Is There an Optimal Number of Alternatives for Likert-Scale Items? Study I

ARTICLE *in* EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT · OCTOBER 1971

Impact Factor: 1.15 · DOI: 10.1177/001316447103100307

Jacob Jacoby
New York University

Is There an Optimal Number of Alternatives for Likert Scale Items? Study I: Reliability and Validity

Michael S. Matell and Jacob Jacoby

Educational and Psychological Measurement 1971 31: 657
DOI: 10.1177/001316447103100307

Educational and Psychological Measurement
1971, 31, 657–674.

IS THERE AN OPTIMAL NUMBER OF ALTERNATIVES FOR LIKERT SCALE ITEMS? STUDY I: RELIABILITY AND VALIDITY

MICHAEL S. MATELL[1] and JACOB JACOBY
Purdue University

Given that rating scales are so widely used in the social sciences, both as research tools and in practical applications, determination of the optimal number of rating categories becomes an important consideration in the construction of such scales. As Garner (1960) pointed out, the basic question is whether for any given rating instrument there is an optimum number of rating categories, or at least a number of rating categories beyond which there is no further improvement in discrimination of the rated items. Garner and Hake (1951), Guilford (1954), and Komorita and Graham (1965) indicated that if we use too few rating categories, our scale is obviously a coarse one, and we lose much of the discriminative powers of which the raters are capable. Conversely, we could also grade a scale so finely that it is beyond the rater's limited powers of discrimination.

Ghiselli (1948) and Guilford (1954) contended that the optimal number of steps is a matter for empirical determination in any situation, and suggested that there is a wide range of variation in refinement around the optimal point in which reliability changes very little. Guilford felt that it may be advisable in some favorable situations to use up to 25 scale deviations. Ghiselli suggested that either reliability of measurement or ease of rating be used as a basis for the empirical determination of the optimal number of steps. Factors which affect the optimal number of rating categories, or at

[1] Now at the Procter and Gamble Company, Cincinnati, Ohio.

least the number beyond which there will be no further improvement in discrimination, are, according to Garner (1960), clearly a function of the amount of discriminability inherent in the items being rated. He suggested that there can be no single number of categories appropriate for all rating situations.

Champney and Marshall (1939) reported that under favorable rating conditions the practice of limiting rating scales to five or seven points may often give inexcusably inaccurate results. They suggested that the optimal number of steps is a function of the conditions of measurement. They also considered that, unless it could be shown that for a particular task either accuracy is not desirable or discrimination beyond seven points cannot be attained, it may be appropriate to use 18- to 24-step rating scales. Both Champney and Marshall and Guilford suggested that when the rater is trained and interested, the optimal number of steps may be in the 20-point range. The literature, however, contains few descriptions of scales employing such large numbers of rating categories.

Jahoda, Deutsch, and Cook (1951) and Ferguson (1941) opined that the reliability of a scale increases, within limits, as the number of possible alternative responses is increased. Cronbach (1950) suggested that there is no merit to increasing the reliability of an instrument unless its validity is also increased at least proportionately. He concluded that "it is an open question whether a finer scale of judgment gives either a more valid ranking of subjects according to belief, or scores more saturated with valid variance (p. 22)." Earlier, Symonds (1924), in contrast to Cronbach, contended that the problem of determining the number of steps to utilize is primarily one of reliability. He implied that optimal reliability is obtained with a 7-point scale. If more than seven steps are utilized, increases in reliability would be so small that it would not pay for the extra effort involved. However, if the raters are untrained or relatively disinterested, maximal reliability will be reached with fewer steps. Champney and Marshall (1939) suggested that nine steps in a rating scale produce the maximal reliability for trained raters. Contrary to the above suggestions, results of empirical investigations by Bendig (1954) and Komorita (1963) indicated that reliability is independent of the number of scale points employed. Komorita concluded that utilization of a dichotomous scale would not significantly decrease the reliability of the information obtained when compared to that obtained from a multi-step scale.

Whether an increase in the number of scale points is associated with an increase in reliability, and how many scale points should be employed beyond which there would be on further meaningful increase in reliability, are both empirical questions. Studies addressed to these reliability questions have typically employed a measure of internal consistency (either split-half stepped up by the Spearman-Brown Prophecy Formula or Kuder-Richardson Formula 20). Utilization of a stability (test-retest) measure appears to be nonexistent. It should be apparent that both reliability coefficients—internal consistency *and* stability—must be assessed if meaningful and complete answers to the questions posed are to be provided.

Moreover, studies dealing with the number of alternatives problem emphasize reliability as the major, and in some instances, only criterion in the choice of the number of scale points. However, according to both Cronbach (1950) and Komorita and Graham (1965), the ultimate criterion is the effect a change in the number of scale points has on the validity of the scale. An intensive literature search failed to reveal any empirical investigation addressed to this question.

Multi-step Likert-type rating scales provide two components of information—the direction and the intensity of an individual's attitudinal composition. Peabody (1962) concluded that the total scores obtained with any Likert-type scale represent primarily the directional component, and only to a minor degree, the intensity component. Both Peabody (1962) and Cronbach (1950) suggested that differences in the intensity component primarily represent differences in response set tendencies, i.e., tendencies for subjects to use a particular degree of agreement or disagreement toward any attitudinal object regardless of the direction. Cronbach concluded that any increase in test reliability due to response set, in the final analysis, dilutes the test results and lowers its validity.

This investigation was undertaken to answer a fundamental and deceptively simple question: is there an optimal number of alternatives to use in the construction of a Likert-type scale? Of specific concern was whether variations in the number of scale alternatives affected either reliability or validity.

Method

Subjects

Four-hundred and ten undergraduate psychology students enrolled in a large midwestern university participated in this experiment. The procedure first involved selecting adjective statements for each scale point ($n = 40$), then determining the inter-rater reliability on those statements selected ($n = 10$), and, lastly, conducting the experiment proper ($n = 360$) in which 20 subjects were assigned to each of the 18 different Likert scale formats. Different samples of students attending classes in general introductory psychology, introductory applied psychology, industrial psychology, and consumer psychology were used for each segment of the study.

Scale Construction and Instruments

Anchoring verbal statements to Likert scale points has usually been conducted on an intuitive basis. In the present investigation the statements were determined empirically. Forty subjects already familiar with the technique of paired comparisons were presented 17 different statements, ranging from "I am uncertain" to "I infinitely agree," in paired comparison format, and asked to select the statement from each pair "which indicated greater agreement." A total of 136 comparisons were made by each subject. (There is no reason to believe that the results would have been any different had the instructions specified disagreement rather than agreement.) The information derived from this procedure served as the basis for selecting those statements used to construct the 18 (i.e., 2- to 19-point) Likert-type rating formats. Criteria for the selection of a statement were: (a) that it have a minimal number of reversals (less than five out of a possible 40), and (b) that it be approximately equidistant (where possible) from the statement preceding and following it. Ten of the original 17 statements came closest to meeting those criteria. These 10 statements were then presented to a new group of 10 students who were instructed to rank them in the order of increasing disagreement. (The purpose of the disagreement instructions with these subjects, in contrast to the agreement instructions given the earlier 40 subjects, was simply to insure that the relative intensity of the descriptive adjectives remained invariant, i.e., was unaffected by the direction of the statement.) An average rank-order correlation coefficient was then computed to determine the inter-rater reliability.

The instrument used in the experiment proper was a modified Allport-Vernon-Lindzey Scale of Values (1960), containing 60 items. Eighteen different versions, in which the number of alternatives for each item ranged from a 2-point to a 19-point format, were constructed, using the ten adjective descriptors obtained in the first part of the study. The criterion for the construction of each format was that each scale point be approximately equidistant from the ones preceding and following it.

Procedure

The experimenter entered the testing room and proceeded to distribute the rating booklets. Arrangement of the booklets was in such an order that the first subject received a 2-point rating scale, the second received a 3-point rating scale, and so on, until the eighteenth subject received a 19-point rating scale booklet. This procedure was repeated until all subjects had obtained rating booklets. For test-retest purposes, subjects were asked to record their names, course name and number, time and place of meeting, and instructor's name on top of their rating booklets. The subjects were then instructed to open their booklets, read the instructions, record the time, rate the 60 statements, and then record the time at completion of the task. The rating instructions were the same for all the booklets, except that every block of 20 subjects used a different scale to rate the statements. Subjects did not know they were using different rating scales.

After completing the modified Study of Values, the subjects proceeded to fill out an attached criterion measure. Statements in the criterion measure explicitly spelled out what each subscale on the Study of Values was designed to measure, as defined by its test manual. Using a graphic rating scale, each subject was asked to rate the present importance of each of the six value areas in his life.

Three weeks after the first administration, and with the assistance of the identification data provided at the first session, each subject was contacted and received another rating booklet identical to the first. Upon completion, the purpose of the experiment was explained and questions were answered.

Data obtained from the premeasure were analyzed to determine the internal consistency reliability (Cronbach's alpha, 1951) and concurrent validity. Both measures, pre- and post, were used to assess the test-retest reliability, predictive validity, and the reliability of the criterion measure for attentuation-correction purposes.

A Fisher *Z* transformation (Fisher, 1921) was undertaken to convert all reliability and validity coefficients in order to insure normality. These transformations were then analyzed by a single classification analysis of variance procedure to determine whether there were significant differences in reliability and validity as a function of rating format. Each of these analyses was segmented by the six value areas in the modified Study of Values.

Following data collection, the responses to each item of the modified Study of Values were converted to dichotomized or trichotomous measures. All even-numbered formats were dichotomized at the center. Responses to the left of center were scored "agree," while those to the right were designated "disagree." The odd-numbered formats were trichotomized, yielding the categories of "agree," "uncertain," and "disagree." The resultant reliability and validity coefficients were then determined for each original and collapsed rating format and subsequently transformed into Fisher *Z*'s. The standard error of the difference between the original and collapsed set of *Z*'s was computed and then divided into the difference between the original and reduced *Z* coefficients. This procedure, a critical ratio, allowed us to determine whether the original correlations were significantly different from those obtained by collapsing these many-stepped formats to dichotomous or trichotomous measures.

Results

Table 1 summarizes the results of the adjective selection procedure and presents for each statement the proportion of "greater agreement" judgments made by the subjects. Employing the criteria of minimal reversals and approximate equidistance from preceding and succeeding statements, the 10 statements finally selected, together with their scale value, are graphically presented in Figure 1. To ascertain the consistency (inter-rater reliability) with which these statements were ranked, 10 additional subjects proceeded to rank them. An average rank-order correlation coefficient of .99 was obtained, indicating an extremely high degree of

TABLE 1

Scale Values of the Intensity Ratings for the Original Set of Statements

Statement	Proportion of "greater agreement"
I am uncertain	.00
I am uncertain, but probably agree	.08
I hardly agree	.17
I scarcely agree	.20
I minutely agree	.22
I vaguely agree	.29
I barely agree	.30
I slightly agree	.41
I moderately agree	.45
I pretty much agree	.53
I strongly agree	.63
I intensely agree	.74
I immensely agree	.76
I extremely agree	.76
I absolutely agree	.92
I infinitely agree	.94
I unlimitedly agree	.94

agreement among raters as to the rank associated with each statement.

Tables 2 through 5 present the internal-consistency reliability, test-retest reliability, concurrent validity, and predictive validity coefficients (the latter two corrected for criterion attenuation) for each of the 18 rating formats hexacotimized by each of the Allport-Vernon-Lindzey value areas. Table 6 presents the results of analyses

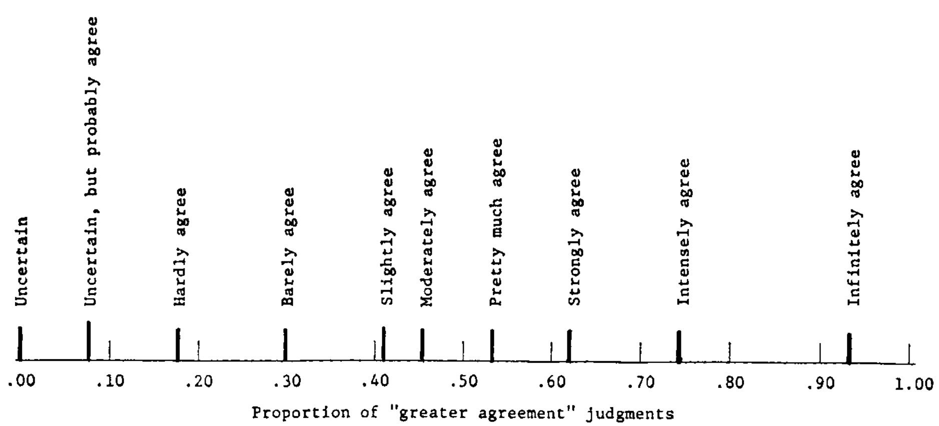

Figure 1. Graphic representation of the scale values of the selected statements.

of variance computed for each value area to assess the extent to which there was a relationship between the rating formats and the reliability and validity measures. This table displays the *F* ratio for each criterion and value area, indicating whether the relationship found was significant and, if so, to what extent. Examination of Tables 2 through 6, as well as visual inspection of graphs charted from the data contained in these tables, reveals that there is no systematic relationship between predictive validity, concurrent validity, internal-consistency reliability, and test-retest reliability and the number of steps in a Likert-type rating scale. This lack of a systematic relationship was replicated for each of the six value areas encompassed in the modified Allport-Vernon-Lindzey Study of Values.

Table 7 presents the reliability and validity vectors for the 18 original and collapsed rating formats. Figures 2, 3, and 4 graphically display the test-retest reliability, concurrent validity, and predictive validity coefficients for the original and reduced rating formats, respectively. It is apparent that a large degree of overlap exists among each of the three pairs of figures. There appears to be only minimal differences between the reliability and validity vectors based upon the original rating formats and those obtained by collapsing these

TABLE 2

Internal Consistency Reliability Coefficients for Each Rating Format Hexacotomized by Value Area

Format	Theoretical	Political	Economic	Aesthetic	Religious	Social
2	.43	.63	.69	.82	.50	.48
3	.57	.79	.74	.63	.73	.64
4	.62	.64	.63	.61	.85	.73
5	.49	.49	.66	.59	.70	.63
6	.63	.59	.50	.63	.79	.66
7	.63	.56	.26	.63	.88	.81
8	.82	.54	.77	.74	.79	.79
9	.69	.06	.71	.55	.72	.58
10	.66	.50	.46	.67	.83	.91
11	.43	.05	.83	.56	.76	.72
12	.57	.59	.58	.67	.79	.83
13	.50	.53	.70	.59	.61	.60
14	.50	.59	.34	.56	.74	.66
15	.52	.71	.53	.63	.67	.73
16	.64	.52	.66	.66	.70	.69
17	.81	.81	.60	.74	.77	.73
18	.30	.36	.36	.49	.65	.80
19	.62	.24	.69	.64	.79	.87

TABLE 3

Test-Retest Reliability Coefficients for Each Rating Format Hexacotomized by Value Area

Format	Theoretical	Political	Economic	Aesthetic	Religious	Social
2	.64	.99	.99	.99	.99	.98
3	.62	.90	.71	.71	.84	.70
4	.61	.81	.85	.91	.86	.86
5	.78	.81	.63	.87	.89	.83
6	.73	.62	.31	.78	.68	.87
7	.89	.89	.74	.93	.91	.80
8	.92	.81	.83	.94	.88	.88
9	.75	.79	.89	.75	.84	.82
10	.75	.67	.79	.73	.89	.71
11	.15	.76	.86	.89	.82	.84
12	.61	.73	.47	.85	.84	.91
13	.58	.81	.80	.88	.86	.76
14	.47	.65	.58	.71	.78	.79
15	.65	.77	.85	.75	.79	.69
16	.83	.82	.89	.80	.83	.82
17	.64	.75	.85	.61	.69	.82
18	.61	.50	.80	.45	.68	.75
19	.78	.49	.66	.85	.74	.65

TABLE 4

Concurrent Validity Coefficients for Each Rating Format Hexacotomized by Value Area

	Concurrent Validity Coefficients											
Format	Theoretical		Political		Economic		Aesthetic		Religious		Social	
2	.10	.16*	.01	.02*	.03	.04*	.08	.09*	.11	.14*	.43	.62*
3	.03	.05	.70	.89	.45	.51	.28	.35	.62	.67	.46	.58
4	.27	.40	.23	.29	.47	.53	.32	.51	.63	.66	.48	.58
5	.05	.13	.07	.08	.37	.39	.45	.54	.86	.87	.52	.60
6	.44	.50	.08	.09	.62	.66	.67	.82	.66	.73	.19	.26
7	.40	.59	.03	.03	.14	.18	.68	.76	.71	.87	.19	.24
8	.43	.52	.57	.60	.65	.75	.40	.43	.78	.81	.50	.58
9	.27	.46	.04	.05	.72	.76	.38	.44	.59	.67	.26	.28
10	.36	.46	.01	.02	.13	.15	.41	.48	.55	.60	.68	.90
11	.26	.36	.39	.43	.72	.86	.19	.35	.64	.76	.33	.41
12	.18	.23	.06	.06	.41	.48	.53	.67	.31	.36	.61	.79
13	.18	.22	.11	.15	.32	.42	.63	.72	.60	.65	.44	.53
14	.20	.24	.04	.06	.14	.16	.55	.75	.62	.66	.15	.18
15	.34	.45	.30	.35	.46	.56	.41	.51	.78	.88	.45	.49
16	.28	.40	.00	.01	.69	.91	.30	.41	.51	.55	.74	.83
17	.81	.93	.54	.72	.33	.51	.33	.40	.16	.17	.22	.27
18	.26	.38	.30	.49	.04	.04	.42	.63	.71	.89	.52	.67
19	.51	.60	.02	.04	.05	.07	.64	.71	.24	.30	.66	.86

* The asterisked columns have been corrected for criterion attenuation.

TABLE 5

Predictive Validity Coefficients for Each Rating Format Hexacotomized by Value Area

Format	Theoretical		Political		Economic		Aesthetic		Religious		Social	
2	.12	.20*	.10	.12*	.01	.01*	.11	.12*	.06	.06*	.50	.73*
3	.10	.13	.54	.68	.55	.62	.49	.62	.49	.54	.07	.08
4	.23	.35	.44	.55	.48	.54	.33	.51	.61	.64	.61	.74
5	.29	.72	.04	.05	.45	.48	.56	.67	.85	.86	.15	.17
6	.39	.44	.10	.11	.55	.59	.61	.74	.75	.83	.11	.15
7	.01	.02	.05	.06	.37	.49	.70	.77	.76	.94	.07	.09
8	.42	.51	.51	.54	.62	.71	.55	.59	.88	.90	.56	.64
9	.05	.09	.04	.06	.81	.84	.44	.51	.58	.66	.20	.22
10	.41	.53	.02	.02	.48	.56	.32	.37	.63	.70	.18	.24
11	.43	.59	.31	.34	.43	.41	.10	.19	.46	.55	.31	.38
12	.52	.66	.07	.08	.40	.46	.42	.54	.43	.50	.24	.31
13	.41	.50	.25	.35	.14	.30	.41	.46	.61	.66	.41	.50
14	.36	.43	.07	.10	.24	.27	.49	.67	.57	.61	.37	.45
15	.22	.29	.36	.41	.64	.77	.34	.43	.61	.69	.43	.47
16	.46	.66	.03	.06	.64	.85	.52	.70	.63	.68	.65	.74
17	.78	.90	.01	.01	.06	.09	.28	.33	.31	.33	.30	.37
18	.24	.34	.18	.29	.03	.04	.36	.54	.44	.55	.39	.50
19	.54	.63	.38	.60	.16	.22	.60	.67	.31	.38	.31	.40

* Corrected for attenuation.

formats to dichotomous and trichotomous measures. Three critical ratios, computed to determine whether these validity and reliability vectors differed, resulted in nonsignificance (Table 8), demonstrating that, regardless of the number of steps originally employed to collect the data, conversion to dichotomous or trichotomous measures does not result in any significant decrement in reliability or validity. Therefore, provided that an adequate number of items are contained on the inventory, increasing the precision of measurement does not eventuate in greater reliability or validity vectors.

Discussion and Conclusions

The evidence from the present study led us to conclude that both reliability and validity are independent of the number of scale points used for Likert-type items. Both internal consistency and stability measures were obtained. The average internal consistency reliability across all areas was .66, while the average test-retest reliability was .82. Both reliability measures, test-retest and internal consistency, were found to be independent of the number of scale points. This finding is consistent with those reported by Bendig (1954), Komorita (1963), Komorita and Graham (1965), and Peabody (1962), contrasts with findings by Symonds (1924) and

TABLE 6

Summary Table of Reliability and Validity Coefficients by Value Area

Criterion	Value Area											
	Theoretical		Political		Economic		Aesthetic		Religious		Social	
	F ratio	*P*	*F* ratio	*P*	F ratio	*P*	*F* ratio	*P*	*F* ratio	*P*	*F* ratio	*P*
Test-Retest Reliability	3.74	.005	23.61	.001	18.96	.001	4.82	.001	5.36	.001	7.39	.001
Internal Consistency Reliability	2.71	.010	7.66	.001	2.62	.025	0.80	NS	4.32	.001	3.69	.005
Concurrent Validity*	2.71	.010	4.65	.001	4.89	.001	2.87	.01	15.17	.001	12.76	.001
Predictive Validity*	6.11	.001	2.40	.025	12.50	.001	2.58	.025	5.39	.001	1.72	.100

* Corrected for criterion attenuation.

TABLE 7

Reliability and Validity Coefficients for the Original and Reduced Rating Formats

	Test-Retest Reliability		Concurrent Validity		Predictive Validity	
Rating Format	Original Format	Collapsed Format	Original Format	Collapsed Format	Original Format	Collapsed Format
2	.99	.99	.43	.43	.51	.51
3	.70	.70	.47	.47	.07	.07
4	.86	.83	.49	.55	.62	.73
5	.83	.82	.52	.41	.15	.04
6	.88	.80	.19	.28	.12	.19
7	.80	.84	.20	.20	.08	.19
8	.88	.84	.51	.03	.56	.07
9	.82	.78	.26	.42	.21	.22
10	.72	.82	.68	.47	.19	.05
11	.85	.82	.34	.47	.32	.51
12	.92	.88	.62	.64	.24	.27
13	.77	.66	.44	.16	.42	.11
14	.68	.67	.15	.20	.38	.44
15	.70	.65	.45	.40	.44	.37
16	.82	.71	.74	.67	.66	.71
17	.82	.80	.22	.04	.30	.33
18	.75	.62	.52	.36	.39	.40
19	.65	.70	.66	.75	.31	.43

Note.—All values are based upon the social scale.

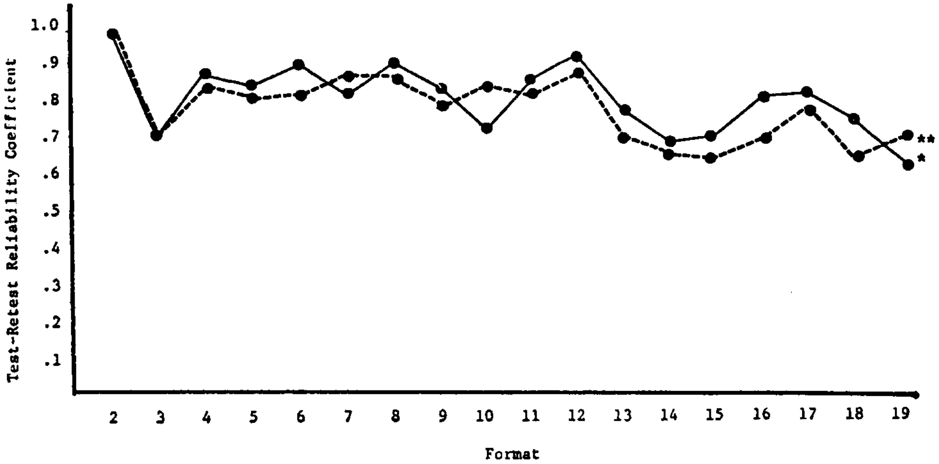

Figure 2. The test-retest reliability coefficients for the original and collapsed rating formats.

* Original Rating Format.
** Collapsed Rating Format.

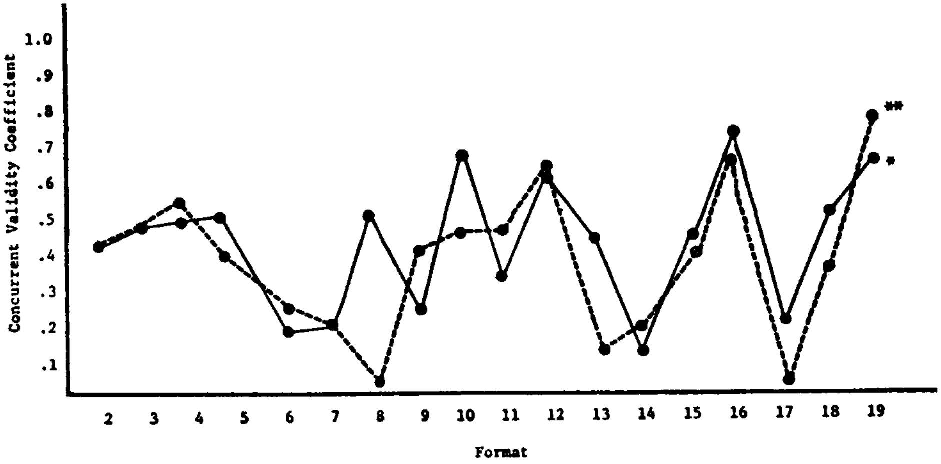

Figure 3. The concurrent validity coefficients for the original and collapsed rating formats.

* Original Rating Format.
** Collapsed Rating Format.

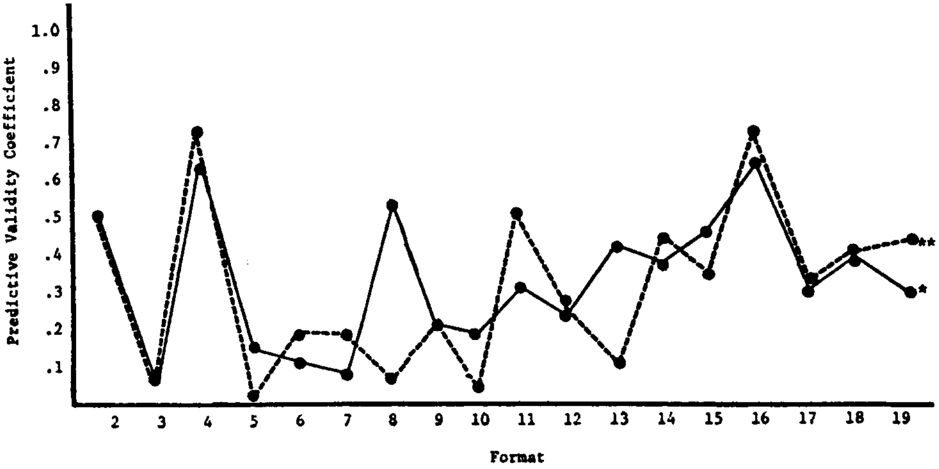

Figure 4. The predictive validity coefficients for the original and collapsed rating formats.

* Original Rating Format.
** Collapsed Rating Format.

TABLE 8

Summary Table of the Tests of Significance on the Reliability and Validity Coefficients for the Original and Collapsed Rating Formats

Criterion	Original Format	Collapsed Format	Critical Ratio	P
Test Retest Reliability	.82	.78	1.47	NS
Concurrent Validity	.45	.40	.80	NS
Predictive Validity	.34	.33	−.13	NS

Champney and Marshall (1939), and contests opinions proffered by Jahoda, Deutsch and Cook (1951) and Ferguson (1941) to the effect that the reliability of a scale will increase as the number of scale points increase. Based upon the evidence adduced thus far, it would seem that reliability should not be a factor considered in determining Likert-scale rating format, as it is independent of the number of scale steps employed.

Cronbach (1950) claimed that if utilization of finer rating scales, as opposed to coarser ones, increases the reliability of measurement, then this increase should be attributed to the addition of response set variance and not to any increase in the refinement of measurement. Response set was not found to be a factor affecting the intensity component; finer rating scales did not yield an increase in the reliability of measurement over coarser ones. The extent to which response set bias influenced the directional component of the responses is unknown.

With respect to Cronbach's (1959) and Komorita and Graham's (1965) contention that validity should be the ultimate criterion, as far as the authors can determine, this study is the first to attempt to assess the relationship between validity and number of scale points. As with reliability, validity was found to be independent of the number of scale points contained in the rating scale. This finding remains even after correcting the predictive and concurrent validity coefficients for criterion attenuation. Moreover, the same results were obtained for each of the areas on the modified Study of Values. We can conclude, therefore, that when considering the number of steps to employ in a Likert scale rating format, validity need not be considered because there is no consistent relationship between it and the number of scale steps utilized.

In an attitude survey there usually is no manifest criterion present since behavior is not necessarily a function of attitudes. The choice, in this study, of whether to use either an internal measure (i.e., correlation of each item with total score, less that item) or an external measure was made in favor of the latter. The internal measure has no intrinsic relationship to external reality, while the external criterion does. Directing attention to the obtained validity vectors, we note that, while not consistently high or low, in most cases they compare quite favorably with the bulk of those reported in the literature. Ghiselli (1955), in a comprehensive review of both published and unpublished studies, found that the range of average validities for psychological predictors was in the .30's and low .40's. An average of .50 was a distinct rarity. The average concurrent validity coefficient (corrected for attenuation) in the current study, across all formats and value areas, was .53. The average predictive validity (again corrected for attenuation) was .51.[2]

Komorita and Graham (1965), in discussing studies by Komorita (1963) and by Bendig (1954), stated that "if it is a valid generalization (i.e., independence of reliability and number of scale steps), the major implication is that, because of simplicity and convenience in administration and scoring, all inventories and scales ought to use a dichotomous, 2-point scoring scheme (p. 989)." Peabody's (1962) results indicated that composite scores, consisting of the sum of scores on bipolar, 6-point scales, mainly reflect direction of response and are only minimally influenced by intensity of response. He concluded from this that there is justification for scoring bipolar items dichotomously according to direction of response. This investigation provided empirical evidence in support of these assumptions.

The lack of any significant differences in reliability and validity stemming from the utilization of a particular format, or from collapsing a many-stepped format into a dichotomous or trichotomous measure, led to the conclusion that total scores obtained with Likert-type scales, as both Peabody and Cronbach have suggested, represent primarily the directional component and only

[2] Concurrent and predictive validity vectors, uncorrected for criterion attenuation, were .42 and .40, respectively.

to a minor degree the intensity component. Therefore, of the three components contained in a Likert-type composite scale score—direction, intensity, and error—the directional component accounts for the overwhelming majority of the variance.

It has been demonstrated that regardless of the number of steps originally employed to collect the data, conversion of these many-stepped response scales to dichotomous or trichotomous measures does not result in any significant decrement in reliability or validity. Therefore, increasing the precision of measurement does not eventuate in greater reliability or validity vectors, provided that an adequate number of items are contained in the inventory.

One ramification of this finding, if substantiated, would be greater flexibility in the adoption of a given format for a given predictor, criterion, and subject. Since there appears to be independence between reliability and validity vectors and rating format, desirable practical consequences might be obtained from allowing the subject to select the rating format which best suits his needs. This might result in the highly favorable consequence of increasing the subject's motivation to complete the scale. Conversely, if the respondent is not satisfied with a particular rating format, regardless of the reason, the possibility exists that deleterious effects might result from the unsatisfactory rating format-respondent interaction. This interaction could eventuate in a decrement in interest and/or reduced motivation to continue the rating procedure or to complete any remaining parts of the measurement process.

Indeed, it is even conceivable that the subject could record his own responses (open-ended) to each item, without a previously prepared rating format being provided. Subsequently, these subject-produced responses could be transformed to dichotomous or trichotomous measures. Such a strategy might be used to secure the cooperation of individuals who typically are difficult to obtain. By catering to the idiosyncracies of these individuals or, for that matter, any group of respondents, and allowing them to respond in any manner they desire, besides obtaining greater cooperation from these individuals, we might also be able to increase the return rate of our instruments.

A final consideration is the comparison of such data with data which were previously collected with different rating formats. To overcome this problem, previously collected data could be collapsed into dichotomous or trichotomous measures. This reduction in the precision of measurement, as demonstrated in this research, would not lead to any deleterious effects vis-a-vis reliability or validity. The resultant response distributions, originally based upon different rating formats, could then be directly compared since they would now all be projected from the same base measure.[3]

A basic question appears to be whether the utilization of fine rating scales increases the refinement of measurement over that which is obtained with coarse dichotomous or trichotomous scales. The overwhelming consistency of results of this study, in addition to those obtained by Peabody (1962), Komorita and Graham (1965), and Bendig (1954), strongly suggests a negative answer to this question.

The primary practical implication of this study is that investigators would be justified in scoring attitude items dichotomously (or trichotomously), according to direction of response, after they have been collected with an instrument that provides for the measurement of the intensity component along with the directional component.

Further research should now be conducted to determine whether the present findings can be generalized beyond the Likert-type scale to different types of scales (e.g., Osgood's Semantic Differential, Thurstone-type scales, graphic rating scales, etc.) and for other purposes (e.g., the rating of behavior, personality, industrial work performance, etc.). It should also be determined whether the conclusions are generalizable to different subject populations defined by such parameters as level of education or ability, and by psychological, experiential, demographic, and ecological characteristics.

REFERENCES

Allport, G. W., Vernon, P. E., and Lindzey, G. *Study of values.* Boston: Houghton Mifflin Company, 1960.

[3] To compare dichotomous and trichotomous measures with each other, the "agree" and "disagree" response categories could be given the weights of one and three, respectively. The remaining "uncertain" response category on the trichotomous format would then be weighted two.

Bendig, A. W. Reliability and the number of rating scale categories. *Journal of Applied Psychology,* 1954, 38, 38–40.

Champney, H. and Marshall, H. Optimal refinement of the rating scale. *Journal of Applied Psychology,* 1939, 23, 323–331.

Cronbach, L. J. Further evidence on response sets and test design. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1950, 10, 3–31.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika,* 1951, 16, 297–334.

Ferguson, L. W. A study of the Likert technique of attitude scale construction. *Journal of Social Psychology,* 1941, 13, 51–57.

Fisher, R. A. On the "probable error" of a coefficient of correlation. *Metron,* 1921, 1, Part 4, 1–32.

Garner, W. R. Rating scales, discriminability, and information transmission. *Psychological Review,* 1960, 67, 343–352.

Garner, W. R. and Hake, H. W. The amount of information in absolute judgments. *Psychological Review,* 1951, 58, 446–459.

Ghiselli, E. E. and Brown, C. W. *Personnel and industrial psychology.* New York: McGraw-Hill, 1948.

Ghiselli, E. E. *The measurement of occupational aptitude.* Berkeley: University of California Press, 1955.

Guilford, J. P. *Psychometric methods.* New York: McGraw-Hill, 1954.

Jahoda, M., Deutsch, M., and Cook, S. W. (Eds.) *Research methods in social relations.* New York: Dryden Press, Inc., 1951.

Komorita, S. S. Attitude content, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology,* 1963, 61, 327–334.

Komorita, S. S. and Graham, W. K. Number of scale points and the reliability of scales. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1965, 4, 987–995.

Likert, R. A technique for the measurement of attitudes. *Archives of Psychology,* 1932, No. 140.

Peabody, D. Two components in bipolar scales: Direction and extremeness. *Psychological Review,* 1962, 69, 65–73.

Symonds, P. M. On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology,* 1924, 7, 456–461.