Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung

Bernd Rohrmann

Auf der Basis psychometrischer und psycholinguistischer Überlegungen sowie praktischer Erfahrungen aus der Feldforschung mit Interviews wurden 4 fünfstufige *Antwortskalen* für Urteile auf den Dimensionen Häufigkeit, Intensität, Wahrscheinlichkeit und Zustimmung entwickelt. Um die Skalenstufen verbal eindeutig verankern zu können, waren zuvor 100 *Graduierungsbegriffe* in zwei Skalierungsexperimenten (1966, 1976) auf inhaltliche Wertigkeit und Prägnanz untersucht worden.

Empirische Erfahrungen – u. a. aus zwei Befragungen von Interviewern (1975, 1977) – unterstützen das Konzept, für die empirische Sozialforschung mit Bevölkerungsstichproben numerisch *und* sprachlich gegliederte Antwortskalen mäßiger Stufenzahl einzusetzen.

Based on psychometric and psycholinguistic considerations as well as on practical experiences with questionnaires a concept for standardized rating scales was developed. It provides 5-point-scales defined by numerical and verbal and graphic graduations.

To enable definite verbal labeling of scale intervals 100 words (mainly adverbs) were scaled with respect to their grading values and ambiguity.

Several scales on intensity, frequency, probability and evaluation aspects were set up; empirical evidence confirms their utility especially in field research with ‚non-sophisticated' respondents.

1. Problemstellung

1.1 Antwortskalen in der empirischen Sozialforschung

In der sozialwissenschaftlichen Forschung von Psychologie, Soziologie, Pädagogik, Medizin usw. werden Antwortschemata benötigt, mit denen Fragen graduiert beantwortet, Objekte graduiert eingestuft, Eigenschaften graduiert beurteilt werden können. Derartige Skalen – die in der Literatur (z. B. Guilford, 1954, Kap. 11; Clauss, 1968; Hennig, 1975) u. a. als Urteilsskalen, Einschätzskalen bzw. rating scales oder (subjective) estimate scales bezeichnet werden – seien hier Antwortskalen genannt.

Sie sollen quantitative Beschreibungen des Ausprägungsgrades von Merkmalen bzw. Sachverhalten in Situationen erbringen, wo einerseits qualitative/kategoriale Aussagen nicht hinreichend sind, andererseits aber das ‚Messen' in einem Bewertungsakt des urteilenden Individuums besteht.

Derartige Quantifizierungen werden umso wichtiger, je stärker sozialwissenschaftliche Daten einer analytischen Auswertung durch statistische Methoden zugeführt werden sollen. Entsprechende Bedeutung haben die Antwortskalen in Beurteilungs- oder Beobachtungssituationen und vor allem in Fragebögen verschiedenster Art.

1.2 Psychometrische Kriterien für Antwortskalen

Durch Antwortskalen sollen zumindest „ordinale", möglichst aber „kardinale" Meßergebnisse erzielt werden. In diesem Fall müssen u. a. die vorgegebenen Skalenstufen gleichabständige Intervalle repräsentieren und vor allem von den Antwortenden auch so interpretiert und realisiert werden.

Auf den dabei stattfindenden psychometrischen Prozeß und die zugrundegelegte Meß- bzw. Skalentheorie soll hier nicht näher eingegangen werden (siehe dazu Stevens, 1951; Togerson, 1958; Sixtl, 1967; Coombs et al., 1970; Gutjahr, 1972; Orth, 1974).

Bezogen auf Zweck und Anwendung von Antwortskalen und die statistische Verarbeitung der resultierenden Daten interessiert vorrangig das Kriterium der Äquidistanz; theoretische und praktische Anforderungen wie etwa Differenzierungsleistung, Eindimensionalität, Zuverläßigkeit oder Verständlichkeit kommen hinzu.

Antwortskalen präsentieren eine Graduierung des Kontinuums, auf dem geurteilt werden soll. Sie enthalten im allgemeinen eine bestimmte Anzahl von Stufen, die numerisch (durch Zahlenreihen) und/oder verbal (z. B. durch Graduierungspartikel) und/oder grafisch (durch Aufteilung von Linien oder Flächen) symbolisiert und visualisiert werden (meist in kombinierter Form).

Die Zahl der Stufen – üblicherweise zwischen 3 und 11 – muß einerseits auf das Differenzierungsvermögen des Urteilers und andererseits den Differenzierungsbedarf des Forschers abgestimmt sein.

1.3 Generelle Beurteilungsdimensionen

Primäre Beurteilungsdimensionen, die in nahezu allen Kontexten benötigt werden, sind vor allem „Intensität" und „Häufigkeit". Als globale Dimensionen ist besonders die „Bewertung" von Vorgängen oder Aussagen (gut-schlecht, richtig-falsch, usw.) zu nennen, die besonders für die Einstellungsmessung (Thurstone, 1929; Likert, 1932; Edwards, 1957; Dawes, 1972) und die Diagnostik (vgl. Remmers, 1963 bzw. Tent, 1970; Tailor et al., 1968; Cranach & Frenz, 1969; Langer & Schulz, 1974) wichtig ist; dabei wird oft auf Häufigkeits- oder Intensitäts-Graduierungen aufgebaut (vgl. Cohen et al., 1969), was auch für die Beurteilungsdimension „Wahrscheinlichkeit" gilt.

1.4 Zielsetzung der Untersuchung

Obgleich immer wieder betont wird, daß nur angemessen skalierte Variablen theoretische Modelle prüfbar machen und den Anforderungen der benötigten statistischen Analysetechniken genügen (z. B. Tack, 1969 oder Ziegler, 1972), wird in Darstellungen der Skalierungsverfahren (z. B. Torgerson, 1958; Sixtl, 1967; Scheuch & Zehnpfennig, 1974) zwar die Verrechnung von Reaktionen/Urteilen/Antworten in vielfältigster Weise behandelt, auf die Skalenvorlagen für die primäre Datengewinnung aber meist nur geringes Augenmerk gelegt. Eine nähere Behandlung des Antwortskalenproblems fehlt ebenso in den meisten Werken zur Methodik der empirischen Sozialforschung (z. B. Selltiz et al., 1959; Noelle, 1963; König, 1965; Galtung, 1969; Lindzey & Aronson, 1969; Atteslander, 1971; Friedrichs, 1973; König, 1974; Holm, 1975).

Sixtl (1967, p. 145) bezeichnet es geradezu als Charakteristikum von Rating-Verfahren, „daß eine empirische Eichung des Ausprägungsgrades, den die Aussagen oder verbalen Kennzeichnungen repräsentieren, unterbleibt". Dies ist um so unbefriedigender, als in der Soziologie – wo das Interview als „Königsweg" der empirischen Sozialforschung gilt (König 1965, p. 27) – und der Psychologie Ratingskalen eine zentrale Bedeutung haben (einen konkreten Beleg gibt z. B. Dawes (1972, p. 96), demzufolge im Jahrgang 1970 des Journal of Personality and Social Psychology 60% aller empirischen Arbeiten auf Ratingskalen basierten).

Betrachtet man die tatsächliche Handhabung von Antwortskalen in der Forschung, so wird ein Dilemma deutlich: einerseits werden besonders in der angewandten Forschung vielfach inadäquate (z. B. völlig willkürlich verbalisierte) Antwort-„Skalen" benutzt, und andererseits sind viele der in Laborexperimenten eingesetzten, zumeist für Studenten gemachte Antwortskalen (z. B. rein numerische Skalen von 0 bis 10 oder 0 bis 100) viel zu abstrakt oder differenziert, um für weniger vorgebildete Probanden bzw. außeruniversitäre Bevölkerungsstichproben geeignet zu sein. Besonders bei Befragungen in der Feldforschung zeigt sich, daß viele Befragte mit quantitativen Antwortvorgaben beträchtliche Schwierigkeiten haben und qualitative Antworten bevorzugen.

Ziel dieser Untersuchung war deshalb die Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung, die psychometrischen Mindestansprüchen genügen (nämlich annähernd äquidistant zu messen und hinreichend zu differenzieren) und darüber hinaus inhaltlich und grafisch so gestaltet sind, daß sie auch von intellektuell weniger gewandten Befragten/Probanden/Versuchspersonen verstanden und angemessen gehandhabt werden können. Dies scheint am besten erreichbar, wenn die Abstufungen der Antwortskalen numerisch *und* verbal charakterisiert werden, um metrischen und kommunikativen Intentionen genügen zu können, und zwar so, daß der Bedeutungsgehalt der benutzten Wörter – z. B. „etwas", „ziemlich", „sehr", usw. – mit den benutzten Zahlen korrespondiert. Um solche Zuordnungen empirisch begründet treffen zu können, sind geeignete

Graduierungsbegriffe in mehreren Skalierungsstudien untersucht worden, und zwar für die in 1.3 genannten vier Urteilsdimensionen: Häufigkeit, Intensität, Wahrscheinlichkeit und Bewertung (von Aussagen).

Die bisherigen Ergebnisse sollen im folgenden mitgeteilt werden, um für praktische Anwendungen verfügbar zu sein, auch wenn die theoretischen Probleme – insbesondere psychometrischer Art – nicht immer befriedigend gelöst sind.

1.5 Ablauf der empirischen Studien

In der Ausgangsstudie von 1966 wurden 77 Graduierungsbegriffe zu 4 Dimensionen skaliert, wobei die Versuchspersonen einer Bevölkerungsstichprobe entstammten; 1976 ist diese Studie I an einer ähnlichen Gruppe sowie an einer studentischen Stichprobe mit 65 Begriffen (darunter 23 neuen) wiederholt und erweitert worden, um die 1966 entwickelten Antwortskalen zu überprüfen (Studie II). Informelle und systematische Befragungen von Interviewern, die mit Anwortskalen vertraut sind, ergaben ergänzende Daten. Abb. 1 gibt eine Übersicht zu den empirischen Schritten.

Alle Erhebungen wurden im Rahmen sozialpsychologischer Felduntersuchungen über Lärm durchgeführt (Irle & Rohrmann, 1968; Rohrmann, 1975; Finke, Guski & Rohrmann, 1977).

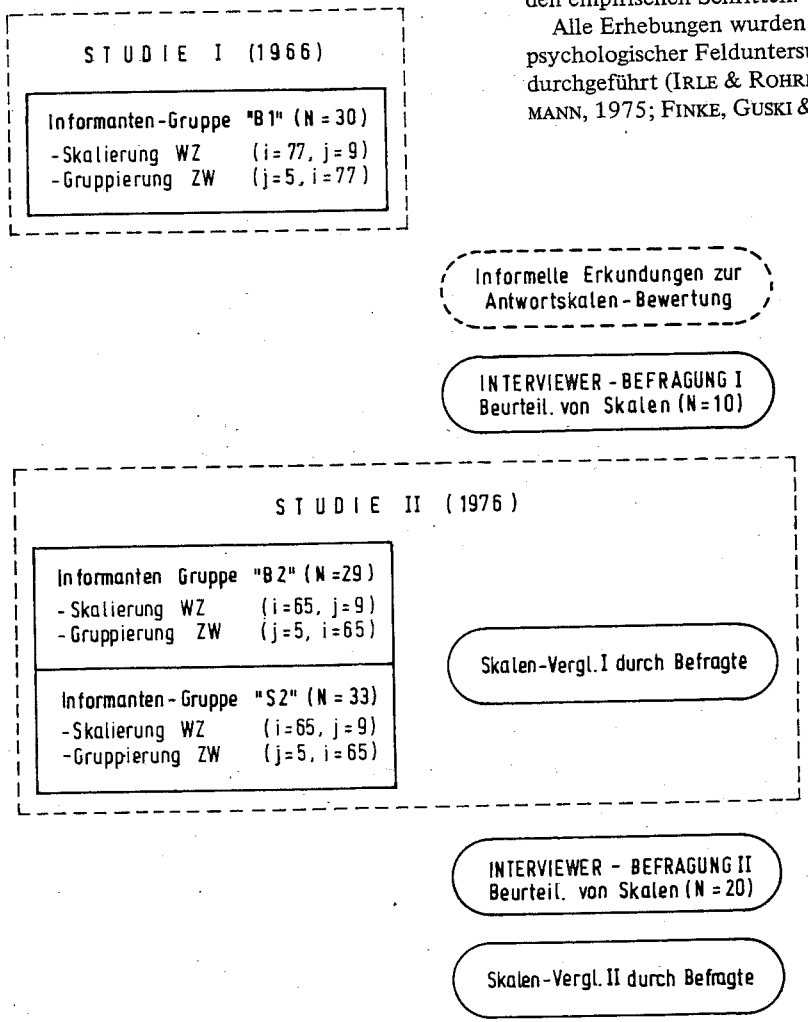


Abb. 1: Studie E.s.A. – Übersicht zu den empirischen Schritten.

2. Studie I (1966): Skalierung von Graduierungsbegriffen

2.1 Psycholinguistischer Ansatz

Die (umgangs-)sprachlichen Graduierungsbegriffe – z.B. „nie" oder „oft" für die Dimension Häufigkeit, „etwas" oder „sehr" für die Intensität, „keinesfalls" oder „vielleicht" für die Wahrscheinlichkeit, usw. – sind als Ausdrücke von Abstufungen spontan verständlich, haben aber keine definite Bedeutung. Für Antwortskalen sind solche Begriffe am besten geeignet, die von unterschiedlichen Personen gleichartig interpretiert werden, deren Bedeutung zumindest in der interessierenden Dimension möglichst ähnlich ist. Dazu scheint nötig, daß die denotative Bedeutung eines Wortes (Bühlers Signalfunktion bzw. die „kriterialen" – kritisch entscheidenden – (Wort-)Eigenschaften Carrolls; vgl. Hörmann, 1970, p. 205) prägnant und seine konnotative Bedeutung (etwa im semantischen Raum Osgoods) arm ist.

Zu untersuchen ist also, welchen Skalenwert die verfügbaren Graduierungsbegriffe auf dem zu gliedernden Kontinuum haben oder welche Graduierungsbegriffe vorgegebene Skalenwerte am besten ausfüllen; beides kann durch Skalierungsprozeduren geschehen, in denen die Reaktionen mehrerer Urteiler auf die verbalen Stimuli zu eindimensionalen Anordnungen der Beurteilungsobjekte verarbeitet werden.

Allerdings sollten die Informanten zwei Voraussetzungen erfüllen, nämlich zum einen hinreichende (linguistische) Kompetenz und zum anderen ähnliche (Sprach-)Performanz wie jene Personen aufweisen, bei denen die zu entwikkelnden Skalen später angewendet werden sollen.

Liste der 1966 und 1976 skalierten Graduierungsbegriffe

Dimension H Häufigkeit:

oft (a), häufig (b), immer (d), gelegentlich (e), manchmal (g), kaum (j), selten (k), einigemal (l), nie (n), sehr oft (q), sehr selten (r);

(nur 1966:) vielmals (c), mehrfach (f), ein paar Mal (h), wenig (i), oftmals (m), niemals (o), immerzu (p);
(nur 1976:) ziemlich selten (w), ab und zu (x), ziemlich oft (y), meistens (z).

Dimension I Intensität:

ziemlich (a), sehr (b), außerordentlich (e), völlig (h), etwas (i), mittelmäßig (j), einigermaßen (k), nicht (l), wenig (m), kaum (n), gar nicht (o), annähernd (r);

(nur 1966:) ungemein (c), besonders (d), überaus (f), ganz (g), ein wenig (p), schwerlich (q);
(nur 1976:) halbwegs (w), überwiegend (y), teilweise (z).

Dimension W Wahrscheinlichkeit:

gewiß (a), zweifellos (b), vielleicht (d), wahrscheinlich (e), kaum (g), keinesfalls (i), sicher nicht (j), sehr wahrscheinlich (l), mit Sicherheit (m), möglicherweise (q), eventuell (r);

(nur 1966:) sicher (c), unter Umständen (f), schwerlich (h), wohl nicht (k), unter keinen Umständen (n), vermutlich (o), unwahrscheinlich (p);
(nur 1976:) wahrscheinlich nicht (w), ganz sicher (x), ziemlich wahrscheinlich (y), wenig wahrscheinlich (z).

Dimension B Bewertung von Aussagen:

sehr richtig (b), ziemlich richtig (c), etwas richtig (e), annähernd richtig (f), etwas falsch (n), annähernd falsch (o), sehr falsch (s), ziemlich falsch (t);

(nur 1966:) völlig richtig (a), starke Zustimmung (d), schwache Zustimmung (g), mehr richtig als falsch (h), unentschieden (i), ungewiß (j), neutral (k), teils richtig teils falsch (l), weder Zustimmung noch Ablehnung (m), schwache Ablehnung (p), mehr falsch als richtig (q), völlig falsch (r), starke Ablehnung (u), richtig (v), falsch (w);
(nur 1976:) teils/teils (x), etwas dagegen (y), etwas dafür (z), stimmt nicht (A), stimmt wenig (D), stimmt mittelmäßig (G), stimmt ziemlich (H), stimmt sehr (K), trifft wenig zu (M), trifft ziemlich zu (L), trifft gar nicht zu (P), trifft völlig zu (U).

2.2 Sammlung und Vorauswahl von Graduierungsbegriffen

Zunächst wurden mehrere hundert Wörter aus den entsprechenden Duden-Bänden, aus vorhandenen Antwortskalen von Fragebögen, Beobachtungsbögen usw. gesammelt. Aus einer Grobauswahl und Vorsortierung gemäß den 4 vorgegebenen Dimensionen (s. o.) durch 3 Urteiler ergaben sich 4 Serien von 18 + 18 + 18 + 23 = 77 Wörtern, die aus der vorstehenden Liste zu ersehen sind.

2.3 Skalierungskonzept

Um die geeignetsten Begriffe zur Graduierung von Antwortstufen zu finden, wurden 2 Skalierungsansätze benutzt; der erste geht von den Begriffen, der zweite von vorgegebenen numerischen Stufen aus.

2.3.1 Beschreibung von Wörtern durch Zahlen

Im ersten Ansatz (hier Skalierung „WZ" genannt) werden dem Urteiler alle Wörter einer Serie gleichzeitig vorgelegt, der sie entsprechend dem Ausprägungsgrad (von Intensität, Häufigkeit usw.), den sie ausdrücken, in eine 9-stufige „equal appearing interval scale" (THURSTONE, 1928, 1929) von −4 bis +4 einsortieren soll, sodaß den Wörtern Zahlen (nach Transformation zwischen 1 und 9) zugeordnet werden können. (Das Verfahren wurde zuerst von THORNDIKE 1910 benutzt; vgl. GUILFORD, 1954, p. 204ff).

Aus der Auswertung der Antwortverteilungen aller Urteiler lassen sich dann Kennwerte für die Wertigkeit und Prägnanz der Begriffe gewinnen. (Auf methodische Probleme bei diesem Verfahren soll erst bei der Darstellung von Studie II eingegangen werden.)

Um zu große Abstraktheit zu vermeiden und den Urteilern die Aufgabenstellung zu verdeutlichen, sind die zu skalierenden Begriffe im Zusammenhang mit Beispielfragen – nämlich als mögliche Antworten auf diese – eingeführt worden.

2.3.2 Beschreibung von Zahlen durch Wörter

Der zweite Skalierungsansatz (hier Gruppierung „ZW" genannt) geht von einer fünfstufigen, mit den Zahlen −2/−1/0/+1/+2 gekennzeichneten Skala aus; die Informanten sollen je Dimension aus der jeweiligen Begriffserie jene 5 Wörter auswählen, die ihrem Sprachempfinden nach die treffendste Graduierung darstellen, zu den vorgegebenen Zahlen also am besten passen.

Aus den Wahlhäufigkeiten ergeben sich wesentliche Entscheidungshilfen für die Gruppierung von Begriffen zu Antwortskalen.

2.4 Datenerhebung

Die Skalierungen wurden 1966 im Rahmen einer Feldstudie von 2 x 30 Personen durchgeführt (Informanten-Gruppe „B1", Quotenstichprobe aus der Bevölkerung eines Hamburger Außenbezirks; Altersgruppen 21–60 Jahre, aus externen Gründen nur Frauen, 50% Volksschülerinnen; vgl. IRLE & ROHRMANN, 1968).

Auf die möglichen Vorteile studentischer Informanten (z. B. hinsichtlich Instruktionsverständnis und Urteilskonsistenz) wurde also zugunsten einer besseren Annäherung an die primär interessierende Bezugsgruppe – ‚normale' Bevölkerung – verzichtet. Die Erhebung der Urteile geschah in Form eines voll standardisierten Interviews, in dem je Urteiler jeweils 4 von den insgesamt 8 Skalierungsaufgaben (4 Begriffsserien, 2 Skalierungstechniken) in andere Fragen eingekleidet waren.

Die Abfolge der Serien und die benutzten Beispielsätze wurden zwischen den beiden Teilgruppen und innerhalb jedes Erhebungsbogens systematisch variiert. Die Begriffe selbst wurden auf Kärtchen – in jeweils zufällig gemischter Form – vorgelegt und mußten in 5 bzw. 9 Fächer einsortiert werden.

2.5 Ergebnisse

Die resultierenden Daten erbrachten die folgenden statistischen Resultate (Tab. 1A–1D).

Tab. 1A: Häufigkeitsbegriffe – Resultate aus Studie I

Informanten: „B1"	Skalierung WZ				Gruppierung ZW				
	Mittlere Skalenwerte				genannt für Stufe				
Begriffe	M	Md	Mdn	s	−2	−1	0	+1	+2
n nie	1.1	1	1.0	0.25	!!				
o niemals	1.1	1	1.0	0.35					
r sehr selten	2.0	2	1.9	0.85					
j kaum	2.5	2	2.3	0.94					
k selten	2.7	3	2.8	0.88		!			
i wenig	3.1	3	3.0	1.39					
e gelegentlich	4.5	5	4.6	1.53			!		
g manchmal	4.6	5	4.8	1.33			!		
l einigemal	5.0	6	5.6	1.66					
h ein paar Mal	5.2	6	5.3	1.62					
f mehrfach	6.6	7	6.8	1.28					
m oftmals	6.9	7	6.9	1.02					
b häufig	7.2	8	7.5	1.07					
a oft	7.3	7	7.3	0.94				!!	
c vielmals	7.3	8	7.6	1.41					
q sehr oft	8.1	8	8.2	0.82					
p immerzu	8.6	9	9.0	0.85					
d immer	8.6	9	9.0	0.85					!!

Tab. 1B: Intensitätsbegriffe – Resultate aus Studie I

Informanten: „B1"	Skalierung WZ				Gruppierung ZW				
	Mittlere Skalenwerte				genannt für Stufe				
Begriffe	M	Md	Mdn	s	−2	−1	0	+1	+2
o gar nicht	1.0	1	1.0	0.18	!!!				
l nicht	1.4	1	1.0	0.63	!				
m wenig	2.8	2	2.5	1.22		!			
n kaum	3.1	2	2.8	1.48		!			
q schwerlich	3.4	3	3.2	1.79					
p ein wenig	3.7	2	3.4	1.57					
i etwas	4.7	6	5.1	1.71					
j mittelmäßig	5.1	5	5.1	0.85			!!		
r annähernd	5.2	6	5.4	1.42					
k einigermaßen	5.4	6	5.6	1.33					
a ziemlich	6.7	8	7.5	1.80				!	
d besonders	7.8	8	7.8	0.89					
g ganz	8.0	9	8.3	1.07					
f überaus	8.2	9	8.5	1.10					
h völlig	8.2	9	9.0	1.01					
c ungemein	8.4	9	9.0	0.86					
b sehr	8.5	9	9.0	0.78					!!
e außerordentlich	8.7	9	9.0	0.92					!!!

Tab. 1C: Wahrscheinlichkeitsbegriffe – Resultate aus Studie I

Informanten: „B1"	Skalierung WZ				Gruppierung ZW				
	Mittlere Skalenwerte				genannt für Stufe				
Begriffe	M	Md	Mdn	s	−2	−1	0	+1	+2
n unter keinen Umständen	1.0	1	1.0	0.18	!!				
i keinesfalls	1.2	1	1.0	0.48	!				
j sicher nicht	2.0	1	1.7	1.11					
p unwahrscheinlich	2.1	2	2.0	0.97					
g kaum	2.6	2	2.4	1.07		!			
h schwerlich	2.9	3	2.9	1.05					
k wohl nicht	3.0	2	2.8	1.27					
d vielleicht	5.4	5	5.2	0.77			!		
f unter Umständen	5.4	5	5.4	1.25					
q möglicherweise	5.5	5	5.5	0.86					
r eventuell	5.7	5	5.6	1.16					
o vermutlich	5.7	6	5.9	1.27					
e wahrscheinlich	6.5	6	6.5	1.20					
l sehr wahrscheinlich	7.9	8	8.0	0.83					!
b zweifellos	8.1	9	8.4	1.04					
a gewiß	8.2	9	8.4	1.02					
c sicher	8.5	9	9.0	0.63					
m mit Sicherheit	8.5	9	9.0	0.63					!!!

Tab. 1D: Bewertung von Begriffen – Resultate aus Studie I

Informanten: „B1"	Skalierung WZ				Gruppierung ZW				
	Mittlere Skalenwerte				genannt für Stufe				
Begriffe	M	Md	Mdn	s	−2	−1	0	+1	+2
r völlig falsch	1.0	1	1.0	0.10	!!				
s sehr falsch	1.2	1	1.0	0.59					
u starke Ablehnung	1.4	1	1.0	0.62					
w falsch	1.8	1	1.7	0.87					
t ziemlich falsch	2.3	3	2.4	0.80					
o annähernd falsch	3.0	3	3.0	0.81		!			
p schwache Ablehnung	3.4	4	4.0	0.77					
q mehr falsch als richtig	3.5	4	3.5	0.94					
n etwas falsch	3.8	4	3.7	1.38					
k neutral	5.0	5	5.0	0.00					
l teils richtig teils falsch	5.0	5	5.0	0.26					
j ungewiß	4.8	5	5.2	0.94					
m weder Zust. noch Abl.	4.9	4	5.0	0.37			!		
i unentschieden	5.3	5	5.3	0.98			!!		
e etwas richtig	6.0	6	6.0	0.67					
g schwache Zustimmung	6.3	6	6.0	0.60					
h mehr richtig als falsch	6.5	6	6.4	0.82					
f annähernd richtig	6.6	7	6.7	0.81					!
c ziemlich richtig	7.0	7	7.0	0.67					
v richtig	8.1	9	8.3	1.06					
d starke Zustimmung	8.5	9	9.0	0.68					
b sehr richtig	8.7	9	9.0	0.54					!
a völlig richtig	8.9	9	9.0	0.40					!!

2.5.1 Wertigkeit und Prägnanz der Begriffe

Die Skalenposition und damit die inhaltliche Wertigkeit der Begriffe ergibt sich aus Maßen der zentralen Tendenz, ihre Prägnanz aus Streuungsmaßen.

In Tab. 1A bis 1D (linker Teil) sind entsprechende Ergebnisse aus der Skalierung „WZ" zusammengestellt, nämlich arithmetisches Mittel, Median und Modus (wobei die – streng genommen vorzuziehenden – Mdn-Werte nur unwesentlich von den M-Werten abweichen) sowie die Standardabweichungen für die 77 untersuchten Begriffe. Die Betrachtung der Daten zeigt, daß die Urteilsstreuung zu den Enden des Kontinuums hin abnimmt, extreme Begriffe also überwiegend prägnanter erscheinen (allerdings ist daran wohl auch ein „ceiling"-Effekt beteiligt), doch gilt dies auch für einige Begriffe, die explizit eine Mittelposition ausdrücken sollen.

Abb. 2 verdeutlicht das am Beispiel von 5 Häufigkeitsbegriffen: Während „nie" und „immer" von nahezu allen Informanten gleich interpretiert werden, wird ein unbestimmter Begriff wie „gelegentlich" in 5 verschiedene Urteilskategorien eingeordnet.

Zur Bewertung der Streuungsdaten kann man diese mit der maximalen Variabilität vergleichen: Bei Gleichverteilung der Urteile ergäbe sich s = 2.63.

Daß die Streuungen teilweise relativ groß sind, weist allerdings nicht nur auf die Unbestimmtheit der Graduierungswörter, sondern ist auch Ausdruck der Unsicherheit der Informanten gegenüber dem geforderten Urteilsprozeß. Tatsächlich hatten einige der angesprochenen Personen intellektuelle Schwierigkeiten (teils mit den Begriffen an sich, teils mit der Art der Skalierung), weil sie verständlicherweise mit solchen linguistischen Bewertungen nicht vertraut waren; in einigen Fällen mußte die Befragung deshalb abgebrochen werden.

2.5.2 Wahlhäufigkeit von Antwortstufen

Die Fünfergruppierungen sind so ausgewertet worden, daß je vorgegebener Zahlenstufe gezählt wurde, wie häufig ihr bestimmte Begriffe zugeordnet wurden. In den Tab. 1A bis 1D (rechte Hälfte) sind die meistgenannten Begriffe markiert, wobei „!" mindestens ¼, „!" mindestens ⅓ und „!!!" mindestens ½ aller Nennungen bedeutet.

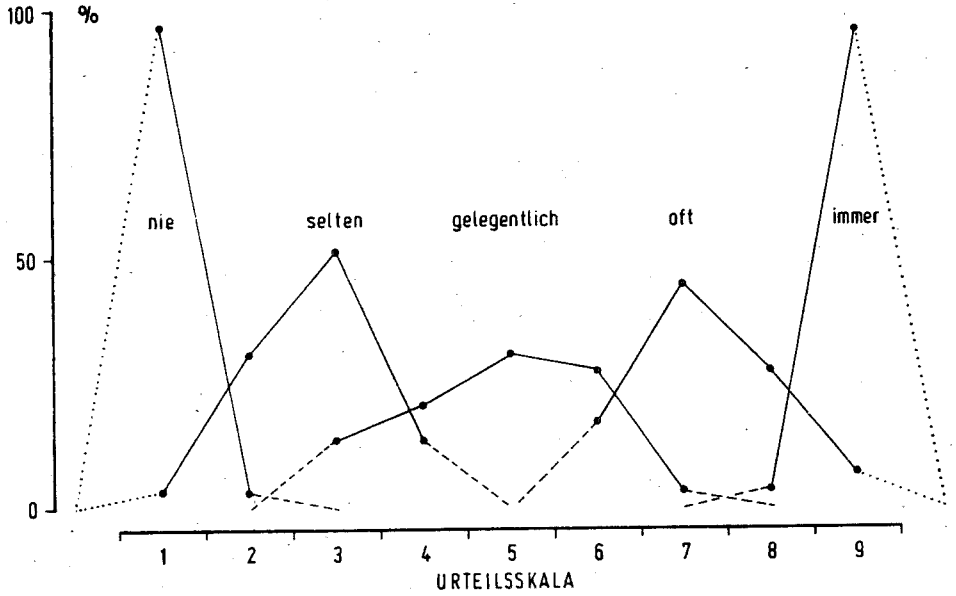


Abb. 2: Urteilsverteilungen für 5 Häufigkeitsbegriffe (Informantengruppe „B1", N = 29).

Zwar divergieren die Gruppierungen sehr stark (es standen ja auch 18 bzw. 23 Alternativen zu Gebote), doch heben sich bestimmte Begriffe – besonders zur Bezeichnung von „–2" und „+2" heraus.

3. Ausarbeitung und Anwendung von Antwortskalen

3.1 Entwicklung von Antwortskalen für Fragebögen

Das Ziel war, für die Anwendung in einem Feldforschungsprojekt möglichst eindeutige und verständliche Skalen zu entwickeln und diese darum numerisch *und* verbal *und* grafisch zu kennzeichnen. Dazu mußten vor allem 3 Entscheidungen gefällt werden: über die Zahl der Stufen, deren Bezeichnung und die formale Gestaltung.

3.1.1 Numerische Abstufung

Bei der Festlegung der Stufenzahl geht es um einen Kompromiß zwischen zwei entgegengesetzten Zielen, nämlich so fein wie möglich zu messen und andererseits nicht mehr Graduierungen vorzugeben, als den Differenzierungsmöglichkeiten und -wünschen der Befragten oder Urteiler entspricht. Aus psychologischer Sicht (z. B. CRONBACH, 1964; NUNNALLY, 1970) oder informationstheoretischer Sicht (z. B. MILLERS „magical number 7 plus or minus 2", 1956) liegen zwischen 5 und 9 Stufen nahe; in demoskopischen Umfragen werden – im Blick auf das bevorzugte Antwortverhalten der Zielgruppe – häufig nur 2stufige (Ja/Nein) oder 3stufige Antwortvorgaben verwendet.

Aufgrund eigener praktischer Efahrungen – vor allem in der Feldforschung – und informeller Erkundungen bei den Informanten der Studie I erschien eine 5stufige Antwortskala als die beste Lösung. Sie wurde zunächst durch die Zahlen –2/–1/0/+1/+2 benannt; zur neutralen Kategorie kommen also je zwei Grade von „Ja" oder „Nein" bzw. „Gegebensein" oder „Nichtgegebensein" hinzu. Aus psychometrischen und praktischen Gründen – Vorzeichenverwechslungen und Ablochprobleme – ist allerdings später die numerische Abstufung 1/2/3/4/5 benutzt worden, die zudem eher als äquidistant erlebt wird (vgl. dazu GUILFORD 1954, S. 264) und darum die Metrik der Begriffe besser verdeutlichen dürfte.

3.1.2 Verbale Bezeichnung

Die Auswahl der Wörter für die Antwortstufen stellt das eigentliche Problem dar. Da es zunächst um Äquidistanz geht, kamen z. B. Begriffe mit den Skalenpositionen 1/3/5/7/9 oder 2/3.5/5/6.5/8 in Frage.

Für die Häufigkeitsbegriffe – vgl. Tab. 1A – hieße das im ersten Fall „nie/wenig/einigemal/oftmals/(immer)" und im zweiten Fall „sehr selten/(wenig)/einigemal/mehrfach/sehr oft". Bei der Bewertung von Aussagen – vgl. Tab. 1D – ergäbe sich für die „richtig/falsch"-Begriffe „völlig falsch/annähernd falsch/teils ... teils .../(annähernd richtig)/völlig richtig" bzw. „falsch/mehr falsch als richtig/teils ... teils .../mehr richtig als falsch/richtig".

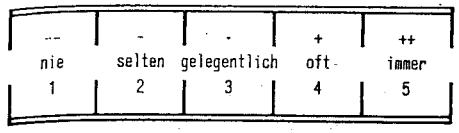
Das erste Muster führt also zu extremen, das zweite Muster zu relativ schwachen ‚Endstufen'. Um einerseits Stufen zu vermeiden, die bei der Beantwortung inhaltlicher Fragen zu selten gewählt werden, weil sie zu extrem sind, und andererseits hinreichende Abstände zwischen den Stufen zu haben, ist (soweit möglich) als Abstufungs-Muster 1.5/3.25/5/6.75/8.5 angestrebt worden.

Die schließliche Entscheidung über die Begriffe sollten 5 Gesichtspunkte berücksichtigen:

- die Skalenpositionen (mit dem ebengenannten Ziel),
- die Prägnanz (vgl. die Streuung in Tab. 1),
- die Geläufigkeit von Wörtern (u. a. aus den Angaben über die Wahlhäufigkeiten in Tab. 1 zu ersehen),
- die Eignung der Begriffe zur Kombination untereinander und mit Adjektiv- oder Verbformen (z. B. widerspricht „überwiegend wahrscheinlich" oder „annähernd wahrscheinlich" dem ‚Sprachgefühl', nicht hingegen „ziemlich wahrscheinlich"; ähnliches gilt bei „oft", usw.),
- die Praktikabilität besonders in mündlichen Befragungen (dies legt kurze Begriffe nahe).

Drei der endgültigen Skalen sind in Abb. 3 wiedergegeben. Dazu einige Erläuterungen:

Skala H: Häufigkeit



Skala I: Intensität

Skala B: Bewertung von Aussagen

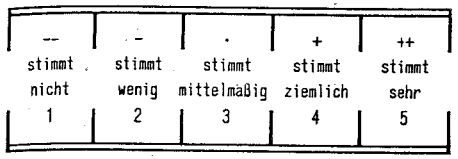


Abb. 3: Wiedergabe der entwickelten Antwortskalen.

Häufigkeit: Den Ausschlag gaben die Daten aus der Gruppierung "ZW" (Tab. 1A). Statt „gelegentlich" wäre auch „manchmal", statt „oft" auch „häufig" geeignet (allerdings sollte ein Begriff, der der Benennung der Dimension – Häufigkeit – entspricht, vermieden werden); in beiden Fällen ist jedoch die Distanz um 2.6 und damit größer als erwünscht. Zu den Extremstufen-Begriffen „nie" versus „sehr selten", „immer" versus „sehr oft" siehe die Diskussion oben.

Intensität: Problematisch, nämlich nicht hinreichend eindeutig, ist „ziemlich", doch war im Skalenbereich zwischen „5" und „8" keine Alternative verfügbar (vgl. Tab. 1B); „wenig" wurde gegenüber „kaum" wegen der geringeren Streuung vorgezogen. Auf die Extreme „außerordentlich", „gar nicht" ist im Blick auf kombinierte Antwortskalen verzichtet worden. Die gewählte Reihe war gleichzeitig zur Kombination mit Adjektiven (z. B. gut, schlecht, wahrscheinlich) oder Verben (z. B. stimmt, gefällt) vorgesehen.

Wahrscheinlichkeit: Von den Skalierungsdaten her (Tab. 1C) kamen die Begriffe „keinesfalls", („kaum" oder „wohl nicht"), „vielleicht", („wahrscheinlich" oder „sehr wahrscheinlich") und „sicher" in Betracht. Die 2. und 4. Stufe sind jedoch durch andere – leider in Studie I nicht direkt skalierte – Begriffe ersetzt worden, nämlich „wahrscheinlich nicht" und „ziemlich wahrscheinlich" (außerdem wurde „sicher" später in „ganz sicher" verändert); diese (eher subjektive) Entscheidung sollte zu einer sprachlich besser abgestimmten Reihe führen.

Bewertung von Aussagen: Aus Tab. 10 ergibt sich eine Bevorzugung der „richtig/falsch"-Reihe in der Form „völlig falsch", „annähernd falsch", („unentschieden" oder „teils richtig/teils falsch"), „annähernd richtig", „völlig richtig"; alternativ können auch „ziemlich" statt „annähernd" und „sehr" statt „völlig" eingesetzt werden. Auch die „Zustimmungs-/Ablehnungs"-Reihe erweist sich als akzeptabel.

Für bestimmte Anwendungen in sozialpsychologischen Befragungen, nämlich die Beantwortung von ‚statements' (besonders solchen in Ich-Form) und die Zuordnung von Attributen zu Beurteilungsobjekten, ist allerdings keine dieser beiden Möglichkeiten gewählt worden: Die „richtig/falsch"-Reihe erweckt beim Befragten zu sehr den Eindruck, es gäbe eine findbare ‚Wahrheit' (wie in einem Wissenstest), während tatsächlich die persönliche Meinung interessiert; die „Zustimmungs-/Ablehnungs"-Reihe ist in der Anwendung im Interview sehr unhandlich, und sie gestattet nicht, die Graduierungsbegriffe in Satzmuster wie „Finden Sie es ..., daß ..." einzubauen.

Als Ausweg ist aus der Kombination des Terminus „Zustimmen" mit den Intensitäts-Graduierungen die Antwortskala „stimmt nicht – ... wenig – ... mittelmäßig – ... ziemlich – ... sehr" gebildet worden (s. Abb. 3).

Die Häufigkeits- und die Intensitätsbegriffe ermöglichen zahlreiche weitere Antworten bzw. Urteilsskalen; außer Kombinationen mit verschiedenen Adjektiven wurden folgende Versionen eingesetzt:

- „paßt nicht ... paßt sehr" für die Zuordnung von Eigenschaftswörtern zu Begriffen (z. B. unipolare Profile);
- „trifft nie zu ... trifft immer zu" für die Beurteilung von Befindlichkeiten (z. B. psychosomatischen Symptomen);
- „sehr dafür – etwas dafür – unentschieden – etwas dagegen – sehr dagegen" zur Bewertung von Standpunkten usw.

Eine wichtige Grundlage solcher Anwendungen liegt darin, daß der Graduierungseffekt entsprechender Partikel bei unterschiedlichen Inhalten gemäß dem „CLIFFschen Gesetz" (CLIFF, 1959) relativ konstant ist.

Linguistisch gesehen handelt es sich dabei meist um Adverben, die „adverbial" (vor Verben) oder „adnominal" (vor Adjektiven bzw. syntaktisch Äquivalenten) stehen, ohne satzgliedwertig zu sein (vgl. ROHRMANN, 1974, p. 187f.).

3.1.3 Grafische Gestaltung

Wie Abb. 3 zeigt, werden die 5 Stufen als gleichgroße Fächer repräsentiert und die Ziffern gleichabständig angeordnet, um auch optisch

auf die Wahrnehmung einer äquidistanten Skala hinzuwirken. (Andere Möglichkeiten werden z. B. bei CHAMPNEY & MARSHALL, 1939; GUILFORD, 1954; CLAUSS, 1968 diskutiert). Zusätzlich sind die Randstreifen farbig markiert worden, um in der Kommunikation mit Befragten usw. die verschiedenen Antwortskalen schnell und eindeutig benennen zu können („rote Skala", „gelbe Skala", usw.).

3.2 Einsatz in Felduntersuchungen

Die dargestellten Antwortskalen sind seit 1966 in zahlreichen Fragebogen-Untersuchungen und Skalierungsexperimenten eingesetzt worden (u. a. IRLE & ROHRMANN, 1968; ROHRMANN, 1973; BUCHTA & KASTKA, 1974; SCHÜMER-KOHRS & SCHÜMER 1974; ROHRMANN, 1976; KASTKA, 1977; EICHNER, 1977; GUSKI et al., 1978; FROGNER, in Vorb.).

Im Falle von *mündlichen* Interviews liegt dem Befragten die jeweilige Antwortskala als Karte vor, während ihm die Fragen (oder statements oder Eigenschaftslisten usw.) teils vorgelesen, teils als gesonderte Liste vorgelegt werden; der Befragte kann wahlweise mit den Begriffen oder den Zahlen antworten. Der Interviewer codiert im allgemeinen direkt die Zahlen.

Als Beispiel sei eine Instruktion für eine Zufriedenheits-Bewertung aus einem Fragebogen über Umweltbelastungen angeführt:

„Ich möchte Sie nun fragen, wie zufrieden Sie mit einer Reihe von Dingen sind, die ihre Wohnsituation betreffen, z. B. mit Ihrer Wohnung.
Gelbe Skala vorlegen!
Uns interessiert bei diesen Fragen nicht einfach, *ob* Sie zufrieden sind oder nicht, sondern wir möchten es etwas genauer wissen, nämlich: in welchem *Grade* Sie zufrieden oder nicht zufrieden sind. Wir haben deshalb auf dieser *gelben* Skala 5 Stufen der Zufriedenheit vorgesehen
(zeigen):
nicht (1), wenig (2), mittelmäßig (3), ziemlich (4), sehr (5).
Bitte wählen Sie von diesen 5 Stufen zunächst diejenige Antwort aus, die Ihrer Zufriedenheit mit Ihrer Wohnung entspricht.
(Skala u. U. wiederholt erklären). Liste x vorlegen und durchgehen:
(„und wie zufrieden sind Sie mit ...")."

Bei *schriftlichen* Befragungen o. ä. werden die Antwortstufen teils abgesetzt, teils in den Fragenwortlaut integriert (in jedem Fall aber grafisch herausgehoben) dargeboten. –

Wieweit sich die entwickelten Antwortskalen empirisch bewährt haben, wird in Abschnitt 5 näher erörtert werden.

4 Studie II (1976): Replikation der Item-Skalierungen

4.1 Zweck der Studie

Die erste Skalierungstudie von 1966 ist 1976 aus mehreren Gründen wiederholt und dabei ausgeweitet worden,

- um zu prüfen, ob die zunächst gefundene Charakteristik der Begriffe replizierbar ist;
- um die Beurteilungen von ‚normaler' Bevölkerung mit jenen einer vorgebildeten Gruppe – hier Studenten der Soziologie/Psychologie – vergleichen zu können;
- um die ursprüngliche Begriffsserie inhaltlich ergänzen zu können.

Bei dieser Überprüfung, Aktualisierung und Erweiterung der Befunde von 1966 (die im Rahmen einer sozialpsychologischen Feldstudie möglich wurde und teils auf diese bezogen war) sollte aus Gründen der Vergleichbarkeit das Hauptexperiment in möglichst gleichartiger Form repliziert werden.

4.2 Neuauswahl der Begriffe

Zunächst wurden einige Begriffe hinzugenommen, die in bereits eingesetzten (eigenen oder fremden) Antwortskalen vorkommen, aber 1966 nicht einbezogen waren (dies betrifft besonders die Ausdrücke zur Bewertung von Aussagen); außerdem wurden mehr Kombinationen mit bestimmten Intensitätsstufen verwendet. Da mehr als 15–20 Begriffe je Dimension für die Informanten kaum noch überschaubar sind, wurden einige entbehrlich scheinende Begriffe der ursprünglichen Serien gestrichen. (Vgl. im einzelnen die obenstehende Liste.)

4.3 Datenerhebung

Studie II wurde in zwei Teilstudien mit verschiedenen Beurteilergruppen aufgeteilt.

4.3.1 Urteiler-Gruppe B2

Als Informanten dienten 29 Personen einer Quotenstichprobe (Bevölkerung Hamburger Stadtbezirke, Altersgruppen 21–61 Jahre, 52% Frauen, 55% Volksschüler), die im Zusammenhang mit einer Lärmuntersuchung (FINKE et al., 1977) zur Mitarbeit gewonnen werden konnten.

Die Urteile wurden analog der Studie I in Form eines vollstandardisierten Interviews erhoben, wobei aber diesmal alle Informanten alle Begriffe zu beurteilen hatten. Die eigentliche Aufgabenstellung und der technische Ablauf (Vorlage von Kärtchen, wechselnde Abfolge der Aufgaben usw.) wurden weitgehend repliziert (vgl. 2.4), doch sind die einzelnen Instruktionen verändert worden, wo dies zum besseren Verständnis notwendig erschien.

4.3.2 Informantengruppe S2

Wie Studie I gezeigt hatte, stellt der geforderte Urteilsprozeß eine relativ große Anforderung (und für manche Informanten wohl auch eine Überforderung) dar, was für einen Teil der Urteilsstreuung und einige Urteilsinkonsistenzen mitverantwortlich sein dürfte. Deshalb wurde 1976 zum Kontrast zusätzlich eine akademisch vorgebildete Gruppe (mit psychometrischen Grundkenntnissen) herangezogen, nämlich Studenten.

Bei inhaltlich identischer Aufgabenstellung geschah die Erhebung in Form eines vollstandardisierten Gruppenexperiments mit verlesenen Instruktionen und schriftlicher Beantwortung (doch wurden die Begriffe ebenfalls auf Kärtchen vorgelegt und waren auf entsprechenden Tafeln einzusortieren). Teilnehmer waren 33 Hamburger Studenten (Ort: Seminar für Sozialwissenschaften; Durchschnittsalter: 23 Jahre; 60% weiblich; Herkunft aus verschiedenen Bundesländern).

4.4 Ergebnisse der Item-Skalierung

Es resultierten 8 Datensätze; die statistischen Ergebnisse zur Skalierung und Gruppierung der Begriffe sind in den Tab. 2A bis 2D für die Informantengruppen „B2" und „S2" zusammengefaßt (wobei die Wiedergabe von Mdn und Md entbehrlich schien).

Betrachtet man die Resultate zu Wertigkeit und Prägnanz der Begriffe, so wird wiederum eine relativ große Urteilshomogenität (im Einklang mit linguistischen Erwartungen) deutlich; bei den meisten Begriffen ist die Urteilsstreuung relativ gering (s um 1.0 oder geringer), und mehrere Begriffe (solche am Skalenende) wurden sogar von allen Informanten (teils wegen des genannten „ceiling"-Effekts) identisch eingestuft.

Vergleicht man die beiden Gruppen „B2" und „S2", so wird die größere Homogenität der studentischen Informanten deutlich (wofür deren im Schnitt besseres Aufgabenverhältnis – und eventuell die Vorkenntnis über Skalen – sicher mitbestimmend war).

Die Skalenwerte fallen weitgehend ähnlich aus (maximale Differenz 0.9), doch sind einige Unterschiede aufgrund der geringen Urteilsstreuung signifikant (in der Tab. mit * bezeichnet; Signifikanzniveau: $p = 0.05$; t-Tests oder bei Varianzinhomogenität U-Test). Dies zeigt sich auch in den Rangplätzen der Begriffe: die Rangkorrelationen zwischen den Mittelwerten aus „B2" und „S2" betragen für die 4 Begriffsserien 0.98/0.97/1.00/0.99 (SPEARMANS Rho; $p \leqslant 0.01$).

Der Einfluß der Informanten-Gruppe auf die Hierarchie der Begriffe ist also insgesamt nicht sehr groß und offenbar nicht systematisch (in linguistischer Hinsicht), so daß die Resultate im Prinzip zusammengefaßt interpretiert werden können.

Vor der Anwendung der erzielten Skalenwerte bleiben zwei psychometrische Fragen zu klären, nämlich ob die Urteile unidimensional sind und auf einer äquidistanten Skala liegen. Beides ist für die hier vorliegenden konkreten Daten nicht ohne weiteres mit statistischen Mitteln prüfbar.

Zur *Dimensionalitätsfrage* würden sowohl Faktorenanalysen der errechenbaren Korrelationsmatrizen zwischen den Begriffen wie die Anwendung etwa des

Tab. 2A: Häufigkeitsbegriffe – Resultate aus Studie II

Informanten:	„B2"		„S2"		„B2" + „S2"				
	Skalierung WZ				Gruppierung ZW				
Begriffe	M	s	M	s	−2	−1	0	+1	+2
n nie	1.0	0.00	1.0	0.00	!!!				
r sehr selten	2.0	0.94	1.8	0.39					
j kaum	2.6	1.35	2.9	0.68					
k selten	2.8	0.74	3.1	0.66		!!			
w ziemlich selten	2.8	0.71	3.1	0.83					
x ab und zu	4.1	1.06	4.5	0.67			!		
g manchmal	4.7	1.14	4.5	0.62					
e gelegentlich	4.7	1.16	4.5	0.62			!		
l einigemal	4.9	0.82	4.6	1.06					
b häufig	6.6	0.94	6.9	0.83				!	
y ziemlich oft*	7.1	0.88	6.5	0.67					
a oft*	7.3	1.07	6.7	0.65				!	
z meistens	7.5	1.09	7.8	0.50					
q sehr oft	8.1	0.67	8.0	0.43					
d immer	8.9	0.58	9.0	0.00					!!!

Tab. 2B: Intensitätsbegriffe – Resultate aus Studie II

Informanten:	„B2"		„S2"		„B2" + „S2"				
	Skalierung WZ				Gruppierung ZW				
Begriffe	M	s	M	s	−2	−1	0	+1	+2
o gar nicht	1.0	0.00	1.0	0.00	!!!				
l nicht	1.5	0.51	1.5	0.57					
m wenig	2.6	0.68	2.7	0.57		!			
n kaum	2.8	0.64	2.5	0.57		!!			
i etwas	3.7	1.22	3.7	0.96					
k einigermaßen	4.9	1.22	4.7	1.13					
w halbwegs	5.0	0.73	5.0	0.61					
z teilweise*	5.0	1.03	4.4	0.87					
j mittelmäßig	5.2	1.07	5.1	0.66			!!		
r annähernd	5.4	1.40	5.3	1.33					
a ziemlich	6.0	1.14	6.4	0.90				!	
y überwiegend	6.9	0.95	7.2	0.68				!!	
b sehr	8.2	0.76	8.1	0.74					
h völlig	8.5	0.57	8.7	0.57					!!!
e außerordentlich	8.6	0.74	8.6	0.60					!

KRUSKAL-Verfahrens der multidimensionalen Skalierung (1964) auf entsprechende Matrizen von Ähnlichkeitsmaßen zu ungeeigneten Resultaten führen. Der Grund liegt darin, daß übliche parametrische wie nichtparametrische Zusammenhangsmaße die hier entscheidenden Mittelwertsunterschiede nicht reflektieren (die verrechnete Varianz bzw. Kovarianz also im wesentlichen Zufallsstreuung ausdrückt); bildet man aber (ex post) Distanzmaße auf der Basis der Urteilsmittelung, ist die Dimensionalität nicht mehr prüfbar.

Offensichtlich hätte es eines anderen Skalierungskonzeptes bedurft, um außer den primär interessierenden Skalenpositionen auch valide (Un-)Ähnlichkeitskoeffizienten bestimmen zu können (wobei die verfügbaren psychometrischen Techniken – z. B. im Sinne eines Paarvergleichs nach THURSTONE – allerdings wegen der Vielzahl der notwendigen Vergleichsurteile einen hohen Erhebungsaufwand erfordern, der bei 65 oder 77 Begriffen in dieser Studie nicht realisierbar war).

Ungeachtet dieses statischen Problems ist die Annahme der Mehrdimensionalität für die hier vorliegenden

Tab. 2C: Wahrscheinlichkeitsbegriffe – Resultate aus Studie II

Informanten:	„B2"		„S2"		„B2" + „S2"				
	Skalierung WZ				Gruppierung ZW				
Begriffe	M	s	M	s	−2	−1	0	+1	+2
i keinesfalls	1.1	0.26	1.0	0.00	!!!				
j sicher nicht	1.3	0.48	1.3	0.53	!!				
g kaum	2.4	0.87	2.7	0.80		!			
w wahrscheinlich nicht	2.8	1.43	2.3	0.59					
z wenig wahrscheinlich	2.8	0.60	2.9	0.90		!!			
r eventuell	4.6	1.02	4.5	0.91			!		
d vielleicht	4.9	1.07	4.7	0.74			!!		
q möglicherweise	5.4	1.09	4.9	0.96					
e wahrscheinlich*	5.7	0.92	6.5	0.75				!	
y ziemlich wahrscheinlich	6.3	1.42	6.7	0.84				!!	
l sehr wahrscheinlich*	7.4	0.98	7.9	0.50					
a gewiß	8.1	0.79	8.4	0.71					
b zweifellos*	8.4	0.83	8.8	0.44					
m mit Sicherheit	8.4	0.90	8.8	0.50					!
x ganz sicher*	8.4	0.91	8.9	0.29					!!

Tab. 2D: Bewertung von Aussagen – Resultate aus Studie II

Informanten:	„B2"		„S2"		„B2" + „S2"				
	Skalierung WZ				Gruppierung ZW				
Begriffe	M	s	M	s	−2	−1	0	+1	+2
s sehr falsch	1.1	0.44	1.3	0.47	!				
A stimmt nicht	1.3	0.67	1.3	0.48	!				
P trifft gar nicht zu	1.3	0.77	1.0	0.00	!!				
t ziemlich falsch	2.8	0.90	2.6	0.61					
o annähernd falsch	2.9	0.83	2.8	0.71					
M trifft wenig zu	2.9	0.94	3.0	0.73		!			
D stimmt wenig	3.1	1.22	2.9	0.88					
n etwas falsch*	3.3	0.84	4.1	1.10					
y etwas dagegen	3.6	1.02	4.0	1.05					
x teils-teils	4.9	0.67	5,1	0,29			!!!		
G stimmt mittelmäßig	5.1	0.74	5.5	0.83					
e etwas richtig	5.4	1.30	5.6	1.20					
z etwas dafür	5.7	1.04	5.7	0.96					
H stimmt ziemlich*	6.7	1.32	7.3	0.64					
f annähernd richtig*	6.6	1.15	7.2	0.71					
c ziemlich richtig	7.1	0.69	7.4	0.66					
L trifft ziemlich zu	7.1	0.75	7.3	0.63					
b sehr richtig	8.6	0.57	8.8	0.44					
K stimmt sehr	8.6	0.57	8.6	0.50					
ü trifft völlig zu	8.9	0.24	8.9	0.24					!!!

4 Begriffsserien wenig plausibel, weil die begrifflichen Zieldimensionen sehr eindeutig sind und die Art der Vorauswahl kaum Chancen dafür eröffnet. (Soweit andere Studien über Graduierungsbegriffe – z. B. COHEN et al., 1969; WEGENER, 1976 – darauf überhaupt eingehen, ergeben sich keine Hinweise auf Mehrdimensionalität.)

Was nun die *Äquidistanzfrage* betrifft, so wird für Ergebnisse von Skalierungen mit dem Verfahren der „equal appearing intervals" teils eine leichte Verzerrung an den Skalenenden – jedenfalls im Vergleich zu Paarvergleichsresultaten – unterstellt. (Dies wurde überwiegend an ‚statements' zur Attitüden-Messung – vgl. z. B. KELLEY et al., 1955 – untersucht; SIXTL

(1967, p. 254) kommt zu dem Schluß, daß Urteiler keine gleichen Intervalle realisieren). Als Weg zu präziseren Skalenwerten mit Intervallskalen-Niveau haben SAFFIR 1937 (in Anlehnung an ein Konzept THURSTONEs) und andere (vgl. GUILFORD, 1954, p. 223 ff.; SIXTL, 1967, p. 240 ff.) das Verfahren der „successive categories" entwickelt. Die Berechnung derartiger „nachträglich bestimmter Abstände" (probeweise wurden ein Verfahren nach GUILFORD, 1954 und ein in VELDMAN, 1967 dargestelltes Programm gerechnet) führt allerdings zu keinen wirklich brauchbaren Ergebnissen. Die Ursache dafür liegt in den fehlenden Voraussetzungen: Die Urteilsverteilungen sind zumindest für die extremeren Begriffe weder normal noch varianzhomogen, und einige Begriffe erzeugen überhaupt keine Urteilsstreuung (vgl. Tab. 1 und 2). Schließt man diese Begriffe aber aus, kann keine Gesamtskala mehr errechnet werden. Von der Anwendung der Methode (deren Metrik ja an die Variabilität der Urteile gebunden ist) mußte deshalb zunächst abgesehen werden.

Auch der z. B. von HOFSTÄTTER & WENDT (1966) gegangene Weg, für die in Antwortskalen bereits eingesetzten Begriffe über eine Normalisierungstransformation ‚ex post' treffende Skalenwerte zu finden, erscheint unbefriedigend: er beruht ja auf der – nicht prüfbaren – Annahme, die ‚wahre' Verteilung der entsprechenden Antworten sei normal. Wendet man aber die Prozedur auf anormale (links- oder rechtsschiefe, bimodale) Verteilungen an, die tatsächliches Urteilsverhalten (etwa soziale Normen oder Polarisierungen) reflektieren, müßte eine Normalisierung zu falschen Skalenwerten führen.

Auf eine Weiterverarbeitung bzw. Reanalyse der Daten über das hier benutzte THURSTONE-Konzept hinaus ist darum vorerst verzichtet worden; es kommt hinzu, daß bei den zu treffenden Entscheidungen (vgl. 4.7) gerade für die kritischen Endkategorien wie „nie" und „immer" oder „nicht" und „sehr" ohnedies kaum Alternativen zu Gebote standen. Dies ergibt sich auch aus der Auswertung der Gruppierungen (für Studie II im rechten Teil von Tab. 2 wiedergegeben; „B2" und „S2" zusammengefaßt) und bei der Berücksichtigung der Prägnanz der Begriffe als Entscheidungskriterium.

4.5 Vergleich mit den Resultaten aus Studie I

Vergleicht man die Resultate von Studie I und II anhand der Informanten-Gruppen „B1" und „B2", so ergeben sich überwiegend gleiche Bewertungen der Graduierungsbegriffe (Differenzen meist unter 0.5, max. 1.1; vgl. Tab. 1 und 2); die Rangreihen der in beiden Studien skalierten Begriffe sind nahezu identisch (SPEARMANS Rho = 0.98, 0.99, 0.96, 1.00 für die 4 Serien; $p \leq 0.01$). In Abb. 4A (Häufigkeit) und 4B (Intensität) sind für die 5 in Antwortskalen eingegangenen Begriffe die Resultate aus allen 3 Datensätzen (B1, B2, S2) gegenübergestellt.

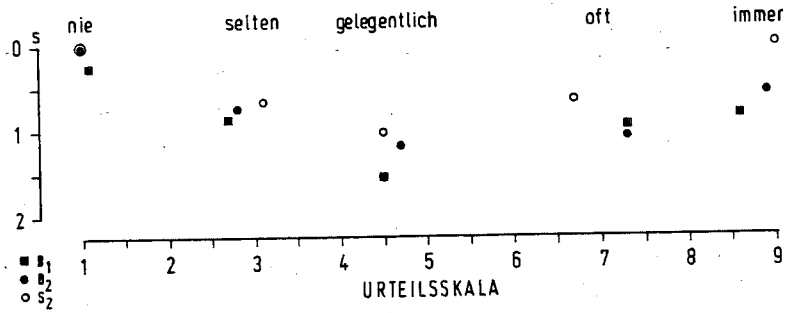


Abb. 4A: Vergleich der mittleren Skalenwerte (M) und Streuungen (s) für Häufigkeitsbegriffe laut 3 Informantengruppen.

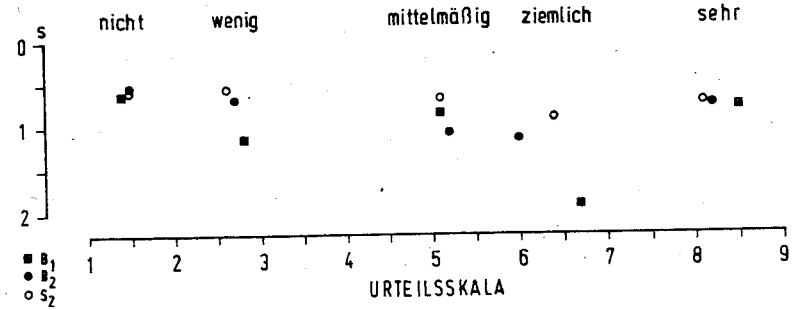


Abb. 4B: Vergleich der mittleren Skalenwerte (M) und Streuungen (s) für Intensitätsbegriffe laut 3 Informantengruppen.

(Die W- und B-Begriffe sind nicht dargestellt, weil sie in Studie I nur teilweise untersucht worden waren.) Die Grafiken zeigen, daß die Informanten-Gruppen von 1966 und 1976 zu weitgehend gleichartiger Auffassung über diese Begriffe neigen (ausgenommen „ziemlich" mit signifikantem Unterschied zwischen B1 und B2), wobei jedoch die Distanzen zwischen den Skalenwerten bei der Häufigkeitsserie eher befriedigen können als bei der Intensitätsserie (in der „wenig" und besonders „ziemlich" 1976 zu niedrig liegen).

Erwähnenswert ist schließlich, daß auch die Unterschiedlichkeit der Serien in Studie I und II – nur etwa die Hälfte der Begriffe ist ja gleichgeblieben – keinen wesentlichen Einfluß hat (vgl. Tab. 1 und 2); die Beurteilung der Begriffe scheint also in diesem Falle einigermaßen kontextstabil zu sein.

An dieser Stelle sei ferner auf die wenigen Befunde aus anderen Untersuchungen hingewiesen, die seit Studie I publiziert wurden.

Bei einer Überprüfung des „CLIFFschen Gesetzes" (s. o.) erhielt KRISTOF (1966, p. 26) Daten zur Modifizierungswirkung von 9 Adverbien; 3 davon gehören zu den hier untersuchten Begriffen („etwas" = 0.47, „ziemlich" = 0.86, „sehr" = 1.45). Diese Kennwerte sind zwar numerisch nicht vergleichbar, deuten aber auf eine etwas niedrigere Position von „ziemlich", als aus Studie I resultierte.

SIXTL, (1967, p. 315) ließ Studenten indefinite Mengenangaben – Begriffe zwischen „nichts" und „alles" – in Prozent ausdrücken. Von den mitgeteilten 10 Wörtern zählt nur „wenig" (Wert 15.2,,%") zu den hier behandelten Begriffen (ohne allerdings als Intensitätsgraduierung gemeint zu sein).

In einer Arbeit über Verhaltensbeschreibungen nennen COHEN et al. (1969, p. 14) Skalenwerte für 11 Häufigkeitsadverbien, die auf der Basis sukzessiver Kategorien bestimmt wurden; 5 davon („selten" = 2.3, „gelegentlich" = 5.2, „manchmal" = 5.7, „häufig" = 11.1, „oft" = 13.3) sind vergleichbar. Sieht man von der andersartigen (nicht näher erläuterten) Metrik ab, so scheint die Relation der 5 Begriffe zueinander ähnlich zu sein wie in Tab. 1A bzw. 2A.

ESSER (1970, p. 214) führt Skalenwerte und Urteilsstreuungen von 33 Graduierungsbegriffen an, die allerdings sehr unterschiedlichen Beurteilungsebenen entstammen (darunter „selten" mit „Z" = 14, „manchmal" = 43, „meistens" = 78, „wenig" = 9, „etwas" = 17, „mittelmäßig" = 46, „teilweise" = 50; übrige Begriffe nicht vergleichbar).

Immerhin fordert ESSER – im Gegensatz zu anderen Publikationen über Schätzskalen oder Antwortschemata – explizit, „daß die verbalen Benennungen geeicht sein müssen, damit hinreichend gesichert ist, daß sie auch den numerischen Wert repräsentieren, dem sie zugeordnet sind" (1970, p. 14; ähnlich: CLAUSS, 1968).

Graduierungen im Zusammenhang mit der Bewertungs-Dimension „gut – schlecht" (18 Stufen von „uneingeschränkt gut" bis „uneingeschränkt schlecht") untersuchte WEGENER (1976), wobei er die Skalenwerte mit Hilfe einer Ähnlichkeitsskalierung und eines nichtmetrischen MDS-Verfahrens gewann: Kriterium sind die Ausprägungen auf der 1. Dimension. (Ähnlichen Inhalten – like-/dislike-Ausdrücken – galt ferner eine ältere Arbeit von JONES & THURSTONE, 1955, doch ging es auch dort nicht um die Graduierungspartikel selbst.)

Insgesamt ergibt sich, daß zwar Studie I und II trotz einiger Änderungen von Ablauf und Inhalt gut korrespondierende Ergebnisse ergeben, der Bezug zu Resultaten anderer Autoren aber schon wegen der zu großen Methodenunterschiede (und der zu geringen Menge einbezogener Begriffe) kaum möglich ist.

4.6 Ergänzende empirische Befunde

4.6.1 Geläufigkeit von Graduierungsbegriffen

In Studie II (sowie gelegentlich innerhalb anderer explorativer Studien) wurden beiden Informanten-Gruppen offene Fragen gestellt, die Antworten mit Graduierungsbegriffen nahelegen; dies zielte darauf ab, besonders geläufige und im spontanen Gespräch bevorzugte Wörter herauszufinden. Als Häufigkeitsbegriffe wurden dabei „oft", „selten" (oder Kombinationen mit diesen Wörtern) sowie „immer" und „nie" am meisten benutzt; als Intensitätsbegriffe am häufigsten „sehr", „ziemlich", „mittelmäßig" und „(gar) nicht". Da diese und weitere genannte Begriffe denen der entwickelten Antwortskalen entsprechen, scheint die – durchaus angestrebte – Affinität zur ‚normalen' Umgangssprache hinreichend gegeben zu sein.

4.6.2 Vergleich von Skalentypen durch Urteiler

Die Informanten von Studie II sowie 32 Teilnehmer eines anderen Skalierungsexperiments wurden gebeten, ihre Präferenzen gegenüber 5- und 11stufigen Antwortskalen und solchen mit oder ohne Verbalisierungen zu nennen

Skala X: Bewertung (bipolar 11stufig)

außerordentlich gut	+5
sehr gut	+4
gut	+3
ziemlich gut	+2
mehr gut als schlecht	+1
mittelmäßig	0
mehr schlecht als gut	-1
ziemlich schlecht	-2
schlecht	-3
sehr schlecht	-4
außerordentlich schlecht	-5

Skala Y: Graduierung (unipolar 11st.)

überhaupt nicht	0
I	1
II	2
III	3
IIII	4
IIIII	5
IIIIII	6
IIIIIII	7
IIIIIIII	8
IIIIIIIII	9
IIIIIIIIII außerordentlich	10

Abb. 5: Zwei Beispiele für 11stufige Antwortskalen.

und zu begründen; dazu wurden als Beispiele die 2 in Abb. 5 wiedergegebenen Skalen und die 5stufige Intensitätsskala aus Abb. 3 vorgegeben.

Dabei ergab sich, dass etwa zwei Drittel der studentischen Informanten (S2) Variante X und etwa ein Viertel Variante Y gegenüber der Skala I bevorzugten; demgegenüber sprechen sich etwa zwei Drittel der Gruppe B2 (‚normale' Bevölkerung) für die 5stufige Skala und etwa ein Drittel für Variante X aus; auch die zusätzliche Gruppe B3 präferierte I vor X und Y. (Als Hauptgründe werden im einen Fall bessere Differenzierungsmöglichkeiten und im anderen Fall größere Übersichtlichkeit und leichtere Entscheidung beim Antworten angeführt.)

Tab. 3: Präferenz von Antwortskalen

	Informanten-Gruppe:		
Bevorzugte Skala:	„S2"	„B2"	„B3"
X (11stufig, verbal)	22	7	11
Y (11stufig, abstrakt)	8	2	7
I (5stufig, verbal)	3	20	14

Diese – natürlich weder hinsichtlich der Fallzahl noch der Beispiele repräsentativen – Daten deuten darauf hin, daß verbal gestützte Skalen mehrheitlich bevorzugt werden, und daß vielstufige Skalen – die für Studenten überwiegend problemlos sind – von Befragten der ‚außerakademischen' Bevölkerung weniger akzeptiert werden.

4.6.3 Konsistenz von Begriffsbeurteilungen

Angesichts der individuellen Unterschiede in der Beurteilung der Begriffe ist zu fragen, wieweit die für Antwortskalen ausgewählten Graduierungen in der Skalierungsaufgabe WZ zumindest ordinal konsistent beurteilt wurden. Ein einfaches Kriterium dafür ist der Anteil der Informanten, der die jeweils 5 Begriffe einer Serie in anderer Abfolge auf der 9stufigen Urteilsskala lokalisierte, als den mittleren Skalenwerten entspricht.

Das Resultat einer entsprechenden Reanalyse der Daten aus Studie II gibt Tab. 4.

Tab. 4: Anzahl inkonsistenter Beurteilungen ausgewählter Begriffe

Informanten	„B2" (n = 29)				„S2" (n = 32)			
Antwortskala	H	I	W	B	H	I	W	B
Paritäten	6	7	3	9	0	6	0	4
Vertauschungen . .	1	4	4	4	1	2	1	2

Durchschnittlich traten bei 7% der Informanten Vertauschungen auf (Beispiel: 1, 4, 3, 6, 9 als Reaktion auf die Begriffe der H-Skala) und bei 14% Paritäten (Beispiel: 1, 3, 5, 8, 8); beides gleichzeitig bzw. mehr als eine Inkonsistenz je Reihe geschahen in etwa 1% der Fälle. Dabei ist zu beachten, daß bei Begriffsserien mit extremen Endstufen, Skala H und W, wie zu erwarten konsistentere Skalenwert-Folgen resultieren als bei den enger gestuften Serien I und B. (Entsprechend ergeben sich z. B. bei einer alternativen Serie „gar nicht/wenig/mittelmäßig/überwiegend/völlig" nur halb soviele Inkonsistenzen wie bei der bisherigen I-Skala „nicht ... sehr".)

Insgesamt verdeutlicht diese Auswertung, daß Graduierungsbegriffe, die in den Mittelwerten eindeutig gestuft sind, von einzelnen Informanten keineswegs ebenso konsistent erlebt werden; in diesem Sinne wäre jene Antwortskala am besten, deren Verbalisierungen am seltensten inkonsistent skaliert wurden.

4.7 Konsequenzen für die Gestaltung der Antwortskalen

Bezogen auf das Feldforschungsprojekt, innerhalb dessen Studie II realisiert wurde, war zu entscheiden, ob die 1966 entwickelten Antwortskalen weiterhin akzeptiert werden können oder in der Verbalisierung geändert werden müssen.

Dazu ergibt sich aus den dargestellten Daten (vgl. auch Abb. 4), daß die gewählten Begriffe auch gemäß der Resultate von Studie II die anläßlich Studie I formulierten Kriterien (vgl. 3.1) einigermaßen erfüllen. (Ebenso wurden in der Gruppierungsaufgabe 1966 und 1976 überwiegend dieselben Begriffe bevorzugt als Stufen-Verbalisierung eingesetzt). Soweit es speziell die Abstände der Skalenwerte betrifft, sind die Unterschiede „B1"-„B2" weitgehend insignifikant und insofern ohne wesentliche Konsequenzen für die Auswahl der Begriffe; andererseits wird die angestrebte Äquidistanz auch bei den neuen Wertereihen gemäß Studie II verfehlt (vgl. z. B. die Distanzen zwischen den 5 Häufigkeitsbegriffen: 1.6–1.8–2.6–1.7 bzw. 1.8–1.9–2.6–1.6; intendiert: 1.75), und die Begriffe „wenig" und „ziemlich" erweisen sich wiederum als problematische Lösung für Stufe 2 und 4 einer fünfstufigen I-Skala.

Weil es im konkreten Anwendungsfall von 1976 besonders auf Vergleichbarkeit mit vorangegangenen Projekten ankam (es handelte sich um die 4. Untersuchung zur selben Problemstellung), erschien es insgesamt vertretbar, die 1966 erarbeiteten Antwortskalen (Abb. 3) vorerst beizubehalten. Allerdings sind für verschiedene Serien auch einige Alternativen plausibel, die kurz diskutiert werden sollen (vgl. jeweils die Tab.):

Häufigkeit (H-Skala):

Problematisch ist eigentlich nur die Besetzung der Mittelkategorie durch „gelegentlich" (M = 4.7/4.5), doch ist die einzige Alternative, „einigemal" (M = 4.9/4.6) in der Prägnanz nicht günstiger (und vielleicht – anders als „selten" oder „oft" – etwas zu konkret). Weiter wäre „häufig" eine gute Alternative zu „oft"; dagegen spricht nur, daß die Skalenbenennung und eine bestimmte Stufe namensgleich sind, was vermieden werden sollte. (Am Rande sei auf den interessanten Sachverhalt hingewiesen, daß „ziemlich selten" und „selten" sowie „ziemlich oft" und „oft" gleich, „sehr selten" und „sehr oft" aber anders bewertet werden: da „ziemlich" als Intensitätsgraduierung selbst im Bereich zwischen Skalenmitte und Extrem liegt, bleibt es als Zusatz zu „selten" oder „oft" offenbar wirkungslos.)

Resümee zu den Häufigkeitsskalen: „nie – selten – gelegentlich – oft – immer" erscheint – auch im Blick auf die Begriffspräferenzen der Informanten – als beste Lösung; wollte man die absoluten „nie" und „immer" an den Skalen-Enden vermeiden, so wären „sehr selten" und „sehr oft" die Alternative.

Intensität (I-Skala):

Die Stufen 2 und 4 (angestrebte Skalenwerte: 3.25 und 6.75) sind unbefriedigend besetzt, wobei für „wenig" gar keine und für „ziemlich" nur eine sprachlogisch problematische Alternative – „überwiegend" – zur Verfügung steht (vgl. 3.1.2; „ziemlich" ist besser kombinierbar). Ein Ausweg wäre, Stufe 1 gewissermaßen zurückzuverlegen und durch den sehr prägnanten Begriff „gar nicht" zu besetzen (evtl. auch „sehr" durch „völlig"; vgl. 4.6.3).

Resümee: Entweder die ursprüngliche Skala (bzw. – soweit linguistisch möglich – „annähernd" statt „ziemlich") oder „gar nicht-wenig-mittelmäßig-überwiegend-völlig".

Wahrscheinlichkeit (W-Skala):

Stufe 2 ist ungünstig besetzt, doch scheint die zunächst konzipierte Skala (von der ja 3 Begriffe erst in Studie II skaliert werden konnten) noch am ehesten akzeptabel.

Resümee: „keinesfalls-wahrscheinlich nicht-vielleicht-ziemlich wahrscheinlich-ganz sicher".

Bewertung von Aussagen (B-Skala):

Die Daten von Studie II gestatten eine bessere Abwägung zwischen den Alternativen. Sie zeigen zunächst, daß die ‚Neukonstruktion' „stimmt nicht-wenig-mittelmäßig-ziemlich-sehr" (vgl. 3.1.2) leidlich befriedigende Abstände ergibt (im Mittel: 1.7-2.3-1.7-1.6). Dies gilt jedoch ebenso für die „trifft zu"-Reihe (graduiert durch „gar nicht/wenig/teils-teils/ziemlich/völlig" und die „richtig-falsch"-Reihe (graduiert je mit „annähernd/

sehr", und z.B. „teils-teils" in der Mitte); hingegen ist die „dafür/dagegen"-Reihe ungeeignet.

Resümee: Bei Gleichwertigkeit der Alternativen sollte die Entscheidung von der Art des Beurteilungsgegenstands (Eigenschaften, Verhaltensbeschreibungen, Bewertungen usw.) abhängig gemacht werden.

Die bisherige Erörterung bezog sich auf 5stufige Antwortschemata; wollte man z.B. 4, 7 oder 9 Stufen haben, ließe sich aus den angeführten Daten ebenfalls eine entsprechende Auswahl treffen. (Angesichts der Urteilsstreuungen – und damit der ‚Überlappungen' der Begriffe – ergeben sich allerdings bei vielstufigen Skalen Probleme; auch unter diesem Gesichtspunkt erscheint die 5stufige Lösung als sinnvoller Kompromiß.)

Tab. 1 bzw. 2 lassen sich im Übrigen auch dazu heranziehen, in der Literatur vorgefundene Antwortskalen zu überprüfen, und sie belegen, wie unbefriedigend willkürlich verbalisierte Stufungen ausfallen können.

Dazu nur 3 Beispiele: Einer Graduierung von „gerne" durch „gar nicht/nicht/etwas/sehr" (HOLM, 1975, S. 72) entsprechen die sehr ungleich abständigen Skalenwerte 1.0/1.4/3.7/8.5 aus Tab. 1. Die Serie „nie/selten/oft/häufig" (aus einer demoskopischen Umfrage) ist gemessen an den Skalenwerten 1.1/2.7/7.3/7.2 ordinal nicht mehr eindeutig. Beim Versuch einer 7stufigen Graduierung durch „außerordentlich/sehr/ziemlich/einigermaßen/wenig/nicht sehr/kaum/eigentlich überhaupt nicht" (aus einer Hamburger Diplomarbeit) ist durch das fehlplazierte „kaum" keine sinnvolle Abfolge der Antwortstufen mehr gegeben.

Verbalisierungen sollten also zumindest so eindeutig hierarchisch geordnet sein, daß sie als gleichabständig intendiert aufgefaßt werden können.

5. Empirische Erfahrungen zu den Antwortskalen

5.1 Linguistische Probleme beim Einsatz von Antwortskalen

Bereits angesichts der Begriffsauswahl (vgl. 3.1.2) waren psycholinguistische Gesichtspunkte bei der Stufenbenennung angesprochen worden. Die praktischen Erfahrungen aus der bisherigen Anwendung seien kurz zusammengefaßt.

Es erwies sich, daß sich speziell die Intensitäts-Skala sehr gut mit verschiedensten Inhalten und sprachlichen Formen – Adjektiven (z.B. wahrscheinlich), Partizipien (z.B. störend oder gestört), Adverben (z.B. gern) und finiten Verben (z.B. stimmt) – verknüpfen läßt; dies ermöglicht, in Fragebögen, Beurteilungsbögen usw. sehr häufig dieselbe Antwortskala einsetzen zu können und damit den Lernaufwand für die Probanden zu vermindern. (Für bestimmte Begriffe, etwa das oben diskutierte Beispiel „überwiegend", ist diese Kombinierbarkeit hingegen nicht immer gegeben.) Schwierigkeiten ergeben sich allerdings bei unipolaren Benennungen bipolarer Dimensionen, z.B. „gut": „nicht gut", „wenig gut" usw. bezeichnen sicher nicht die negativsten Bewertungen, und der Bezug zu „schlecht" bzw. „etwas schlecht" o.ä. ist ungewiß. Also sollten keine derartigen Graduierungen von Wörtern vorgenommen werden, die nur als Teilbereich eines größeren Kontinuums (hier „schlecht-gut") interpretiert werden.

Ein andersartiges Problem kann bei der Häufigkeitsskala auftreten, daß nämlich die Ausprägungen „nie" und „immer" bei dem zu messenden Sachverhalt empirisch nicht vorkommen oder logisch unsinnig sind (z.B.: „Wie häufig gehen Sie ins Kino?"). In solchen Fällen ist die in 4.7 genannte Alternativskala sinnvoller.

Hinsichtlich der konkurrierenden Skalenvarianten zur Bewertung von Aussagen (vgl. die Diskussion in 3.1.2) ergab die bisherige Anwendung, daß die „stimmt"-Serie eindeutig am vielseitigsten einsetzbar war, insbesondere als Antwortschema für Statement- oder Symptomkataloge (speziell in Ich-Form), Eigenschaftszuordnungen, Verhaltensbeschreibungen usw., d.h., nicht nur für Tatsachen-, sondern besonders für Befindlichkeitsfragen. Für eben diese wäre ja die Urteilsebene „richtig-falsch" psycholinguistisch nicht angemessen.

Ein weiterer Gesichtspunkt ist, ob die ausgewählten Graduierungsbegriffe auch direkt in einen Satz integriert werden können („stört der Autoverkehr Sie nie, selten, gelegentlich, oft oder immer?", „fühlen Sie sich nicht, wenig, mittelmäßig, ziemlich oder sehr belästigt?" usw., d.h. interne Antwortvorgabe bei geschlossenen Fragen). Gerade bei halbstandardisierten

Interviews, die einerseits noch einigermaßen Gesprächscharakter haben und andererseits abgestufte Reaktionen erbringen sollen, kann dies wichtig sein. (Die I- und H-Skala sind in dieser Hinsicht sehr günstig, die bisher entwickelte W-Skala hingegen ungeeignet.)

5.2 Bewertung der Antwortskalen durch Interviewer

Aus zwei Interviewer-Befragungen (1975, n = 10; 1976, n = 20; standardisierte schriftliche Umfrage und mündliche Diskussion) ergibt sich zu den entwickelten Skalen u.a.:

- Etwa ein Fünftel der Befragten hat auch bei wiederholter Instruktion durchgängige Schwierigkeiten, abgestuft und auf der geforderten Antwortskala zu reagieren.
- Etwa drei Viertel der Befragten antworten im Wesentlichen verbal mit den Begriffen der vorgelegten Antwortskala und etwa ein Viertel numerisch (beides ist freigestellt).
- Etwa ein Drittel lernt im Laufe eines Interviews die Stufen bzw. Begriffe einer Antwortskala auswendig, wenn diese oft vorkommt und nicht zu häufig mit anderen Skalen wechselt.
- Insgesamt bewerten die Interviewer die ‚Praktikabilität' der Antwortskalen in der Feldforschung sowohl hinsichtlich der Verbalisierungen wie der Stufenzahl als gut (Erfahrungsbasis: etwa 1000 Interviews).

Als vorteilhaft wurde ferner angeführt, daß die durchgängige Verwendung und eigenständige Repräsentation der – möglichst wenigen – Antwortskalen den Erklärungsaufwand mindert („bitte antworten Sie wieder mit der roten Skala!") und erleichtert, durchgängig graduierte Antworten zu erhalten.

5.3 Bewertung der Antwortskalen durch Befragte

Die Aussagen von Befragten ergeben, daß verbal gekennzeichnete Antwortskalen eindeutig bevorzugt werden (wie schon aus Studie II resultierte; vgl. 4.6.2). Die erwähnten Schwierigkeiten im Skalengebrauch liegen offenbar nicht so sehr an den Verbalisierungen, die gut (und viel schneller als die numerischen Graduierungen) verstanden werden, sondern an der für viele ungewohnten Situation, überhaupt graduierend und mit fremdbestimmten Begriffen – statt mit „ja/nein" bzw. ihren eigenen Worten – zu antworten. Durch die (intendierte) Geläufigkeit der eingesetzten Graduierungsbegriffe scheint dieses Problem gemindert zu werden.

5.4 Resultierende statistische Charakteristika

Den Informanten aus Skalierungsstudie II („B2", „S2") sowie einer Fluglärmstudie (ROHRMANN, 1976; n = 397) wurden zur Beantwortung der selben Frage, einer Selbsteinschätzung der Lärmempfindlichkeit), zwei Skalentypen vorgegeben: Skala I (Abb. 3) und die 11stufige Skala Y (Abb. 5). Für diese spezielle Variable resultierten jeweils annähernd normal verteilte Antworten auf der fünfstufigen Skala I und sehr breit verteilte, tendenziell bimodale Verteilungen auf Skala Y (bei etwa verdoppelter Streuung wurden die Stufen um 3 und um 7 am besten ausgenutzt). Die Korrelationen zwischen den je 2 Wertereihen liegen um 0.90, so daß man die Messung durch Skala I als befriedigend betrachten kann. (Argumente zugunsten mäßiger Stufenzahlen ergeben sich u.a. auch aus Experimenten von GREEN & RAO, 1970 und MANTELL & JACOBY, 1971.)

Ob 5stufige Antwortskalen generell ‚bessere' oder ‚schlechtere' Antwortverteilungen erbringen, kann dennoch nicht sicher gesagt werden.

So ergab sich in eigenen Untersuchungen häufig, daß „nie" oder „nicht" die meistgewählten Stufen waren – wenngleich weniger aus psychometrischen Gründen und eher deshalb, weil der interessierende Sachverhalt gar nicht oder nur sehr selten gegeben war bzw. auftrat; in diesem Fall wird man keine Normalität erwarten können. (Tatsächlich ist nicht nur hier schwierig zu entscheiden, ob eine unerwartete oder ‚unerwünschte' Verteilungsform an der benutzten Skala liegt, wie häufig unterstellt und selten bewiesen wird, oder an der ‚tatsächlichen' Charakteristik der gemessenen Variable.)

Immerhin ist denkbar, daß der innerhalb einer Skala schwierige Übergang vom „Nichtgegebensein" (das ja nicht graduierbar ist), zu Abstufungen des Ausprägungsgrads eines

Merkmals bei den hier bevorzugten verbalisierten unipolaren Skalen eher zu ‚Brüchen' führt als bei einer abstrakten Skala der Art „−2/−1/0/+1/+2".

Wollte man sich dagegen sichern, daß die ausgewählten Wörter das zu gliedernde Kontinuum nicht exakt abstufen, so könnte man statt der Werte 1 bis 5 in der Auswertung auch die ermittelten Skalenwerte aus Studie I + II verwenden. Wie entsprechende Analysen ergaben, ist die Auswirkung auf Interkorrelationen zu anderen Variablen (besonders bei Einzelfragen, die ohnedies zu Summenwerten oder Faktorwerten weiterverrechnet werden) allerdings nicht wesentlich. (Ähnliches berichtet z. B. auch LABOWITZ, 1967.)

Gegen die ex-post-Gewichtung der Antwortstufen gibt es im übrigen einen gewichtigen Einwand: es ist gar nicht gesichert, daß die jeweiligen Urteiler ausschließlich auf die verbalen Teilreize reagieren; sofern sie sich an den numerischen oder graphischen Reizkomponenten – deren Äquidistanz weit weniger in Frage steht – orientieren, wäre die angesprochene Korrektur fragwürdig.

Auswirkungen auf die Reliabilität sind ebenfalls denkbar, aber anhand der eigenen Daten nicht prüfbar (vgl. jedoch z. B. FINN, 1972).

6. Diskussion und Kritik

6.1 Resümee zu den bisherigen Erfahrungen

Die bisherigen Erfahrungen mit dem dargestellten Antwortskalen-Konzept zeigen, daß – vor allem im Umgang mit Bevölkerungsstichproben – die verbal gegliederte Form sinnvoll und hilfreich ist, und daß zur Benennung bestimmter Skalenstufen hinreichend verständliche und prägnante Begriffe in der Sprache zur Verfügung stehen, also auch verbale Antwortskalen quasi geeicht werden können.

Neben den unmittelbaren Vorteilen bei der Datengewinnung und Datenverarbeitung kommt ein Vorteil in der Dateninterpretation hinzu, der zumindest in der „problemorientierten Forschung" (IRLE, 1975) wesentlich sein kann: Reaktionen auf den vorgeschlagenen Antwortskalen haben eine konkrete Bedeutung: daß jemand sich z. B. von einer bestimmten Umweltbelastung „ziemlich" oder „sehr" beeinträchtigt fühlt, daß X% öffentlichen Maßnahmen dagegen „nicht" oder „wenig" vertrauen usw., ist in gewissem Grade auch absolut interpretierbar, während z. B. die Zuweisung des abstrakten Werts „4" oder „2" oder auch eines z-Werts nur eine relative Information vermittelt. Die Antwortskalen ermöglichen Aussagen dieser Art (die besonders im Umgang mit Außenstehenden – außerhalb von Universität oder Forschungsinstituten – wichtig sein können), ohne auf die Vorteile einer graduierten Skala verzichten und rein qualitativ ‚messen' zu müssen.

Wieweit die dargestellten Antwortskalen freilich Intervallskalenqualität aufweisen, ist nach den bisherigen Studien noch nicht sicher zu sagen; beim Anlegen eines strengen Maßstabs ist dies offenbar nicht gegeben. (Es kommt hinzu, daß Urteiler den benutzten Begriffen bei verschiedenen Beurteilungsgegenständen möglicherweise unterschiedliche Bedeutung zumessen; ob die zugleich vorgegebenen numerischen Graduierungen derartiges wettmachen können, ist mit den verfügbaren Daten nicht klärbar.)

Wie man Einstufungen auf der verbalisierten Antwortskala verrechnen will, hängt wohl von der generellen Entscheidung über den Einsatz von metrischen Verfahren – wie Varianz-, Faktoren-, Diskriminanz-, Regressionsanalyse usw. – bei derartigen, u. U. nur ordinalen, Daten ab. (Auf die kontroverse Diskussion dazu soll hier nicht eingegangen werden; vgl. aber LABOWITZ, 1970 und 1972; WILSON, 1971; SCHEUCH & ZEHNPFENNIG, 1974 und KIM, 1975). Durch die zunehmende Verbreitung nonmetrischer Analysetechniken wird diese Frage vermutlich an Bedeutung verlieren.

6.2 Unzulänglichkeiten des Untersuchungsansatzes

Zumindest in zweifacher Hinsicht ist das dargestellte Vorgehen unzulänglich:

- die empirische Basis ist zu schmal (unmittelbar von der Fallzahl her und mittelbar hinsichtlich der repräsentierten Bevölkerungsgruppen und insbesondere Regionen);
- die herangezogene Skalierungsprozedur ist für weitergehende statistische Analysen (insbesondere mehrdimensionale Skalierung) wenig geeignet;
- das erörterte Äquidistanzproblem kommt hinzu.

Bezogen auf die eigentliche Zielsetzung – für zwei konkrete problemorientierte Untersuchungen die benötigten und praktikablen Antwortskalen zu entwickeln – wiegt sicher der erste Punkt schwerer: Immerhin sind linguistische Unterschiede zwischen den deutschen Sprachgebieten gegeben, die sich auch auf die Bewertung der Graduierungspartikel auswirken mögen; hier wäre eine bessere Generalisierbarkeit sehr wünschenswert.

Darüberhinaus müßte auch der (wegen der Einbindung in bestimmte Feldforschungsprojekte bisher eher begrenzte) methodologische Aufwand vermehrt werden, um zu psychometrisch besser abgesicherten Ergebnissen zu kommen.

6.3 Konsequenzen für weitere Studien

Abschließend seien einige Überlegungen zur inhaltlichen und methodischen Fortführung der Untersuchung von Antwortskalen erörtert. Um die Anwendungsmöglichkeiten zu verbessern, könnten weitere verbalisierte Skalen mit mehr als 5 – insbesondere 7 – Stufen entwickelt werden.

Da Lösungen der Art „gar nicht-wenig-etwas-mittelziemlich-sehr-völlig" nicht sehr befriedigend sind (vgl. Tab. 2B), ist es wahrscheinlich besser, die Gegenbegriffe einer Polarität mit je 3 Begriffen zu graduieren, Beispiel: „sehr/ziemlich/etwas" zustimmen/ablehnen, und eine neutrale Kategorie wie „mittel" oder „unentschieden" dazwischen zu setzen. Welche Skalenstufen allerdings bei solchen bipolaren Lösungen angemessen sind, und wie die vorliegenden Ergebnisse auf diesen Fall anwendbar sind, bedarf noch näherer theoretischer und empirischer Klärung. (Dabei ist auch zu berücksichtigen, wieweit bipolare Skalen hinreichend eindeutig messen; vgl. KAPLAN, 1972; TROMMSDORFF, 1975, p. 87f.).

Ebenso soll die Verflechtung der Intensitäts- oder Häufigkeitsstufen mit verschiedenen Erlebnisdimensionen noch genauer untersucht werden (eine entsprechende Studie zur Messung und Bewertung von „Belästigungs"-Graden ist derzeit in Arbeit).

Auch die Auswirkungen der Antwortskalen auf empirische Daten müssen noch systematischer quantifiziert werden; so können z. B. konkurrierende Varianten (numerisch versus verbal versus numerisch + verbal) in zufällig bestimmten Teilstichproben von Untersuchungen eingesetzt werden. Dabei könnte auch den Wechselwirkungen der Wörter und Zahlen – von denen ja ein „vereindeutigender Einfluß" (HENNIG, 1975, p. 357) erwartet wird – nachgegangen werden.

Wesentlich erscheint ferner die Anwendung neuerer Skalierungs- und Analysetechniken für die Begriffsbewertung selbst. Dazu kommen etwa das „three-mode"-Konzept (TUCKER, 1964) oder die Bestimmung von Distanzmaßen (statt Korrelationen; vgl. SCHLOSSER, 1976) und Weiterverarbeitung mit KRUSKALS MDS bzw. ähnlichen Verfahren (siehe WEGENER, 1976) in Betracht.

Dies ist für eine Teilmenge der bisher untersuchten 100 Begriffe – besonders die in eigenen Antwortskalen schon eingesetzten – vorgesehen (wobei der primäre Beurteilungsakt besser als bisher auf die teils erheblichen Urteilsschwierigkeiten von Informanten aus Bevölkerungsstichproben abgestimmt werden soll; wie die zu entwickelnden Antwortschemata, so sollten auch die Skalierungsverfahren verstehbar und zumutbar sein).

Schließlich wird es bei der Weiterentwicklung von Antwortskalen sinnvoll sein, die verschiedenen Anwendungszwecke in der sozialwissenschaftlichen Forschung zu berücksichtigen: Bei der Handhabung durch Befragte/Probanden/Versuchspersonen – was hier im Vordergrund stand – gelten u. U. andere Kriterien als bei der Benutzung etwa in Beobachtungs- oder Codierungssituationen durch den Forscher selbst; das grundsätzliche Ziel – mit bestimmten Ansprüchen zu messen – ist freilich in beiden Fällen das gleiche.

Literatur

ATTESLANDER, P. 1975⁴. Methoden der empirischen Sozialforschung. Berlin/New York.

BUCHTA, E. & KASTKA, J. 1974. Erfassung der Lärmbelästigung durch Verkehr mittels physikalischer Messungen und sozialwissenschaftlicher Erhebungsmethodik. Manuskript (Institut für Hygiene, Universität Düsseldorf).

CHAMPNEY, H. & MARSHALL, H. 1939. Optimal refinement of the rating scale. Journal of Applied Psychology 23, 323–331.

CLAUSS, G. 1968. Zur Methodik von Schätzskalen in der empirischen Forschung. Probleme und Ergebnisse der Psychologie 26, 7–52.

CLIFF, N. 1959. Adverbs as multipliers. Psychological Review 66, 27–44.
COHEN, R., REY, E.-R. & SIXTL, F. 1969. Die Kombination von „Häufigkeit" und „Intensität" im diagnostischen Urteil. Psychologische Forschung 33, 9–20.
COOMBS, C.H., DAWES, R.M. & TVERSKY, A. 1970. Mathematical psychology – an elementary introduction. Englewood Cliffs (N.Y.).
CRANACH, M. v. & FRENZ, H.G. 1969. Systematische Beobachtung. S. 269–331 in: Graumann.
CRONBACH, L.J. 1964[2], 1970[3]. Essentials of psychological testing. New York.
DAWES, R.M. 1972. Fundaments of attitude measurement. New York.
DFG-Forschungsbericht 1974. Fluglärmwirkungen – eine interdisziplinäre Untersuchung über die Auswirkungen des Fluglärms auf den Menschen. 3 Bde., Boppard.
DUDEN. 1959. Rechtschreibung (Der große Duden, Bd. 1). Weinheim. (Vgl. auch Grebe, 1959; Müller, 1972.)
EDWARDS, A. 1957. Techniques of attitude scale construction. New York.
EICHNER, K. 1977. Die Entstehung von sozialen Normen. Arbeitsbericht für die DFG. Hamburg.
ESSER, U. 1970. Skalierungsverfahren. S. 184–242 in: Friedrich.
FINKE, H.O., GUSKI, R. & ROHRMANN, B. 1977. Konzeption und erste Ergebnisse einer interdisziplinären Untersuchung zum Problem Großstadtlärm (Projekt „Betroffenheit einer Stadt durch Lärm"). Kampf dem Lärm 24, 128–132.
FINN, R.H. 1972. Effects of some variations in rating scale characteristics on the means and reliabilities of rating. Educational and Psychological Measurement 32, 255–265.
FREDERIKSEN, N. & GULLIKSEN, H. 1964. Contributions to mathematical psychology. New York.
FRIEDRICH, W. 1970. Methoden der marxistisch-leninistischen Sozialforschung. Berlin.
FRIEDRICH, W. & HENNIG, W. (Eds.) 1975. Der sozialwissenschaftliche Forschungsprozeß. Berlin.
FRIEDRICHS, J. 1973. Methoden empirischer Sozialforschung. Reinbek.
FROGNER, E. 1978. Aggressives Sozialverhalten – eine sportsoziologische Untersuchung. In Vorb. (Hamburg).
GAGE, N.L. 1963. Handbook of research on teaching. Chicago.
GALTUNG, J. 1969[2]. Theory and method of social research. Oslo.
GRAUMANN, C. 1969. Sozialpsychologie, Theorien und Methoden (Handbuch der Psychologie, Band 7/1). Göttingen.
GREBE, P. Dudenredaktion 1969[1], 1973[3]. Duden – Grammatik der deutschen Gegenwartssprache (Der große Duden, Bd. 4). Mannheim.
GREEN, P.E. & RAO, V.R. 1970. Rating scales and information recovery – how many scales and response categories to use? Journal of Marketing 34, 33–39.
GUILFORD, J.P. 1954[2]. Psychometric methods. New York et al.
GUSKI, R., WICHMANN, U., ROHRMANN, B. & FINKE, H.-O. 1978. Konstruktion und Anwendung eines Fragebogens zur sozialwissenschaftlichen Untersuchung der Auswirkungen von Umweltlärm. Zeitschrift für Sozialpsychologie 9, 50–65.
GUTJAHR, W. 1972. Die Messung psychischer Eigenschaften. Berlin.
HENNIG, W. 1975. Schätzskalen. S. 345–367 in: Friedrich & Hennig.
HOFSTÄTTER, P.R. & WENDT, D. 1966. Quantitative Methoden der Psychologie. München.
HOLM, K. (Ed.) 1975. Die Befragung (Bd. 1). München.
HÖRMANN, H. 1967[1], 1970[2]. Psychologie der Sprache. Berlin/Heidelberg.
INGENKAMP, K.-H. (Ed.) 1970. Handbuch der Unterrichtsforschung. Weinheim.
IRLE, M. 1975. Lehrbuch der Sozialpsychologie. Göttingen.
IRLE, M. & ROHRMANN, B. 1968. Gesamtbericht über die Hamburger Voruntersuchung zum DFG-Projekt Fluglärmforschung der sozialpsychologischen Sektion, unveröffentlicht. Mannheim/Hamburg.
JONES, L.V. & THURSTONE, L.L. 1965. The psychophysics of semanties – an experimental investigation. Journal of Applied Psychology 39, 31–66.
KAMINSKI, G. 1976. Umweltpsychologie – Perspektiven, Probleme, Praxis. Stuttgart.
KAPLAN, K.J. 1972. On the ambivalence-indifference problem in attitude theory and measurement. Psychological Bulletin 77, 361–372.
KASTKA, J. 1976. Untersuchungen zur Belästigungswirkung der Umweltbedingungen Verkehrslärm und Industriegerüche. S. 187–224 in: Kaminski.
KELLEY, H.H., HOVLAND, L.I., SCHWARZ, M. & ABELSON, R.P. 1955. The influence on judges attitudes in three methods of attitude scaling. Journal of Social Psychology 42, 147–158.
KÖNIG, R. (ed.) 1965[4]. Das Interview. Köln/Berlin.
KÖNIG, R. (Ed.) 1967[1], 1974[3]. Handbuch der empirischen Sozialforschung.
KIM, J.-O. 1975. Multivariate analysis of ordinal variables. American Journal of Sociology 81, 261–298.
KRISTOF, W. 1966. Das Cliffsche Gesetz im Deutschen. Psychologische Forschung 29, 22–31.
KRUSKAL, J.B. 1964. Nonmetric multidimensional scaling – a numerical method. Psychometrika 29, 115–130.
LABOVITZ, S. 1967. Some observations on measurement and statistics. Social Forces 46, 151–160.
LABOVITZ, S. 1970. The assignment of numbers to rank order categories. American Sociological Review 35, 515–524.
LABOVITZ, S. 1972. Statistical usage in sociology – sacred cows and ritual. Sociological Methods and Research 1, 13–38.
LANGER, I. & SCHULZ v. THUN, F. 1974. Messung komplexer Merkmale in Psychologie und Pädagogik. München.

LIKERT, R. 1932. A technique for the measurement of attitudes. Archives of Psychology 22, 1–55.
LINDZEY, G. & ARONSON, E. (Eds.) 1968. The handbook of social psychology (Vol. II, research methods). New York.
MANTELL, M.S. & JACOBY, J. 1971. Is there an optimal number of alternatives for Likert scale items? Educational and Psychological Measurement 31, 657–67.
MILLER, G.A. 1956. The magical number seven, plus or minus two – some limits on our capacity for processing information. Psychological Review 63, 81–97.
MÜLLER, W./Dudenredaktion 1972. Duden – Sinn- und sachverwandte Wörter und Wendungen (Der große Duden, Bd. 8). Mannheim.
NOELLE, E. 1963. Umfragen in der Massengesellschaft. Reinbek.
NUNNALLY, J.L. 1970. Introduction to psychological measurement. New York.
OSGOOD, C.G., SUCI, G.J. & TANNENBAUM, P.H. 1957. The measurement of meaning. Urbana.
REMMERS, R. 1963. Rating methods in research on teaching. S. 329–378 in: Gage.
ROHRMANN, B. Methodologische Erfahrungen aus einer interdisziplinären Felduntersuchung zur Wirkung von Lärm auf den Menschen; Beitrag zum 17. Deutschen Soziologentag, Sektion Methoden, Kassel 2.11.1974.
ROHRMANN, B. 1974. Psychometrische und textstatistische Studien zu syntaktischen Variablen. Hamburg.
ROHRMANN, B. 1976. Die Störwirkung des Flugbetriebs an Landeplätzen – eine empirische Studie. Kampf dem Lärm 23, 6–11.
SAFFIR, M. 1937. A comparative study of scales constructed by three psychophysical methods. Psychometrika 2, 179–198. (Zit. nach Guilford, 1954).
SCHLOSSER, O. 1976. Einführung in die sozialwissenschaftliche Zusammenhangsanalyse. Reinbek.
SCHÜMER-KOHRS, A. & SCHÜMER, R. 1974. Der sozialwissenschaftliche Untersuchungsteil. Kap. 4 in: DFG-Forschungsbericht Fluglärmwirkungen.
SCHEUCH, E.K. & ZEHNPFENNIG, H. 1974. Skalierungsverfahren in der Sozialforschung. S. 97–204 in: König.
SELLTIZ, C., JAHODA, M., DEUTSCH, M. & COOK, S.W. 1959[2]. Research methods in social relations. New York et al.
SIXTL, F. 1967. Meßmethoden in der Psychologie. Weinheim.
STEVENS, S.S. 1951. Mathematics, measurement, and psychophysics. S. 1–49 in: Stevens.
STEVENS, S.S. 1951. Handbook of experimental psychology.
TACK, W. 1969. Mathematische Modelle in der Sozialpsychologie. S. 232–265 in: Graumann.
TAYLOR, J.B., PTACEK, M., GRIFFIN, C. & COYNE, L. 1968. Rating scales as measures of clinical judgement. Educational and Psychological Measurement 28, 747–766.
TENT, L. 1970. Schätzverfahren in der Unterrichtsforschung. S. 863–899 in: Ingenkamp.
THURSTONE, L.L. 1928. Attitudes can be measured. American Journal of Sociology 33, 529–554.
THURSTONE, L.L. 1929. Theory of attitude measurement. Psychological Bulletin 36, 222–241.
THURSTONE, L.L. & CHAVE, E.J. 1929. The measurement of attitudes. Chicago.
THURSTONE, L.L. 1959. The measurement of values. Chicago.
TORGERSON, W.S. 1958. Theory and methods of scaling. New York.
TROMMSDORFF, V. 1975. Die Messung von Produktimages für das Marketing – Grundlagen und Operationalisierung. Köln et al.
TUCKER, L.R. 1964. The extension of factor analysis to three-dimensional matrices. In: Fredriksen & Gulliksen.
VELDMAN, D.J. 1967. Fortran programming for the behavioral sciences. New York/London.
WEGENER, B. 1976. Ein Vergleich multidimensionaler Skalierungsalgorithmen, Teil I. ZUMA-Manuskript; Mannheim August.
WILSON, T. 1971. Critique of ordinal values. Social Forces 49, 432–444.
ZIEGLER, R. 1972. Theorie und Modell. München/Wien.