*Ersch. zuerst in: Bericht über den 37. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1990. Bd. 1. Göttingen: Hogrefe, 1990, S. 506-507*

# ZUR LEISTUNGSFÄHIGKEIT VON RASCH MODELL UND KLASSISCHER TESTTHEORIE FÜR DIE INFERENZSTATISTISCHE EVALUATION VON TESTSCORES

*Kempf, Wilhelm (Konstanz)*

Entgegen der verbreiteten Auffassung von der universellen Anwendbarkeit der klassischen Testtheorie (KT) ist der von ihr angegebene Vertrauensbereich des True-Scores

(1) $KONF_{\gamma}$ $\{xvt - zkrit\sigma(F.t) \leq \tau vt \leq xvt + zkrit\sigma(F.t)\}$

an drei Voraussetzungen geknüpft: (A1) an jene empirischen Zusatzannahmen, welche für die Reliabilitätsschätzung erforderlich sind, aus der sich der Standardmeßfehler errechnet, (A2) an die näherungsweise Übereinstimmung des Standardmeßfehlers $\sigma(F.t)$ mit der tatsächlichen Standardabweichung des Meßfehlers der Person $\sigma(Fvt)$ und (A3) an die Normalverteilung des Meßfehlers.

Untersucht man, wie der Testscore Xvt aus den einzelnen Antworten Avi einer Vp auf die k Items eines Tests zusammengesetzt ist, so zeigt sich, daß bei binären Items und unter der verbreiteten Scoreformel $Xvt = \Sigma i\ Avi$ , nur die letzte dieser Voraussetzungen als einigermaßen unproblematisch gelten kann. Wegen des zentralen Grenzwertsatzes kann Verteilung des Meßfehlers bei großem k und mittlerem $\tau vt$ durch die $N[O, \Sigma i\ pvi(1-pvi)]$ approximiert werden. Annahme (A1) ist dagegen hoch restriktiv und stellt strenge Anforderungen an die Testkonstruktion, während Voraussetzung (A2) wegen der Abhängigkeit der Fehlervarianz $\sigma^2(Fvt) = \Sigma i\ pvi(1-pvi)$ vom True Score $\tau vt$ jedenfalls verletzt ist. Folglich kann die KT die Konfidenzgrenzen des True-Scores nicht einmal im Item-Sampling-Paradigma (ISP) korrekt wiedergeben. Die Breite des Vertrauensbereiches wird von der KT im mittleren Scorebereich unterschätzt und für großes oder kleines $\tau vt$ überschätzt. Hier nehmen die Konfidenzgrenzen zudem unsinnige Werte an, welche außerhalb des Wertebereichs der Scorevariable liegen. Wie schlecht die Ergebnisse ausfallen, hängt von der Homogenität der Eichstichprobe ab, aus welcher die Reliabilität bestimmt wurde: je heterogener sie ist, desto stärker wird $\sigma(Fvt)$ im mittleren Scorebereich unterschätzt. Bei Anwendung des Rasch Modells (RM) ergibt sich der Vertrauensbereich des True Scores durch Einsetzen der Konfidenzgrenzen des Personenparameters in $E(Xvt) = \Sigma i\ fi(\zeta)$. Dabei werden die Konfidenzgrenzen im ISP durch das RM (mit $\sigma i = 1$) ebenfalls nicht perfekt, aber doch mit sehr guter Näherung rekonstruiert.

Entsprechend führt auch der Signifikanztest für den Unterschied zweier Testscores mittels der Prüfgröße $z = (Xvt-Xwt)/[\sigma(F.t)\sqrt{2}]$ bereits im ISP zu einer eklatanten Uberbewertung der Scoredifferenzen durch die KT, während die Prüfgröße

*Konstanzer Online-Publikations-System (KOPS)*
*URL: http://www.ub.uni-konstanz.de/kops/volltexte/2008/6399/*
*URN: http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-63990*

$z = (\zeta v - \zeta w)/\sqrt{[I_t(\zeta v)^{-1} + I(\zeta w)^{-1}]}$ des RM die korrekten Verhältnisse deutlich besser wiederzugegen vermag.

Indem sie von unabhängigen Testscores und damit von unabhängigen Itemstichproben ausgehen, sind beide Methoden nur im ISP anwendbar und bewirken im Fixed-Test-Paradigma (FTP) eine noch krassere Überbewertung der Scoredifferenzen. Ein angemessenes Prüfverfahren stellt im FTP der Test von McNemar dar, dessen Anwendung zudem an keine Modellannahmen gebunden ist. Geltung des RM ist allerdings notwendig und hinreichend für die spezifische Objektivität des Tests, so daß die Verteilung der Prüfgröße ausschließlich vom Unterschied zwischen den Vpn abhängt.

Kempf, W.: Ein pragmatischer Ansatz in der statistischen Theorie psychologischer Testscores. Universität Konstanz (im Druck).