SYSTEMATISCHE VERZERRUNGEN IN DER BEURTEILUNG VON PERSONEN: FAKTUM ODER FIKTION?

Peter Borkenau und Alois Angleitner

Summary (Systematic distortions in judgments about persons: Fact or fiction?): Shweder & D'Andrade (1980) report low relationships between the structures of rating and behavior matrices. In constructing their behavior matrices, however, they forced their subjects to assign each observed behavior to only one category. Thus, no correlations due to behavior overlap were possible. It is very probable, however, that when making global trait-ratings, subjects use same behaviors in making several judgments. An own study is presented wherein behavior matrices allowing and not allowing for item overlap were systematically compared. The behavior matrix allowing for overlap was more similar to the corresponding rating and semantic similarity matrix. It is argued that correlations between trait-ratings result from two sources: Covariations between behaviors and item overlap.

1. Problemstellung

Spätestens vor 100 Jahren wurde erstmalig folgende These vertreten: "Each character term has a separate shade of meaning, while each shares a large part of its meaning with some of the rest" (Galton, 1884, S. 181). Diese These einer Bedeutungsüberlappung zwischen Traitbegriffen könnte möglicherweise die von Shweder & D'Andrade (1980) präsentierten Befunde erklären: Diese berichten von einer nur geringen Entsprechung der Struktur von Verhaltensdaten einerseits sowie Ratings und semantischen Ähnlichkeitsbeziehungen der Trait-beschreibenden Begriffe andererseits. Hingegen wiesen die globalen Ratings und die semantischen Ähnlichkeiten bezüglich ihrer jeweiligen Struktur hohe Korrespondenzen auf. Shweder & D'Andrade werten diese Befundlage als Beleg für ihre Systematische Verzerrungshypothese, der zufolge es im Laufe der Zeit zu einer systematischen Verzerrung beobachteter Kovariationen in Richtung der semantischen Ähnlichkeiten komme. Bisher nicht untersucht wurde hingegen eine alternative Erklärung, welche wir als Systematische Überlappungshypothese bezeichnen wollen. Dieser zufolge resultieren Interkorrelationen zwischen globalen Trait-Ratings unter anderem daraus, daß identisches Verhalten als indikativ für mehrere Traits angesehen wird. Die von Shweder & D'Andrade präsentierten Befunde lassen für eine solche Deutung insofern Raum, als im Zuge der Erstellung der Verhaltensmatrix beobachtete Handlungen stets einer Trait-Kategorie zuzuordnen waren, wobei die Kategorien mitunter ein hohes Maß semantischer Ähnlichkeit aufwiesen (z.B. "stimmt nicht zu" und "zeigt Antagonismus"). Ziel der hier dargestellten Studie war der Nachweis, daß bei einer Berücksichtigung überlappender Verhaltenselemente zwischen Trait-Kategorien die resultierende Struktur der Verhaltensfrequenzen in sehr viel höherem Maße der Struktur paralleler globaler Selbstratings korrespondiert.

2. Methode

Die Untersuchung erstreckte sich auf die Trait-Kategorien "Reserviertheit", "Submissivität", "Dominanz", "Streitsucht" und "Freundlichkeit". Diese Begriffe wurden von 28 Beurteilern hinsichtlich ihrer semantischen Ähnlichkeiten eingeschätzt, und zwar im vollständigen Paarvergleich und auf 7-stufigen Ratingskalen, welche von +3 (= identische oder sehr ähnliche Bedeutung) bis -3 (= sehr starker oder direkter Gegensatz) reichten. Die über die 28 Beurteiler gemittelten Einschätzungen sind in Spalte 1 der Tabelle 1 aufgeführt. Zur Erstellung einer Matrix globaler Trait-Ratings wurden einer Stichprobe von N = 55 Studenten der Universität Bielefeld 7-stufige Ratingskalen vorgegeben, welche sich auf die fünf oben erwähnten

Traits bezogen. Die Pbn sollten sich diesbezüglich selbst einschätzen. Die Ratingskalen waren unipolar formuliert und die Pole als 1 (= trifft überhaupt nicht zu) bzw. 7 (= trifft vollkommen zu) gekennzeichnet. Die Reihenfolge der Vorgabe der fünf Ratingskalen wurde für die verschiedenen Probanden permutiert, um systematische Positionseffekte auszubalancieren. Die Interkorrelationen zwischen diesen Ratings sind in der zweiten Spalte der Tabelle wiedergegeben.

Zwecks Erstellung von Verhaltensmatrizen wurden die erwähnten 55 Pbn um Verhaltensberichte gebeten. Zu diesem Zweck wurde das von Buss & Craik (1983) in deren Untersuchungen verwendete Itemmaterial übersetzt. Dieses besteht aus jeweils 100 Handlungen pro Traitkategorie. Die Items sind spezifischer formuliert als in gängigen Fragebogen, da sie unmittelbar Verhaltensfrequenzen erfassen sollen. So lautet etwa ein typisches Item "ich feilschte um Preise". Die Pbn sollen dazu angeben, ob sie dieses Verhalten schon einmal gezeigt haben und, wenn ja, wie oft. Als Alternativen stehen "selten", "manchmal" und "häufig" zur Verfügung. Die Reihenfolge der Itemvorgabe war für jeden Pbn unterschiedlich, um auch hier Reihenfolge-Effekte auszubalancieren. Jeder Pb bearbeitete alle 500 Items.

Gemäß der zentralen Hypothese der Untersuchung kam der Kodierung der Itembeantwortungen die Schlüsselrolle zu: Verfahren, welche Überlappungen ausschlossen, waren systematisch mit solchen zu vergleichen, welche Überlappungen erlauben. Ersteres wurde erreicht, indem zwei Rater jede der 500 Verhaltensweisen einer und nur einer Kategorie zuordneten, also entweder der Reserviertheit oder der Dominanz usw. Ein Auswertungsschlüssel, welcher überlappende Verhaltenselemente zuläßt, wurde hingegen folgendermaßen erstellt: Zwei Beurteilern wurden alle 500 Handlungen jeweils fünfmal per Terminal präsentiert, wobei die Itemreihenfolge in allen Fällen unterschiedlich war. Die Rater wurden instruiert, jeweils auf einer siebenstufigen Ratingskala anzugeben, ein wie gutes Beispiel für Reserviertheit, Dominanz, usw. das jeweilige Verhalten sei. Die Skala reichte von 1 (= sehr schlechtes Beispiel für ...) bis 7 (= sehr gutes Beispiel für ...). Die Wahlen waren mittels der Tastatur des Terminals zu markieren. Zunächst war solchermaßen die Prototypikalität aller 500 Handlungen für eine erste Trait-Kategorie einzuschätzen. Die folgenden 500 Ratings galten der 2. Kategorie usw., wobei beiden Beurteilern die Kategorien in unterschiedlicher Reihenfolge präsentiert wurden.

3. Resultate

In einem ersten Auswertungsschritt galt es, die Beurteiler -Übereinstimmungen zu ermitteln. Die gemittelten Ratings der semantischen Ähnlichkeitsbeziehungen wiesen eine nahezu perfekte Reliabilität von .96 auf (Intraclass-Korrelation). Die über die zwei Rater gemittelten Prototypikalitätsratings wiesen gemäß ihrer Interkorrelation und Aufwertung nach der Spearman-Brown Formel eine Reliabilität von durchschnittlich .62 auf. Waren die 500 Handlungen einer und nur einer Kategorie zuzuordnen, so resultierte ein Kontingenzkoeffizient von C = .46.

Die Verhaltensberichte der 55 Pbn wurden nunmehr folgendermaßen gescort: Die Angaben bezüglich der Verhaltensfrequenzen wurden in eine Skala transformiert, welche von 0 (= Handlung nicht ausgeführt) bis 3 (= Handlung schon oft ausgeführt) reichte. Trait-Scores, welche überlappende Verhaltenselemente ausschließen, wurden sodann wie folgt ermittelt: Die jeweilige Frequenzangabe ging mit einem Gewicht von 1 in einen Trait-Score ein, wenn der erste Beurteiler die Handlung dieser Trait-Kategorie zugeordnet hatte. Hatte er sie hingegen einer anderen Kategorie subsumiert, so blieb sie bei der Berechnung des Trait-Scores unberücksichtigt. Jede

Handlung ging entsprechend in die Berechnung nur eines der fünf Summenwerte ein, während sie bei den übrigen vier keine Berücksichtigung fand. Die resultierenden fünf Summenscores wurden schließlich noch durch die Summe aller fünf Trait-Scores des jeweiligen Probanden dividiert, um eine Korrektur für einen interindividuell unterschiedlichen Gebrauch der Frequenzskala im Sinne eines Response Set zu schaffen. Die so ermittelten Trait-Scores wiesen eine mittlere Validität von .43 gegenüber dem parallelen globalen Selbstrating auf. Die Interkorrelationen der solchermaßen ermittelten Trait-Scores finden sich in der dritten Spalte der Tabelle 1.

Tab. 1: Semantische Ähnlichkeitsmatrix, Matrix globaler Trait-Ratings und Verhaltensmatrizen ohne (Spalte 3) sowie mit (Spalte 4) überlappenden Verhaltenselementen (Erläuterungen im Text)

Trait-Paar		(1)	(2)	(3)	(4)
reserviert	- dominant	-0.75	-.26	-.50	-.61
reserviert	- submissiv	-0.07	.44	.07	.41
reserviert	- streitsüchtig	-0.68	-.22	-.28	-.53
reserviert	- freundlich	-0.29	-.14	-.28	-.26
dominant	- submissiv	-2.68	-.22	-.74	-.84
dominant	- streitsüchtig	1.25	.44	.58	.67
dominant	- freundlich	-0.04	.13	-.40	-.12
submissiv	- streitsüchtig	-1.86	-.12	.50	-.78
submissiv	- freundlich	-0.14	-.31	.31	.35
streitsüchtig	- freundlich	-2.25	-.14	-.75	-.47

Die Spalten 1 und 2 korrelieren miteinander zu .59, die Spalten 1 und 3 zu .57 und die Spalten 2 und 3 zu .42. Es zeigt sich somit eine etwas höhere Korrespondenz zwischen der Verhaltensmatrix und den beiden übrigen Matrizen als von Shweder & D'Andrade berichtet. Dies ist möglicherweise durch den Umstand bedingt, daß wir mit selbstberichtetem Verhalten arbeiteten.

Werden nun diese Entsprechungen deutlich höher, wenn eine Handlung in die Berechnung mehrerer Trait-Scores eingeht? Um dies zu überprüfen, wurden die Trait-Scores so ermittelt: Jede Verhaltensfrequenz wurde mit dem Prototypikalitätsrating multipliziert, das sie bezüglich der jeweiligen Trait-Kategorie erhalten hatte. Sodann wurden die resultierenden Produkte über die 500 Handlungen hinweg aufsummiert. Dieser Score wurde zwecks Kontrolle von Response Set Varianz abschließend noch durch die Summe aller mit 1 gewichteten Verhaltensfrequenzen der jeweiligen Person dividiert. Die so ermittelten Trait-Scores korrelierten im Mittel zu .42 mit dem jeweils parallelen globalen Trait-Rating. Mithin blieb die Validität der Traitscores unbeeinflußt von der Art der Gewichtung der Verhaltensfrequenzen. Deutliche Verbesserungen zeigten sich jedoch hinsichtlich der Struktur dieser Verhaltensmatrix, die in Spalte 4 wiedergegeben ist: Sie korreliert mit den semantischen Ähnlichkeitsbeziehungen (Spalte 1) zu .84 und mit den Rating-Interkorrelationen (Spalte 2) zu .69. Beide Koeffizienten sind statistisch signifikant. Mithin zeigen sich deutlich höhere Korrespondenzen zwischen Verhaltensmatrix einerseits und Rating- und Ähnlichkeitsmatrix andererseits, wenn in der Verhaltensmatrix überlappende Elemente zugelassen sind. Ergänzende Analysen zeigten eine wesentliche Ursache für diesen Befund: Je semantisch ähnlicher sich zwei Kategoriennamen waren, desto höher waren die Prototypikalitätsratings für diese Kategorien korreliert (r = .89). Auch waren die Interkorrelationen der Trait-Ratings desto höher, je stärker die einschlägigen Prototypikalitätsratings kovariierten (r = .53). Insgesamt sprechen somit die Befunde für die These, daß Interkorrelationen zwischen globalen Trait-Ratings auf zwei Quellen beruhen: Kovariationen zwischen Verhaltensfrequenzen und überlappenden Verhaltenselementen bei einander ähnlichen Traits.

Literatur

Buss, D.M. & Craik, K.H. Act prediction and the conceptual analysis of personality scales: Indices of act density, bipolarity, and extensity. Journal of Personality and Social Psychology, 1983, 45, 1081 - 1095.

Galton, F. Measurement of character. Fortnightly Review, 1884, 42.

Shweder, R.A. & D'Andrade, R.G. The systematic distortion hypothesis. R.A. Shweder (Ed.), Fallible judgment in behavioral research. San Francisco: Jossey-Bass, 1980.

AUF DER SUCHE NACH PERSONEN MIT EIGENSCHAFTEN: UNTERSUCHUNGEN ZUR RESTRIKTION DES EIGENSCHAFTSMODELLS AUF UNTERGRUPPEN VON PERSONEN, VERHALTENSWEISEN UND SITUATIONEN[1)]

Manfred Amelang, Claudia Kobelt und Albrecht Frasch

Summary (In search of persons with traits: Studies on the confinement of the trait model to subgroups of persons, behaviors, and situations): Following the approach of Bem & Allen (1974) the results of a study with N = 173 subjects show that reliability is very low for measures of cross-situational variability, trait observability and appropriateness of trait-descriptive terms. Peer ratings of these measures and combinations of peer- and self-ratings moderate the validity coefficients of trait self-ratings against external criteria in a consistent manner.

So erfolgsversprechend der Forschungsansatz von Bem & Allen (1974) zur Einschränkung des Eigenschaftsmodells auf Untergruppen von Personen schien, so wenig konnten die positiven Befunde der ersten Stunde später repliziert werden (Mischel & Peake, 1982; Chaplin & Goldberg, 1984; Amelang & Borkenau, 1984).

Einer der Gründe dafür liegt möglicherweise in einer nur geringen Reliabilität der selbsteingeschätzten Variabilitätstendenz als einer Dimension von absoluter Schlüsselfunktion. Diesem potentiellen Ursachenkomplex sollte im Rahmen der vorliegenden Untersuchung nachgegangen werden.

Außerdem interessierte die Frage, inwieweit neben Selbsteinschätzungen der transsituativen Varabilitätstendenz auch Fremdurteile dazu geeignet sind, die Validität von Persönlichkeitsskalen zu moderieren und ob eine Kombination beider Datenquellen zu einheitlicheren Resultaten führt.

Unsere Erhebung bezog zu diesen beiden Problemkreisen darüber hinaus auch die Beobachtbarkeit von Traits und deren individuelle Angemessenheit zur Beschreibung der jeweiligen Persönlichkeit als Klassifikationsgesichtspunkte zur Identifikation von prädizierbaren Personen mit ein.

Schließlich gingen wir der Frage nach, ob Maße der transsituativen Variabilität, der Beobachtbarkeit und Angemessenheit nicht nur die Validität globaler Selbsteinschätzungen, sondern auch die Gültigkeit des Freiburger Persönlichkeitsinventars (FPI, Fahrenberg, Selg & Hampel, 1973) moderieren.