Zeitschrift für Differentielle und Diagnostische Psychologie 1980, 1, Heft 2, S. 127 – 136

Itemüberlappung zwischen Persönlichkeitsfragebogen als Problem für Validitätsschätzungen*

Alois Angleitner & Franz-Josef Löhr Universität Bielefeld

Zusammenfassung

Die meisten gängigen Persönlichkeitsfragebogen sind durch inhaltsgleiche Items miteinander verflochten. Die Korrelationen zwischen Skalen aus solchen Fragebogen sind im Hinblick auf die Konstruktvalidität als überhöht anzusehen, da wiederholte Messungen gleicher Items der Reliabilität zuzurechnen sind. Die Verflechtung von 16 Fragebogen mit insgesamt 1800 Items wurde systematisch erfaßt und zusammenfassend beschrieben. Einige Möglichkeiten zur rechnerischen Korrektur von Validitätskoeffizienten werden dargestellt. Abschließend wird die weithin unklare inhaltliche Beziehung zwischen Items und Traits als Ursache mangelnder Validität diskutiert.

Summary

Most of our personality inventories overlap, i.e. contain some roughly equivalent items. Correlations between scales stemming from these inventories are too high as measures of construct validity, because repeated measurements of items with equivalent content account for reliability, not for validity. Equivalent items from 16 German inventories containing about 1800 items were counted in a systematic way. Some suggestions for computing corrected validity coefficients are presented. Finally, the discussion points out that to a large extent the relationship between items and traits is not yet clarified with respect to the level of content, and that this accounts for the lack of validity.

Wer sich mit Persönlichkeitsfragebogen (Fbn) beschäftigt und mit mehreren dieser Instrumente arbeitet, stellt sehr bald fest, daß in den verschiedenen Fbn teilweise inhaltsgleiche Items enthalten sind. Dieser Sachverhalt weist darauf hin, daß entweder die Autoren verschiedener Fbn voneinander „abgeschrieben" haben, bzw. sich von bereits vorliegenden Fbn inspirieren ließen, oder daß mit bestimmten Konzepten (z. B. Neurotizismus, Extraversion) schon ganz bestimmte implizite Vorstellungen darüber verbunden sind, durch welche Iteminhalte diese Konzepte am besten repräsentiert werden können.

In den Handanweisungen zu den einzelnen Fbn sind jedoch nur unpräzise Angaben zur Herkunft der Items enthalten (vgl. die Übersicht bei ANGLEITNER 1976), so daß die Verflechtung der Fbn auf dem Itemebene nur durch einen systematischen Vergleich des Wortlautes der einzelnen Items aufgedeckt werden kann. Einen solchen Vergleich führten die Verfasser für die gängigsten deutschen Fragebogen durch.

*Dieser Artikel entstand im Rahmen des Forschungsprojekts „Persönlichkeitsfragebogen" mit finanzieller Unterstützung der DFG.
Die Verfasser danken den Studentischen Mitarbeitern G. JEPSEN, A. LOHMANN, B. ROSENBAUER, M. SANDERS, P. WERITZ, A. BÖCKMANN u. P. ZAAR für ihre Mitarbeit bei den Itemvergleichen.

1. Problematik inhaltlich gleicher Items

Die Wiederholung inhaltlich gleicher Items führt bei vielen Probanden, die eine Fragebogenbatterie bearbeiten, zu Verunsicherung („Was habe ich da denn vorhin angekreuzt?") oder Motivationsverlust („Wieso taucht die Frage denn schon wieder auf?").

Das Gefühl, ähnlichen Fragen schon früher begegnet zu sein und jetzt nicht mehr zu wissen, wie man sich denn früher entschieden hat, führt zusammen mit anderen Faktoren (große Anzahl von Fragen; Zwang, zwischen zwei Alternativen sich entscheiden zu müssen; Verärgerung über „dumme Fragen") zu erheblichen aversiven Reaktionen beim Beantworten. Abgesehen von allen anderen Einwänden gegen die Fragebogenmethode dürfte allein dieser Umstand die Aussagekraft von Fragebogenergebnissen schmälern.

Vor allem haben gleichlautende Items jedoch innerhalb der Persönlichkeitstheorie Auswirkungen auf Versuche einer empirischen Konstruktvalidierung. Zum einen ist denkbar, daß ein- und dasselbe Item in verschiedenen Fbn unterschiedlichen Skalen zugeordnet und damit zur Messung verschiedener Konstrukte benutzt wird. Zum anderen sind die Meßwerte gleichbenannter Skalen, die eine Untermenge gleichlautender Items besitzen, miteinander konfundiert; Korrelationen zwischen solchen Skalen – gewöhnlich als Validitätskoeffizienten (concurrent validity) interpretiert – enthalten daher einen Anteil, der unter dem Aspekt der Reliabilität, nicht der Validität zu interpretieren ist. Ein gegebener Validitätskoeffizient ist in diesem Falle also überhöht.

Die Konstruktvalidität ist allerdings nicht nur von theoretischem Interesse; vielmehr stellt die Korrelation einer Skala mit bereits existierenden, gleichbenannten Skalen ein wesentliches Argument der Fb-Konstrukteure dar, wenn der Anwendungsbereich eines Fb abgesteckt werden soll. Dadurch wird die Konstruktvalidität auch für die diagnostische Praxis bedeutsam.

1.1 Itemüberlappung unter dem Aspekt des item-sampling

Persönlichkeitsbezüge oder traits gelten allgemein als hypothetische Konstrukte, die empirisch nicht unmittelbar, sondern nur über bestimmte Manifestationen im Verhalten erfaßbar sind. Wir können dabei in unserem Zusammenhang die Frage ausklammern, ob hypothetische Konstrukte als real existent oder als bloße begriffliche Fiktionen anzusehen sind.

Fragebogenitems sind verbale Repräsentationen einzelner Aspekte von traits. Die Zusammenstellung von Items zu einer Persönlichkeitsskala läßt sich somit auffassen als Stichprobenziehung aus der Menge der möglichen verbalen Repräsentationen aller Aspekte eines bestimmten traits, dem Item-Universum. Auf dem Hintergrund des Item-sampling-Modells läßt sich die Überschneidung von Persönlichkeitsskalen durch ein Mengendiagramm veranschaulichen (Abb. 1).

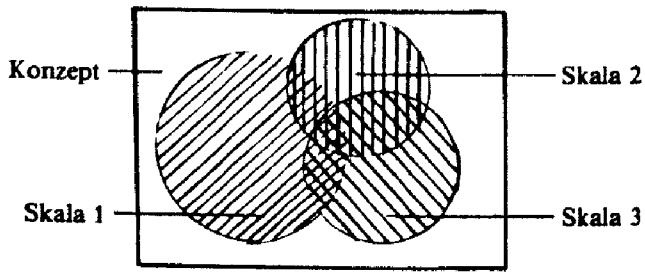


Abbildung 1: Darstellung der verbalen Repräsentation eines Konzeptes durch drei Skalen mit teilweise überschneidenden Items als Mengendiagramm

1.2 Reliabilität oder Validität?

Die bisher erfolgte Gleichsetzung von Items mit trait-Aspekten, die zugleich eine Betrachtung von Skaleninterkorrelationen (Übereinstimmungsgrad *verschiedener* Manifestationen desselben traits) unter dem Validitäts-Gesichtspunkt ermöglicht, wird allerdings bei der Einbeziehung von Parallelskalen wieder problematisch. Parallelskalen, etwa die Extraversionsskalen des EPI-A und des EPI-B weisen keine gemeinsamen Items auf, beanspruchen aber offen-

bar die Erfassung derselben Aspekte des Persönlichkeitszuges „Extraversion" und werden nicht unter Validitäts-, sondern unter Reliabilitäts-Gesichtspunkten miteinander verglichen.

Ohne auf die Frage „Reliabilität oder Validität?" noch weiter eingehen zu müssen, können wir hier zunächst festhalten, daß die gleichsinnige Beantwortung inhaltlich gleicher Items in zwei verschiedenen Skalen auf keinen Fall für die (konvergente) Validität in Anspruch genommen werden kann. Dabei bedeutet „inhaltlich gleich", daß beide Items sich nicht nur auf denselben trait-Aspekt beziehen, sondern diesen auch am selben „Beispiel" darstellen.

Eine solche restriktive Definition von inhaltlicher Äquivalenz hat den Vorteil, gegen unterschiedliche Interpretationen der Iteminhalte bzw. trait-Aspekte relativ immun zu sein, wenngleich sich Interpretationsprobleme nicht völlig ausschließen lassen (man denke z. B. auch an mögliche Kontexteffekte, die u. U. selbst bei äußerlich identischen Items zu verschiedenen Interpretationen führen können.). Andererseits wird eine Korrektur von Validitätskoeffizienten bei restriktiver Auslegung des Äquivalenzbegriffs eher konservativ ausfallen, d. h. die „wahre" Validität dürfte noch unter dem um einen entsprechenden Betrag verminderten Wert liegen (s. u. Punkt 4).

2. Empirischer Vergleich

Die empirische Untersuchung hat zum Ziel, die Itemverflechtung von Fragebogen und Skalen anhand überprüfbarer Kriterien zu erfassen und durch die Häufigkeit inhaltsgleicher Items in einem Fragebogen- bzw. Skalenpaar zu quantifizieren. Darüber hinaus soll für einige Beispiele die relative Bedeutung von Itemüberschneidungen durch das Verhältnis gemeinsamer Items zur Vereinigungsmenge der Items illustriert werden.

2.1 Methode

Der Vergleich wurde jeweils für die Items von 10 Fbn für Erwachsene (GT, MMQ, MPI, EPI-A, EPI-B, FPI, PIT, 16PF-A, 16PF-B, MMPI) und von acht Fbn für Kinder und Jugendliche (KAT, AFS, FS5-10, INK, HANES-I, HANES-II, EPI-A, EPI-B) vorgenommen. Dazu wurde der Text jedes Items zunächst maschinenlesbar gespeichert. Der Versuch, den Vergleich mit Hilfe nichtnumerischer Programme vollständig maschinell durchzuführen, scheiterte am Umfang des Materials (ca. 1800 Items) und am Fehlen eines zur Fragebogenanalyse geeigneten semantischen Systems. Jedoch wurde mit Hilfe eines Textanalyse-Programms für jeden Fb ein alphabetisches Verzeichnis aller darin vorkommenden Wörter mit Angabe der Häufigkeit und der Herkunftsitems erstellt.

Diese Wortlisten wurden dann für jedes Fb-Paar miteinander verglichen; dabei wurden nur bedeutungstragende Wörter und deren Synonyme berücksichtigt. Tauchten solche Wörter in den beiden verglichenen Wortlisten auf, so wurden die betreffenden Items im vollständigen Wortlaut nachgeschlagen und gegebenenfalls notiert. Auf diese Weise entstanden für alle Fragebogenpaare vorläufige Listen gleicher bzw. ähnlicher Items.

Anschließend wurden die Erfahrungen aus dem ersten Vergleichsdurchgang zusammengefaßt und als Instruktion schriftlich festgehalten[1]. Anhand dieser Instruktion wurden die vorläufigen Listen von Itempaaren erneut von zwei Beurteilern unabhängig voneinander durchgegangen. Jedes Itempaar wurde nach seiner Bedeutungsähnlichkeit beurteilt. Die Beurteilungen wurden verglichen und Abweichungen diskutiert; wurde keine Einigkeit erzielt, so wurden die Itempaare einem dritten Beurteiler vorgelegt, dessen Entscheidung dann ausschlaggebend war (bei insgesamt 49 Itempaaren erforderlich). Itempaare, bei denen die inhaltliche Ähnlichkeit als nicht ausreichend eingeschätzt wurde, wurden gestrichen.

Bei der Diskussion der nicht übereinstimmend eingestuften Itempaare zeigte sich einmal mehr, daß eine Reihe von Items auch von Personen mit ähnlichen Voraussetzungen (beide Beurteiler waren Psychologie-

[1] Die Instruktion findet sich im Wortlaut bei ANGLEITNER & LÖHR (1978), Anhang A.

Studenten) inhaltlich verschieden interpretiert wird, wobei jeder von der Richtigkeit seiner Interpretation subjektiv überzeugt ist.

Zur Verdeutlichung der angelegten Kriterien folgen einige Beispiele. Neben den Wort für Wort identischen Items wurden auch Itempaare wie das folgende als völlig inhaltsgleich betrachtet: – „Haben Sie häufig Kopfschmerzen?" – „Ich habe häufig Kopfschmerzen." –

Die völlig inhaltsgleichen Items wurden bei der Auswertung besonders berücksichtigt (s. u.).

Itempaare wie

– „Ich ziehe das Handeln dem Pläneschmieden vor." – „Planen Sie lieber, als daß Sie handeln?" – (mit entgegengesetzter Polung) wurden nicht als „völlig inhaltsgleich", sondern als „inhaltlich ähnlich" bezeichnet. Als Beispiel für ein Itempaar, bei dem die Ähnlichkeit als nicht ausreichend eingeschätzt wurde, sei das folgende aufgeführt:

– „Ich tue vieles, was ich hinterher bereue." – „Denken Sie nach einer brenzlichen und kritischen Situation gewöhnlich, daß Sie etwas hätten tun sollen, was Sie unterlassen haben?" – Die inhaltliche Übereinstimmung dieser beiden Items reicht nicht aus, da das erste ganz allgemein gehalten ist, während sich das zweite nur auf bestimmte (nämlich kritische) Situationen bezieht.

2.2 Ergebnisse[2]

2.2.1 Item-Überlappungen zwischen den Fragebogen

Tabelle 1 zeigt die Itemverflechtung der Fragebögen für Kinder und Jugendliche. Die linke untere Hälfte der Tabelle enthält zu jedem Fragebogenpaar die Anzahl sämtlicher inhaltsgleicher Items, während die obere rechte Hälfte nur diejenigen Items enthält, welche entweder Wort für Wort identisch sind oder nur ganz geringfügige Abweichungen aufweisen („völlig inhaltsgleiche" Items, s. o.). Zum Beispiel weisen EPI-A und HANES I insgesamt 10 inhaltlich gleiche Items auf (Spalte 1/Zeile 5), von denen 3 Items (fast) völlig identisch sind (Zeile 1/Spalte 5). In der Diagonale ist für jeden Fragebogen die Gesamt-Itemzahl angegeben.

Am ausgeprägtesten sind die Überschneidungen zwischen EPI, HANES und INK, also Fragebogen nach der Konzeption von EYSENCK. Dies gilt auch, wenn man nicht die absoluten Zahlen betrachtet, sondern den Anteil inhaltlich gleicher Items relativ zur Gesamt-Itemzahl[3].

Die Ergebnisse zu den Fragebogen für Erwachsene sind in Tabelle 2 dargestellt. Auffällig ist, daß der (in der Tabelle nicht aufgeführte) Gießen-Test keine und die 16 PF-Formen A und B nur sehr wenige gemeinsame Items mit anderen Fragebogen aufweisen. Beeindruckend sind die hohen absoluten Zahlen inhaltlich gleicher Items in EPI und MMPI (41) bzw. PIT und MMPI (54). Bezieht man allerdings das Verhältnis der gemeinsamen Items zur jeweiligen Gesamtzahl ein, so dominieren wiederum die Ähnlichkeiten zwischen den EYSENCKschen Fragebogen EPI, MPI und MMQ (mit Ausnahme des Vergleichs MMQ – MPI).

2.2.2 Item-Überlappungen zwischen den Skalen

Das Ausmaß der Itemverflechtung zwischen den Skalen der Fbn für Kinder und Jugendliche ist in Tabelle 3 dargestellt. In vier Fällen liegen Item-Überlappungen zwischen Skalen vor, die unterschiedliche Konstrukte zu erfassen vorgeben, und zwar zwischen 1. Ängstlichkeit (KAT) und Extraversion (EPI-A), 2. Extraversion (INK-A und HANES I) und Schulunlust (AFS) sowie 3.

---

[2] Eine nach Fragebogenpaaren geordnete Gegenüberstellung des Wortlautes der inhaltlich gleichen Items findet sich bei ANGLEITNER & LÖHR (1978).

[3] Die Gesamtzahl der Items eines Fb-Paares ist hier zweckmäßigerweise als Vereinigungsmenge zu definieren, d. h. als Gesamtzahl aller inhaltlich verschiedenen Items (vgl. Abb. 1); rechnerisch: Summe der Itemzahlen beider Fbn minus Anzahl der gemeinsamen Items.

Tabelle 1: Inhaltlich gleiche Items in 8 Fragebogen für Kinder und Jugendliche.

	EPI-A	EPI-B	KAT	INK(A)	HANES I	HANES II	AFS	FS 5-10
EPI-A	57		0	3	3	7	0	
EPI-B		57		3	9	5	0	
KAT	2		19	0	0	0		0
INK(A)	9	6	2	41	15	7	0	
HANES I	10	11	2	15	36		0	
HANES II	11	11	2	7		32	0	
AFS	2	1		2	3	3	50	0
FS 5-10			1				2	38

Oberhalb der Diagonale: Inhaltlich gleiche Items.
Unterhalb der Diagonale: Inhaltlich gleiche und ähnliche Items.
Diagonale: Gesamtzahl der Items jedes Fragebogens.

Tabelle 2: Inhaltlich gleiche Items in 9 Fragebogen für Erwachsene.

	16 PFA	16 PFB	EPI-A	EPI-B	MMPI	MPI	MMQ	FPI	PIT
16 PFA	187	2			0			1	
16 PFB	2	187	0		0	0	0	2	0
EPI-A		2	57	0	0	4	4	3	0
EPI-B		1		57	2	3	2	6	1
MMPI	2	1	5	5	566	1	1	14	22
MPI		2	14	9	2	48		9	0
MMQ		2	13	7	9		56	8	1
FPI	3	7	12	11	41	13	13	212	4
PIT		1	3	6	54	2	4	9	214

Oberhalb der Diagonale: Inhaltlich gleiche Items.
Unterhalb der Diagonale: Inhaltlich gleiche und ähnliche Items.
Diagonale: Gesamtzahl der Items jedes Fragebogens.

Neurotizismus (HANES 2) und sozialer Erwünschtheit (AFS)[4].

Für die erfaßten Neurotizismus- vs. Ängstlichkeitsskalen lassen sich Item-Überlappungen bei 8 von insgesamt 20 Skalenpaaren feststellen; allerdings gelten die Konstrukte „Ängstlichkeit" und „Neurotizismus" auch auf der theoretischen Ebene nicht als unabhängig. Mit maximal 2 Items pro Skalenpaar ist das Ausmaß der Überschneidungen relativ gering. Der Anteil inhaltsgleicher Items an der Vereinigungsmenge der Items eines Skalenpaares beträgt maximal 5,40% (zur Berechnung vgl. Fußnote 3).

Die meisten der in Tab. 3 dargestellten Item-Überlappungen beziehen sich jedoch auf gleichbenannte Skalen. Bei den folgenden Angaben zu den überhaupt möglichen Überschneidungen wurde berücksichtigt, daß Überschneidungen zwischen Skalen desselben Fragebogens aufgrund der Konstruktion von vornherein ausgeschlossen sind (EPI, HANES, AFS).

Die erfaßten Neurotizismusskalen sind – bei Berücksichtigung der obigen Einschränkung – allesamt durch inhaltlich gleiche Items miteinan-

[4] Wortlaut der Items:
zu (1): „Sind Sie im allgemeinen ohne Sorge?" (EPI-A, Nr. 3) – „Ich mache mir fast immer irgendwelche Sorgen." (KAT, Nr. 6)
zu (2): „Bist Du gern mit anderen zusammen?" (INK-A, Nr. 3 sowie HANES I, Nr. 4) – „Oft möchte ich am liebsten ganz allein für mich sein." (AFS, Nr. 20)
zu (3): „Ich bin nie schlecht gelaunt." (AFS, Nr. 24) – „Bist Du öfters schlechter Laune?" (HANES II, Nr. 13)

Tabelle 3: Anzahl inhaltlich gleicher (obere Dreiecksmatrix) sowie inhaltlich gleicher und ähnlicher (untere Dreiecksmatrix) Items in den Skalen der Kinderfragebogen. (Die Diagonale enthält die Gesamtzahl der Items jeder Skala)

		EPI-A			EPI-B			INK (A)		HANES I			HANES II		AFS			KAT	FS 5-10
		E	N	L	E	N	L	E	N	E_1	E_2	N_1	N_2	L	MA	PA	SE	SU	Angst	Angst
EPI-A	E	24						2		1	1								0	
	N		24						1			1	5		0				0	
	L			9										2						
EPI-B	E				24			1		1	3									
	N					24			2			5	3							
	L						9							2			0			
INK (A)	E	3			3			19		2	5							0		
	N		6			3–			22			8	7		0				0	
HANES I	E_1	3			1–			2		8								0		
	E_2	2			3			5+			8									
	N_1		5			7			8+			20			0	0			0	
HANES II	N_2		7			7			7				20				0		0	
	L			4+			4+							12			0			
AFS	MA		2						1			1			15					
	PA											1				15				0
	SE						1						1!	2			10			
	SU							1!		1!								10		
KAT		1!	1						2			2	2						19	0
FS 5-10																2+			1	38

„+": Maximaler Wert eines Konstruktbereichs
„–": Minimaler Wert eines Konstruktbereichs
„!": Überlappung unterschiedlicher Konstruktbereiche

der verflochten. Die Anzahl der gemeinsamen Items beträgt zwischen 3 und 8 Items; ihr Anteil an der Vereinigungsmenge der jeweiligen Skalenpaare liegt zwischen 7% und 23,5%. Im Extraversionsbereich sind ebenfalls alle in Frage kommenden Skalen miteinander verflochten, allerdings in geringerem Maße (zwischen 1 und 5 gemeinsame Items, entsprechend einem Anteil von 3,2% bis 22,7%). Auch die Kontrollskalen („Lügen" und „soziale Erwünschtheit") sind bis auf eine Ausnahme (AFS – SE vs. EPI-A – L) durch Überlappungen gekennzeichnet. Die Zahl der gemeinsamen Items beträgt maximal 4, was jedoch angesichts relativ kurzer Skalen in diesem Bereich einem Anteil von 23,5% entspricht. Bei den aufgeführten Ängstlichkeitsskalen („manifeste Angst" und „Prüfungsangst" im AFS sowie „Kinder-Angst-Test" und „Fragebogen für Schüler") weisen nur 2 von 5 untersuchten Skalenpaaren 1 bzw. 2 (entsprechend 4%) gemeinsame Items auf.

Bei der Betrachtung der oberen Dreiecksmatrix in Tab. 3 fällt auf, daß die nach den EYSENCKschen Vorstellungen konstruierten Skalen auch verhältnismäßig viele völlig inhaltsgleiche Items enthalten.

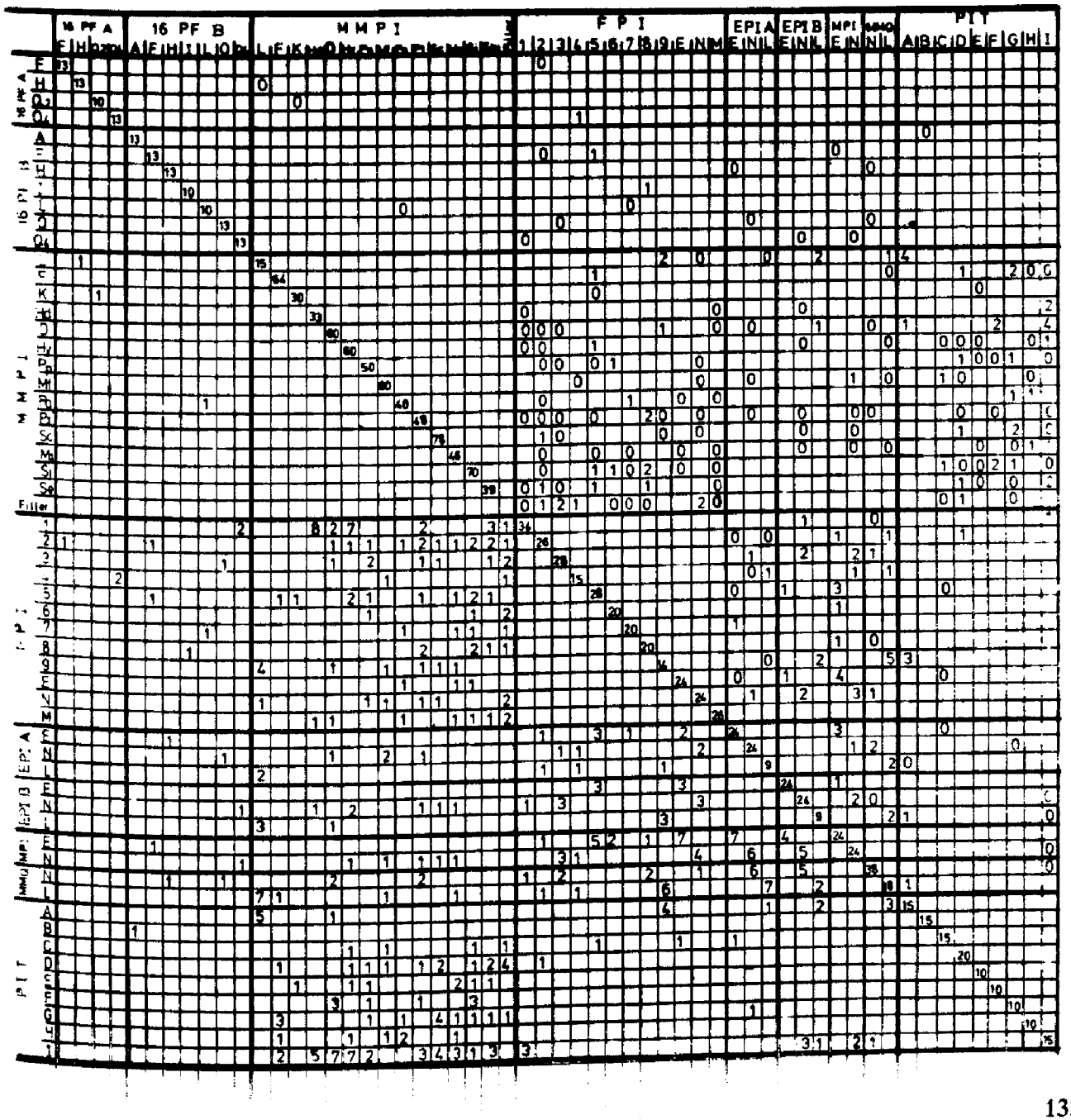
Die Ergebnisse des Itemwortlautvergleichs der Fbn für Kinder und Jugendliche unterstützen

unsere Hypothese, daß hohe konvergente Validitäten, wie sie etwa für den Neurotizismusbereich immer wieder festgestellt wurden, durch Item-Überlappungen größeren Ausmaßes zwischen den Skalen mitbedingt sind, die das betreffende Konstrukt (hier: Neurotizismus) erfassen sollen.

Für die Skalen der Erwachsenen-Fragebogen sind die gefundenen Item-Überlappungen in Tabelle 4 zusammengefaßt. Eine detaillierte Beschreibung der dokumentierten Verflechtungen ist wegen der Vielzahl der Skalen unmöglich; die Interpretation wird durch die Mehrfachauswertungen einzelner Items im MMPI und den Zusatzskalen des FPI (E, N, M) zusätzlich erschwert.

Eine Inspektion von Tab. 4 zeigt auffallende Unterschiede hinsichtlich der Anteile ähnlicher Items in den verschiedenen Skalen: Alle nicht aufgeführten Skalen weisen keinerlei Item-

Tabelle 4: Anzahl inhaltlich gleicher bzw. ähnlicher Items in den Skalen der Erwachsenenfragebogen.
Oberhalb der Diagonale: Inhaltlich gleiche Items.
Unterhalb der Diagonale: Inhaltlich gleiche und ähnliche Items.
Diagonale: Gesamtzahl der Items jeder Skala.



Überlappung auf; dagegen besitzen z. B. die FPI-Skala 1 und die MMPI-Skala Hd 8 gemeinsame Items, was – bezogen auf die Vereinigungsmenge – einem Anteil von 13,6% entspricht. Relativ hohe Ähnlichkeit besteht auch zwischen einzelnen Extraversionsskalen, etwa MPI-E und FPI-E (7 gemeinsame Items, entsprechend 17,1% Anteil an der Vereinigungsmenge). Erwähnenswert ist vielleicht auch noch, daß vor allem bestimmte „bewährte" Lügenitems bei den Fb-Autoren offensichtlich sehr beliebt sind, zumal wenn man die absoluten Häufigkeiten in Tab. 4 mit den relativ niedrigen Itemzahlen dieser Skalen (neben L-Skalen auch PIT-A und FPI 9) in Beziehung setzt. So repräsentieren die 7 gemeinsamen L-Items des MMPI und MMQ ca. 27% der Vereinigungsmenge, die 7 gemeinsamen L-Items des MMQ und des EPI-A sogar 35%.

3. Korrektur des Validitätskoeffizienten

Der systematische Wortlautvergleich bestätigt die Vermutung, daß verschiedene Persönlichkeitsfragebogen bzw. -skalen in teilweise erheblichem Umfang miteinander verflochten sind. Werden mehrere Fragebogen zur Erfassung eines Konstruktes eingesetzt, so erhält man in den meisten Fällen keine unabhängigen Messungen, sondern Werte, die partiell als Ergebnis einer wiederholten Messung anzusehen sind. Will man auf den Einsatz von Fragebogen oder auf einschlägige vorliegende Korrelationsstudien in der empirischen Persönlichkeitsforschung schon nicht verzichten, so müssen zumindest die Validitätskoeffizienten entsprechend korrigiert werden.

Mit Hilfe der klassischen Testtheorie könnte man diese Korrektur etwa folgendermaßen vornehmen:

Man geht von der bekannten Tatsache aus, daß die Unreliabilität zweier gegebener Skalen den Validitätskoeffizienten vermindert. Sollen nun die überlappenden Items eines Skalenpaares nur in einer der beiden Skalen für die Korrelation berücksichtigt werden, so werden die Skalen verkürzt und ihre Reliabilität damit vermindert. Durch entsprechendes Einsetzen in die Spearman-Brown-Formel und die Minderungskorrektur-Formel (correction for attenuation) erhält man folgende Schätzung für den korrigierten Validitätskoeffizienten r'_{12}:

$$r'_{12} = r_{12} \cdot \frac{\sqrt{(N_1 - G_1) \cdot (N_2 - G_2)}}{\sqrt{(N_1 - r_{11} G_1) \cdot (N_2 - r_{22} G_2)}}$$

Diese Korrekturformel kann angewandt werden, wenn die unkorrigierte Skaleninterkorrelation r_{12} und die Skalenreliabilitäten r_{11} und r_{22} bekannt sind; außerdem benötigt man die Skalenlänge N_1 und N_2, die Zahl der gemeinsamen Items G und eine Regel zur Aufteilung dieser Items auf die Skalen (z. B. nach Zufall), so daß $G_1 + G_2 = G$ wird.

Die beschriebene Umrechnung bietet sich an, wenn mit Validitätskoeffizienten gearbeitet werden soll, die der Literatur entnommen werden. Hat man dagegen – etwa bei eigenen Arbeiten – die Rohdaten zur Verfügung, so erhält man natürlich genauere Ergebnisse, wenn man die inhaltlich gleichen Items bereits bei der Auszählung der Skalenrohwerte der Probanden berücksichtigt, indem man diese Items für eine der beiden Skalen nicht mitzählt. Es wäre sicherlich auch interessant, Validitätskoeffizienten zu vergleichen, die am selben Datenmaterial einmal durch nachträgliche Korrektur und zum anderen durch Korrelation „bereinigter" Skalenwerte gewonnen werden.

Als weitere Möglichkeit könnte man mit Hilfe von Monte-Carlo-Studien den durch Itemüberlappung bedingten Anteil an den Korrelationen zwischen Skalen zu eruieren versuchen. Dabei denken wir daran, für die gemeinsamen Items vollständige Beantwortungsstabilität anzunehmen, für die nicht überlappenden Items von Zufallsbeantwortungen auszugehen. Man erhält auf diese Weise eine maximale Schätzung des artefiziellen Validitätsanteils.

4. Diskussion

In der vorliegenden Arbeit haben wir uns im empirischen Teil darauf beschränkt, die Itemüberschneidungen zwischen Fragebogen bzw.

Skalen systematisch festzustellen und summarisch zu beschreiben. Neben den aufgeführten Möglichkeiten zur Korrektur von Validitätskoeffizienten lassen sich die Ergebnisse (s. auch ANGLEITNER & LÖHR 1978) für eine Reihe weiterer Fragestellungen verwerten:

So könnte man die Trennschärfen inhaltlich gleicher Items mit den Trennschärfen von Items vergleichen, die nur in einer Skala Verwendung finden. Dies würde zumindest Anhaltspunkte dafür liefern, warum bestimmte Items besonders „beliebt" sind.

Ferner könnte man den hier vorgenommenen paarweisen Vergleich von Fragebogen auf eine Synopse mehrerer Fragebogen ausdehnen. Dabei würde man wahrscheinlich „Item-Archetypen" entdecken, die nach unserer intuitiven Kenntnis nicht selten Trivialitäten beinhalten werden, z. B. „Ich bin ein lebhafter Mensch" für Extraversion, „Sind Sie häufig nervös?" für Neurotizismus oder „Ich sage nicht immer die Wahrheit" für Lügenskalen.

Wir wollen nun noch einige grundsätzliche Probleme der Fragebogenkonstruktion aufgreifen, die bei den inhaltlich gleichen Items besonders deutlich werden. Versucht man z. B., bei der Anwendung der Formel zur Validitätskorrektur (s. o.) nicht nach Zufall, sondern nach inhaltlichen Kriterien zu entscheiden, welche Skala ein Item behalten bzw. verlieren soll, so wird man vergeblich nach einer ausreichend präzisen Ableitung der Iteminhalte aus einer begrifflichen Analyse des zu erfassenden Konstrukts suchen. Als Konsequenz der primär empirisch orientierten Konstruktionsstrategien wird nämlich in der Regel wenig Wert darauf gelegt, wie ein Itempool zustandekommt. Die zunächst nur sehr vagen Konzepte werden erst nachträglich durch empirische Relationen zwischen den Items (Faktorenladung, Trennschärfe, Konsistenz) näher bestimmt. Mit einem anderen Itempool oder einer anderen Analysestichprobe kommt man bei diesem Vorgehen zwangsläufig zu unterschiedlichen Konzeptbestimmungen, die wegen der fehlenden theoretisch-begrifflichen Vorarbeit lediglich über statistische Zusammenhänge (bzw. Divergenzen), nicht jedoch auf der theoretischen Ebene miteinander verglichen werden können.

Die hier geforderte theoretische Vorarbeit entspricht der Forderung nach „substantiver Validität", die von LOEVINGER (1957) in die Diskussion eingeführt und bislang nur von JACKSON (1970) bei der Entwicklung der „Personality Research Form" (PRF) explizit berücksichtigt wurde. Durch diese rational-begriffliche Vorarbeit wird allerdings die Anwendung statistischer Überprüfungen und Absicherungen der Skalen nicht überflüssig (vgl. JACKSONS sequentielle Strategie); die Konstruktion von Fragebogen wird aber durch den Verzicht auf das „vorschnelle Operationalisieren" eines Konzepts erheblich langwieriger und schwieriger.

Im Gegensatz zu HERRMANN (1973) sind wir nicht der Auffassung, daß sich durch ein empirizistisches Vorgehen im Laufe des Forschungsprozesses eine Erhellung der Konzepte quasi von selbst ergibt. Vielmehr werden durch ein solches Vorgehen immer mehr Konzepte bzw. Konzeptbeschreibungen generiert. Ein Vergleich derartiger Konzepte im Sinne konvergenter bzw. diskriminanter Validierungsversuche stellt letztlich im wesentlichen einen Vergleich subjektiver sprachlicher Bennenungsvorlieben der Konzepturheber dar.

Die Berücksichtigung der substantiven Validität kann zwar sicherlich die Aussagekraft von Fragebogenergebnissen verbessern. Jedoch können die mit der Fragebogenmethode prinzipiell verbundenen Begrenzungen dadurch auch nicht überwunden werden, etwa die Tatsache, daß es sich um rein verbale Verfahren handelt (PETERSON 1965; FISKE 1974). Darüber hinaus ist die Konzeptualisierung interindividueller Differenzen durch ein trait-Konzept keineswegs unproblematisch (FISKE 1973), und selbst, wenn man eine solche Konzeptualisierung akzeptiert, so bleibt doch offen, wie die von verschiedenen Theoretikern eingebrachten traits letztlich begründet werden.

Literatur

ANGLEITNER, A.: Methodische und theoretische Probleme bei Persönlichkeitsfragebögen mit einer ausführlichen Analyse deutschsprachiger Persönlich-

keitsfragebogen. Unveröff. Habilitationsschrift, Univ. Bonn, 1976 (Textband, Tabellenband).

ANGLEITNER, A. & LÖHR, F. J.: Inhaltlich gleiche Items in deutschen Persönlichkeitsfragebogen – Ergebnisse eines systematischen Vergleichs und dessen Bedeutung für die Validitätsproblematik. Berichte aus dem psychologischen Institut der Universität Bonn, 1978, Nr. 21.

BECKMANN, D. & RICHTER, H. E.: Gießen-Test (GT). Ein Test für Individual- und Gruppendiagnostik. Handbuch. Bern: Huber, 1972.

BUGGLE, F. & BAUMGÄRTEL, F.: Hamburger Neurotizismus- und Extraversionsskala für Kinder und Jugendliche, HANES, K. J. Göttingen: Hogrefe, 1972.

CATTELL, R. B. & EBER, H. W.: Specimen set for the Sixteen Personality Factor Questionnaire „16 PF", experimental edition der deutschen Form A und B. Champaign, III.: Institute for Personality and Ability Testing, 1962.

EGGERT, D.: EYSENCK-Persönlichkeits-Inventar. E-P-I. Handanweisung für die Durchführung und Auswertung. Göttingen: Hogrefe, 1974.

EYSENCK, H. J.: Das „Maudsley Personality Inventory" (MPI). Göttingen: Hogrefe, 1959.

EYSENCK, H. J.: Maudsley-Persönlichkeitsfragebogen, 2. verbesserte Auflage. Göttingen: Hogrefe, 1964.

FAHRENBERG, J., SELG, H. & HAMPEL, R.: Das Freiburger Persönlichkeitsinventar FPI – Handanweisung, 2. stark erweiterte Auflage. Göttingen: Hogrefe, 1973.

FISKE, D. W.: The limits for the conventional science of personality. *Journal of Personality* 1974, 42, 1 – 11.

FISKE, D. W.: Can a personality construct be validated empirically? *Psychological Bulletin* 1973, 80, 89 – 92.

GÄRTNER-HARNACH, V.: Fragebogen für Schüler FS 5-10. In: K. H. INGENKAMP (Hrsg.): Deutsche Schultests. Weinheim: Beltz Testgesellschaft, 1973.

HERRMANN, Th.: *Persönlichkeitsmerkmale.* Stuttgart: Kohlhammer, 1973.

JACKSON, D. N.: Personality Research Form Manual. Goshen, N. Y.: Research Psychologists Press, 1967.

JACKSON, D. N.: A sequential system for personality scale development. In: C. D. SPIELBERGER (Ed.): *Current topics in clinical and community psychology,* Vol. 2, New York: Academic Press, 1970, 61 – 96.

LOEVINGER J.: Objective tests as instruments of psychological theory. *Psychological Reports* 1957, 3, 635 – 694.

MITTENECKER, E. & TOMAN, W.: Der P.I.-Test. Beihefte zur Wiener Zeitschrift für Philosophie, Psychologie, Pädagogik, Heft 1, Wien: A. Sexl, 1951.

NISCHAN, C.: Der INK-Persönlichkeitsfragebogen zur Erfassung von Introversion und Neurotizismus bei Kindern und Jugendlichen im Alter von 9 bis 16 Jahren. In: HELM, J., KASIELKE, E. & MEHL, J. (Hrsg.): Neurosendiagnostik. Beiträge zur Entwicklung klinisch-psychologischer Methoden. Berlin: VEB Deutscher Verlag der Wissenschaften, 1974, 228 – 250.

PETERSON, D. R.: Scope and generality of verbally defined personality factors. Psychological Review, 1965, *72,* 48 – 59.

SPREEN, O.: MMPI-Saarbrücken. Handbuch zur deutschen Ausgabe des Minnesota Multiphasic Personality Inventory. Bern: Huber, 1963.

THURNER, F. & TEWES, U.: Der Kinder-Angst-Test (KAT). Ein Fragebogen zur Erfassung des Ängstlichkeitsgrades von Kindern ab 9 Jahren. Göttingen: Hogrefe, 1969 (2. verbesserte Auflage 1972).

WIECZERKOWSKI, W., NICKEL, H., JANOWSKI, A., FITTKAU, B. & RAUER, W.: Angstfragebogen für Schüler (AFS). Braunschweig: Georg Westermann Verlag, 1974.

Prof. Dr. A. Angleitner
Dipl.-Psych. F.-J. Löhr
Fakultät für Psychologie und Sportwissenschaften
Universität Bielefeld
Postfach 8640
4800 Bielefeld 1