Psychologische Beiträge, Band 18, 1976, S. 431–442

Analyse dreistufig zu beantwortender Fragebogenitems

von Dieter Heller, Bayreuth, und Hans-Peter Krüger, Nürnberg

Zusammenfassung, Summary, Résumé

In Fragebögen tauchen häufig ternäre Antwortmöglichkeiten auf (z.B. ja – nein – weiß nicht). Für eine Analyse solcher Items werden Kennwerte vorgeschlagen. Dabei wird ein der Schwierigkeit analoger Popularitätsindex eingeführt, der in p_+ den Anteil der Ja-Antworten angibt. Aus p_+ und p_-, dem Anteil der Nein-Antworten, wird ein Aktualitätsindex definiert, der in den Grenzen zwischen 0 und 1 variiert und die Eliminierung wenig aktueller Fragen gestattet. Zur Bestimmung der Trennschärfe werden Punkt-Vierfelder- oder phi-Koeffizienten angegeben, die aus der Zerlegung des χ^2 einer 3x2-Tafel in eine 2x2-Tafel und in eine auf die Unbestimmtheitsantwort zurückgehende Restkomponente entstehen. Ebenso werden Koeffizienten für Itemkorrelationen aus 3x3-Tafeln und Indizes für die Reliabilität angegeben.

Analysis of tri-fold questionnaire items

In this paper the question is raised how to proceed in item-analysis if the answers may be given in three ways: "yes", "no" or "don't know". In defining and calculating parameters the following suggestions are made: define the difficulty index of an item as the proportion of the number of "yes" answers p_+ within a random sample of N Ss. Define an index of actuality as a function of p_+ p_- where p_- is the proportion of the number of "no" answers such that min. $f(p_+p_-) = 0$ and max. $f(p_+p_-) = 1$ and eliminate items having low actuality indices. Define item-discriminating power (item-test correlation) by a point four-fold or phi coefficient determined by partitioning the chi square of a 3x2 contingency table into 2x2 contingency plus a residual due to "don't know" answers. In a similar way define item-intercorrelations on the basis of 3x3 contingency tables and reliability indices. *(L. Canders)*

Analyse d'éléments de questionnaire à réponse triple

Les questionnaires comportent fréquemment des questions à réponse triple (p. ex. oui – non – sans réponse). On propose des valeurs permettant d'analyser semblables éléments de questionnaire. On institue à cet effet un index de popularité analogue à l'index de difficulté, et fournissant en p_+ la proportion des réponses positives. A partir de p_+ et de p_-, proportion des réponses négatives, on définit un index d'actualité variant entre 0 et 1 et permettant l'élimination des questions inactuelles. Pour définir le pouvoir séparateur, on indique des coefficients φ ou à

quatre points résultant de la division du χ^2 d'une table de 3x2 en une table de 2x2 et en un reste résultant de la réponse incertaine. On indique en outre des coefficients d'intercorrélations d'items tirés de tables de 3x3, ainsi que des indices de réliabilité.

(J. Chanel)

1. Testaufgaben

Die auf der klassischen Testtheorie fußende Itemanalyse von Leistungstest unterscheidet im Regelfall nur zwischen richtig und nicht-richtig (falsch oder unbeantwortet) bei Testaufgaben. Sie tut dies unter der Annahme, daß ausgelassene Aufgaben wie falsch beantwortete zu bewerten sind, vielleicht mit dem nur psychologisch zu wertenden Unterschied, daß eine Auslassung eine weniger kritische Art der Unfähigkeit, eine Aufgabe zu beantworten ist als eine Falschantwort. Eine vierte Kategorie nicht richtig beantworteter Aufgaben wären die nicht in Angriff genommenen Aufgaben; diese werden jedoch im Rahmen einer Aufgabenanalyse deshalb nicht als eigenständige Antwortkategorie bewertet, weil die Itemanalyse unter Power-Bedingungen so erfolgt, daß alle Aufaben von allen Pbn der Analysenstichprobe in der praktisch unlimitierten Testzeit beantwortet bzw. in Angriff genommen werden können (vgl. LIENERT 1969, S. 75 ff.).

In praxi gilt für Aufgaben eines Leistungstests, daß man nur zwischen richtig und nicht-richtig beantworteten Aufgaben unterscheidet.

2. Fragebogenitems

Betrachtet man nun statt Aufgaben in Leistungstest Items von Fragebögen, so findet man dort neben Beantwortung in Schlüsselrichtung – ja bei positiv-, nein bei negativ formulierten Fragen – und Nichtbeantwortung in Schlüsselrichtung oft eine dritte Antwortmöglichkeit (z.B. weiß nicht, vielleicht, sowohl als auch) als Unbestimmtheitsantwort vor. Hier stellt sich die Frage: kann man – wie bei Leistungstests – die Unbestimmtheitsantwort auf eine Frage wie die Nichtinangriffnahme einer Aufgabe als Nichtbeantwortung in Schlüsselrichtung ansehen? Offenbar könnte die Unbestimmtheitskategorie entweder (a) als Stufe einer ordinalen Beantwortungsskala aufgefaßt werden (wie bei (1) niemals (2) selten (3) oft) oder (b) als Kategorie einer hybriden (nominal-ordinal) Beantwortungsskala mit zwei ordinalen Skalenpunkten („ja" und „nein") und einem nominalen Skalenpunkt („trifft nicht zu").

keine Antwort auf die Frage, wie man bei Fragebogen (P- und E-Skalen) mit dreistufigen Antworten eine Itemanalyse durchführt, also Schwierigkeits- und Trennschärfe-Indizes sowie Iteminterkorrelationen berechnet.

Im folgenden soll versucht werden, eine solche Analyse zu finden. Bevor jedoch im Einzelfall mit der Itemuntersuchung begonnen wird, ist zu fragen:

3. Sind die Unbestimmtheitsantworten von Pbn- und Itemparametern abhängig?

Der Untersucher sollte sich zuerst vergewissern, ob die „weiß nicht"-Antwort (1) abhängig ist von den Items und (2) von den Pbn. Zur Demonstration einer solchen Abhängigkeitsanalyse verwenden wir Daten aus dem Kinder-Angst-Test von THURNER (1969), der an einer 7. Volksschulklasse (N = 43 Schüler) mit den Antwortmöglichkeiten „ja – nein – weiß nicht" durchgeführt wurde. Zur Vereinfachung der Darstellung wurde aus der Rangreihe der Schüler nach ihren Gesamttestscores nur jeder 3. Schüler genommen und aus den nach Häufigkeit der Ja-Antworten geordneten n = 19 Items neun nach Zufall selegiert.

Die entstehende 15x9-Matrix wurde so umsortiert, daß in den Zeilen steigende Gesamtwerte, in den Spalten steigende Ja-Antworten zu finden sind. Bei Gleichheit von Ja-Antworten gab die höhere Zahl der „Nein" den Ausschlag. So entstand Tabelle 1.

Wir haben nun diese Matrix nach Zeilen und Spalten wie in Tab. 1 trichotomiert und in den neun Untermatrizen die Häufigkeit der „weiß nicht"-Antworten ausgezählt. So entstand Tab. 2, wobei das untere Tertil $Ä_-$ Schüler mit niederen Angstwerten, $Ä_0$ Schüler mit mittleren und $Ä_+$ mit hohen enthält; analog sind I_- unpopuläre Items und I_+ populäre Antworten.

Das Ergebnis ist deutlich: mit zunehmenden Angstwerten der Pbn und zunehmender Popularität der Items steigen die Häufigkeiten der „weiß nicht"-Urteile, was psychologisch sehr sinnvoll als ein Ausweichen oder als „Abwehr" interpretiert werden kann.

Zur Beurteilung der Abhängigkeit kann ein χ^2-Test auf Diagonalsymmetrie (BOWKER 1948; siehe dazu auch LIENERT 1973, 200 f.) berechnet werden, in den die Differenzen der symmetrisch um die Nebendiagonale der Matrix in Tab. 2 gelagerten Zellenhäufigkeiten eingehen.

Bei einem solch eindeutigen Ergebnis ist nichts dagegen einzuwenden, die „weiß nicht"-Antworten (zumindest als halben Punkt) in Schlüsselrichtung zu bewerten und die Itemanalyse wie bisher durchzuführen. Bei größeren Datenmatrizen kann das vorgeschlagene Verfahren

Tabelle 1

ITEMS nach Popularität geordnet

Pbn (Namen) nach Angstscores geordnet	3	9	10	14	6	4	12	7	15	
A	-	-	-	-	-	-	-	-	-	0
B	-	-	-	-	-	-	-	0	-	0
C	-	-	-	-	-	-	-	0	-	0
D	-	-	-	-	-	-	-	-	+	1
E	-	-	-	-	-	-	-	-	+	1
F	-	-	-	-	-	-	-	+	0	1
G	-	-	-	-	-	-	0	-	+	1
H	-	-	+	-	-	-	-	+	-	2
I	-	-	-	-	-	-	+	0	+	2
K	-	-	-	0	+	0	+	-	-	2
L	-	-	-	-	0	+	+	0	0	2
M	-	-	+	-	-	0	0	+	-	2
N	-	+	-	+	-	-	-	+	+	4
O	0	0	0	+	0	+	0	+	+	4
P	-	+	-	-	+	+	+	+	+	6
	0	2	2	2	2	2	4	5	7	28
	unteres			mittleres			unteres			

Test-Score-Tertil: unteres / mittleres / oberes

Popularitätstertil

Tabelle 2

Gesamttestwert	Items I_-	I_0	I_+	
$Ä_-$	0	0	2	2
$Ä_0$	0	2	3	5
$Ä_+$	3	3	4	10
	3	5	9	17

$$\chi^2 = \frac{(0-3)^2}{0+3} + \frac{(0-4)^2}{0+4} + \frac{(0-3)^2}{0+3} =$$

$$= 3 + 4 + 3 = 10$$

mit $3(3-1)/2 = 3$ FG

auch auf der Ebene der einzelnen Items durchgeführt werden. Eine solche Absicherung der „weiß nicht"-Antwort vor der Aufgabenanalyse vermag später die berechneten Kennwerte in das richtige Licht zu stellen. Darauf soll jedoch hier verzichtet werden. Im folgenden werden wir eine Itemanalyse unter Berücksichtigung der Unbestimmtheitsantwort[1] vorschlagen.

4. Schwierigkeits- bzw. Popularitätsindex

Im Unterschied zu Leistungsaufgaben sollten Persönlichkeits- oder Einstellungsfragen nicht als leicht oder schwierig, sondern – wie sich mehr und mehr im Fachjargon einbürgert – als „populär" (leicht, häufig in Schlüsselrichtung beantwortet) und „unpopulär" (schwer, selten in Schlüsselrichtung beantwortet) bezeichnet werden. Entsprechend dem Schwierigkeitsindex p ist der Popularitätsindex einer Frage durch den Anteil der in Schlüsselrichtung gegebenen Antworten definiert, wenn es sich um binäre bzw. ja-nein-Antworten handelt (wobei in der Regel je zur Hälfte ja- und nein-Antworten die Schlüsselrichtung vertreten, um Response- bzw. Antwortstereotype auszubalancieren).

Wie definiert man nun die Popularität einer Frage mit 3 Antwortmöglichkeiten? Handelt es sich um eine ordinale Antwortskala, dann könnte eine Skalierung – etwa via T-Transformation (vgl. LIENERT 1962) – erfolgen und der Skalenmittelpunkt als Popularitätsindex p definiert werden.

Diese Definition widerspräche jedoch der gängigen Vorstellung von p als einem Prozentrang- oder Riditwert. Der einfache Weg, der die im Einzelfall oft schwierig zu beantwortende Frage, ob eine ordinale Antwortskala vorliegt, überflüssig macht, besteht darin, die Unbestimmtheitskategorie einfach außer acht zu lassen und p als einen Anteil der Schlüsselantworten unter den Bestimmtheitsantworten zu definieren:

$$(2) \qquad p_+ = N_{i+}/N \,.$$

Darin bedeuten N_{i+} die Zahl derjenigen unter den N Pbn einer Analysenstichprobe, die das Item (i) in Schlüsselrichtung – oder wie wir der Einfachheit halber sagen wollen – mit „Ja" beantwortet haben und $N = N_+ + N_0 + N_-$ die Gesamtzahl aller Pbn der Analysenstichprobe. Analog

1 Selbstverständlich ist auch eine Verrechnung ordinaler Daten nach angegebenen Formeln möglich, jedoch werden die vorgeschlagenen Kennwerte in der Regel zu konservativ ausfallen, da die dritte Kategorie (bei Hybridskalen z.B. „weiß nicht", bei Ordinalskalen z.B. „manchmal") als nicht differenzierend (und so kennwertsenkend) verrechnet wird, obwohl sie auf Ordinalniveau durchaus eine eigene Aussagekraft hat.

läßt sich für jedes Item ein Dispopularitätsindex $p_- = N_{i-}/N$ und ein Unbestimmtheitsindex $N_{i0}/N = p_0$ definieren, in welchen N_{i-} die Zahl der Nein-Antworten und N_{i0} die Zahl der unbestimmten Antworten bezeichnet.

Aktualitätsindex

Setzt man die 3 Indizes p_{i+}, p_{i0} und p_{i-} zueinander in Beziehung, so läßt sich nach dem Vorbild des Aktualitätsmaßes (vgl. HOFSTÄTTER 1963, S.136) ein Aktualitätsindex definieren, der nach geeigneter Normierung in den Grenzen zwischen Null und 1 variiert.

$$(3) \qquad a_i = \frac{4\sqrt{p_{i+} \cdot p_{i-}}}{2 - p_{i0}} = \frac{4\sqrt{p_{i+} \cdot p_{i-}}}{1 + p_{i+} + p_{i-}}$$

Eine Frage i ist inaktuell, wenn sie kein Spannungsverhältnis zwischen Ja- und Nein-Antworten aufbaut, wenn also entweder $p_{i+} = 0$ (was die Frage disqualifizieren würde), oder wenn $p_{i-} = 0$, was plausibler erscheint; in beiden Fällen ist $a_i = 0$. Andererseits ist eine Frage höchst aktuell, wenn $p_{i+} = p_{i-} = 1/2$ und $p_{i0} = 0$, in welchem Fall $a_i = 1$ wäre.

Zur Verdeutlichung: Frage 10 in Tab. 1 hat mit einem $p_+ = .13$ und einem $p_- = .80$ ein $a = .675$, Frage 4 mit $p_+ = .20$ und $p_- = .67$ ein $a = .782$ und Frage 15 mit $p_+ = .47$ und $p_- = .40$ ein $a = .926$.

Der Aktualitätsindex a_i eines Items i ist ein Itemkennwert, der auch im Rahmen der Itemanalyse von 3-stufigen Fragen berücksichtigt werden sollte: Ideal ist hohe Aktualität, die jedoch nur bei fehlender Unbestimmtheit erreicht wird, desiderabel ist mittlere Aktualität, die bei mäßiger Unbestimmtheit zu erwarten ist und indesiderabel ist niedrige Aktualität, wie sie bei hoher Unbestimmtheit eintritt. An dieser Stelle wird wieder deutlich, was bereits in der Einleitung gesagt wurde: die vorgeschlagenen Kennwerte sind bezogen auf Hybridskalen und reagieren nicht auf eine eigene inhaltliche Bedeutung der dritten Kategorie. Interpretiert werden kann bei Aktualität a_i eines Items i als die Höhe der Motivation, ein Item eindeutig zu beantworten oder als Grad der „Abwehr" solches zu tun. Items, die nicht ein empirisch zu bestimmendes Minimum an Aktualität aufweisen, sollten im Rahmen der Itemselektion eliminiert werden. Der Aktualitätsindex macht keine Voraussetzungen im Hinblick auf die Frage, ob die Unbestimmtheitsantwort auf der durch die Ja-Nein-Antworten aufgespannten Dimension liegt oder nicht!

Aus Abbildung 1 ist der Aktualitätsindex direkt abzulesen, wenn die beiden Anteile p_+ und p_- bekannt sind. Man sucht sich auf der Abszisse sein p_+, auf der Ordinate sein p_- und findet im Kreuzungspunkt der beiden

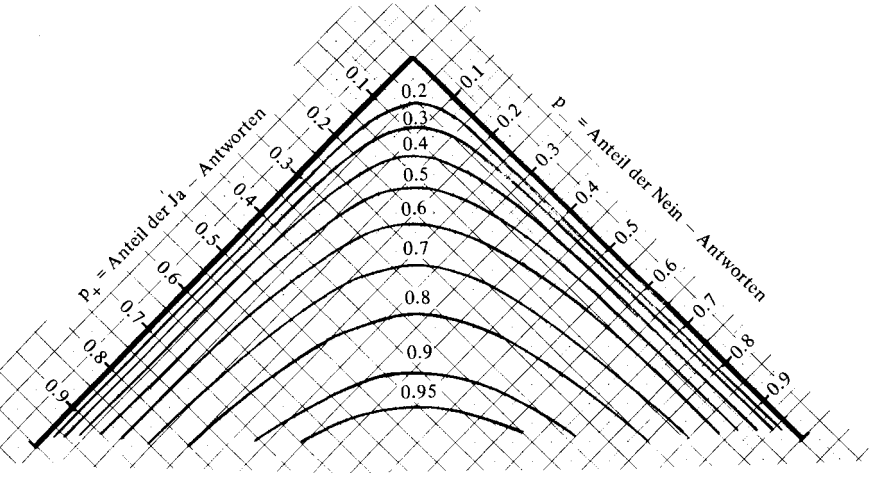

Abb. 1

Lote a. Die Abbildung enthält die Indizes für a = 0.1 (0.1) 0.9 und zusätzlich .95.

5. Der Trennschärfen- und Gültigkeitsindizes

Im Fall binärer Itembeantwortung, wie sie bei Leistungstests die Regel ist, wird die Trennschärfe eines Items i so bestimmt, daß die Stichprobe der N Pbn in zwei Hälften – eine mit supramedianen, die andere mit submedianen Testrohwerten – unterteilt und eine Vierfeldertafel erstellt mit den Zeileneingängen i+ (= Item i richtig beantwortet) und i– (= Item falsch oder nicht beantwortet) und den Spalteneingängen X+ (= Zugehörigkeit zur oberen Hälfte) und X– (= Zugehörigkeit zur unteren, submedianen Hälfte) der N Pbn (vgl. LIENERT 1967, 102 ff.). Wie definiert man nun die Trennschärfe eines Items bei ternärer Itembeantwortung?

Es läge nahe, die Unbestimmtheitsantwort einfach unberücksichtigt zu lassen, doch würde dies im Extremfall dazu führen, daß ein Item als trennscharf identifiziert würde, obschon es nur von einem verschwindend kleinen Teil der N Pbn mit Ja oder Nein beantwortet wurde. Das zeigt Tabelle 3 mit N=100 Pbn

Tabelle 3

	X+	X-	
i+	4=a	1=b	6
i0	45=e	45=f	90
i-	1=c	4=d	5
	50	50	100=N

$$r_{pf} = \frac{(ad - bc)}{\sqrt{(a+b)\,(c+d)\,(a+c)\,(b+d)}} \qquad r_{pf} = \frac{(4{\cdot}4-1{\cdot}1)}{\sqrt{5{\cdot}5{\cdot}5{\cdot}5}} = +0{,}60$$

Der Punkt-Vierfelder- oder Phi-Koeffizient $r_{pf} = 0{,}60$ übertrifft hier bei weitem einen aus der gesamten Tafel zu berechnenden PEARSONschen Kontingenzkoeffizienten $C^2 = \chi^2/(\chi^2+N)$, der nahe bei Null liegt und ein intuitiv zutreffenderes Bild von der „wahren" Trennschärfe dieses Items mit so niedriger Popularität und ebenso niedriger Aktualität vermittelt.

Um diese Dilemma aufzulösen, wird ein Punkt-Vierfelder-Korrelationskoeffizient als Trennschärfeindex vorgeschlagen, der sich aus einer Vierfelder-χ^2-Zerlegung einer 3x2-Feldertafel nach KIMBALL (1954) und KASTENBAUM (1960) herleitet:

$$(4) \qquad r^*_{pf} = \frac{N\,(ad - bc)}{\sqrt{N(a+b)\,(c+d)\,(a+b+c+d)\,(a+c+e)\,(b+d+f)}}$$

Darin bedeuten a,b,c,d die entsprechend bezeichneten Felder der Tabelle 3 und e,f die beiden Felder der mittleren Zeile bzw. deren Frequenzen. Durch Einsetzen erhalten wir einen Trennschärfeindex

$$r^*_{pf} = \frac{100(4{\cdot}4 - 1{\cdot}1)}{\sqrt{100(5)\,(5)\,(10)\,(50)\,(50)}} = +0{,}19$$

der seiner Größenordnung nach sogleich als plausibel erscheint, weil er indirekt die Felder der mittleren Zeile der 3x2 Feldertafel in Tabelle 3 mit einbezieht[2].

Gleiche Überlegungen wie für den Trennschärfeindex ergeben sich für Definition und Berechnung eines Gültigkeitsindex, der daraus resultiert,

2 Die Herleitung erfolgt aus der Beziehung $r^2_{pf} = \chi^2_{pf}/N$, wobei $\chi^2_{pf} = N^2(a_1 b_2 - a_2 b_1)^2 / ABn_1 n_2(n_1+n_2)$ bei KIMBALL (1954, S.453) definiert ist, mit A und B als den Spaltensummen und n_1 und n_2 als den 2 kritischen, die Vierfeldertafeln konstituierenden Zeilensummen!

daß man anstelle der Stichprobenhalbierung nach dem Median der Testrohwerte (Summe der in Schlüsselrichtung beantworteten Fragen) die Analysenstichprobe nach einem Außenkriterium (z.B. klinische versus normale Stichprobe) unterteilt. Auch hier entsteht eine 3x2-Feldertafel mit dem einzigen Unterschied, daß deren Spaltensummen nicht, wie bei der Medianhalbierung, numerisch gleich sein müssen.

Verschiedentlich werden Trennschärfe- und Gültigkeitsindizes nach der Methode von FLANAGAN (1931) ermittelt: Sie besteht darin, daß man die Analysenstichprobe nicht dichotomiert (nach ihrem Median), sondern trichotomiert, und zwar so, daß die beiden extremen Drittel je 27 % mit hohen und niedrigen Ja-Werten (Rohwerten) enthalten und das mittlere Drittel die restlichen 46 % der Pbn mit mittleren Ja-Werten enthält. Wie man in diesem Fall, der eine 3x3-Feldertafel konstituiert, zur Definition eines r^*_{pf}-analogen Trennschärfeindex gelangt, zeigen die Überlegungen des nächsten Abschnittes.

6. Iteminterkorrelationen

Die Gültigkeit von Persönlichkeits- und Einstellungsfragebögen wird ebenso wie die klinischen Symptomskalen heutzutage meist über Faktorenanalysen, d.h. über Interkorrelationen der n Items eines Testinstrumentes abgeschätzt, und im Sinne der Konstruktvalidität interpretiert. Im klassischen Fall binärer Itembeantwortung berechnet man die Vierfelder-Interkorrelationen, meist als Phi-Koeffizienten definiert, erstellt eine Korrelationstafel und extrahiert eine begrenzte Zahl von Faktoren, die man nach geeigneter Rotation als Validitätskonstrukte interpretiert. Wie hat man nun im Fall ternärer Itembeantwortung vorzugehen, wie vor allem aus den durch Gegenüberstellung je zweier Items gewonnenen 3x3-Feldertafeln Interkorrelationskoeffizienten zu gewinnen?

Wie bei der Trennschärfenanalyse, so soll auch hier darauf verzichtet werden, eine ordinal skalierte Antwortabstufung anzunehmen, so daß Verallgemeinerungen des Phi-Koeffizienten von einer 2x2- auf eine 3x3-Feldertafel nicht in Betracht gezogen werden dürfen! Unter dieser Vorannahme erscheint als einziger Ausweg wieder die χ^2-Zerlegung der 3x3-Feldertafel in eine Komponente, die der Phi-Korrelation entspricht und einer Restkomponente, die nicht interessiert. KIMBALL (1954) hat auch für diesen Fall ein spezielles Kalkül entwickelt, dessen Korrelationsäquivalent wie folgt notiert, wenn i und j die beiden Items sind:

$$(5) \qquad r^*_{pf} = \frac{N_{j-}(N_{i-}a - N_{i+}c) - N_{j+}(N_{i-}b - N_{i+}d)}{\sqrt{(N_{j+})(N_{j-})(N_{i+})(N_{i-})(N_{j+}+N_{j-})(N_{i+}+N_{i-})}}$$

Die in Formel 5 verwendeten Symbole sind gemäß Tabelle 4 definiert:

Tabelle 4

		Item j			
		+	0	−	
	+	a = 15	g = 16	b = 5	N_{i+} = 36
Item i	0	e = 30	h = 8	f = 10	N_{i0} = 48
	−	c = 4	k = 2	d = 10	N_{i-} = 16
		N_{j+}	N_{j0}	N_{j-}	N = 100
		49	26	25	

Bei gegebenen je 2 Randsummen sind die restlichen beiden Randsummen und durch a,b,c,d alle übrigen Besetzungszahlen (e,f,g,h,k) bestimmt, wie aus den fiktiven Häufigkeiten der Tabelle 4 hervorgeht. Der Phi-Koeffizient aus der 3x3-Feldertafel ergibt sich danach zu

$$r^{*}_{pf} = \frac{25(16\cdot 15 - 36\cdot 4) - 49(16\cdot 5 - 36\cdot 10)}{\sqrt{(49)\,(25)\,(36)\,(16)\,(49+25)\,(36+16)}} = +0{,}31$$

Dieser adjustierte Phi-Koeffizient ist angesichts der bivariaten Häufigkeitsverteilung in Tabelle 4 wesentlich realistischer als ein „extremaler" Phi-Koeffizient von $r_{pf} = (15\cdot 10 - 5\cdot 4)/\sqrt{20\cdot 14\cdot 19\cdot 15} = +0{,}46$, der als Interkorrelationskoeffizient zwischen den Items i und j weit überhöht erscheint.

Soll das Item i nicht mit dem Item j, sondern mit den trichotomierten Testrohwerten im Sinne eines Trennschärfeindex korreliert werden, verfährt man analog. Geht man nach FLANAGAN vor, werden die N_{j+}/N=27 % der Pbn als „+", die mittleren N_{j0}/N=46 % als „0" und die unteren N_{j-}/N=27 % mit „−" bezeichnet und nach Formel 5 ausgewertet. Bei Tertilen wie in unserer Tabelle 1 vereinfacht sich Formel 5 zu

$$r^{+}_{pf} = \frac{N_{i-}(a-b) - N_{i+}(c-d)}{\sqrt{(2N/3)\,(N_{i+})\,(N_{i-})\,(N_{i+}+N_{i-})}}$$

7. Itemreliabilität und Popularitätswandel

Dasselbe Problem wie bei der Interkorrelation zweier ternär abgestufter Items ergibt sich bei der Re-Test-Reliabilitätsberechnung für ein- und dasselbe Item i: Man korreliert dann eine Erhebung zum Zeitpunkt 1

und eine Erhebungswiederholung zu einem späteren Zeitpunkt 2. Statt i und j als Zeilen- und Spalteneingängen sind sodann 11 und 12 zu setzen, um die Retestreliabilität (Stabilität) eines Items durch einen Reliabilitätskoeffizienten r_{ij}^{+} abzuschätzen.

Bei der Reliabilitätsbeurteilung stellt sich noch ein besonderes Problem, das in der klassischen binären Beantwortung als Schwierigkeitsänderung auftritt. Man betrachte etwa Tabelle 5.

Tabelle 5

		i2 +	i2 −	
i1	+	40	10	50
	−	30	20	50
		70	30	100

$$r_{ii} = \frac{40 \cdot 20 - 10 \cdot 30}{\sqrt{50 \cdot 50 \cdot 70 \cdot 30}} = +0{,}22$$

Man erkennt sogleich, daß das Item i bei der Wiederholung leichter geworden ist, denn sein Schwierigkeitsindex ändert sich von $p_{i1} = 50/100 = 0{,}50$ nach $p_{i2} = 70/100 = 0{,}70$, welche Änderung durch den Symmetrietest von McNEMAR (1947) zu erfassen ist: $\chi^2 = (30-10)^2/(30+10) = 10{,}0$ mit 1 Fg.

In analoger Weise lassen sich Popularitätsänderungen erfassen, wenn man in einer 3x3-Feldertafel die Frequenzen der komplementären Antwortkonfigurationen 1+/2− und 1−/2+ bzw. die beiden Popularitätsindizes miteinander vergleicht.

Die Beachtung von Popularitätsänderungen ist deshalb von Bedeutung, weil die Itemreliabilität durch sie mit beeinflußt wird.

Will man ein Maß für den Grad der Popularitätsänderung eines Items innerhalb einer gegebenen Zeitspanne von t_1 bis t_2 definieren, so empfiehlt sich gemäß $phi^2 = \chi^2/N$ auch mit McNEMARS χ^2 zu verfahren. Da N = b + c, ist dieses so definierte Phi nach Vorzeichenwechsel

$$\phi_{12} = \frac{c - b}{c + b} = \frac{30-10}{30+10} = +0{,}50\,.$$

Dieser Koeffizient mit FECHNERS Korrelationsmaß (vgl. Biometr. Wörterbuch, 1968, S.166) identisch; sein positives Vorzeichen bezeichnet eine Popularitätssteigerung von der ersten zur zweiten Vorgabe des betreffenden Items.

Auch dieser – in der klassischen Testtheorie noch nicht berücksichtigte Itemkennwert – seine Schwierigkeits- oder Popularitätsänderung bei wiederholter Vorgabe – mag ein Selektionskriterium darstellen, insoweit, als unter Idealbedingungen ein $\phi_{12} = 0$, also Popularitätsstabilität erwartet werden sollte. Nach möglichen Gründen eines Popularitätswandels bei wiederholter Vorgabe eines Fragebogens wird bei all jenen Items zu forschen sein, die einen substantiellen Wandel in der einen oder anderen Richtung erkennen lassen.

Literatur

Biometrisches Wörterbuch (Red. G. H. ZSCHOMMLER) Bd. I und II. Berlin: VEB Deutscher Landwirtschaftsverlag 1968.

BOWKER, A. H.: A test for symmetrie in contingency tables. J. Am. Stat. Ass. 43 (1948), 572–574.

FLANAGAN, J.C.: General considerations in the selection of test items and a short method of estimating the productmoment-coefficient from the data at the tails of the distribution. J. Educ. Psychol. 30 (1939), 674.

HOFSTÄTTER, P. H.: Einführung in die Sozialpsychologie. 3. Auflage. Stuttgart 1963.

KASTENBAUM, M. A.: The separation of molecular compounds by countercurrent dialysis: a stochastic process. Biometrika 47 (1960), 69–77.

KIMBALL, A. W.: Short-cut formulas for the exact partition of χ^2 in contingency tables. Biometrics 10 (1954), 452–458.

LIENERT, G. A.: Testaufbau und Testanalyse. 3. Aufl., Weinheim 1969.

– Über die Anwendung von Variablen-Transformationen in der Psychologie, Biometr. Z. 4 (1962), 145–181.

– Verteilungsfreie Methoden in der Biostatistik. 2. Auflage. Meisenheim 1973.

McNEMAR, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12 (1947), 153–157.

THURNER, F., u. U. TEWES: Der Kinder-Angst-Test. Göttingen 1969.

Dieter Heller
FB Erziehungswissenschaften
Universität Bayreuth
Geschwister-Scholl-Platz 3
8580 Bayreuth

Dr. Hans-Peter Krüger
Seminar für Psychologie
FB Erziehungs- und Kulturwissenschaften
der Universität Erlangen-Nürnberg
Regensburger Straße 160
8500 Nürnberg