die umgekehrte Beziehung darauf deutet, daß die beiden Items zwei verschiedenen Skalen angehören. Man kann nun für jedes Itempaar (j,m) die Größe

$$f_{jm} = \sum_{r=2}^{k-2} d_{jm}^{(r)}$$

berechnen, wobei $d_{jm}^{(r)} = 1$ wenn $\hat{\Delta}_{rjm} > \Delta_{rjm}$, 0 sonst. Ferner bestimmt man die Größe f_{jm}^{*} die wie folgt definiert ist: $f_{jm}^{*} = 1$ wenn f_{jm} sehr große Werte annimmt und $f_{jm}^{*} = 0$ für sehr kleine Werte von f_{jm}. Sehr groß heißt k-2-c und sehr klein 2+c, wobei c abhängig von der Anzahl der verschiedenen Scoregruppen zu wählen ist, um Zufallsschwankungen in den $\hat{\Delta}$ zu berücksichtigen; c wird umso kleiner sein je größer k wird.

Trägt man nun die f_{jm}^{*} in eine (kxk)-Dreiecksmatrix ein so könnte man beispielsweise folgendes Bild erhalten:

	1	2	3	4	5	6	7	8	9	10
1										
2	1									
3	0	0								
4	0	0	0							
5	1	1	0	0						
6	0	0	1	0	0					
7	0	0	1	0	0	1				
8	0	0	1	0	0	1	1			
9	0	0	1	0	0	1	1	1		
10	1	1	0	0	1	0	0	0	0	

(Abbildung 8.1)

Ein Muster wie in Abb. 8.1 weist darauf hin, daß der aus zehn Items bestehende Fragenbogen aus zwei verschiedenen

Subtests zusammengesetzt ist wobei Skala 1 die Items: 1,2,5,10 und Skala 2 die Items: 3,6,7,8,9 umfaßt. Item 4 gehört keinem Subtest an und könnte aus dem Fragebogen eliminiert werden. Ordnet man die zu einer Subskala gehörenden Items nebeneinander an, so erhält man das folgende Bild:

	1	2	5	10	4	3	6	7	8	9
1										
2	1									
5	1	1								
10	1	1	1							
4	0	0	0	0						
3	0	0	0	0	0					
6	0	0	0	0	0	1				
7	0	0	0	0	0	1	1			
8	0	0	0	0	0	1	1	1		
9	0	0	0	0	0	1	1	1	1	

(Abbildung 8.2)

Hier sind die beiden Subskalen recht klar erkennbar. In der Praxis werden natürlich selten so eindeutige Muster aufzufinden sein, vor allem werden in der Matrix verschiedene Positionen nicht mit 0 oder 1 besetzt sein. Dennoch scheint dieses Verfahren besser geeignet etwaige Strukturen der dargestellten Art zu erkennen als dies aus größeren Zahlenmengen möglich ist, besonders wenn die Anzahl der Items und somit der Rohscoregruppen groß wird.

Graphische Analysemethoden, wie sie in diesem Abschnitt dargestellt wurden, können insgesamt als brauchbare Hilfsmittel aufgefaßt werden, die Struktur bestimmter Daten unter dem Rasch-Modell zu untersuchen. Eventuelle Verletzungen der

Modellannahmen können erkannt werden und auch über die Art der Nichtkonformität lassen sich Aufschlüsse erzielen. Solcherart generierte Hypothesen sollten dann aber dem Instrumentarium inferenzstatistischer Verfahren ausgesetzt werden, da graphische Analysemethoden zwar wertvolle Hinweise geben nicht aber exakte Wahrscheinlichkeitsaussagen über bestimmte Annahmen liefern können.

LITERATUR

ANDERSEN,E.B.: Conditional Inference and Models for Measuring. Kopenhagen 1973(a)

ANDERSEN,E.B.: A Goodness-of-Fit Test for the Rasch-Model. Psychometrika 38,p.123-140,1973(b)

BIRNBAUM,A.: Some Latent Trait Models and their Use in Inferring an Examinee's Ability. In: LORD,F.M. und NOVICK,M.R.(Hrg.): Statistical Theories of Mental Test Scores. Reading,Ma.,p.397-479,1968

COX,D.R.: The Analysis of Binary Data. Methuen, London 1970

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

FISCHER,G.H.,SCHEIBLECHNER,H.H.: Algorithmen und Programme für das probabilistische Testmodell von Rasch. Psych. Beitr. 12, S.23-51,1970

HARKNESS,W.L.: Properties of the Extended Hypergeometric Distribution. Ann.Math.Stat.,p.938-945,1965

HATZINGER,R.,DITTRICH,R.: Using GLIM for Computing the Rasch-Model and the Corresponding Multiplicative Poisson-Model.Res.Mem.178,Inst.f.Adv.Stud.,Vienna 1981

LAZARSFELD,P.F.: The Logical and Mathematical Foundation of Latent Structure Analysis. In: STOUFFER,S.A. et al.: Measurement and Prediction. Princeton Univ. Press,1950(a)

LAZARSFELD,P.F.: Some Latent Structures (ib.),1950(b)

LEHMANN,E.L.: Testing Statistical Hypotheses. Wiley,N.Y.,1959

MARTIN-LÖF,P.: Statistika Modeller. Inst.f.Försäkringsmat. och mat. Stat.,Univ.Stockholm,1973

MOLENAAR,I.M.: Some Improved Diagnostics for Failure of the Rasch-Model. Heyman Bull. Psych. Inst. Univ. Groningen, HB-80-482-EX,1981

NEYMAN,J.,SCOTT,E.L.: Consistent Estimates Based on Partially Consistent Observations. Econometrika, 16,p.1-32,1948

RAO,C.R.: Linear Statistical Inference and its Application. Wiley,N.Y.,1973

RASCH,G.: Probabilistic Models for some Intelligence and Attainment Tests. Kopenhagen,1960

RASCH,G.: On General Laws and the Meaning of Measurement in Psychology. Univ.Calif.Press,1961

ROUSSAS,G.G.: A First Course in Mathematical Statistics. Reading,Ma.,1973

TJUR,T.: A Connection between Rasch's Item Analysis Model and a Multiplicative Poisson Model. Inst.Math. Stat.,Prepr.6,Univ.Copenhagen,1981

VAN DEN WOLLENBERG,A.L.: The Rasch-Model and Time-Limit Tests. Stichting Studentenpers Nijmegen,1979

VAN DEN WOLLENBERG,A.L.: Two new Tests for the Rasch-Model. Vakgr.Psych.Math.,80 MA 01, Kat.Univ.Nijmegen,1980

VAN DEN WOLLENBERG,A.L.: The Q2 Test for the Rasch-Model Revisited. Vakgr.Psych.Math.,81 MA 04, Kat.Univ. Nijmegen,1981