ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/223194474

Are Likert scales unidimensional?

Article *in* Social Science Research · September 1987

DOI: 10.1016/0049-089X(87)90003-2

CITATIONS
39

READS
207

2 authors, including:

Magnus Stenbeck
Karolinska Institutet

76 PUBLICATIONS **1,748** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project SIMSAM INFRA View project

Project SIMSAM INFRA View project

All content following this page was uploaded by Magnus Stenbeck on 02 April 2018.

The user has requested enhancement of the downloaded file.

SOCIAL SCIENCE RESEARCH 16, 245–259 (1987)

Are Likert Scales Unidimensional?

OTIS DUDLEY DUNCAN AND MAGNUS STENBECK

University of California, Santa Barbara

Likert's method of scoring his categories can be justified by the Rasch Rating-Scale model if the categories produce unidimensional responses. Tests involving eight items from the 1956–1958–1960 American Panel Study suggest that unidimensionality often is not achieved with the Likert categories strongly agree, agree, disagree, and strongly disagree. Combining the first two and the last two of these categories can be justified under a two-dimensional, Content-by-Intensity, Rasch model, provided that the two components are not confounded in the response process. Tests using the same data suggest that such confounding may occur fairly often. Thus the methods commonly used for analyzing Likert data may not be as generally applicable as research workers have assumed. Validation of measurement involving rating scales should be an integral part of any research employing the Likert design. © 1987 Academic Press, Inc.

One of the surprising results from modern investigations in item response theory is that the "method of summated ratings" suggested by Likert (1932) over half a century ago can, under certain conditions, be given a rigorous justification as a consequence of the unidimensional Rasch Rating-Scale model (Andrich 1982; 1985; Wright and Masters 1982; Masters 1985). It is important to note that this result requires only that the Likert categories, such as strongly agree (SA), agree (A), disagree (D), and strongly disagree (SD), comprise an ordinal scale. Specifically, the assignment of integer scores to response categories does *not* require the assumption of "equal distances" separating pairs of adjacent categories. But an ordinal scale, on the usual understanding of that somewhat fuzzy idea, is by assumption unidimensional: it grades a population according to degrees of some single quality that is shared by all persons and items to be scaled. Thus, the categories "can have one and only one order from 'least' to 'most' " (Wright and Douglas 1986, p. 5). These authors,

This research was supported by National Science Foundation Grant SES-8411359. We are indebted to the Inter-University Consortium for Political and Social Research for making available the data from the 1956–58–60 American Panel Study, but neither that organization nor the original investigators are responsible for our tabulations, analyses, and interpretations. Requests for reprints should be sent to O. D. Duncan, Department of Sociology, University of California, Santa Barbara, CA 93106.

0049-089X/87 $3.00

Copyright © 1987 by Academic Press, Inc.
All rights of reproduction in any form reserved.

like Andrich (1985), note that when a polytomous response has $m + 1$ categories there are potentially as many as m distinct dimensions or traits on which persons and items may be differentiated. But they stop short of the logical next step, which is to provide systematic comparisons between uni- and multidimensional models for actual data sets, together with tests appropriate for distinguishing between them. We take an empirical approach here, thereby complementing the formal developments in the works just cited.

Although Likert in his own research used various question forms (J. Converse, 1984), in contemporary usage "Likert item" generally refers to a question with the adverb–verb combinations SA, A, D, and SD (or something of the sort) as precoded response categories which are provided to the respondent; we shall follow that usage. But it should be emphasized that the issue of unidimensionality arises with any form of categorical rating scale (e.g., good, fair, poor or A, B, C, D, F) or any set of ostensibly ordered response alternatives (declare war, send military aid, keep out of the conflict).

Our illustrative data come from the 1956–1958–1960 American Panel Study in the series of national election studies of the Survey Research Center. We consider eight items recently analyzed by Taylor (1983) and Brody (1986) in contributions building on the seminal suggestions of P. Converse (1964) concerning the structure of belief systems. In most of this work, the issue confronted here was bypassed inasmuch as analyses were carried out with collapsed response categories SA + A and D + SD. Our first formulation of a latent trait model may shed some light on the validity of this prevalent practice; but to focus on that issue specifically we also offer a second formulation which provides a definite criterion for recognizing when it will produce misleading results.

We have three-wave panel data for the eight questions below:

Q1: The government should leave things like electric power and housing for private businessmen to handle.

Q2: The government in Washington ought to see to it that everybody who wants to work can find a job.

Q3: If cities and towns around the country need help to build more schools, the government in Washington ought to give them the money they need.

Q4: If Negroes are not getting fair treatment in jobs and housing, the government should see to it that they do.

Q5: The government in Washington should stay out of the question of whether white and colored children should go to the same school.

Q6: This country would be better off if we just stayed at home and did not concern ourselves with problems in other parts of the world.

Q7: The United States should keep soldiers overseas where they can help countries that are against communism.

Q8: The United States should give economic help to the poorer countries of the world even if those countries can't pay for it.

Each of these questions was preceded by a filter question which, in effect, invited the respondent with "no opinion" not to choose any of the response alternatives; these included "not sure, it depends" in addition to SA, A, D, SD. We have omitted responses other than these four, thereby reducing the (unweighted) sample size from over 1000 to several hundred (depending on the question and the combination of years being analyzed). In all our analyses we are comparing the responses of the panel of respondents to the same question administered on two or three occasions. Thus, we are not greatly concerned with the usual issue in scale analysis, which is to determine whether different items are indicators of the same latent trait(s). However, because of the 2- or 4-year interval between interviews, an appreciable number of respondents may have made nonnegligible shifts in their locations on the latent trait(s) pertaining to a single item, in which case that same item may indeed appear to measure different things at different times. The tests to be presented may lead to rejection of our multidimensional models in this case, although we know nothing about the power of such tests.

ANALYSIS OF 4 × 4 TABLES

To begin, we examine the illustrative 4 × 4 cross-classification shown in Table 1, elaborating on the exposition by Hout, Duncan, and Sobel (in press) of the relationship of the Rasch measurement model to this kind of table. With four response categories, the Rasch model may involve as many as three latent traits. However, the number of waves in a panel design affects our ability to detect the operation of more than two of them. This statement will become clearer as we sketch the derivation of the model for the multiway contingency table that is implied by the Rating-Scale formulation of the Rasch model for an individual response.

The top portion of Table 2 succinctly specifies the form of the Rasch model by showing how the parameters for individuals (x_i, y_i, z_i) and those for items (α_t, β_t, γ_t) enter into the response probability. (In the panel design, with a single item administered several times, "item" becomes a label for the "occasion.") The translation of the scheme of exponents into formulas for the probabilities is given in the lower portion of the table. Specification of the Rasch model is completed by invoking the postulate of local independence, which holds that, conditional on (i,t), responses to two or more items (responses on two or more occasions) are independent. This allows us to write the joint probability for any designated sequence of responses as the product of the probabilities given

TABLE 1

Response to Q4 in 1958 by Response in 1960, Observed Counts and Estimates of Expected Frequencies under Three-Trait Model (QSym) and Rating Scale (One-Trait) Model

	1960				
1958	SA	A	D	SD	Total
SA	345	82	10	21	458
A	76	45	4	18	143
D	12	16	8	16	52
SD	20	17	9	57	103
Total	453	160	31	112	756
			Expected, QSym		
SA	345*	83.78	7.88	21.34	458*
A	74.22	45*	6.62	17.16	143*
D	14.12	13.38	8*	16.50	52*
SD	19.66	17.84	8.50	57*	103*
Total	453*	160*	31*	112*	756*
			Expected, Rating Scale		
SA	345*	79.20	13.06	18.64	455.90
A	78.80	41.01	12.03	19.59	151.43
D	12.93	11.97	4.01	12.53	41.44
SD	18.36	19.40	12.47	57*	107.23
Total	455.09	151.58	41.57	107.76	756

* Constrained to be the same as observed count by the conditional maximum likelihood estimation procedure.

in Table 2. The implications of the Rasch model for the aggregate cross-classification of responses (e.g., Table 1) are made explicit by the calculation of expected frequencies for all possible sequences of responses, obtained by summing the joint probabilities over the index i ($1 \leqslant i \leqslant n$, where n individuals comprise a sample from the population in which the Rasch model holds). The outcome is depicted schematically in Table 3, which illustrates a method that may be readily (albeit tediously) extended to three or more items (occasions). Consider, for example, the cell of the 4 × 4 cross-classification pertaining to persons responding A in 1958 and SD in 1960, that is, to the response pattern 24. According to Table 2, the probability of responding A in 1958 is

$$\alpha_1 \gamma_1 x_i z_i / D_{1i}$$

and the probability of SD in 1960 is

$$\alpha_2^3 \beta_2 \gamma_2 x_i^3 y_i z_i / D_{2i}.$$

TABLE 2

Specification of Response Probabilities in the Multitrait Rasch Model with Four Response Categories (Rating Scale Formulation)

Response	Exponents of Parameters		
U	α_t x_i	β_t y_i	γ_t z_i
1 (SA)	0	1	0
2 (A)	1	0	1
3 (D)	2	0	0
4 (SD)	3	1	1
	Response Probability[a]		
1	$\beta_t y_i / D_{ti}$		
2	$\alpha_t \, \gamma_t x_i z_i / D_{ti}$		
3	$\alpha_t^2 x_i^2 / D_{ti}$		
4	$\alpha_t^3 \beta_t \, \gamma_t x_i^3 y_i z_i / D_{ti}$		

[a] Where D_{ti} equals the sum of the numerators of the four response probabilities.

TABLE 3

Parameterization of the Multitrait Model for the 4 × 4 Cross-Classification

	Response		Person parameters[a]			Item parameters[b]					
Cell No.	1958	1960	x_i	y_i	z_i	α_1	α_2	β_1	β_2	γ_1	γ_2
1	1	1	0	2	0	0	0	1	1	0	0
2	1	2	1	1	1	0	1	1	0	0	1
3	1	3	2	1	0	0	2	1	0	0	0
4	1	4	3	2	1	0	3	1	1	0	1
5	2	1	1	1	1	1	0	0	1	1	0
6	2	2	2	0	2	1	1	0	0	1	1
7	2	3	3	0	1	1	2	0	0	1	0
8	2	4	4	1	2	1	3	0	1	1	1
9	3	1	2	1	0	2	0	0	1	0	0
10	3	2	3	0	1	2	1	0	0	0	1
11	3	3	4	0	0	2	2	0	0	0	0
12	3	4	5	1	1	2	3	0	1	0	1
13	4	1	3	2	1	3	0	1	1	1	0
14	4	2	4	1	2	3	1	1	0	1	1
15	4	3	5	1	1	3	2	1	0	1	0
16	4	4	6	2	2	3	3	1	1	1	1

[a] Entries become subscripts of S parameters.

[b] Entries become exponents of item (occasion) parameters.

Hence the probability of the sequence is

$$\alpha_i \alpha_2^3 \beta_2 \gamma_1 \gamma_2 x_i^4 y_i z_i^2 / D_{1i} D_{2i}.$$

The expected frequency of this sequence in a sample of n persons is

$$\alpha_1 \alpha_2^3 \beta_2 \gamma_1 \gamma_2 S_{412},$$

where S_{412} is simply an abbreviation for

$$\sum_{i=1}^{n} x_i^4 y_i z_i^2 / D_{1i} D_{2i}.$$

Note that the triple subscript of S_{412} merely recapitulates the sequence of exponents of x_i, y_i, and z_i (in that order) and that this sequence will be the same for all response patterns involving the same combination of responses irrespective of order. In the example, both the response pattern 24 and pattern 42 give rise to parameter S_{412}, although the two responses are distinguished by the exponents of their item parameters (compare cells 8 and 14 in Table 3).

One final convention must be mentioned. In a Rasch model, it is necessary to adopt an arbitrary unit for the scale on which all (strictly positive) person and item parameters are measured. It is convenient here to set all parameters pertaining to the first occasion to unity, so that $\alpha_1 = \beta_1 = \gamma_1 = 1.0$. This is tantamount to deleting the three columns pertaining to these parameters from Table 3 and modifying the formulas in the preceding paragraph accordingly. (In Tables 1 and 3, the first occasion is taken to be 1958; in cross-classifications involving the 1956 response, it will be convenient to regard that year as the "first" one.)

Inspection of Table 3 shows that there is a distinct S parameter for each unique response combination, or 10 such parameters in all. (Note that the pattern of exponents of person parameters is the same for cells 2 and 5, 3 and 9, 4 and 13, 7 and 10, 8 and 14, and 12 and 15.) Moreover, we see that all 10 combinations are distinguished by the first two subscripts; the third subscript is therefore redundant. Hence, we cannot detect individual differences produced by the third latent trait when the design includes only two items (occasions). Indeed, the same remark holds for the three-way table produced with three items (or a single item administered on three occasions), where we have 20 response combinations (regarding, for example, response patterns 234, 243, 324, 342, 423, 432 as belonging to the single response combination conveniently designated 234). In the design with four items (occasions), there are, altogether, 35 response combinations (accounting for 256 response patterns), and we find only one pair of response combinations, 1333 and 2224, whose first two subscripts are the same (S_{610} and S_{614}), so that the third subscript is meaningful. Even so, the design does not provide any information that can be used to reject the hypothesis $z_i = z_0$, a constant that serves to distinguish between persons giving one of the responses 1333, 3133, 3313, or 3331 and those responding 2224, 2242, 2422, or 4222. When there are five items (occasions), we obtain 36 response combinations (accounting for $4^5 = 1024$ distinct response patterns) and consideration of the exponents of x_i and y_i alone produces no less than 52 distinct S parameters; we must consider the exponent of z_i as well to obtain all 56. Clearly, the design of the present study, with two- and three-wave panel cross-tabulations, will not permit testing of hypotheses concerning the absence of a third dimension of variation among individuals, and even the five-wave design will not have great statistical power for this purpose. The two- and three-trait models do differ, however, in regard to the absence or presence of the parameter(s) pertaining to items.

Our interpretation of the unidimensional—or, more descriptively, the single latent trait—model is that the person parameters y_i and z_i are replaced by constants, say $y_i = y_0$ and $z_i = z_0$ for all i. The entries under y_i and z_i in Table 3 are then interpreted as exponents of y_0 and z_0. There are only 7 S parameters, $S_0, S_1, \ldots, S_6$ (defined by the exponent of x_i), the two constants y_0 and z_0, and the item parameter α_2 in the model for the 4 × 4 cross-classification implied by the Rasch Rating-Scale model. Hence we have $df = 16 - 7 - 2 - 1 = 6$.

By a similar argument, we interpret the two-trait model as one in which the person parameter z_i is replaced with a constant z_0, and there are two item parameters, α_2 and β_2. But we find that z_0 is actually redundant, inasmuch as 10 S parameters are distinguished by the exponents of x_i and y_i, so that there are no further distinctions among response combinations to be captured by z_0. Hence the model has $df = 16 - 10 - 2 = 4$.

The three-trait model, therefore, with 10 S parameters and item parameters α_2, β_2, and γ_2, differs from the two-trait model by only 1 *df*. As Hout et al. (in press) demonstrate, this model with $df = 3$ is equivalent to the model of quasisymmetry (QSym), which is well known in the literature on log linear models for categorical data (see e.g., Caussinus 1965; Bishop, Fienberg, and Holland 1975).

In Table 1 we show the estimates of expected frequencies in one 4 × 4 cross-classification under the QSym and Rating-Scale models. From these estimates we obtain the likelihood-ratio chi-square statistics L^2(QSym) = 2.77 ($df = 3$) and L^2(Rating-Scale) = 15.80 ($df = 6$, $P = .01$). Hence the data may be deemed consistent with the former model but not the latter. Altogether, we have 24 such tables for the 1956 by 1958, 1956 by 1960, and 1958 by 1960 cross-classifications for each of the eight items. Using either Pearson's chi-square statistic X^2 or the likelihood-ratio statistic L^2, we reject QSym (at the conventional .05 level) for only 2 of the 24 tables; for Q1, 1958 by 1968, we obtain $L^2 = 8.53$ ($P = .04$) and for

Q7, 1956 by 1958, $L^2 = 17.84$ ($P < .001$). These results, implying that the 1956, 1958, and 1960 responses to any of the eight questions reflect the same set of person parameters, suggest that 1956–1960 was a period of relative stability of popular attitudes concerning the issues tapped by Qs 1–8.

But the term "stability" must be interpreted rather strictly in this context. The QSym model does allow for unrestricted changes in the aggregate marginal response distributions and—provided the table is one with symmetric associations—attributes them to shifts across years in the locations of the *questions* on the latent trait continua rather than to shifts in the distribution of respondents in regard to the latent traits. The "stability" therefore pertains strictly to respondents' dispositions.

Although respondents' attitudes were apparently stable, the Likert response categories do not produce indicators that are unequivocally unidimensional. We make this inference by assessing the fit of the (unidimensional) Rating-Scale model and by comparing it with the fit of the multitrait model (QSym). In Table 1 we have seen a particular case in which the data are not well described by the stronger model. The same kind of result—failure of the data to fit the Rating-Scale model—is obtained for exactly half of our 24 tables, including the 2 which also fail to fit the QSym model. In fact, only for Q5 do we obtain an acceptable fit to the Rating-Scale model in all three of the pairwise comparisons of years. For 14 of the 24 tables we find a significantly better fit for the QSym model than for the Rating-Scale model. (A more detailed presentation of results for three-way tables will be given below.)

ANALYSIS OF THREE-WAY TABLES

We have already sketched the derivation of models for the 4 × 4 × 4 cross-classifications, inasmuch as it involves a direct extension of the procedure given in detail in Tables 2 and 3. The unitrait Rasch Rating-Scale model implies a log linear model for this 64-cell contingency table with 10 S parameters defined by exponents ("scores") of x_i, the two constants y_0 and z_0, and the two item parameters α_2 and α_3. Hence $64 - 10 - 2 - 2 = 50$ df. Recognition of a second latent trait (y_i) requires us to distinguish 20 S parameters, one for each possible combination of responses. The constant z_0 then becomes wholly redundant; but there are two more item parameters, β_2 and β_3. Hence $df = 64 - 20 - 4 = 40$. Finally, the three-trait model involves two additional item parameters, γ_2 and γ_3, so that $df = 38$. This model is equivalent to the straightforward generalization of QSym which has been termed "total quasisymmetry" by Sobel (1986). The operation of the third latent trait is recognizable only as it is reflected in the item parameters. The three models are connected by a chain of implications, inasmuch as the Rating-Scale model is a special case of total QSym. Hence in assessing the applicability of

TABLE 4
Likelihood Ratio χ^2 Statistics (L^2) for fit of Alternative Models to 4 × 4 × 4 Cross-Classifications

	L^2 for model fit			Differences in L^2		
Question	1-dim (i)	2-dim (ii)	3-dim (iii)	(i) – (ii)	(i) – (iii)	(ii) – (iii)
1	60.43	39.01	36.71	21.42*	23.72*	2.29
2	74.62*	47.54	43.35	27.07*	31.26*	4.19
3	62.58	36.64	33.24	25.95*	29.34*	3.39
4	64.61	50.02	42.78	14.60*	21.83*	7.23*
5	65.20	43.72	42.69	21.48*	22.51*	1.03
6[a]	110.87*	59.30*	54.42*	51.58*	56.45*	4.88
7	65.86	46.33	43.28	19.52*	22.57*	3.05
8	78.43*	48.85	45.93	29.58*	32.50*	2.92
df	50	40	38	10	12	2

[a] For (iii), Pearson's statistic X^2 = 52.15 (P = .06).
* P < .05

these models we may not only examine the statistics pertaining to the fit of each but also compare the respective fits by considering differences in their L^2 statistics in the light of differences in their *df*. The results are summarized in Table 4. The general pattern is clear. While the Rating-Scale model enjoys a reasonably good fit to five of the eight questions, for all eight questions the three-dimensional model provides a fit whose superiority is statistically significant—see Table 4, column (i) – (iii). There is clear evidence, therefore, that people and items are heterogeneous in ways not captured by the view of the Likert categories SA, A, D, SD as a simple ordinal scale. Results summarized in the last column of Table 4 seemingly support the two-dimensional model in preference to the 3-dimensional model in most analyses. But it is well to recall here that the three-wave design is intrinsically incapable of detecting the third dimension insofar as differences among individuals are concerned.

ALTERNATIVE FORMULATION

As we noted earlier, much of the hitherto reported work with the 1956–1958–1960 panel data involved a reduction of the four Likert categories to two, combining SA with A and D with SD. No special justification of this procedure was given by the investigators using these data, but occasionally one does find an explicit rationale such as the one given by Reiss (1967, pp. 220–221), whose data were obtained by giving respondents a choice of six categories: strong agree, medium agree, slight agree, slight disagree, medium disagree, and strong disagree.

> The questions were treated as dichotomous . . . . It was possible to trichotomize several of the twelve items in the scale and still meet all the Guttman requirements. However, for thc sake of simplicity, all questions were dichotomized into *agree* or *disagree* regardless of the intensity of such feelings. In research it is still advisable to use the six-way choice after each question, for without such an elaborate choice some respondents feel that they are not able to elaborate fully their beliefs in this area. Such a concession to respondent satisfaction is a small price to pay for cooperation.

Also relevant is the body of work on the "intensity component" of attitudes in connection with the *Studies in Social Psychology in World War II,* where the investigators often asked the soldiers first for a positive or negative reaction on some matter and then asked a separate question like "How strongly do you feel about this?" (Suchman 1950, p. 219). Presumably this design feature was motivated by the assumption that the Likert format is vulnerable to a confounding of the Content and Intensity components, whereas the assumption of the procedure used by Reiss and many others is that there is no such confounding.

To investigate this issue empirically with the data studied here requires a model that focuses on the conceptual distinction between the two components. As Table 5 illustrates, we can modify the previous formulation so as to replace the first latent trait with one that governs only the propensity to agree or disagree (whether strongly or otherwise). Hence w_i pertains to the strength of the individual's disposition to say "disagree" (the Content component) and y_i pertains to the Intensity of her or his feelings about the issue. Then z_i is clearly the parameter which produces an association of the two when the response is elicited. If Content and Intensity are really distinct dimensions and if the design of the Likert question does not artifactually produce a confounding of them, then $z_i = 1$, that is, the model reduces to one with only a pair of latent traits. Another possibility is that for all respondents it is, say, more acceptable to express oneself "strongly" in agreement with a proposition than in disagreement. This could be reflected in a model with $z_i = z_0$, a constant that is the same for all persons.

In the analysis, therefore, we are especially interested in comparing the two-trait model (with and without the constant z_0) with the three-trait model (with person parameter z_i). The derivation of log linear models for two-way and three-way cross-classifications is accomplished in the manner already illustrated for the Rating-Scale formulation. Hence, it should suffice to summarize the main features of the alternative models. For the 4 × 4 table, the two-trait (Content and Intensity) model has 9 S parameters, defined by "scores" on w_i and y_i, and the two item parameters δ_2 and β_2. Hence, without the constant z_0, the model has $df = 16 - 9 - 2 = 5$. If the constant is included and allowed to change by reason

TABLE 5
Specification of Response Probabilities in the Multitrait Rasch Model with Four Response Categories (Content by Intensity Formulation)

	Exponents of parameters		
Response	δ_t	β_t	γ_t
U	w_i	y_i	z_i
1 (SA)	0	1	0
2 (A)	0	0	1
3 (D)	1	0	0
4 (SD)	1	1	1
	Response probability[a]		
1	$\beta_t y_i / D_{ti}$		
2	$\gamma_t z_i / D_{ti}$		
3	$\delta_t w_i / D_{ti}$		
4	$\delta_t \beta_t \gamma_t w_i y_i z_i / D_{ti}$		

[a] Where D_{ti} equals the sum of the numerators of the four response probabilities.

of the parameter γ_2, we have $df = 3$ and the model is equivalent to QSym, as in the first formulation. The two-wave design cannot actually detect the third trait as a variable differentiating individuals, but it can detect an association of Content and Intensity which may be the same for all persons. The association between Content and Intensity in the parameterization of response probabilities in Table 5 is given by the odds ratio

$$P_{4i}P_{2i}/P_{1i}P_{3i},$$

where $P_{gi} = Pr(U=g|i)$. Evaluating this ratio in terms of model parameters we have

$$(\delta_t\beta_t\gamma_t w_i y_i z_i/D_{ti})\,(\gamma_t z_i/D_{ti})/(\beta_t y_i/D_{ti})\,(\delta_t w_i/D_{ti}) = \gamma_t^2 z_i^2 .$$

Hence, even if $z_i = z_0$ (but $z_0 \neq 1$) and $\gamma_t = 1.0$, Content and Intensity will be associated *for each respondent,* so that the two dimensions will not be separable. The implications of this situation for the 4 × 4 cross-classification may be brought out by rearranging the table into a format (Table 6) suited for analysis as a 16-fold table (see Duncan 1985, especially pp. 116–120, where the situation analogous to the one described here is described in terms of "correlated changes"). In Table 6 we focus on respondents whose responses changed between 1956 and 1958 in regard to both Content (D to A or A to D) and Intensity (S to W or W to S), that is, on the four cells defined by the intersection of the second and third rows and the second and third columns. The cross-product ratio for these cells is

$$\gamma_2^2 S_{112}^2/S_{110}^2,$$

TABLE 6
Parameterization of the Three-Trait Content × Intensity Model for the 4 × 4 Cross-Classification Rearranged as a 16-fold Table

			Intensity[a]			
Content		1956	S	S	W	W
1956	1958	1958	S	W	S	W
D	D		$\delta_2\beta_2\ \gamma_2 S_{222}$	$\delta_2 S_{211}$	$\delta_2\beta_2\ \gamma_2 S_{211}$	$\delta_2 S_{200}$
D	A		$\beta_2 S_{121}$	$\gamma_2 S_{112}$	$\beta_2 S_{110}$	$\gamma_2 S_{101}$
A	D		$\delta_2\beta_2\ \gamma_2 S_{121}$	$\delta_2 S_{110}$	$\delta_2\beta_2\ \gamma_2 S_{112}$	$\delta_2 S_{101}$
A	A		$\beta_2 S_{020}$	$\gamma_2 S_{011}$	$\beta_2 S_{011}$	$\gamma_2 S_{002}$

[a] W = not "strongly."

which, in general, will not equal unity. Now if we replace z_i with z_0, the third subscript of each S parameter is deleted, but each cell now has the additional multiplicative parameter z_0 with an exponent equal to what was hitherto that subscript. With this interpretation of the three-trait model, the association for the four "supercritical" cells (Duncan, 1985, p. 111) becomes

$$\gamma_2^2 z_0^4,$$

and this association vanishes only if $\gamma_2 = z_0 = 1.0$. A test of significance on this association or, equivalently, a test of the hypothesis of row–column independence in the 2 × 2 table comprised by the supercritical cells can be used to resolve this issue separately from the broader issue of the fit of the entirety of the table to the three-trait model.

The test just described was performed for each of our 24 4 × 4 tables (eight questions, three pairs of years), and the hypothesis of independence in the supercritical cells was rejected at the .05 level for exactly one-third of them. The hypothesis is rejected for the comparison of 1956 with 1960 for Qs 1, 4, 6, 7, and 8; for the comparison of 1958 with 1960 for Qs 1 and 8; and for the 1956–58 comparison for Q8. Three questions, 2, 3, and 5, produce no rejections, so that one might tentatively argue that the separation of the two dimensions, Content and Intensity, was achieved for these questions. But the very fact that it breaks down for the majority of questions means that it cannot be taken for granted.

The bearing of these results on the procedure of collapsing over the Intensity categories is brought out by further study of Table 6. The collapsing amounts to using only the row marginal totals of the table to estimate the item parameter contrasting 1958 with 1956 with regard to Content, i.e., to estimate δ_2. But we see that the ratio of the total for the third row to the total for the second row is

$$\frac{\delta_2(\beta_2\gamma_2 S_{121} + S_{110} + \beta_2\gamma_2 S_{112} + S_{101})}{(\beta_2 S_{121} + \gamma_2 S_{112} + \beta_2 S_{110} + \gamma_2 S_{101})} \neq \delta_2.$$

However, if we delete the third subscript of the S parameters and also delete γ_2–stipulating, in effect, that Content and Intensity are uncorrelated for every individual—the expressions in parentheses in the numerator and denominator are the same, so that the ratio of the two marginal totals becomes simply δ_2, as we would want it to be. If independence in the supercritical cells does not hold—as we infer it does not in 8 of our 24 tables—then this ratio no longer estimates δ_2 but some opaque composite of item and person parameters.

Extending the Content-by-Intensity formulation to the case of the three-way (4 × 4 × 4) table, we find that the model with all three latent traits is equivalent to QSym (and therefore equivalent to the three-trait model in the Rating-Scale formulation). If latent trait z_i and item parameters γ_2 and γ_3, all pertaining to the association of Content with Intensity, are deleted from the model, we find that there are 16 S parameters derived from scores on w_i and y_i, as well as the 4 item parameters δ_2, δ_3, β_2, and β_3. Hence the model has $64 - 16 - 4 = 44$ *df*. We see in Table 4 that the three-trait model (QSym) is rejected only for Q6 on the basis of the L^2 statistic for the fit of the model (although that decision would be reversed if Pearson's X^2 statistic were used instead). In none of the eight three-way tables do we find either of the parameters γ_2 and γ_3 as large in absolute value as 1.96 times it's estimated standard error. If we delete these parameters from the model and recognize only the 16 S parameters reflecting traits w_i and y_i, then we have a model with 43 *df* that differs from the two-trait model described above only with regard to inclusion of the parameter z_0. This parameter is significantly large by comparison with its standard error in the estimates pertaining to Qs 1, 6, and 8; and this modification of the two-trait model affords an unsatisfactory overall fit to the data for Qs 5 and 6 as judged by the L^2 statistic.

In evaluating the results just summarized it is well to remember that we are working with just a few hundred cases (the sample sizes vary from 446 for Q8 to 704 for Q6) spread over 64 cells of the three-way cross-classification, while the cells that provide specific information about the comparison of the two- and three-trait models are relatively few in number and not necessarily the most heavily populated ones either. Hence if we get even a few definite indications of a confounding of Content and Intensity in the response process, that should be enough to serve as a warning against the easy assumption that the response categories can be rendered unidimensional by an ex post facto collapsing. We therefore reiterate the caution of Rasch (1966, p. 107), "If the basic model holds for . . . five categories, it is mathematically almost impossible for the three-category model also to hold. Thus the grouping, tempting as it may be, will usually tend to slur the specific objectivity."

CONCLUSION

In opinion research the Likert categories are sometimes preferred over a simple agree/disagree dichotomy on the assumption that expanding the number of response categories enhances the precision of the single item in estimating the individual's location on a single latent continuum. Adding some qualifying adverbs to the basic agree/disagree dichotomy is believed to provide a finer graduation of the measurement instrument. For repeated measurement, or when using several different Likert items to measure a single latent variable, Likert's method of adding integer scores has been used in order to arrive at a total person-score, which is usually regarded as an "interval scale." As in the first argument in favor of the Likert categories, the total score is held to provide better measurement of the location of a person on the latent continuum than the score on a single item because it provides a larger number of distinct observable discrete values. Both assertions rest heavily on the assumption that Likert items produce unidimensional measurement. If, on the contrary. Likert's elaboration of the response categories introduces additional latent variables into the response process and if the measurement of additional variables interferes with the measurement of the original latent trait, then the method threatens to contaminate rather than to improve measurement.

Our empirical results suggest that the conventional Likert categories cannot be assumed to comprise a unidimensional scale in survey inquiries on political opinions. Nor can their shortcoming in this regard necessarily be overcome by collapsing categories SA and A as well as D and SD, since there is disturbing evidence that the distinction between Content and Intensity, which could justify this maneuver, may well be confounded by the Likert design. It remains to be shown, however, that Suchman's device of using separate questions to get at the two components consistently overcomes this problem.

Rating scales are so widely used in investigations of attitudes and other subjective phenomena that one might well be hesitant in generalizing these results beyond the kind of subject matter considered here. But we venture to suggest, nevertheless, that "ordinal scales" are not created by fiat, as though the usual scientific criteria for validating a measurement procedure can be dispensed with in favor of a common sense interpretation of category semantics. It may take a great deal more experimental research and qualitative analysis to develop rating categories that are relatively invulnerable to the pathologies noted here.

REFERENCES

Andrich, D. (1982). Using latent trait measurement models to analyze attitudinal data: A synthesis of viewpoints. In D. Spearritt (Ed.), *The improvement of measurement in education and psychology* (pp. 89–126). Melbourne: Australian Council for Educational Research.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. B. Tuma (Ed.), *Sociological methodology 1985* (pp. 33–80). San Francisco: Jossey–Bass.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA.: MIT Press.

Brody, C. J. (1986). Things are rarely black and white: Admitting gray into the Converse model of attitude stability. *American Journal of Sociology,* **92,** 657–677.

Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux des correlation. *Annales de la Faculté des Sciences, de l'Université de Toulouse,* **29,** 77–182.

Converse, J. M. (1984). Attitude measurement in psychology and sociology: The early years. In C. F. Turner and E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 3–39). Troy, New York: Russell Sage Foundation.

Converse, P. E. (1964). The nature of belief systems in mass publics. In D. E. Apter (Ed.), *Ideology and discontent* (pp. 206–261). London: Free Press of Glencoe.

Duncan, O. D. (1985). New light on the 16-fold table. *American Journal of Sociology,* **91,** 88–128.

Hout, M., Duncan, O. D., and Sobel, M. E. (in press). Association and heterogeneity: Structural models of similarities and differences. In C. C. Clogg (Ed.), *Sociological methodology 1987.*

Masters, G. N. (1985). A comparison of latent trait and latent class analyses of Likert-type data. *Psychometrika,* **50,** 69–82.

Rasch, G. (1966). An individualistic approach to item analysis. In P. F. Lazarsfeld and N. W. Henry (Eds.), *Readings in mathematical social science* (pp. 89–107). Chicago: Science Research Associates.

Reiss, I. L. (1967). *The social context of premarital sexual permissiveness.* New York: Holt, Rinehart & Winston.

Sobel, M. E. (1986). Some models for the multiway contingency table with a one-to-one correspondence among categories. Unpublished manuscript.

Suchman, E. A. (1950). The intensity component in attitude and opinion research. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, and J. A. Clausen, *Measurement and prediction.* Princeton, NJ: Princeton Univ. Press.

Taylor, M. C. (1983). The black-and-white model of attitude stability: A latent class examination of opinion and nonopinion in the American public. *American Journal of Sociology,* **89,** 373–401.

Wright, B. D., and Douglas, G. A. (1986). The rating scale model for objective measurement. (Memorandum No. 35). University of Chicago: MESA Psychometric Laboratory.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

View publication stats