Zuerst erschienen in: Zeitschrift für Differentielle und Diagnostische Psychologie 4 (1992), S. 269-275

Ein pragmatischer Ansatz in der statistischen Theorie psychologischer Testscores

A pragmatic approach to statistical theories of psychological test scores

Wilhelm Kempf
Universität Konstanz

Zusammenfassung: Die vorliegende Arbeit diskutiert die Anwendbarkeit der Klassischen Testtheorie und des Rasch-Modells unter verschiedenen Paradigmen der Testadministration. Wie gezeigt wird, liefert die Klassische Testtheorie in den meisten Fällen ein hochgradig ungenaues Konfidenzintervall für den True-Score. Während die Anwendung des Rasch-Modells in diesen Fällen die Konstruktion eines deutlich besseren Konfidenzintervalls erlaubt, gibt es jedoch auch andere Fälle, in welchen das Rasch-Modell nicht anwendbar ist, während die Klassische Testtheorie durchaus brauchbare Ergebnisse liefert. Dasselbe zeigt sich auch für die statistische Signifikanzprüfung von Scoredifferenzen. Bei Tests, die (wie üblich) durch eine feste Itemauswahl definiert sind, führen jedoch beide Modelle zu einer Überbewertung der Scoreunterschiede. Der Vergleich der True-Scores zweier Vpn sollte in diesem Fall daher statt dessen mit Hilfe des Tests von McNemar erfolgen.

Abstract: The present paper discusses the applicability of Classical Test Theory and of the Rasch-Model under different paradigmas of test administration. As is demonstrated, the confidence interval provided by Classical Test Theory for the true-score, is highly inaccurate in most cases. While application of the Rasch-Model allows for the construction of a superior confidence interval in these cases, there exist other cases, however, where the Rasch-Model does not apply, while Classical Test Theory produces reasonable results. The same holds for the statistical significance of score-differences. If tests are (as usually) defined as fixed selections of test items, however, then both models overestimate the score differences. Comparison of true-scores of different Ss on the same test should, therefore, be based on the well known *Test of McNemar,* instead.

Trotz der schwerwiegenden Kritik, welcher die Klassische Testtheorie seit den 60er Jahren ausgesetzt war, stellt sie in der angewandten Testpsychologie noch heute den gebräuchlichen Standard dar. So enthält z. B. das Handbuch psychologischer und pädagogischer Tests (Brikkenkamp, 1975, 1983) fast ausschließlich solche Testverfahren, die nach der klassischen Testtheorie entwickelt wurden, und Schulze (1990) stellt fest, daß die auf Grundlage der klassischen Testtheorie entwickelten Skalen auch im ZUMA-Handbuch sozialwissenschaftlicher Skalen (1988) immer noch eine konkurrenzlose Position einnehmen. Das Rasch-Modell konnte sich trotz überlegener (statistischer) Theorie in der Testpraxis dagegen bisher nicht entsprechend durchsetzen.

Gründe dafür sind u. a. der verbreitete Glaube an die universelle Anwendbarkeit der Klassischen Testtheorie sowie die Tatsache, daß die Klassische Testtheorie den Bedürfnissen des Testpraktikers stärker entgegenkommt, indem sie z. B. pragmatische Problemlösungen anbietet, wie ein Vertrauensbereich für den Testscore einer Vp angegeben und/oder die Differenz zweier Testscores hinsichtlich ihrer statistischen Signifikanz bewertet werden kann.

Wenn man von der universellen Anwendbarkeit der Klassischen Testtheorie spricht, so hat man dabei aber nur die Modellstruktur der Klassischen Testtheorie im Blick, wonach jeder Vp v und jedem Test t eine Zufallsvariable X_{vt} mit endlichem Erwartungswert τ_{vt} («True Score») und endlicher Varianz $\sigma^2(X_{vt}) = \sigma^2(F_{vt})$ mit $F_{vt} =: X_{vt} - \tau_{vt}$ («Meßfehler») entspricht. Sobald es um die Anwendung der Klassischen Testtheorie (z. B. um die Reliabilitätsbestimmung von Tests) geht, werden jedoch Zusatzannahmen wie z. B. die Parallelität von Tests oder Testteilen erforderlich. Ähnlich wie die Modellannahmen des Rasch-Modells haben diese Zusatzannahmen einen empirischen Gehalt und

sind daher auch grundsätzlich einer empirischen Überprüfung zugänglich (vgl. Lord & Novick, 1968; Meder, Kempf & Boneberg, 1990).

Selbst wenn ein Test den zur Bestimmung seiner Reliabilität erforderlichen Zusatzannahmen genügt, ist die praktische Brauchbarkeit der Klassischen Testtheorie für die inferenzstatistische Evaluation der in dem Test erzielten Scores damit jedoch noch nicht erwiesen. Sowohl das Konfidenzintervall für den True-Score

$$\text{KONF}\{x_{vt}-z_{krit}\sigma(F_{\cdot t}) \leq \tau_{vt} \leq x_{vt}+z_{krit}\sigma(F_{\cdot t})\} \quad (1)$$

als auch der Signifikanztest für den Unterschied zweier Testscores mittels der N(0,1)-verteilten Prüfgröße

$$z = (X_{vt}-X_{wt})/[\sigma(F_{\cdot t})\sqrt{2}] \quad (2)$$

(vgl. Lienert (1969, S. 454) ist an die zusätzlichen Voraussetzungen (a) der Normalverteilung des Meßfehlers und (b) der (zumindest näherungsweisen) Übereinstimmung des Standardmeßfehlers $\sigma(F_{\cdot t})$ mit der tatsächlichen Standardabweichung des Meßfehlers der Person $\sigma(F_{vt})$ gebunden.

Untersucht man die Realitätshaltigkeit dieser Voraussetzung für den Summenscore $X_{vt} = \sum_i A_{vi}$ von Tests aus binären Items ($A_{vi} = 1$ falls «gelöst», 0 sonst), so zeigt sich, daß in diesem Falle lediglich Voraussetzung (a) einigermaßen unproblematisch ist: aufgrund des Zentralen Grenzwertsatzes ist X_{vt} (und damit auch F_{vt}) bei großer Itemanzahl k näherungsweise normalverteilt (da sich X_{vt} und F_{vt} nur durch eine additive Konstante unterscheiden).

Um die Realitätshaltigkeit der Voraussetzung (b) zu beurteilen, ist es erforderlich, das Zustandekommen der Testscores genauer zu untersuchen, wofür zwischen zwei Paradigmen des psychologischen Testens unterschieden wird, die beide von der Existenz eines endlich oder unendlich großen Pools von Testaufgaben ausgehen (vgl. Kempf, 1991).

Im Item Sampling Paradigma wird aus diesem Pool für jede Testung einer Vp eine eigene Zufallsstichprobe von k Testaufgaben gezogen (Ziehen mit Zurücklegen). Die Zufallsvariation der Testleistung einer Vp beruht ausschließlich auf der Zufallsauswahl der von ihr bearbeiteten Testaufgaben. Die Lösungswahrscheinlichkeiten $p_{vi} = p_{v\cdot}$ sind gleich dem Anteil der Testaufgaben des Pools, welche die Vp beherrscht.

Im Fixed Test Paradigma dagegen wird allen Vpn dieselbe Auswahl von Testaufgaben vorgelegt, so daß die Zufallsvariation der Testleistung hier auf der Zufälligkeit des Erfolges beruht, den die Person bei der Bearbeitung der Aufgaben hat, welche sich ihrerseits (z. B. in ihrer Schwierigkeit) unterscheiden können, so daß verschiedene Items $i \neq j$ in der Regel auch verschiedene Lösungswahrscheinlichkeiten $p_{vi} \neq p_{vj}$ besitzen. Konstante Lösungswahrscheinlichkeiten $p_{vi} = p_{v\cdot}$ ergeben sich im Fixed Test Paradigma allenfalls als empirische Zusatzannahme (Itemparallelität im Sinne der Klassischen Testtheorie).

Eine Verbindung zwischen den beiden Paradigmen stellt das Hybrid Modell dar, in dem (wie im Item Sampling Paradigma) für jede Testung einer Vp eine eigene Itemstichprobe gezogen wird, die Zufallsvariation der Testleistung jedoch nicht nur auf der Zufälligkeit der Aufgabenauswahl (wie im Item Sampling Paradigma), sondern auch auf der Zufälligkeit des Bearbeitungserfolges beruht (wie im Fixed Test Paradigma).

Die Lösungswahrscheinlichkeiten $p_{vi} = p_{v\cdot}$ sind dann gleich der mittleren Lösungswahrscheinlichkeit der Items des Pools.

Sind die Bedingungen des Item Sampling Paradigmas oder des Hybrid Modells erfüllt, oder sind die Testitems im Fixed Test Paradigma untereinander alle parallel, so folgt der Testscore einer Vp einer Binomialverteilung mit dem Parameter $p_{v\cdot}$ («Binomialmodell»). Eine eigene statistische Theorie psychologischer Testscores ist in diesen Fällen nicht vonnöten. Es können die bekannten Verfahren auf Grundlage der Binomialverteilung angewendet werden.

Einer eigenen (statistischen) Theorie psychologischer Testscores bedürfen wir erst, wenn dies nicht der Fall ist. Die pragmatische Aufgabe einer solchen Theorie besteht dann darin, den Vertrauensbereich und/oder die kritische Differenz von Testscores auch dann noch bestimmen zu können, wenn die Lösungswahrscheinlichkeiten voneinander verschieden sind, zu welchem Zweck das Rasch-Modell die Lösungswahrscheinlichkeiten als Funktion eines Personen(Fähigkeits)-Parameters θ_v und eines Item(Schwierigkeits)-Parameters σ_i approximiert

$$p_{vi} = \exp(\theta_v - \sigma_i)/(1 + \exp(\theta_v - \sigma_i)), \quad (3)$$

während die Klassische Testtheorie die Zusammensetzung des Scores aus den einzelnen Items nicht weiter reflektiert.

Sobald man diese Zusammensetzung jedoch genauer betrachtet, zeigt sich, daß die von der Klassischen Testtheorie als bloß «vereinfachende Annahme» eingeführte Voraussetzung (b) in der Praxis so gut wie nie erfüllt sein kann, da True-Score

$$\tau_{vt} = \sum_i p_{vi} \qquad (4)$$

und Fehlervarianz

$$\sigma^2 (F_{vt}) = \sigma^2 (X_{vt}) = \sum_i p_{vi} (1-p_{vi}) \qquad (5)$$

nicht unabhängig voneinander sind. Im Item Sampling Paradigma, Hybrid Modell und Fixed Test Paradigma mit parallelen Items (d.h. bei Geltung des Binomialmodells) besteht sogar eine streng funktionale, Abhängigkeit der Fehlervarianz vom True-Score

$$\sigma^2 (F_{vt}) = \tau_{vi} (1-\tau_{vi}/k) \qquad (6)$$

(vgl. Abb. 1), wobei die umgekehrt U-förmige Beziehung zwischen True-Score und Fehlervarianz (bzw. Standardabweichung des Meßfehlers) auch dann noch erhalten bleibt, wenn die Items im Fixed Test Paradigma nicht parallel sind, aber dem Rasch-Modell genügen und eine einigermaßen vernünftige Schwierigkeitsverteilung um einen (absolut oder relativ) «mittleren» Schwierigkeitsgrad aufweisen (vgl. Abb. 2). Der umgekehrt U-förmige Zusammenhang zwischen True-Score und Fehlervarianz verschwindet lediglich dann, wenn der Test ausschließlich aus sehr leichten und sehr schwierigen Items besteht, aber keine Items von mittlerer Schwierigkeitsstufe umfaßt (vgl. Abb. 3).

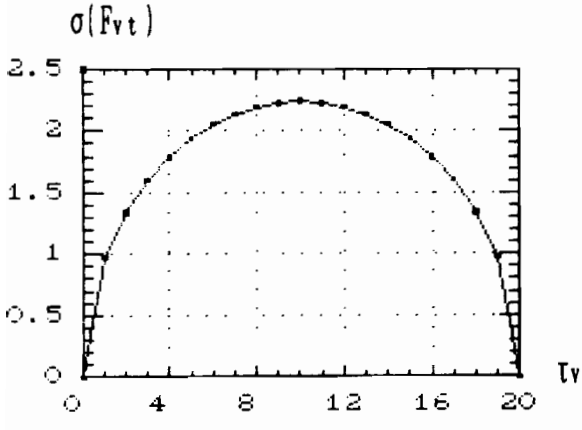

Abbildung 1: Zusammenhang zwischen Standardabweichung des Meßfehlers und True-Score bei einem Test aus k = 20 Items, die alle die selbe Itemschwierigkeit aufweisen (Binomialmodell).

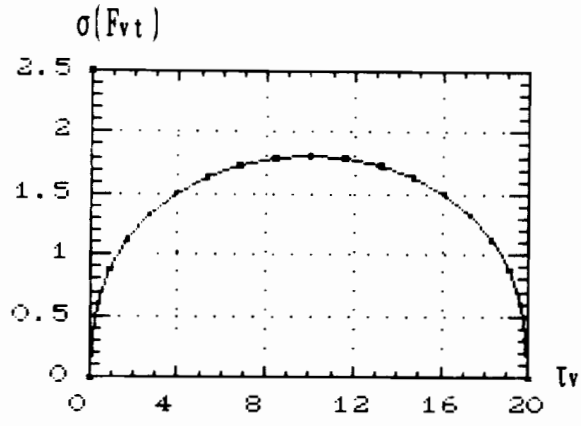

Abbildung 2: Zusammenhang zwischen Standardabweichung des Meßfehlers und True-Score bei einem Test aus k = 20 Items mit standardnormalverteilten Itemschwierigkeiten $-3.29 \leq \sigma i \leq +3.29$ (Rasch-Modell).

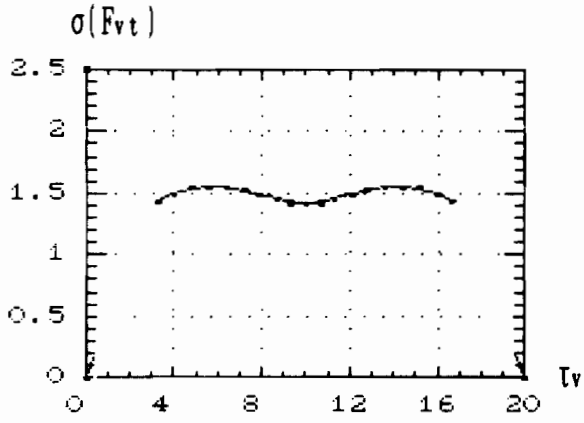

Abbildung 3: Zusammenhang zwischen Standardabweichung des Meßfehlers und True-Score bei einem Test aus k = 20 Items, der keine Items von mittlerem Schwierigkeitsgrad enthält $1.35 \leq |\sigma i| \leq 3.29$ (Rasch-Modell).

Dies hat zur Folge, daß die Klassische Testtheorie die Konfidenzgrenzen des True-Scores nicht einmal bei Geltung des Binomialmodells korrekt wiedergibt. Die Breite des Vertrauensbereichs wird von der Klassischen Testtheorie im mittleren Scorebereich unterschätzt und für großes oder keines τ überschätzt. Für großes und

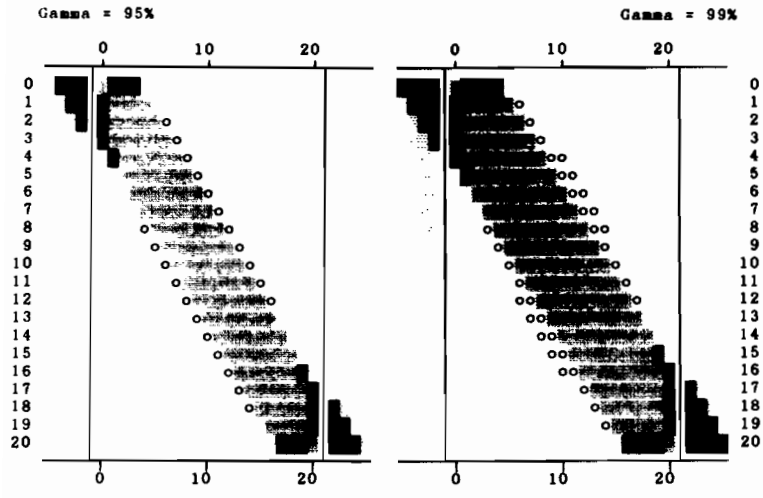

Abbildung 4: Reproduktion der Konfidenzgrenzen der Binomialverteilung durch die Klassische Testtheorie.
▒ ... von Klassischer Testtheorie korrekt wiedergegebene Werte des Konfidenzintervalls
○ ... im Konfidenzintervall der Klassischen Testtheorie fälschlich nicht enthaltene Werte
■ ... im Konfidenzintervall der Klassischen Testtheorie fälschlich enthaltene Werte

kleines τ nehmen die Konfidenzgrenzen zudem unsinnige Werte an, die außerhalb des Wertebereichs der Scorevariablen liegen (vgl. Abb. 4). Wie schlecht die Ergebnisse ausfallen hängt zudem von der Homogenität der Eichstichprobe ab, aus welcher die Reliabilität bzw. der Standardmeßfehler des Tests bestimmt wurde. Je heterogener sie ist, desto stärker wird die Meßgenauigkeit des Tests im mittleren Scorebereich überschätzt.

Im Rasch-Modell kann ein Konfidenzintervall für den True-Score einer Vp durch Einsetzen der Konfidenzgrenzen für den Personenparameter in die Gleichung

$$\tau_{vt} = E(X_{vt}) = \sum_i \exp(\theta_v - \sigma_i)/(1 + \exp(\theta_v - \sigma_i)) \quad (7)$$

gewonnen werden. Dadurch werden die exakten Konfidenzgrenzen der Binomialverteilung zwar ebenfalls nicht perfekt, aber im Vergleich zur Klassischen Testtheorie doch mit sehr guter Näherung wiedergegeben (vgl. Abb. 5).

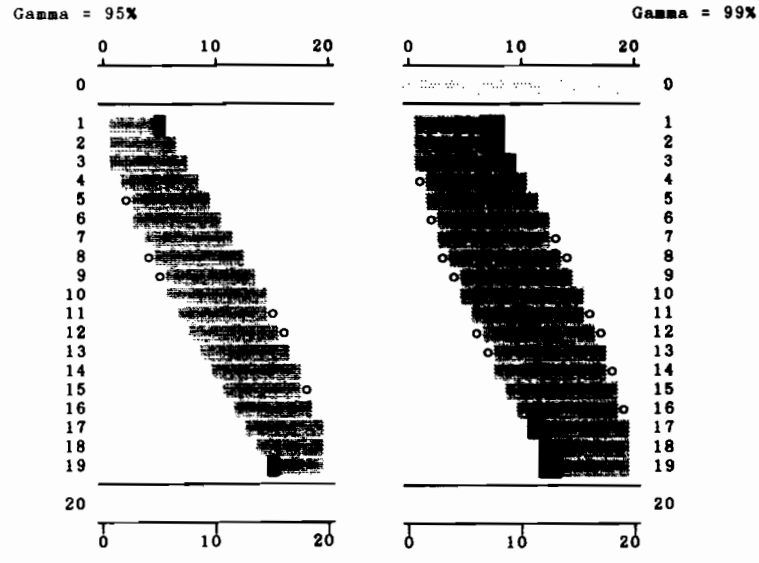

Abbildung 5: Reproduktion der Konfidenzgrenzen der Binomialverteilung durch das Rasch-Modell.
▒ ... vom Rasch-Modell korrekt wiedergegebene Werte des Konfidenzintervalls
○ ... im Konfidenzintervall des Rasch-Modells fälschlich nicht enthaltene Werte
■ ... im Konfidenzintervall des Rasch-Modells fälschlich enthaltene Werte

Entsprechend führt auch der Signifikanztest für den Unterschied zweier Testscores mittels der Prüfgröße (2) sowohl im Item Sampling Paradigma als auch im Hybrid Modell zu einer eklatanten Überbewertung der Scoredifferenzen durch die Klassische Testtheorie (vgl. Abb. 6), während das Rasch-Modell die korrekten Verhältnisse mittels der (für $k \to \infty$) asymptotisch N(0,1)-verteilten Prüfgröße

$$z = (\theta_v - \theta_w) / ([I_t(\theta_v)^{-1} + I_t(\theta_w)^{-1}])^{1/2} \quad (8)$$

schon bei relativ kleinem k ziemlich genau wiedergibt (vgl. Abb. 7). $I_t(\theta_v)$ bezeichnet dabei die

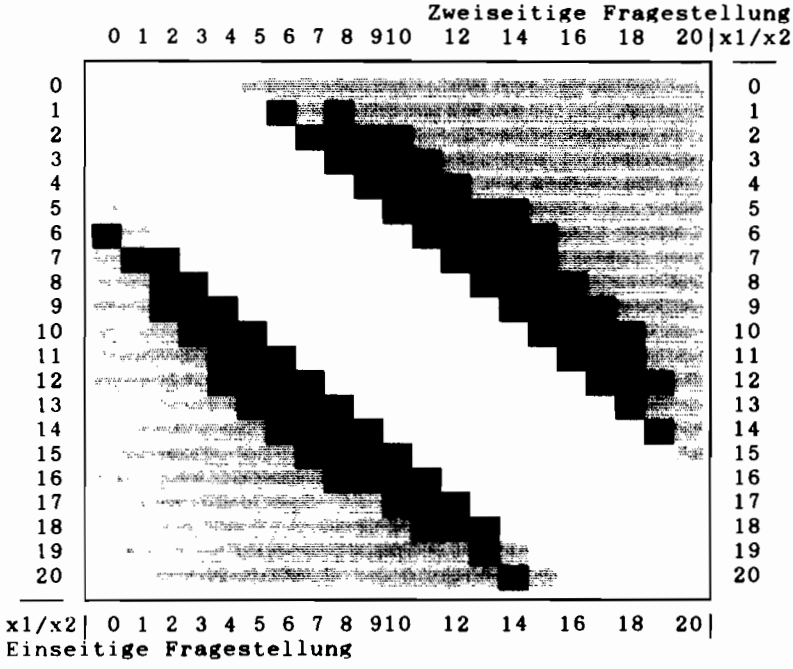

Abbildung 6: Statistische Signifikanz des Unterschiedes zweier Testscores im Binomialmodell
■ ... von Klassischer Testtheorie überbewertet

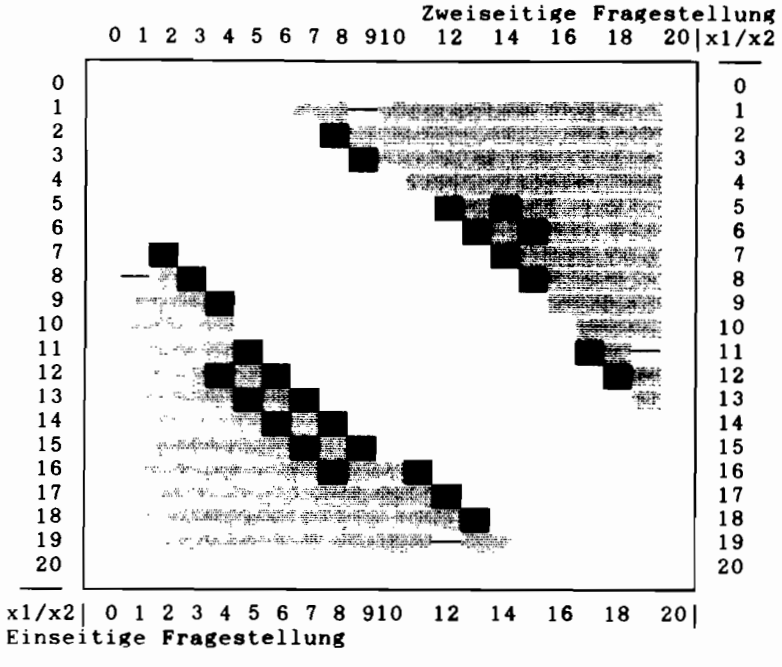

Abbildung 7: Statistische Signifikanz des Unterschiedes zweier Testscores im Binomialmodell
■ ... von Rasch-Modell überbewertet
– ... von Rasch-Modell unterbewertet

von Fischer und Scheiblechner, 1970, angegebene Informationsfunktion des Tests.

Zu einer noch eklatanteren Fehleinschätzung der Scoredifferenzen führen beide Prüfgrößen (2 *und* 8) im Fixed Test Paradigma. Da die Scorevariablen X_{vt} und X_{wt} dort aus abhängigen Itemstichproben stammen und daher positiv kovariieren, wird die Varianz der Maßzahldifferenzen

$$\sigma^2(X_{vt}-X_{wt})=\sigma^2(X_{vt})+\sigma^2(X_{wt})-2\sigma(X_{vt},X_{wt}) \quad (9)$$

durch die in (2) und (8) enthaltene Annahme, wonach $\sigma^2(X_{vt}-X_{wt}) = \sigma^2(F_{vt}) + \sigma^2(F_{wt})$, kraß überschätzt.

Dies hat zur Folge, daß die statistische Signifikanz von Scoreunterschieden ausgerechnet im Fixed Test Paradigma, das ja mit der üblichen Testpraxis übereinstimmt, eklatant überbewertet wird (vgl. Tab. 1).
Ein angemessenes Prüfverfahren für Scoreunterschiede stellt im Fixed Test Paradigma der sogenannte Test von McNemar (McNemar, 1947; vgl. auch Bortz, Lienert & Boehnke, 1990, S. 160) dar, dessen Anwendung zudem keinerlei Modellannahmen voraussetzt. Geltung des Rasch-Modells ist zwar notwendig und hinreichend für die spezifische Objektivität (Rasch, 1968) des Tests von McNemar. Die Anwendbarkeit des Tests ist jedoch nicht daran gebunden.

Tabelle 1: Im Test von McNemar und in der Klassischen Testtheorie für statistische Signifikanz erforderliche Scoredifferenzen am Beispiel der Form A des Subtests WA des IST–70 nach Amthauer (1973) mit k= 20 Items, einer Standardabweichung von $\sigma(X_t) = 2.89$ (siehe Amthauer, 1973, S. 32) und einer Testwiederholungsreliabilität von $r_{tt} = 0.63$ (siehe Amthauer, 1973, S. 28).

Anzahl der von genau einer der beiden Vpn gelösten Items	Test von McNemar			
	einseitige Fragestellung		zweiseitige Fragestellung	
	α = 5%	α = 1%	α = 5%	α = 1%
5	5	-	-	-
6	6	-	6	-
7	7	7	7	-
8	6	8	8	8
9	7	9	7	9
10	8	10	8	10
11	7	9	9	11
12	8	10	8	10
13	7	11	9	11
14	8	10	10	12
15	9	11	9	11
16	8	12	10	12
17	9	11	9	13
18	8	12	10	12
19	9	11	11	13
20	10	12	10	14

	Klassische Testtheorie			
	einseitige Fragestellung		zweiseitige Fragestellung	
	α = 5%	α = 1%	α = 5%	α = 1%
	5	6	5	7

Zusammenfassung

Zusammenfassend läßt sich sagen, daß der Eindruck von der universellen Anwendbarkeit der Klassischen Testtheorie auf einer krassen Fehleinschätzung beruht. Tatsächlich gibt es nur einen Spezialfall, in dem die Klassische Testtheorie – zumindest für das Konfidenzintervall des True-Scores – zuverläßliche Ergebnisse liefert. Dieser Spezialfall ist aber insofern von Interesse, als er immer dann eintritt, wenn alle Lösungswahrscheinlichkeiten entweder sehr klein (nahe Null) oder sehr groß (nahe Eins) sind. Dies entspricht der – für viele Tests durchaus angemessenen – psychologischen Modellvorstellung, daß jede Testaufgabe von einer Person grundsätzlich beherrscht wird oder nicht, daß bei der Bearbeitung der Testaufgaben jedoch (mit einer gewissen geringen Wahrscheinlichkeit) Zufallsfehler und Zufallserfolge eintreten können.

Indem eine Schwierigkeitsordnung der Aufgaben dabei nicht vorausgesetzt werden muß (so daß verschiedene Vpn ganz verschiedene Aufgaben beherrschen mögen – z. B. weil sie ganz verschiedene Bildungsvoraussetzungen mitbringen), handelt es sich hier zudem um einen Spezialfall, der im Rahmen des Rasch-Modells nicht behandelt werden kann.

Die Frage nach der Anwendbarkeit der Klassischen Testtheorie und/oder des Rasch-Modells erweist sich derart nicht als ein Kriterium dafür, welches denn «die bessere Testtheorie» sei, sondern sie erweist sich als die Frage nach den konkreten psychologischen Merkmalen eines Tests, auf welchen testtheoretische Modelle Anwendung finden sollen.

Literatur

Amthauer, R. (1973). *Intelligenz-Struktur-Test, IST–70. Handanweisung für die Durchführung und Auswertung.* Göttingen: Hogrefe.

Bortz, J., Lienert, G. A., Boehnke, K. (1990). *Verteilungsfreie Methoden in der Biostatistik.* Berlin: Springer.

Brickenkamp, R. (1975). *Handbuch psychologischer und pädagogischer Tests.* Göttingen: Hogrefe.

Brickenkamp, R. (1983). *Erster Ergänzungsband zum Handbuch psychologischer und pädagogischer Tests.* Göttingen: Hogrefe.

Fischer, G. H., Scheiblechner, H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch. *Psychologische Beiträge, 12.*

Kempf, W. (1991). Angewandte psychologische Testtheorie: Beiträge zur Kritik der Leistungsfähigkeit der Klassischen Testtheorie für die Praxis. In D. Frey, (Hrsg.), *Bericht über den 37. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1990,* Band 2. Göttingen: Hogrefe.

Lienert, G. A. (1969). *Testaufbau und Testanalyse.* Weinheim: Beltz.

Lord, F., Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, Massachusetts: Addison-Wesley.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated propertions of percentages. *Psychometrika, 12,* 153–157.

Meder, G., Kempf, W., Boneberg, I. (1990). Eine empirische Untersuchung der Restriktivität der klassischen Testtheorie anhand dreier Subtests des IST–70. In D. Frey (Hrsg.), *Bericht über den 37. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1990,* Band 1. Göttingen: Hogrefe.

Rasch, G. (1968). *A mathematical theory of objectivity and its consequences for model construction.* Proceedings of the European Meeting on Statistic, Econometrics and Management Science, Amsterdam, 2–7 september 1968.

Schulze, G. (1990). *Alltagsästhetik, Milieustruktur und Erlebnismarkt. Eine kultursoziologische Untersuchung der Bundesrepublik Deutschland.* Projektbericht für die Deutsche Forschungsgemeinschaft. Anhang II: Meßtheoretische Überlegungen und Skalen. Bamberg.

Zuma (1988) – Handbuch Sozialwissenschaftlicher Skalen. Bonn: IZ.

Anschrift: Prof. Dr. W. Kempf-Palmbach, Universität Konstanz, Fakultät für Wirtschaftswissenschaften und Statistik, Postfach 5560, 7750 Konstanz 1