Zuerst erschienen in: Bericht über den 37. Kongress der Deutschen Gesellschaft für Psychologie in Kiel 1990 (1991), 2, S. 584-587

ANGEWANDTE PSYCHOLOGISCHE TESTTHEORIE: BEITRÄGE ZUR KRITIK DER LEISTUNGSFÄHIGKEIT DER KLASSISCHEN TESTTHEORIE FÜR DIE PRAXIS

Wilhelm Kempf (Konstanz)

Trotz überlegener Theorie konnte sich das Rasch-Modell (RM) in der Testpraxis bisher nicht gegen die Klassische Testtheorie (KT) durchsetzen. Gründe dafür sind u.a. der verbreitete Glaube an die universelle Anwendbarkeit der KT, sowie die Tatsache, daß die KT 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 auf statistische Signifikanz hin bewertet werden kann.

Wenn man von der universellen Anwendbarkeit der KT spricht, so hat man dabei aber nur die Modellstruktur der KT 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 KT geht, werden jedoch Zusatzannahmen (wie z.B. die Parallelität von Tests oder Testteilen) erforderlich, die ähnlich wie die Modellannahmen des RM einen *empirischen* Gehalt haben und daher auch grundsätzlich einer empirischen Überprüfung zugänglich sind.

So muß für parallele Tests t und j in jeder Personenpopulation

(1) $E(X_{\cdot t}) = E(X_{\cdot j})$

(2) $\sigma^2(X_{\cdot t}) = \sigma^2(X_{\cdot j})$

(3) $\rho(X_{\cdot t},Y) = \rho(X_{\cdot j},Y)$ für jede beliebige Variable Y

erfüllt sein. Für τ-äquivalente Tests ist nur die erste dieser Bedingungen gefordert (so daß $E(X_{\cdot t}-X_{\cdot j}) = 0$), während für essentiell τ-äquivalente Tests nur noch gefordert ist, daß

(4) $E(X_{\cdot t}-X_{\cdot j}) = c_{tj}$

eine von der Auswahl der Vpn unabhängige Konstante ist.

Wie ein Blick in die Handbücher der gängigen Tests zeigt, wird eine empirische Prüfung der Zusatzannahmen der KT jedoch kaum je vorgenommen. Obwohl divergierende Ergebnisse der verschiedenen Methoden der Reliabilitätsbestimmung darauf hinweisen, daß die Anwendungsvoraussetzungen für zumindest einige dieser Methoden verletzt sind, wird dem zumeist kein weiteres Augenmerk geschenkt.

Wie Meder, Kempf & Boneberg anhand der beiden "Parallelformen" (A und B) der Subtests WA, ZR und WÜ des IST-70 aufzeigten, sind die Zusatzannahmen der KT zwar ähnlich restriktiv wie die Modellannahmen des RM. Die Beziehungen (1-4) liefern jedoch nur wenig trennscharfe "Modelltests" dafür: Obwohl signifikante Abweichungen zwischen Paralleltest- und Split-Half-Reliabilität (Spearman-Brown bzw. Cronbach-α) auf mangelnde Parallelität von Form A und Form B der Subtests ZR und WÜ schließen ließen, konnten Abweichungen von der Parallelität der Testformen bei einem Stichprobenumfang von n=108 Vpn weder durch Mittelwertsvergleich noch durch Varianzvergleich, noch durch Vergleich der Korrelationen (r_{XY}) mit der durchschnittlichen Abiturnote der Vpn diagnostiziert werden.

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

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

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

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

ist an die zusätzlichen Voraussetzungen (a) der Normalverteilung des Meßfehlers und (b) der Übereinstimmung des Standardmeßfehlers $\sigma(F_{\cdot t})$ mit der tatsächlichen Standardabweichung des Meßfehlers der Person $\sigma(F_{vt})$ gebunden.

Kempf hat die Realitätshaltigkeit dieser Voraussetzungen für den Summenscore $X_{vt} = \Sigma_i A_{vi}$ von Tests aus binären Items ($A_{vi}=1$ falls "gelöst", 0 sonst) untersucht und gezeigt, 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.

Um die Realitätshaltigkeit der Voraussetzung (b) zu beurteilen, wurde das Zustandekommen der Testscores genauer untersucht und zwischen 2 Paradigmen des psychologischen Testens unterschieden, die beide von der Existenz eines endlich oder unendlich großen Pools von Testaufgaben ausgehen.

Im *Item-Sampling-Paradigma* (ISP) 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* (FTP) dagegen, wird allen Vpn die selbe 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 $p_{vi}=p_v$. ergeben sich im FTP allenfalls als *empirische* Zusatzannahme der *Itemparallelität* im Sinne der KT.

Eine Verbindung zwischen den beiden Paradigmen stellt das *Hybrid-Modell* (HM) dar, in dem (wie im ISP) für jede Testung einer Vp eine eigene Itemstichprobe gezogen wird, die Zufallsvariation der Testleistung jedoch nicht nur auf der Zufälligkeit der Aufgabenauswahl (wie im ISP) sondern auch auf der Zufälligkeit des Bearbeitungserfolges beruht (wie im FTP). Die Lösungswahrscheinlichkeiten $p_{vi}=p_v$. sind dann gleich der mittleren Lösungswahrscheinlichkeit der Items des Pools für die Vp.

Sind die Bedingungen des ISP oder des HM erfüllt, oder sind die Testitems im FTM untereinander alle parallel, so folgt der Testcore einer Vp einer Binomialverteilung mit dem Parameter p_v. ("Binomialmodell", BM). Eine eigene statistische Theorie psychologischer Testscores ist dann nicht vonnöten. Es können die bekannten Verfahren auf Grundlage der Binomialverteilung angewendet werden.

Einer eigenen 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 RM die Lösungswahrscheinlichkeiten als Funktion eines Personen(Fähigkeits)Parameters ζ_v und eines Item(Schwierigkeits)-Parameters σ_i approximiert

$$(7) \quad p_{vi} = \exp(\zeta_v - \sigma_i)/(1 + \exp(\zeta_v - \sigma_i)) ,$$

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

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

$$(8) \quad \tau_{vt} = \Sigma_i p_{vi}$$

und Fehlervarianz

$$(9) \quad \sigma^2(F_{vt}) = \sigma^2(X_{vt}) = \Sigma_i p_{vi}(1-p_{vi})$$

nicht unabhängig voneinander sind. Im ISP, HM und FTP mit parallelen Items (d.h. bei Geltung des BM) besteht sogar eine streng funktionale Abhängigkeit der Fehlervarianz vom True-Score

$$(10) \quad \sigma^2(F_{vt}) = \tau_{vi}(1-\tau_{vi}/k) ,$$

wobei die umgekehrt U-förmige Beziehung zwischen True-Score und Fehlervarianz (bzw. Standardabweichung des Meßfehlers) auch dann noch erhalten bleibt, wenn die Items im FTP *nicht* parallel sind, *aber* dem RM genügen *und* eine einigermaßen vernünftige Schwierigkeitsverteilung aufweisen. Er verschwindet lediglich dann, wenn der Test ausschließlich aus sehr leichten und sehr schwierigen Items besteht, aber keine Items von mittlerer Schwierigkeitsstufe umfaßt.

Dies hat zur Folge, daß die KT die Konfidenzgrenzen des True-Scores nichteinmal bei Geltung des BM korrekt wiedergibt. Die Breite des Vertrauensbereichs wird von der KT im mittleren Scorebereich unterschätzt und für großes oder kleines τ überschätzt.

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

$$(11) \quad \tau_{vt} = E(X_{vt}) = \Sigma_i \exp(\zeta_v-\sigma_i)/(1+\exp(\zeta_v-\sigma_i))$$

gewonnen werden. Dadurch werden die korrekten Konfidenzgrenzen des BM zwar ebenfalls nicht perfekt aber im Vergleich zur KT doch mit sehr guter Näherung wiedergegeben.

Entsprechend führt auch der Signifikanztest für den Unterschied zweier Testscores mittels der Prüfgröße (6) sowohl im ISP als auch im HM zu einer Überbewertung der Scoredifferenzen durch die KT, während das RM die korrekten Verhältnisse mittels der asymptotisch N(0,1)-verteilten Prüfgröße

(12) $z = (\zeta_v - \zeta_w)/\sqrt{[I_t(\zeta_v)^{-1} + I_t(\zeta_w)^{-1}]}$

schon bei relativ kleinem k ziemlich genau wiedergibt.

Zu einer noch eklatanteren Fehleinschätzung der Scoredifferenzen führt die Prüfgröße (6) im FTP. Da die Scorevariablen X_{vt} und X_{wt} dort aus *abhängigen* Itemstichproben stammen und daher *positiv kovariieren*, wird die Varianz der Maßzahldifferenzen

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

durch die in (6) enthaltene Annahme, wonach $\sigma^2(X_{vt} - X_{wt}) \approx 2\sigma^2(F_{\cdot t})$ selbst dann kraß überschätzt, wenn $\sigma^2(F_{vt}) \approx \sigma^2(F_{wt}) \approx \sigma^2(F_{\cdot t})$ einigermaßen zutrifft.

Dies hat zur Folge, daß die KT die statistische Signifikanz von Scoreunterschieden ausgerechnet im FTP, das ja mit der üblichen Testpraxis übereinstimmt, eklatant überbewertet.

Ein angemessenes Prüfverfahren für Scoreunterschiede stellt im FTP der *Test von McNemar* dar, dessen Anwendung zudem keinerlei Modellannahmen voraussetzt. Geltung des RM ist zwar notwendig und hinreichend für die *spezifische Objektivität* des Tests von McNemar. Die Anwendbarkeit des Tests ist jedoch nicht daran gebunden.

Zusammenfassend läßt sich sagen, daß der Eindruck von der universellen Anwendbarkeit der KT auf einer groben Fehleinschätzung beruht. Tatsächlich gibt es kaum einen Spezialfall, in dem die KT wirklich zuverlässige Ergebnisse liefert.