.... COMMUNICATIONS AND RESEARCH NOTES

CONTENTS

Three-Point Likert Scales Are Good Enough

JACOB JACOBY and MICHAEL S. MATELL*

The basic question about any given rating instrument is whether or not it has an optimum number of response categories or at least a number beyond which there is no further improvement in discrimination between the rated items. Determining the optimum response categories is especially important in constructing the ubiquitous Likert-type scale [13], which is often used in collecting attitudinal and image data in marketing and public opinion research. Too few response categories result in too

* Jacob Jacoby is Associate Professor of Psychology, Purdue University, and Michael S. Matell is Staff Psychometrician, Ivorydale Technical Center, Procter and Gamble Co., Cincinnati.

coarse a scale and loss of much of the raters' discriminative powers. Conversely, too fine a scale may go beyond the raters' limited powers of discrimination [6, 10].

Ghiselli and Brown [8] and Guilford [10] 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.

Green and Rao, working with simulated data and using reproducibility of the original data configuration as their criterion, presented evidence to indicate that 6- to 7-point scales are optimal, especially if several different instruments are employed concurrently as in a test battery [9]. However, while data recovery may be a significant consideration in some types of measurement problems (e.g., multidimensional scaling), it is neither the only criterion nor probably the most important one for most marketing research problems which employ Likert-type scales. Many would argue that reliability and validity are more basic considerations.

Table 1

INTERNAL CONSISTENCY AND TEST-RETEST RELIABILITY COEFFICIENTS FOR EACH RATING FORMAT HEXACOTOMIZED BY VALUE AREA

Format	*Theoretical*	*Political*	*Economic*	*Aesthetic*	*Religious*	*Social*
Internal consistency						
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
Test-retest						
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

Empirical investigations by Bendig [2] and Komorita [11] indicated that reliability is independent of the number of response categories 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 multistep scale. However, these studies were based only on internal consistency measures, while both types of reliability coefficients—internal consistency and stability (test-retest)—must be assessed if meaningful and complete answers to the questions posed are to be provided.

Moreover, most of the psychometric literature dealing with the number-of-alternatives problem emphasizes reliability as the major (and in some instances, only) criterion in the choice of the number of scale points. However, the ultimate criterion is the effect a change in the number of scale points has on the validity of the scale [3, 12]. An intensive literature search failed to reveal any empirical investigation addressed to this question.

Accordingly, this investigation was undertaken to answer a fundamental and deceptively simple question: considering reliability and validity, is there an optimal number of alternatives to use in the construction of a Likert-type scale?

METHOD AND PROCEDURE

The subjects were 360 undergraduates at Purdue University enrolled in general introductory psychology, applied psychology, industrial psychology, and consumer psychology during the fall 1968 semester.

The instrument used was a modified Allport-Vernon-Lindzey Scale of Values [1] containing 60 statements. Eighteen different versions were constructed in which the number of alternatives for each item ranged from a 2-point to a 19-point format. The criterion for each was that each scale point be approximately equidistant from the ones preceding and following it [14].

The first subject received a 2-point rating scale, the second a 3-point scale, and so on, until the eighteenth received a 19-point scale and all subjects had scales. For test-retest purposes, each subject was asked to record his name, course name and number, time and place of meeting, and instructor's name on top of his rating booklet. Rating instructions were the same for all booklets, except that every block of 20 subjects used a different scale. Subjects did not know they were using different rating scales.

After completing the modified Study of Values, subjects completed the attached criterion measure whose statements explicitly spelled out what each subscale was designed to measure, as defined by its test manual. Using a graphic 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 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 internal consistency reliability (Cronbach's alpha [4]) 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 attenuation and correction.

A Fisher *Z*-transformation [5] was used to convert all reliability and validity coefficients for normality. These transformations were then analyzed by a single classification analysis of variance to determine if there were significant differences in reliability and validity as a function of rating format. Each analysis was segmented by the six value areas in the modified Study of Values.

Responses to each item were converted to dichotomous or trichotomous measures. All even-numbered formats were dichotomized at the center; responses to the left of center were scored "agree," those to the right, "disagree." The odd-numbered formats were trichotomized, yielding the categories of "agree," "uncertain," and "disagree." Then the resultant reliability and validity coefficients were determined for each original and collapsed rating format, and subsequently transformed into Fisher *Z*'s. The standard error of the difference

Table 2

CONCURRENT AND PREDICTIVE VALIDITY COEFFICIENTS FOR EACH RATING FORMAT HEXACOTOMIZED BY VALUE AREA

Format	*Theoretical*		*Political*		*Economic*		*Aesthetic*		*Religious*		*Social*	
Concurrent validity												
2	.10	.16[a]	.01	.02[a]	.03	.04[a]	.08	.09[a]	.11	.14[a]	.43	.62[a]
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	.54	.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
Predictive validity												
2	.12	.20[a]	.10	.12[a]	.01	.01[a]	.11	.12[a]	.06	.06[a]	.50	.73[a]
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	.94	.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	.20	.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

[a] Corrected for criterion attenuation

between the original and collapsed sets 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 if original correlations were significantly different from those obtained by collapsing many-stepped formats to dichotomous or trichotomous measures.

RESULTS

Tables 1 and 2 present the respective 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 hexacotomized by each of the Allport-Vernon-Lindzey value areas. Table 3 presents the results of analyses of variance computed for each value area to assess the extent of 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 or not the relationship found was significant and, if so, to what extent. Tables 1 through 3, as well as graphs charted from the data contained in these tables, reveal 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 Study of Values.

Table 4 presents the test-retest reliability and concurrent and predictive validity coefficients for the 18 original and collapsed rating formats. A large amount of overlap is apparent among each of the three pairs of figures. There appear to be only minimal differences between the reliability and validity vectors based upon the original rating formats and those obtained by collapsing these formats to dichotomous and trichotomous measures. Three critical ratios, computed to determine whether these validity and reliability vectors differed, were nonsignificant, 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:

Criterion	*Original format*	*Collapsed format*	*Critical ratio*	*p*
Test-retest reliability	.82	.78	1.47	n.s.
Concurrent validity	.45	.40	.80	n.s.
Predictive validity	.34	.33	−.13	n.s.

Therefore, provided an adequate number of items are contained on the inventory, increasing the precision of measurement does not lead to greater reliability or validity.

DISCUSSION

The evidence indicates that both reliability and validity are independent of the number of scale points used for Likert-type items. The average internal consistency reliability across all areas was .66, while the average test-retest reliability was .82. Both test-retest and internal consistency were independent of the number of scale points, consistent with previous findings [2, 11, 12, 14]. Based upon the evidence adduced thus far, reliability should not be a factor in determining a Likert-type scale rating format, because it is independent of the number of scale steps employed.

As far as we can determine, this study is the first to attempt to assess the relationship between validity and number of alternatives. As with reliability, validity was found to be independent of the number of scale points 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 determining the number of steps 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.

The obtained validity vectors, in which scores on each scale were correlated with the appropriate criterion measures, are not consistently high or low, but in most cases compare quite favorably with those reported in the literature. Ghiselli [7], in a comprehensive review of both published and unpublished predictors, found that the average value was in the .30's and low .40's. An

Table 3

SUMMARY TABLE OF RELIABILITY AND VALIDITY COEFFICIENTS BY VALUE AREA

Criterion	*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[a]	2.71	.010	4.65	.025	4.89	.001	2.87	.001	15.17	.001	12.76	.001
Predictive validity[a]	6.11	.001	2.40	.025	12.50	.001	2.58	.025	5.39	.001	1.72	.100

[a] Corrected for attenuation.

average of .50 was a distinct rarity. The average concurrent validity coefficient (corrected for attenuation) in this study, across all formats and value areas, was .53. The average predictive validity (again corrected for attenuation) was .51.[1]

Komorita and Graham, in discussing studies by Komorita [11] and Bendig [2], stated that [12, p. 989]:

> If this 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, two-point scoring scheme.

Peabody's results indicated that composite scores, consisting of the sum of scores on bipolar, six-point scales, mainly reflect direction of response and are only minimally influenced by intensity of response. He concluded that there is justification for scoring bipolar items dichotomously according to direction of response [14]. This investigation has 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 shows that total scores obtained with Likert-type scales, as both Peabody and Cronbach have suggested, represent primarily the directional component and secondarily the intensity component. 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.

Table 4

RELIABILITY AND VALIDITY COEFFICIENTS FOR THE ORIGINAL AND REDUCED RATING FORMATS[a]

Rating format	*Test-retest reliability*		*Concurrent validity*		*Predictive validity*	
	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	.23	.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

[a] All values are based upon the social scale.

IMPLICATIONS

It has been demonstrated 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. Given that it is not essential to be able to reproduce the original data array [9], greater flexibility can be gained 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 the rating format, it may be desirable (e.g., increased motivation to complete the scale) to allow a subject to select the rating format which best suits his needs. Conversely, if he is not satisfied with a particular rating format, regardless of the reason, deleterious effects may result from using an unsatisfactory rating format. Respondent interaction could reduce interest or motivation to continue rating.

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. Such responses could be transformed to dichotomous or trichotomous measures. This strategy could be used with individuals who might otherwise not respond. By catering to their idiosyncracies and allowing them to respond as they desire, a researcher could obtain greater cooperation and return rates.

A final consideration is the comparison of such data with data collected with different rating formats. Previously collected data could be collapsed into dichotomous or trichotomous measures, which would not lead to any deleterious effects vis à vis reliability or validity. The resultant response distributions, originally based upon different rating formats, could then be directly compared since they would all be projected from the same base measure.[2]

The primary practical implication of this study is that investigators would be justified in scoring Likert-type scale items dichotomously (or trichotomously), according to direction of response, after they have been collected with an instrument that provides for the measurement of direction and several degrees of intensity.

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

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

[2] 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.

scales, graphic rating scales). It should also be determined whether these conclusions are generalizable to different populations defined by such parameters as level of education or ability and by psychological, experiential, and sociodemographic characteristics.

REFERENCES

1. Allport, Gordon W., Philip E. Vernon, and Gardner Lindzey. *Study of Values*. Boston: Houghton Mifflin, 1960.
2. Bendig, A. W. "Reliability and the Number of Rating Scale Categories," *Journal of Applied Psychology*, 38 (February 1954), 38–40.
3. Cronbach, Lee J. "Further Evidence on Response Sets and Test Design," *Educational and Psychological Measurement*, 10 (Spring 1950), 3–31.
4. ———. "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika*, 16 (September 1951), 297–334.
5. Fisher, Ronald A. "On the 'Probable Error' of a Coefficient of Correlation," *Metron*, 1 (Part 4, 1921), 1–32.
6. Garner, Wendell R. and Harold W. Hake. "The Amount of Information in Absolute Judgments," *Psychological Review*, 58 (November 1951), 446–59.
7. Ghiselli, Edwin E. *The Measurement of Occupational Aptitude*. Berkeley: University of California Press, 1955.
8. ——— and Clarence W. Brown. *Personnel and Industrial Psychology*. New York: McGraw-Hill, 1948.
9. Green, Paul E. and Vithala R. Rao. "Rating Scales and Information Recovery—How Many Scales and Response Categories to Use?" *Journal of Marketing*, 34 (July 1970), 33–9.
10. Guilford, J. P. *Psychometric Methods*. New York: McGraw-Hill, 1954.
11. Komorita, Samuel S. "Attitude Content, Intensity, and the Neutral Point on a Likert Scale," *Journal of Social Psychology*, 61 (December 1963), 327–34.
12. ——— and W. K. Graham. "Number of Scale Points and the Reliability of Scales," *Educational and Psychological Measurement*, 4 (November 1965), 987–95.
13. Likert, Rensis. "A Technique for the Measurement of Attitudes," *Archives of Psychology*, 140 (June 1932).
14. Matell, Michael S. and Jacob Jacoby. "Is There an Optimal Number of Alternatives for Likert Scale Items? Study I: Reliability and Validity," *Educational and Psychological Measurement* (in press).
15. Peabody, Dean. "Two Components in Bipolar Scales: Direction and Extremeness," *Psychological Review*, 69 (March 1962), 65–73.