Journal of Applied Psychology
1972, Vol. 56, No. 6, 506–509

IS THERE AN OPTIMAL NUMBER OF ALTERNATIVES FOR LIKERT-SCALE ITEMS?

EFFECTS OF TESTING TIME AND SCALE PROPERTIES

MICHAEL S. MATELL[1] AND JACOB JACOBY

Purdue University

Three criteria one might use to determine the number of Likert scale rating steps to employ are proportion of scale used, testing time, and usage of an "uncertain" category. A modified version of the Florida Scale of Civic Beliefs (Shaw & Wright, 1967, pp. 307–311), in which the number of alternatives for each item ranged from 2 to 19, was administered to 360 *Ss*. Results of these manipulations indicated that for cumulative scores from Likert-type items, proportion of scale used was independent of the number of scale points, while mean testing time increased and usage of the "uncertain" category decreased as the number of rating steps increased. These results are integrated with earlier findings, and implications for Likert-scale construction are described.

Given the ubiquitous presence of Likert-type scales in industrial and consumer research, determination of the optimal number of response alternatives becomes an important consideration in the construction of such scales. Thus far, research has focused primarily on the question of reliability—in particular, internal consistency. Investigations by Bendig (1954) and Komorita (1963) revealed internal consistency to be independent of the number of alternatives employed. More recently, Matell and Jacoby (1971) replicated these findings and also revealed that stability, predictive validity, and concurrent validity of cumulative scores from Likert-type items were also independent of the number of scale points utilized. The question of these properties for individual items is presently being investigated.

Other factors that might be important criteria in determining the number of alternatives to employ are (*a*) proportion of the scale used, (*b*) testing time, and (*c*) whether or not an "uncertain" category is provided.

On an intuitive basis, it would seem less than optimal to provide raters with many response categories if they consistently utilized only a small proportion of these. Inasmuch as examination of the literature reveals no evidence pertaining to this issue, it is the first factor considered here.

A related issue concerns the desirability of including an intermediate response of "doubtful," "undecided," or "no difference" in the measuring instrument. On the one hand, this is sometimes considered inadvisable because it provides too easy and attractive an escape for respondents who are disinclined to express a definite view. On the other, forcing responses into an agree or disagree format is likely to cause difficulty for many respondents. It is also likely to yield results that are less realistic and more misleading than is true when an intermediate reply is provided for.

An early study by Riker (1944) presents evidence to suggest that some individuals who obtain neutral scores on rating scales do not necessarily consider themselves neutral toward the attitude object. A possible explanation might be the coarseness of the instrument employed. For example, the respondent may find that he does not hold a sufficiently strong positive or negative attitude toward an object to endorse the lowest positive or negative option on the continuum of a coarse scale. However, if a finer scale were provided, thereby allowing the *S* to express his attitudes more precisely, utilization of the zero point may markedly decrease. Consequently, the second factor considered was inclusion of an "uncertain" response category. More specifically, it was hypothesized that as the number

[1] Requests for reprints should be sent to Michael S. Matell, who is now at the Procter & Gamble Company, Ivorydale Technical Center, Cincinnati, Ohio 45217.

of steps in a rating scale increases, utilization of the uncertain alternative will decrease.

Testing time was the last factor considered. To the extent that warm-up, fatigue, and boredom are affected by testing time, Kendall (1964) suggested that for multiple choice aptitude test questions ordered according to difficulty, there will be a unique time limit that maximizes the contributions of these various parameters to validity. Kendall (1962) also advised using a longer time limit for *Ss* of higher ability, suggesting that the result would be an increased flexibility in the adoption of a given measure for different conditions.

Another factor that may affect testing time is the number of alternatives per item. Bricker (1955), Behar (1963), and Bevan and Avant (1968) found that the average reaction time for any given task increased reliably with each increase in the number of available alternatives. In general, response time seems to increase in a negatively accelerating monotonic fashion. To determine whether this also applies to Likert scales, the third and last factor examined was testing time.

METHOD

Fifty undergraduate psychology students participated in an experiment to construct 18 Likert-type rating scales varying in the number of steps they contained (from 2 to 19 steps). Verbally anchored adjective statements for each scale point were empirically selected and evaluated. Three-hundred and sixty students (20 per scale) then proceeded to utilize these 18 different scales for the purpose of responding to 40 items in a modified version of the Florida Scale of Civic Beliefs (cf. Shaw & Wright, 1967, 307–311). A more complete description of the method can be found in Matell and Jacoby (1971).

RESULTS

Summary data for the three variables of interest appear in Table 1. With the exception of the two- and three-point formats, it appears that there were no differences in the proportion of the remaining scales utilized. Although the analysis of variance resulted in a significant F ratio ($F = 7.61$, $p < .001$), the Neuman-Keuls procedure [2] revealed that the main contribution to the significant F ratio was produced by one or two formats. Sixteen out of the 18 scales examined did not differ signfiicantly.

TABLE 1

PROPORTION OF SCALE USED, PROPORTION OF "UNCERTAIN" RESPONSES, AND MEAN TESTING TIME FOR EACH RATING FORMAT

Format in pt.	Proportion of scale used	Proportion of "Uncertain" responses	Mean testing time in min.
2	.94	—	7.0
3	.75	.19	7.2
4	.55	—	7.7
5	.64	.21	7.1
6	.63	—	7.8
7	.62	.12	7.8
8	.53	—	9.3
9	.57	.11	7.6
10	.56	—	8.3
11	.56	.05	9.2
12	.57	—	8.9
13	.65	.03	11.0
14	.58	—	9.3
15	.62	.11	9.1
16	.59	—	10.8
17	.62	.04	8.8
18	.58	—	13.4
19	.60	.03	11.9

After the two- and three-point scales were excluded because they provided the rater with only 1 df and 2 df, respectively, and therefore forced him to use almost every point on the scale, a second analysis of variance was performed on the data. The results indicated that there was a nonsignificant relationship between rating formats of 4 to 19 steps and proportion of format utilized ($F = .77$, *ns*).

Examination of the data in Table 1 indicates that as the number of scale steps increases, usage of the "uncertain" category decreases. The 3- and 5-point scales are associated with a greater usage of the "uncertain" category and, when compared to the remaining seven rating formats, account for the significant F ratio ($F = 4.03$, $p < .01$).

[2] The Neuman-Keuls procedure was employed in this instance and in all other analyses because of its property of holding constant the level of significance for each comparison made. The schematic accompanying the Neuman-Keuls procedure is to be interpreted as follows: all formats underlined by a common line do not differ from each other, while those not underlined by a common line differ.

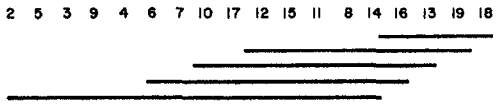

FIG. 1. Summary of the Neuman-Keuls Procedure on testing time.

The mean testing time associated with each format was assessed by analysis of variance ($F = 4.93$, $p < .001$) coupled with the Neuman-Keuls procedure. Significant differences between testing time and rating format were attributable to the greater time taken to rate four formats (13, 16, 18, and 19 steps). There were no differences found between testing times for the remaining 14 rating formats (see Figure 1).

DISCUSSION

Excluding the 2- and 3-point rating formats, it was found that the proportion of the scale utilized among 4 to 19 step formats was invariant (approximately 60%). Since the 2- and 3-point formats provided only 1 df or 2 df, respectively, the rater was forced to utilize almost every point on the scale. This would explain the greater usage of all the points contained in these formats. There were no differences found in the utilization of scale points beyond the 3-point scale. Thus, it appears that there is no a priori reason for favoring one scale over another if the criterion is proportion of scale used.

However, somewhat inconsistent with earlier findings, there is not a considerable positive relationship between number of scale points and testing time. This suggests that when time or time-related factors (e.g., fatigue, warm-up, boredom) are of concern, especially given lengthy tests and test batteries, fewer alternatives should not necessarily be employed. As Bricker (1965) suggested, it is quite possible that practices on multi-step scales might reduce testing time significantly.

Lastly, it was found that as the number of steps increased, the usage of the "uncertain" response category decreases. The 3- and 5-point rating scales had an average of 20% "uncertain" responses, while this category was utilized only 7% of the time on the remaining seven formats. This result is an important factor to be considered in the construction of a rating scale. Particularly if the scale builder desires to minimize usage of the "uncertain" category, he would be advised to use balanced even-numbered scales or scales with many points. The decision would seem to depend on the level of "uncertain" responses one is willing to tolerate.

In conclusion, a summary statement regarding the criteria examined and conclusions reached in this and related cumulative score investigations (as opposed to individual item scores) concerning the number of points to be used on Likert scales is as follows: internal consistency (Bendig, 1954; Komorita, 1963; Matell & Jacoby, 1971), test-retest stability (Goldsamt, 1971; Jones, 1968; Matell & Jacoby, 1971; van der Veen, Howard, & Austria, 1970), concurrent validity (Matell & Jacoby, 1971), predictive validity (Matell & Jacoby, 1971), and proportion of the scale used (this investigation) are independent of the number of response categories provided. Testing time increases minimally with increases in response categories and should only marginally be considered when warm-up, fatigue, and boredom are believed to be relevant factors. Uncertain and neutral response categories tend to be used more often on 3- and 5-point scales, less often on 7- to 19-point scales where such scales have an equal number of positive and negative points. The decision as to their inclusion would seem to depend primarily on the purposes of the research and proclivities of the researcher. Lastly, if a primary consideration is information recovery and reproduction of the original data matrix (an issue of particular concern in multidimensional scaling) especially in situations where several instruments are used to exhaustively sweep the individual's "configuration space," then 6- or 7-point scales would appear to be optimal (Green & Rao, 1970).

REFERENCES

BEHAR, I. On the relation between response uncertainty and reaction time in category judgment. *Perceptual and Motor Skills*, 1963, **16**, 595–596.

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

BEVAN, W., & AVANT, L. L. Response latency, re-

sponse uncertainty, information transmitted and the number of available judgmental categories. *Journal of Experimental Psychology,* 1968, **76**, 394–397.

BRICKER, P. D. The identification of redundant stimulus patterns. *Journal of Experimental Psychology,* 1955, **49**, 73–81.

GOLDSAMT, M. R. Effects of scoring method and rating scale length in extreme response style measurement. Unpublished doctoral dissertation, University of Maryland, 1971.

GREEN, P. E., & RAO, V. R. Rating scales and information recovery—How many scales and response categories to use? *Journal of Marketing,* 1970, **34**, 33–39.

JONES, R. R. Differences in response consistency and subject's preferences for three personality inventory response formats. *Proceedings of the 76th Annual Convention of the American Psychological Association,* 1968, **3**, 247–248. (Summary)

KENDALL, L. M. An investigation of hypotheses regarding the way adjustment of time limits affects validity. Paper read at the 33rd meeting of the Eastern Psychological Association, Atlantic City, 1962.

KENDALL, L. M. The effects of varying time limits on test validity. *Educational and Psychological Measurement,* 1964, **24**, 789–800.

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

MATELL, M. S., & JACOBY, J. Is there an optimal number of Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement,* 1971, **31**, 657–674.

RIKER, B. L. A comparison of methods in attitude research. *Journal of Abnormal and Social Psychology,* 1944, **39**, 24–42.

SHAW, M. E., & WRIGHT, J. M. *Scales for the measurement of attitudes.* New York: McGraw-Hill, 1967.

VAN DER VEEN, F., HOWARD, K. I., & AUSTRIA, A. M. Stability and equivalence scores based on three different response formats. *Proceedings of the 78th Annual Convention of the American Psychological Association,* 1970, **5**, 99–100. (Summary)

WALKER, H. M., & LEV, J. *Statistical inference.* New York: Henry Holt, 1953.

(Received July 16, 1971)