EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXV, NO. 4, 1965

NUMBER OF SCALE POINTS AND THE RELIABILITY OF SCALES

S. S. KOMORITA AND WILLIAM K. GRAHAM
Wayne State University

AMONG the standard scales and inventories in current use, there seems to be relatively little agreement regarding the number of scale points or categories which are used to obtain responses to items. For example, in the California Psychological Inventory (Gough, 1957), a true-false, two-point scale is used; in a Likert-type attitude scale (Likert, 1932), a five-point scale is employed; and in the California F-scale (Adorno, et al, 1950) or the Semantic Differential (Osgood, Suci, and Tannenbaum, 1957), a seven-point scale is employed. Clearly, scales which employ a two-point scale such as agree-disagree, true-false, etc., are much shorter and more convenient to administer and score. If so, why do so many investigators recommend and use the seven-point scale?

One obvious criterion for the choice of the number of scale categories is the ability of subjects to discriminate between the categories. While a scale with few categories may not allow the subject to make full use of his capacity to discriminate, a scale with a large number of categories may be beyond the subject's capacity to discriminate, thus, increasing errors of measurement. Thus, probably the main reason for using a large number of categories is to increase the reliability of the scale. As applied to rating scales, in one of the earliest studies of this problem, Symonds (1924) made a theoretical analysis using Kelley's correction for coarse grouping on the correlation coefficient, and concluded that the optimal number of categories to maximize scale reliability is seven, and that the increase in reliability when more categories are used is negligible. However,

more recent empirical studies on this problem have not supported Symonds' conclusions.

Champney and Marshall (1939) compared the correlation between two forms of a graphic rating scale and measured responses to each form by two methods: using a coarse centimeter scale and a finer millimeter scale. The correlation between the two forms for the millimeter scale was significantly higher than for the centimeter scale, .77 versus .67. They concluded that the optimal number of scale categories is a function of the conditions of measurement and cannot be determined by the method proposed by Symonds. Guilford (1954, p. 291) is in general agreement with this point of view and claims that, "the number 7 recommended by Symonds is usually lower than optimal and it may pay in some favorable situations to use up to 25 scale divisions." On the other hand, two studies by Bendig (1953, 1954) suggest that fewer than seven categories may be justified under certain conditions. In the first of these studies, he found that reliability remained relatively constant for self-rating scales with 3, 5, 7, and 9 categories but significantly decreased for 11 categories. In the second study, Bendig confirmed the results of his first study and found no significant differences in the reliability of rating scales with three to nine categories, but found that a two-point rating scale was significantly lower in reliability than those with three to nine categories. Thus, his results suggest that although there is a definite advantage in using more than a two-point scale, it may be possible under certain conditions to use fewer than seven categories without sacrificing reliability.

In summary, the over-all results of studies concerning the effects of number of scale points on the reliability of rating scales indicate that in some situations *more* than seven categories are optimal, while in other situations, *fewer* than seven categories may be justified. Thus, the effect of number of scale points on reliability seems to vary with the stimulus situation. The major implication, of course, is that the problem is much more complex than simply applying a correction for coarse grouping.

It should be emphasized, at this point, that the studies which have been reviewed dealt primarily with the relationship between number of scale points and the reliability of a *single rating scale.* The purpose of the present study, however, was to determine the effects of number of scale points on the reliability of inventories and test-scales which consist of the *sum of a set of rating scales* (e.g., Likert's

method of summated ratings). In one of the few studies dealing specifically with this problem, Bendig (1954) obtained ratings of food preferences for 20 different foods and pooled the 20 ratings for each subject. Five rating scales with 2, 3, 5, 7, and 9 scale categories were administered to five groups of subjects, each group receiving one of the scales. Using Hoyt's analysis of variance technique as a measure of reliability, the reliabilities of the five scales ranged from .60 to .70, and he concluded, therefore, that test reliability is independent of the number of scale categories.

Similar results were obtained in a study by Komorita (1963). Two forms of a Likert-type attitude scale consisting of 14 items, each with a six-point scale, were administered to a sample of 286 subjects. Responses to the two forms were scored using a two-point scale as well as using the six-point scale. For the six-point scale, the correlation between the two forms was .93, while the correlation for the two-point scale was .91, thus, confirming Bendig's results. In a supplementary study, the same comparison was made for two random samples of three items from the original scales. For the six-point scale, the correlation between the two sets of three items was .83, while the correlation for the two-point scale was .71. Since the difference in reliabilities between the six-point and two-point scales was considerably larger for three items than for the 14 items, it was suggested that if a scale consists of a very small number of items, somewhat better reliability might be obtained if a six or seven-point scale is used instead.

The over-all results of the studies by Bendig and by Komorita indicate that test reliability is independent of number of scale points. If this is a valid generalization, the major implication is that, because of simplicity and convenience in administration and scoring, all inventories and scales ought to use a dichotomous, two-point scoring scheme. Similar conclusions have been made by Peabody (1962) who attempted to analyze scores on bi-polar scales into two components: direction of response and intensity or extremeness of response. His results indicated that composite scores consisting of the sum of scores on bi-polar, six-point scales mainly reflect direction of response, and are minimally influenced by extremeness of response. He concluded, therefore, that there is justification for scoring bi-polar items dichotomously according to direction of response.

There are some reasons to believe, however, on both empirical and

theoretical grounds, that a number of variables may affect the generality of this principle. It is plausible, for example, that the number of items in the scale or the homogeneity of the items in the scale might affect the relationship between test reliability and number of scale points; moreover, there might be an interaction between these variables in their effects on test reliability. It is desirable, therefore, to specify in more detail the conditions under which test reliability is independent of number of scale points. Accordingly, the primary purpose of this study was to determine the effects of two variables, number and homogeneity of items, on the relationship between test reliability and number of scale points.

Method

Scales

To determine the effects of homogeneity of item content, two scales were selected known to vary in homogeneity of content: (a) a relatively homogeneous scale consisting of 24 bi-polar, Semantic Differential adjectives (Osgood, et al, 1957) selected for their high loadings on the evaluative factor (*S*s were asked to rate Gov. George Romney of Michigan), and (b) a random sample of 24 items from the sociability subscale of the California Psychological Inventory (Gough, 1957) whose internal consistency reliability is reported to be approximately .70.

The sociability scale, hereafter referred to as the CPI scale, and the 24 bi-polar adjectives, hereafter referred to as the SD scale, were each presented in two forms. In Form A, the items were presented with a two-category scale, while in Form B, the items were presented with a six-category scale. Except for differences in number of scale categories, the two forms were identical.

Subjects

The *S*s were 260 students enrolled in undergraduate psychology courses. Forms A and B of the CPI were administered to 67 and 56 *S*s, respectively, and Forms A and B of the SD were administered to 67 and 70 *S*s, respectively.

Procedure

The plan of the study was to compare the difference in reliability between the two-point and six-point scales for the SD and CPI

scales. For this purpose, Cronbach's coefficient alpha (1951) was determined for each form. In order to minimize errors of coarse grouping for the two-category interitem correlations, each form was randomly partitioned into four subsets of six items and coefficient alpha was computed with k equal to four.

To determine the effects of number of scale points on reliability as a function of number of items, the reliabilities of the scales for 3, 6, 12, and 36 items were estimated by applying the Spearman-Brown formula to the previously-determined alpha coefficients. As an empirical check on the assumptions of the Spearman-Brown estimates, for each form the mean intercorrelations between the four random subsets of six items was compared with the theoretical Spearman-Brown estimates. In addition, each form was also partitioned into eight random subsets of three items and two random subsets of 12 items (split-half), and the empirical values were compared with the Spearman-Brown estimates.

Results

Table 1 shows the alpha coefficients for each of the forms. It can be seen that the difference between the two-point and six-point SD scales is negligible, while the difference between the two-point and six-point CPI scales is moderately large. This differential effect of number of scale points as a function of the homogeneity of the scale is further demonstrated in Figure 1. This figure shows the

TABLE 1
Alpha Coefficients for Two-Point and Six-Point SD and CPI Scales

	SD		CPI	
	2-pt.	6-pt.	2-pt.	6-pt.
n	67	70	67	56
a	.920	.916	.620	.740

theoretical Spearman-Brown estimates of reliability as a function of number of items as well as the empirical values for 3, 6, and 12 item subscales. For three and six items, the empirical values represent the mean intercorrelations between eight and four random subscales, respectively, while for 12 items, the empirical values represent a single split-half correlation. It can be seen that an excellent fit between theoretical and empirical values was obtained, thus,

justifying the use of the Spearman-Brown formula. Figure 1 also shows that the differences between the two-point and six-point SD scales are negligible across number of items (in fact, the reliabilities for the two-point form are slightly higher than for the six-point form), while the reliabilities of the six-point CPI subscales are *consistently* higher than for the two-point CPI subscales.

Finally, it should be noted that differences between the two-point

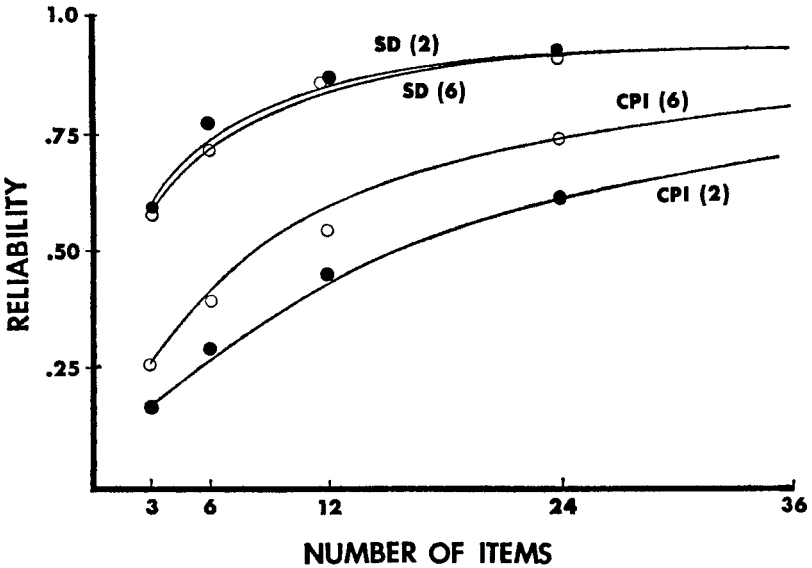

Figure 1. Reliability as a function of number of items (curves represent theoretical Spearman-Brown estimates while points represent empirical values).

and six-point forms, as function of number of items, are relatively constant for *both* the SD and the CPI scales. Althought this result would suggest that number of items in the scale has no effect on the relationship between number of scale points and reliability, theoretically at least, the reliability of the scales should converge to 1.0 as the number of items becomes indefinitely large. Thus, it is quite plausible that if the number of items in the CPI scale were to be made indefinitely large, the *absolute difference* between the two-point and six-point scales would be much smaller. For example, if the CPI scales were to be increased to 240 items, the Spearman-Brown estimates for the two-point and six-point scales would be .942 and .965, respectively, and this difference is considerably smaller than the difference of .620 and .740 for the 24 item scales.

Hence, it is reasonable to assume that if a much larger number of items had been used, the results would have indicated that the effects of number of scale points on reliability would become negligible as the number of items in the scale became indefinitely large.

Discussion and Conclusions

For the SD scale, the results of this study are consistent with the results found by Bendig (1954) and by Komorita (1963), and indicate that with a relatively homogeneous set of items, the reliability of a scale is independent of the number of item scale points. If the items are relatively heterogeneous, however, the results suggest that the reliability of the scale can be increased not only by increasing the number of items but also by increasing the number of item scale points. Just how heterogeneous the scale must be before one can expect a reasonable increase in reliability is a matter for further research. The results of this study suggest that if the reliability of a scale with approximately 25 items and a two-point scale is approximately .60, if the number of scale points is increased to six, one can expect an increase in reliability of about .15.

The major implication of this result is that if there are reasons to believe that the items in a proposed scale or inventory can be expected to be homogeneous, either by item analysis selection of items as in Likert's technique of scaling or by a factor analysis of a set of items as in the Semantic Differential, then a two-category response format such as true-false, agree-disagree, etc., will yield as high a reliability coefficient as a multi-category system. In terms of ease of administration and scoring, therefore, a two-point scheme seems to be preferable to a multi-category scheme.

On the other hand, even with a homogeneous scale, the use of a multi-category system, under certain conditions, may be justified. In using Guttman's scale analysis (1950), for example, an investigator may wish to obtain a separate intensity score as well as a content score using what Suchman has described as the "foldover technique" (1950). Moreover, with a two-point format and an extremely small number of items, the variability of the measure would be quite limited, and it may be desirable to use a multi-category response scheme to increase variability.

It should be strongly emphasized, however, that the above discussion certainly is not meant to imply that the decision to use a small

or large number of scale categories should be based solely on the criterion of increasing reliability. The present study as well as most previous studies on this problem has emphasized reliability as the major criterion in the choice of number of scale categories. The ultimate criterion, of course, is the effect on the validity of the scale. In this connection, a reasonable question one could ask is, "with a relatively heterogeneous scale, (and not with a homogeneous scale), why should an increase in the number of scale points increase the reliability of the scale?"

One possible explanation is that an increase in number of scale points increases the precision of the measuring instrument comparable to measuring height in inches rather than in feet or yards. This explanation, however, does not account for the differential effects for homogeneous scales. A more plausible explanation, therefore, is that some type of response set such as an "extreme response set," (Cronbach, 1946; 1950) may be operating to increase the reliability of heterogeneous scales. If the reliability of the response set component is greater than the reliability of the content component of the scale, the reliability of the scale will be increased by increasing the number of scale points. On the other hand, if the reliability of the response set component is less than or equal to the reliability of the content component, as in a homogeneous scale, the reliability of the scale may not be affected or even decreased. Thus, it is reasonable to assume that increasing the number of scale points permits an extreme response set to be evoked, and the use of a two-point response scale eliminates or minimizes this set.

If this interpretation has any validity, a major implication is that the increase in reliability produced by an increase in number of scale points is a spurious one and may or may not increase the validity of the scale. If the response set component correlates with the criterion, the validity of the scale should be increased by the increase in reliability. However, if the response set component does not correlate with the criterion, the validity of the scale should not be affected despite the increase in reliability. Thus, we have an anomolus relation between reliability and validity as in the attenuation paradox (Loevinger, 1954) where an increase in the reliability of a scale may not have any effect on the validity of the scale. Studies are currently in progress to determine the validity of this interpretation.

REFERENCES

Adorno, T. W., Frenkel-Brunswik, Else, Levinson, D. J. and Sanford, R. N. *The Authoritarian Personality.* New York: Harper, 1950.

Bendig, A. W. "The Reliability of Self-Ratings as a Function of the Amount of Verbal Anchoring and of the Number of Categories on the Scale." *Journal of Applied Psychology,* XXXVII (1953), 38–41.

Bendig, A. W. "Reliability and the Number of Rating Scale Categories." *Journal of Applied Psychology.* XXXVIII (1954), 38–40.

Champney, H. and Marshall, Helen. "Rater's Minimal Discrimination as a Criterion for Determining the Optimal Refinement of a Rating Scale." *Journal of Applied Psychology,* XXIII (1939), 323–331.

Cronbach, L. J. "Response Sets and Test Validity." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, VI (1946), 475–494.

Cronbach, L. J. "Further Evidence on Response Sets and Test Design." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, X (1950) 3–31.

Cronbach, L. J. "Coefficient Alpha and the Internal Structure of Tests."*Psychometrika,* XVI (1951), 297–334.

Gough, Harrison G. *Manual, California Psychological Inventory.* Palo Alto: Consulting Psychologists Press, 1957.

Guilford, J. P. *Psychometric Methods.* (2nd Edition). New York: McGraw-Hill Book Co., Inc. 1954.

Guttman, L. "The Basis of Scalogram Analysis." In S. A. Stouffer, *et al.* (Editors) *Measurement and Prediction.* Princeton, New Jersey: Princeton University Press, 1950.

Komorita, S. S. "Attitude Content, Intensity, and the Neutral Point on a Likert Scale." *Journal of Social Psychology,* LXI (1963), 327–334.

Likert, R. "A Technique for the Measurement of Attitudes." *Archives of Psychology,* 1932, No. 140.

Loevinger, Jane. "The Attenuation Paradox in Test Theory." *Psychological Bulletin,* LI (1954), 493–504.

Osgood, C. E., Suci., C. J., and Tannenbaum, P. H. *The Measurement of Meaning.* Urbana: University of Illinois Press, 1957.

Peabody, Dean. "Two Components in Bipolar Scales: Direction and Extremeness." *Psychological Review,* LXIX (1962), 65–73.

Suchman, E. A. "The Intensity Component in Attitude and Opinion Research." In S. A. Stouffer, *et al* (Editors) *Measurement and Prediction.* Princeton, New Jersey: Princeton University Press, 1950.

Symonds, P. M. "On the Loss of Reliability in Ratings Due to Coarseness of the Scale." *Journal of Experimental Psychology,* VII (1924), 456–461.