REPORT 146

1973

A METHOD FOR EVALUATION OF METRIC PROPERTIES OF RESPONSE SCALES

BO EDVARDSSON

MICHAEL LINDEN

DEPARTMENT OF PSYCHOLOGY • UNIVERSITY OF UPPSALA • SWEDEN

A METHOD FOR EVALUATION OF METRIC PROPERTIES OF RESPONSE SCALES

Bo Edvardsson

Michael Lindén

Abstract

To permit certain statistical operations verbal response categories in questionnaire response scales must be given appropriate numbers. In this study the routines used to assign such numbers are discussed. A method to evaluate the metric properties of response scales, simultaneously being an alternative way of assigning proper numbers, is proposed. For this method, based on subjective estimation, two criteria related to commonality of meaning are suggested. Finally the method is demonstrated in an empirical study on eight different response scales.

Contents

ABSTRACT	
INTRODUCTION	1
METHOD AND CRITERIA	3
AN EMPIRICAL STUDY	6
DISCUSSION	9
REFERENCES	12
TABLE	13-16

Introduction

Response scales of the type introduced by Likert (1932) are today very widely used not only in attitude measurement and opinion surveys but also in measurement of personality traits, interests, and the like.

In the first Likert scales the subjects responded to each attitude statement, choosing one of the following five response categories:

Strongly	Approve	Undecided	Disapprove	Strongly
approve				disapprove

Since then a wide variety of response scales has been used (see e.g. Shaw & Wright, 1967). To permit certain statistical operations the verbal response categories must be assigned appropriate numbers. Particularly when the statistical treatment is based directly on the separate statements this assignment procedure should be regarded as crucial since it influences the measurement level of the response scale. It has been pointed out that computations of e.g. arithmetic means, standard deviations, and product moment correlations are not valid, if it cannot be shown that numbers constitute an interval (or a ratio) scale (Stevens, 1951; Siegel, 1956). Even if this position might be disputed, the use of parametric statistics should be considered more justified the better the measurement properties of the response scale are shown to be.

Likert suggested two methods for assigning numbers to the verbal response categories - the sigma method and the simple method (Likert, 1932; Edwards, 1957). The sigma method is based on the normal distribution. Each category is given

a standard value (z-value) related to the proportion of subjects having chosen that category. The standard values are rounded to integers and a linear transformation is made, usually giving the extreme categories the numbers of 1 and 5. With this procedure the same numbers could obviously be given to two or more categories, and the distance between two adjacent categories could be greater than unit. Further, the numbers are highly dependent on the response distribution in the sample used. If the subjects experience the response categories in the same way and if the hypothesis of normal distribution is accepted, the sigma method, for a certain statement and a certain population, should result in a scale representing the interval level of measurement (at least if the numbers are not rounded to integers). From what has been said the sigma method may be questioned for at least two reasons. First, very little seems to be known regarding the commonality of meaning - that is, do the subjects experience the response categories in the same way, or are there large interindividual differences (cf. Smedslund, 1969, 1970)? Second, even if commonality of meaning is assumed, the validity of the sigma method is no stronger than the validity of the hypothesis of normal distribution, which may be questioned for many attitude statements in many populations.

In the simple method the response categories are ordered and assigned the integers 1, 2, 3, 4, and 5, respectively. Likert showed that the simple method, for the sum total over a number of statements (10 - 15), gave results highly correlated to the results of the sigma method. Thus he preferred

the simple method, and today this method is the conventional one when constructing response scales. If the response categories are experienced in the same way by the scale constructor and by all subjects and if the differences between the response categories are experienced as equal, the numbers assigned by the simple method should be regarded as representing the interval level of measurement. However, many scale constructors seem to choose response categories and assign numbers in a more or less arbitrary way making no check for commonality.

Both the sigma and the simple methods overlook the problem of commonality of meaning, and the validity of the methods should not be taken for granted. As an alternative a subjective estimate method (Torgerson, 1958) is suggested and demonstrated in an empirical study on eight different response scales. For the categories in a response scale the method will give information about the commonality of meaning shared by the subjects, this information being relevant for the measurement level of the response scale.

Method and criteria

The method to be described is based on subjective arrangement of response scale categories. The categories of a scale, ordered by random, are simultaneously presented to a subject. Assuming unidimensionality the subject is instructed to arrange the categories along a linear continuum, including no numbers or units. Both ends are verbally anchored, the left end representing "Greatest degree of disagreement" and

the right end representing "Greatest degree of agreement". The subject is told to put the two most extreme categories, according to his experience, exactly at the ends of the continuum. The other categories are to be placed in such positions that the distances between them represent the differences in meaning experienced by the subject. If no differences are experienced between two or more categories, they should be placed in the same position. After the task is finished the positions are registered, and the distances are taken as representing the relative differences between the response categories as experienced by the subjects.

In relation to this method the following criteria, relevant for the measurement level of a response scale, are suggested.

Subjective order criterion. If a subject places the response categories in the same order as that of the numbers assigned by the scale constructor, a subjective order scale is obtained. This permits intraindividual comparisons over statements. If the proportion of subjects violating the rank order used by the scale constructor is high, the scale should be abandoned or changed. The subjective order criterion is necessary but not sufficient for interindividual comparisons on the order level.

Subjective interval criterion. If a subject places the response categories at points giving distances consistent with the numbers assigned by the scale constructor, a subjective interval scale is obtained. This permits intraindividual comparisons over statements. If there are large interindi-

vidual variations concerning the distances between the response categories, that is concerning the relative differences in meaning, the scale should be changed. If the interindividual variation is acceptable but the distances are very far from those indicated by the numbers, either the numbers or the response categories could be changed. The subjective interval criterion is necessary but not sufficient for interindividual comparisons on the interval level.

As mentioned above the subjective criteria are considered necessary but not sufficient for interindividual comparisons. Probably a response scale can never be proved to hold for interindividual comparisons on the ordinal or interval level since the absolute meanings of the response categories are unknown. It can never be excluded that e.g. "Approve" for one subject means more than e.g. "Strongly approve" for another. However, it may be judged more or less plausible (depending on the constructs used in the response categories, the cultural background of the subjects etc.) that the commonality of meaning for one or more categories should be very high. If it is considered reasonable to define e.g. the extreme categories of a response scale as such psychological universals and if the scale holds for the subjective interval criterion, then the scale obtained should permit interindividual comparisons on the interval level.

Thus from the method described in this section the response scales which do not hold for interindividual compari-

sons on a certain measurement level can be separated from those which may hold. Further, information is obtained as to whether a response scale permits intraindividual comparisons on an ordinal or interval measurement level.

An empirical study

To test the subjective arrangement technique subjects were instructed to indicate by arrows the positions of response categories along a 120 mm continuum, printed on a sheet.

For 97 subjects, sampled at random from a population of psychology students, judgements on eight Swedish response scales (Table 1) were collected by mail questionnaires. Of the eight scales two included six response categories, four included five, and two included four categories. All scales, but response scale 1, have recently been used in attitude research in Sweden. The scales were presented on separate pages and, to prevent transfer effects that might possibly occur, no scales with the same number of response categories were presented in immediate connection. Within this restriction the presentation order of the scales was permuted.

To check equality of intervals and to gain knowledge about the amount of variation caused by perceptual or motor effects, a comparison group of 40 students were instructed to divide 120 mm continua in five, four, and three equal parts, respectively. To get comparison measures for response scale 2 the instruction was to divide the 120 mm

continuum in five parts, the middle one having twice the length of the others. This was done in presence of the experimenter. Since the comparison subjects could be more careful, the variations obtained for this group may underestimate the corresponding variations that will be obtained in the experiment group if equal intervals are experienced. Also the exact variations expected if the 120 mm continua are to be divided into unequal parts are unknown.

The results of the study are summarized in Table 1. The Swedish response categories are translated into English with the original Swedish text within brackets. Obviously, the results presented should be related to the Swedish versions and to the present population. The data are, for matters of comparison, transformed from the 120 mm continua to the scales used by the scale constructors.

 Insert Table 1 about here

Some tentative conclusions about the measurement properties of the response scales can be drawn from the data.

Subjective order criterion. If an operationalization of this criterion is made by permitting e.g. 5 per cent of the population to violate the order used by the scale constructor, it is obvious that response scale 2 should be regarded as insufficient, even if the significance level is set very low. For this response scale, used in the Swedish version of the California F-scale, 19 of the 97 subjects violate the rank order (Table 1, col. 2). Most breaks are caused by the cate-

gory "Have a somewhat different opinion" ("Har något avvikande uppfattning"), which varies from position 2 to position 5. Moreover, if one does not accept ties, that is, when a subject experiences no differences between two or more response categories, response scale 6 should be abandoned. The 15 subjects causing ties in this scale (Table 1, col. 3), experience no differences between the two positive categories and also no differences between the two negative categories. The rest of the scales should pass the subjective order criterion.

Subjective interval criterion. This criterion can be operationalized in many ways. One way is to require that means rounded to integers for the judgements of the experiment group should correspond to the numbers used by the scale constructors (Table 1, col. 7 vs col. 1). Operationalizing it this way response scales 2, 6, and 8 are not acceptable. If, in addition, it is required that the ratio between the largest and the shortest intervals for a certain scale should be about unit (Table 1, col. 6, computed from col. 4), none of the scales is acceptable. The scale that best meets this criterion is response scale 1. The relative commonality of meaning could be operationalized by comparing the interindividual variations for the experiment and the comparison groups. (The standard deviations are found in Table 1, col. 8 resp. 9.) This could be done by the F-test. At the 1 per cent level only one response category, the central category of response scale 3, is not significant, and maybe the operationalization demands too much. The central or neutral categories of the five-category scales, except for response

scale 5, have the least interindividual variations.

For interindividual comparisons at the order and interval measurement levels at least one further criterion should be added - there should be no overlap between the subjects' judgement distributions of different response categories in a certain scale. Overlap should imply that e.g. "Approve" for one subject means more than e.g. "Strongly approve" for another. For the scales presented there is some overlap for the response categories. Response scale 2 is the worst one in this respect and response scale 3 is the best one, having almost no overlap. However, as it cannot be known if the extreme response categories have the same absolute values for all subjects, there is no use in accounting for the overlap. If the subjects do not experience the extreme categories in the same absolute way, this could decrease or increase the degree of overlap.

A comparison of the response scales based on the suggested criteria results in response scale 1 being the best six-category scale and response scale 3 being the best five-category scale. However, none of the scales is considered very good. Thus, from a measurement level point of view, modifications of (some of) the scales studied should be recommended.

Discussion

In this report response scales have been discussed and treated apart from the statements or questions to which they may be related.

There may be some ideas that interactions exist between statements and the response categories related, implying that the differences among the response categories could be judged differently depending on the statements to be answered. However, preliminary observations do not support such ideas. (Of 40 subjects none was able to pick out two statements, among more than 200, for which the experienced differences among the response categories were judged differently.)

More relevant may be to point out that, usually, scale constructors do not present their response categories in random order. As a rule cues are given on how to experience the different categories. Most often the response categories are related to the statements according to the rank order experienced by the scale constructor. Also numbers are sometimes printed together with the verbal response categories. This was done, in a more or less consistent manner, for some of the scales treated in this study. For the worst scale, according to our opinion (response scale 2, used in the Swedish version of the California F-scale), the response categories were printed in two columns and no numbers were printed in the questionnaire. Even if the effects of cues like a certain presentation order or numbers printed together with the verbal categories are not known, any scale including ambiguous verbal (or other) response categories should be avoided.

None of the response scales presented has been regarded as very good. Probably it is possible to find better scales using verbal response categories. However, it could be that

response scales using other symbols, e.g. numbers, supposed to have more commonality of meaning, should be preferred to verbal category scales (cf. Stevens, 1951, p. 28; Guilford, 1954, p. 264). From the measurement level point of view, a number scale, with e.g. the extreme categories verbally anchored, might be a better alternative than a purely verbal category scale.

In addition to measurement level there might be other relevant aspects when choosing response scales. Response frequencies, reliabilities and validities could be influenced by e.g. number of response categories and whether a neutral category is permitted or not.

References

- Edwards, A. L. Techniques of attitude scale construction.
New York: Appleton-Century-Crofts, 1957.
- Guilford, J. P. Psychometric methods. New York: McGraw-Hill,
1954.
- Likert, R. A technique for the measurement of attitudes.
Archives of Psychology, 1932, No. 140.
- Shaw, M. E., & Wright, J. E. Scales for the measurement of
attitudes. New York: McGraw-Hill, 1967.
- Siegel, S. Nonparametric statistics for the behavioral sciences.
New York: McGraw-Hill, 1956.
- Smedslund, J. Meanings, implications and universals: towards
a psychology of man. Scandinavian Journal of Psychology,
1969, 10, 1 - 15.
- Smedslund, J. Circular relation between understanding and logic.
Scandinavian Journal of Psychology, 1970, 11, 217 - 219.
- Stevens, S. S. Mathematics, measurement and psychophysics.
In S. S. Stevens (Ed.), Handbook of experimental psychology.
New York: Wiley & Sons, 1951.
- Torgerson, W. S. Theory and methods of scaling. New York:
Wiley & Sons, 1958.

TABLE 1

13.

Results for experiment group (N=97) and comparison group (N=40). The first column contains the numbers

used by scale constructors.

Response scales	Numbers used	Rank order breaks	Rank order ties	Means exp. group	Means comp. group	Inter-val ratios	Rounded means	St.dev. exp. group	St.dev. comp. group
Response scale 1		3	2			1.51			
Strongly disagree (Tar starkt avstånd)	1			1.00	1.00		1	0.00	0.00
Mainly disagree (Tar avstånd i stort sett)	2			1.79	1.98		2	0.35	0.07
Hesitantly disagree (Tar avstånd med tvekan)	3			2.89	2.97		3	0.41	0.09
Hesitantly agree (Instämmer med tvekan)	4			4.08	4.01		4	0.49	0.11
Mainly agree (Instämmer i stort sett)	5			5.21	5.01		5	0.33	0.07
Strongly agree (Instämmer starkt)	6			6.00	6.00		6	0.00	0.00
Response scale 2		19	4			1.83			
Have a strongly different opinion (Har kraftigt avvikande uppfattning)	1			1.00	1.00		1	0.00	0.00
Have a mainly different opinion (Har i stort sett avvikande uppfattning)	2			2.08	1.97		2	0.73	0.08
Have a somewhat different opinion (Har något avvikande uppfattning)	3			3.82	2.99		4	1.13	0.11
Hesitantly agree (Instämmer med tvekan)	5			4.81	5.03		5	0.68	0.10
Mainly agree (Instämmer i stort sett)	6			6.05	6.04		6	0.43	0.09
Totally agree (Instämmer fullständigt)	7			7.00	7.00		7	0.00	0.00

Response scales

	Numbers used	Rank order breaks	Rank order ties	Means exp. group	Means comp. group	Inter-val ratios	Rounded means	St.dev. exp. group	St.dev. comp. group	14.
Response scale 3		2	0			2.06				
Do not at all agree (Instämmer inte alls)	1			1.00	1.00		1	0.00	0.00	
Mainly do not agree (Instämmer i stort sett inte)	2			1.71	1.98		2	0.42	0.06	
As much goes for as against (Lika mycket talar för som emot)	3			2.98	3.00		3	0.09	0.07	
Mainly agree (Instämmer i stort sett)	4			4.34	4.01		4	0.36	0.07	
Wholly agree (Instämmer helt och hållet)	5			5.00	5.00		5	0.00	0.00	

Response scale 4

		2	2			2.29				
Wholly disagree (Har avstånd helt och hållet)	1			1.00	1.00		1	0.00	0.00	
Hesitantly disagree (Har avstånd med tvekan)	2			2.37	1.98		2	0.36	0.06	
Undecided (Tveksam)	3			2.97	3.00		3	0.21	0.07	
Hesitantly agree (Instämmer med tvekan)	4			3.63	4.01		4	0.43	0.07	
Wholly agree (Instämmer helt och hållet)	5			5.00	5.00		5	0.00	0.00	

Response scales

Numbers Rank Rank Means Means Inter- Rounded St.dev. St.dev.
used order order exp. comp. val. means exp. comp.
breaks ties group group group group group group

Response scale 5

	3	1			2.26		
Do not at all agree (Instämmer inte alls)	1		1.00	1.00		1	0.00 0.00
Hardly agree (Instämmer knappast)	2		2.03	1.98		2	0.50 0.06
Partly agree (Instämmer delvis)	3		3.48	3.00		3	0.47 0.07
Mainly agree (Instämmer i stort sett)	4		4.36	4.01		4	0.31 0.07
Wholly agree (Instämmer helt)	5		5.00	5.00		5	0.00 0.00

Response scale 6

	2	15			3.88		
Absolutely do not agree (Instämmer absolut inte)	1		1.00	1.00		1	0.00 0.00
Do not agree (Instämmer inte)	2		1.43	1.98		1	0.36 0.06
Undecided (Tveksam)	3		2.95	3.00		3	0.24 0.07
Agree (Instämmer)	4		4.58	4.01		5	0.40 0.07
Absolutely agree (Instämmer absolut)	5		5.00	5.00		5	0.00 0.00

Response scales	Numbers used	Rank order breaks	Rank order ties	Means exp. group	Means comp. group	Inter-val ratios	Rounded means	St.dev. exp. group	St.dev. comp. group	15.
Response scale 7		1	0			3.02				
Absolutely wrong (Stämmer absolut inte)	1			1.00	1.00		1	0.00	0.00	
Hardly right (Stämmer knappast)	2			1.79	1.97		2	0.49	0.06	
Mainly right (Stämmer i stort sett)	3			3.45	3.00		3	0.33	0.04	
Absolutely right (Stämmer absolut)	4			4.00	4.00		4	0.00	0.00	
Response scale 8		3	5			3.00				
Not at all interested (Inte alls intresserad)	1			1.00	1.00		1	0.00	0.00	
Somewhat interested (Något intresserad)	2			2.65	1.97		3	0.43	0.06	
Fairly interested (Ganska intresserad)	3			3.20	3.00		3	0.31	0.04	
Very interested (Mycket intresserad)	4			4.00	4.00		4	0.00	0.00	