Zuerst ersch. in: Bericht über den 37. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1990, Band. 1 / hrsg. von Dieter Frey. Göttingen: Hogrefe, 1990, S. 505-510

ANGEWANDTE PSYCHOLOGISCHE TESTTHEORIE: THEORETISCHE, PRAGMATISCHE UND EMPIRISCHE ASPEKTE

Organisation: Kempf, Wilhelm (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. (a) der verbreitete Glaube an die universelle Anwendbarkeit der KT, (b) die vergleichsweise größeren Datenmengen, welche für eine Rasch-Analyse erforderlich sind, sowie (c) 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 hinsichtlich ihrer statistischen Signifikanz bewertet werden kann. Der so begründete Eindruck von der praktischen Brauchbarkeit der KT ändert sich jedoch, wenn man untersucht, (a) an welche Voraussetzungen die von der KT angebotenen Problemlösungen tatsächlich gebunden sind, wenn man (b) die Restriktivität und den psychologischen Gehalt dieser Voraussetzungen denen des RM gegenüberstellt, und wenn man (c) die Qualität der von der KT angebotenen Problemlösungen mit der Qualität jener Lösungen vergleicht, welche sich im Rahmen des RM entwickeln lassen.

Eine solche kritische Gegenüberstellung von RM und KT soll durch die nachfolgenden Beiträge geleistet werden. Den Ausgangspunkt dafür bildet die Unterscheidung zweier Paradigmen des psychologischen Testens, 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 auf der Zufallsauswahl der von ihr bearbeiteten Testaufgaben. Im Fixed-Test-Paradigma (FTP) 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 in ihrer Schwierigkeit (Lösungswahrscheinlichkeit) unterscheiden können.

Sind die Bedingungen des ISP erfüllt, so folgt der Testscore einer Vp einer Binomialverteilung mit dem Parameter pv.: eine eigene statistische Theorie psychologischer Testscores ist dann nicht vonnöten. Dasselbe gilt auch im FTP, wenn die Lösungswahrscheinlichkeiten alle gleich und also die Items parallel im Sinne der KT sind.

Einer eigenen Theorie psychologischer Testscores bedürfen wir erst, wenn dies nicht der Fall ist. Die uns interessierende pragmatische Aufgabe einer solchen Theorie besteht dann darin, den Vertrauensbereich und/oder die kritische Differenz von Testscores auch dann noch bestimmen zu können. Entsprechend zielt auch die Frage nach der Modellanpassung der Daten dann nicht mehr auf einen abstrakten Homogenitätsbegriff, sondern auf die Evaluation der Verläßlichkeit ab, mit welcher dies im konkreten Fall geleistet werden kann.

Wie gezeigt wird, sind die dazu erforderlichen Zusatzannahmen der KT gleichermaßen restriktiv wie die Modellannahmen des RM. Es mangelt jedoch an trennscharfen Modelltests (Mederet al.). Auch setzt die Anwendung des RM in der Testkonstruktion keineswegs voraus, daß bereits jede Vorfassung des Tests an den üblichen großen Personen-

Konstanzer Online-Publikations-System (KOPS)
URL: http://www.ub.uni-konstanz.de/kops/volltexte/2008/6602/
URN: http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-66021

stichproben erprobt wird. Parameterfreie Verfahren (z.B. Kendall's τ) erlauben schon bei relativ kleinem n eine erste Evaluation der Modellanpassung (Röhrenbach), die (über die o.g. Problemstellung hinaus) im RM auch die Vergleichbarkeit der Testscores hinterfragt und so selbst dann wichtige (psychologische) Einsichten zu eröffnen vermag, wenn die Modellannahmen im konkreten Fall zurückgewiesen werden (Piel). Insgesamt erzielt die KT mit einem maximalen Aufwand an formalen Zusatzannahmen die schlechtesten Ergebnisse (Kempf).

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 τvt 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 τvt 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 τvt 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 Überbewertung der Scoredifferenzen durch die KT, während die Prüfgröße

$z = (\zeta v - \zeta w)/\sqrt{[I_t(\zeta v)^{-1} + I(\zeta w)^{-1}]}$ 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).

EINE EMPIRISCHE UNTERSUCHUNG DER RESTRIKTIVITÄT DER KLASSISCHEN TESTTHEORIE ANHAND DREIER SUBTESTS DES IST-70

Meder, Gerhard; Kempf, Wilhelm, & Boneberg, Iris (Konstanz)

Wenn man von der universellen Anwendbarkeit der klassischen Testtheorie (KT) spricht, so hat man dabei nur die Modellstruktur der KT im Blick. Sobald es um die Anwendung der KT (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 (RM) haben diese Zusatzannahmen einen empirischen Gehalt und sind daher auch grundsätzlich einer empirischen Überprüfung zugänglich. So müssen für parallele Tests t und j in jeder Personenpopulation die Beziehungen (1) $E(X.t) = E(X.j)$; (2) $\sigma^2(X.t) = \sigma^2(X.j)$; und (3) $\varsigma(X.t,Y) = \varsigma(X.j,Y)$ für jede beliebige Variable Y erfüllt sein. In einer Untersuchung der "Parallelformen" (A und B) der Subtests WA, ZR und WU des IST-70 an einer Stichprobe von 108 Studierenden der Uni Konstanz erwiesen sich diese Beziehungen jedoch als wenig trennscharfer "Modelltest" für die Zusatzannahmen der KT.

Obwohl signifikante Abweichungen zwischen Paralleltest- (rAB) und Split-Half-Reliabilität (Spearman-Brown bzw. Cronbach-α) auf mangelnde Parallelität von Form A und Form B der Subtests ZR und WÜ schließen lassen (Tab. 1), konnten Abweichungen von der Parallelität der Testformen weder durch Mittelwertsvergleich noch durch Varianzvergleich, noch durch Vergleich der Korrelationen (rXY) mit der durchschnittlichen Abiturnote der Vpn diagnostiziert werden. Bei Subtest WA unterschieden sich dagegen die Mittelwerte der beiden Testformen signifikant voneinander, während die Unterschiede zwischen den insgesamt sehr instabilen Reliabilitätsmaßen statistisch ebensowenig bedeutsam sind wie der Varianzunterschied und der Unterschied zwischen den Kriteriumskorrelationen der beiden Testformen.

Tab. 1	r_{ab}	SP-B	CR-α	m(X)	$s^2(X)$	r_{xy}
WA-A	0.306	0.112	0.112	10.926	4.791	-0.162
WA-B		0.456	0.456	13.759	5.831	0.018
ZR-A	0.581	0.935	0.935	14.593	21.760	-0.182
ZR-B		0.935	0.927	14.593	20.927	-0.228
WÜ-A	0.618	0.870	0.869	12.278	17.608	0.073
WÜ-B		0.887	0.886	12.130	15.094	0.032

Eine Überprüfung der essentiellen τ-Äquivalenz der beiden Testformen anhand der Beziehung (4) E(XA-XB) cAB führte zur Beibehaltung der HO ($t = 0.978$, $df = 106$). Die beiden Testformen des Subtests WA erweisen sich derart zwar nicht als parallel, da Form A schwieriger ist und sich die Testergebnisse in Form A und Form B um generell cAB ≈ -3 Punkte unterscheiden. Sie scheinen aber immerhin die Mindestanforderungen (essentielle τ-Äquivalenz und Homogenität der Varianzen) zu erfüllen, welche zur Berechnung der Paralleltestreliabilität vorauszusetzen sind.

Der Eindruck, wonach die beiden Testformen des Subtests WA (bei zwar geringerer Reliabilität) der KT besser genügen als die anderen beiden Subtests, relativiert sich jedoch, sobald wir die Form der Scoreverteilungen untersuchen: die Scoreverteilung des Subtests WA zeigt bei keiner der beiden Testformen eine signifikante Abweichung von der Binomialverteilung (Form A: $x^2 = 5.314$, $df = 7$; Form B: $x^2 = 8.876$, $df = 6$), welche sich ergibt, wenn der Test zwischen den Vpn überhaupt nicht differenziert, sondern alle Vpn bei allen Items dieselbe Lösungswahrscheinlichkeit besitzen.

Kempf, W., Meder, G.: Theoretische und empirische Untersuchungen zur Restriktivität von Klassischer Testtheorie und Rasch-Modell.

DIE ANWENDUNG DES STOCHASTISCHEN TESTMODELLS VON RASCH AUF DIE FREIBURGER BESCHWERDENLISTE (FBL-K)

Piel, Ewald (Tengen)

Im Rahmen eines DFG-Forschungsprojektes an jungen Erwachsenen wurde die FBL-R von der Arbeitsgruppe Hautzinger an der Universität Konstanz eingesetzt. In die Auswertung mithilfe des Rasch-Modells gingen 807 Testpersonen ein.

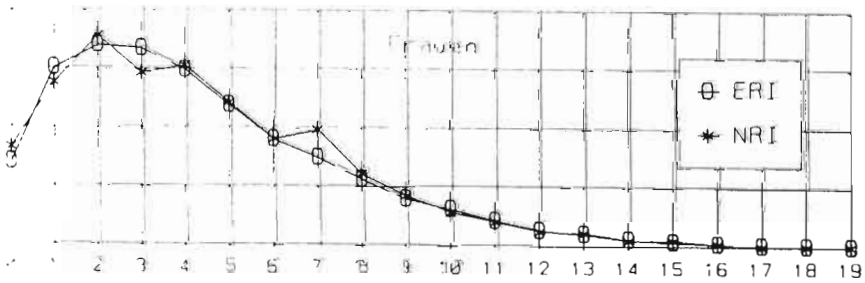
Der Likelihood-Quotienten-Test nach Score-Gruppen zeigt eine signifikante Modellabweichung der Daten an ($x^2 = 48.5$, $df = 19$, $p < .01$). Beim Vergleich der beiden Geschlechter miteinander ergibt sich sogar eine x^2Statistik von 309.3 ($df = 19$, $p < .01$). Die Modellabweichung zwischen den Geschlechtern ist so enorm, daß die Vermutung naheliegt, wesentliche Ursachen der Modellabweichungen seien entweder auf eine der Geschlechtergruppen oder auf die Relation zwischen den beiden Gruppen zurückzuführen. Tatsächlich zeigt die getrennte Auswertung der Daten aus beiden Teilstichproben für die männliche Stichprobe Modellkonformität ($x^2 = 16.7$, $df = 19$, $p = .38$), während sich für die weibliche Population eine signifikante Modellverletzung ergibt ($x^2 = 38.1$, $df = 20$, $p = .01$): Insgesamt mußten 5 Fragen ausgesondert werden, um Modellkonformität der verbleibenden Items zu erzielen ($x^2 = 12.6$, $df = 14$; $p = .53$). Die Ergebnisse weisen darauf hin,

daß die Modellabweichungen jedoch nicht infolge von Iteminhalten zustandekommen, sondern daß sie ein Produkt des Bearbeitungsablaufes sind.

Die Abbildung zeigt, daß die Modellabweichungen vor allem auf serielle Effekte zurückzuführen sind. Darin sind auf der Abszisse die Anzahl r der vorher bejahten Items und auf der Ordinate die Zeilenrandsummen (i=r,...,k) der nri-Matrix aufgetragen, welche die Häufigkeiten angibt, mit welcher ein Item i verneint wird, nachdem vorher r Items bejaht wurden. Zum Vergleich werden die Zeilenrandsummen der eri-Matrix dargestellt, welche die bei lokaler stochastischer Unabhängigkeit erwarteten nri-Werte enthält. Abweichungen zwischen empirischen und erwarteten Häufigkeiten sind Indizien für Transfer-Effekte im Sinne des dynamischen Testmodells nach Kempf. Wie die Abbildung zeigt, gibt es Abweichungen vor allem bei Frauen, die nach r=3 bejahten Fragen erhöhte Klagsamkeit zeigen (weniger Verneinungen als erwartet: Σinri < Σieri) und dies später (nach r=7 Bejahungen) durch verminderte Klagsamkeit (mehr Verneinungen als erwartet: Σinri > Σieri) gleichsam "korrigieren".

Schon von Fahrenberg selbst waren bereits in früheren Untersuchungen Mittelwertsunterschiede zwischen den Geschlechtern festgestellt worden, die als quantitative Unterschiede - etwa in der Klagsamkeit der Probanden - interpretiert werden konnten. Die hier gefundenen Ergebnisse gehen darüber hinaus und zeigen, daß Männer und Frauen hinsichtlich der FBL-K überhaupt nicht direkt verglichen werden können. Da sich die Subpopulationen nicht nur quantitativ, sondern auch qualitativ voneinander unterscheiden, bieten auch die von Baumann & Stieglitz geforderten Repräsentativnormen für Beschwerdelisten nur eine scheinbare Lösung des Problems.

Abbildung zum Beitrag Piel:



Piel, E., Hautzinger, M., Scherbarth-Roschmann, P.: Die Analyse der Freiburger Beschwerdenliste mithilfe des stochastischen Testmodells von Rasch. Diagnostika (im Druck).

EINE ITEMANALYSE DES "TURM VON LONDON"

Röhrenbach, Christine (Konstanz)

Der Turm von London ist eine Variante des Turms von Hanoi, der zur Untersuchung von Problemlöseprozessen häufig Verwendung fand. Diese von Shallice (1982) eingeführte Londoner Variante wird in neuropsychologischen Untersuchungen dazu verwendet, spe-

zifisch kognitive Defizite der Handlungsplanung bei Patienten mit Stirnhirnläsionen nachzuweisen. Von einem Ausgangspunkt soll eine Kugelanordnung dreier verschiedenfarbiger Kugeln auf drei Stäben unterschiedlicher Länge sukzessive so umgeordnet werden, daß ein vorab definiertes Zielmuster entsteht. Für den Turm von London wurden zwei, hinsichtlich ihrer Strukturmerkmale weitgehend parallel aufgebaute Aufgabenreihen (mit je 9 Aufgaben) entwickelt, die sich in dem zu erreichenden Zielmuster unterscheiden. Das Muster am Ausgangspunkt ist bei jeder Aufgabe ein anderes, und das Zielmuster, das in Form einer Abbildung vorgegeben wird, bleibt für je neun Aufgaben dasselbe. Der Proband soll, eine Kugel nach der anderen von Stab zu Stab bewegend, das Ziel erreichen und zwar mit der minimal notwendigen Zahl an Zügen, die der Versuchsleiter bei jeder Aufgabe nennt. Diese variiert zwischen drei bis sechs Zügen. Eine Itemanalyse, mittels des Linearen Logistischen Testmodells, wurde in der Absicht durchgeführt: 1. Die Schwierigkeiten der Aufgaben aus deren Merkmalstruktur zu erklären. 2. Die Homogenität dieser Aufgaben zu prüfen. Aufgrund der geringen Stichprobengröße (n=83 hirngeschädigte Patienten) wurden, anstatt der sonst üblichen Conditional-Maximum-Likelihood-Verfahren, die Strukturparameter mittels Least-Square-Schätzungen auf Grundlage von bedingten Paarvergleichen der Aufgaben berechnet.

Folgende Merkmale definierten die Struktur einer Aufgabe:
- Sind bei der Zielkonfiguration alle Kugeln auf einem Stab?
- Ist der Lösungsweg (z.T.) aus vorausgegangenen Aufgaben bekannt?
- Gibt es mehrere Wege vom Ausgangspunkt zum Ziel?
- Wie lange ist der Lösungsweg: länger als 3/4/5 Züge?
- Wieviele Zugalternativen bestehen am Ausgangspunkt: mehr als 2/3?

Ergebnisse: 1. In einer stufenweisen multiplen Regression der Strukturmerkmalsdifferenzen zwischen Aufgabenpaaren auf die relativen Lösungshäufigkeiten konnte ein Strukturmodell gefunden werden, das zu 81 % die Schwierigkeiten der Aufgaben erklärt. Dabei ließen sich die Gewichte feststellen, mit denen die einzelnen Merkmale zur Schwierigkeit einer Aufgabe beitrugen. In einer ordinalen Modellkontrolle wurde eine Schwierigkeits-Rangordnung der Aufgaben durch Addition der geschätzten Gewichte ihrer Merkmale berechnet und diese mit der Schwierigkeits-Rangordnung aus den tatsächlich erhobenen Lösungshäufigkeiten korreliert. Der Korrelationskoeffizient betrug r = .96.

2. Die Patienten-Stichprobe wurde nach verschiedenen Kriterien in Teilstichproben geteilt, um zu überprüfen, inwieweit die Schwierigkeits-Rangordnungen für die Teilstichproben übereinstimmen. Die Korrelationskoeffizienten lagen zwischen r = .85 bis .95.

Fazit: Mit dem LLTM lassen sich die Merkmale bestimmen, die die Schwierigkeiten der Turm-von-London-Aufgaben konstituieren. Ein Strukturmodell, das 81% der Aufgabenschwierigkeit erklärt, konnte gefunden werden. Die aufgrund dieses Modells geschätzte Schwierigkeits-Randordnung der Aufgaben zeigte hohe Übereinstimmung mit der "empirischen" Rangordnung der Aufgabenschwierigkeiten (r = .96). Die hohen Korrelationen der Schwierigkeits-Rangordnung verschiedener Teilstichproben läßt vermuten, daß die Items in den unterschiedlichen Subpopulationen dasselbe messen, d.h. in diesem Sinne homogen sind.

Röhrenbach,C. (1989). Der Turm von London als Verfahren zur Prüfung praefrontaler Funktionen. Diplomarbeit an der Universität Konstanz.