Psychological Reports, 1999, 84, 178-180. © Psychological Reports 1999

SCALING ISSUES IN FACULTY EVALUATIONS[1]

R. ERIC LANDRUM

Boise State University

Summary.—148 students were asked to evaluate particular survey questions with scales that varied in number of scale points, type of labeling, and labeling of anchors only vs all scale points. No significant difference was found between variations on 4- and 5-point scales, so these data were collapsed. Analysis indicated that students generated scale-point exemplars significantly better for 4-point scales (96.7%) than for 5-point (87.8%) and 10-point (79.3%) scales. These results are discussed with respect to the administration of student evaluations and the care required in selecting a scale type and interpreting the subsequent outcomes.

Often people are requested to fill out surveys with various types of scales used. There is a body of knowledge devoted to this endeavor (e.g., Cliff, 1959), and some of this work is topic specific, such as the study of scale use with students' evaluations of faculty (e.g., Meredith, 1975; Tollefson, Wigington, & McKnight, 1983). It is my contention that for the scale to be understood and used most effectively, a person should be able to understand each point in a survey continuum. For example, when faculty are evaluated by students on a global question such as "Over-all, this instructor is a good teacher," students need to understand each point in the scale. In this instance, sometimes a 1 to 10 scale is used where 1 = strongly disagree and 10 = strongly agree. They key question is 'Can students, using this scale, differentiate between a 6 and a 7'? The unique approach of the present study was that students were asked to generate an exemplar, i.e., example, for each point of the scale in use, providing a new method of studying how people can discriminate between rating scale points.

Previous studies have also examined the role of the rating scale in a person's discrimination amongst scale options. For instance, Andrich and Masters (1988) suggested that "the number of categories should be large enough to take advantage of the judge's capacity to discriminate but not greater" (p. 302). This study provides an empirical technique to assess that capacity. Andrich and Masters (1988) then cited Guilford (1954) who had suggested that, "if we use too few steps, the scale is obviously a course one, and we lose much of the discriminative powers of which raters are capable. On the other hand, we can grade a scale so finely that it is beyond the rater's limited powers of discrimination" (pp. 289-290). Guilford drew generic conclusions about the number of preferred scale points for unipolar and

[1]Address enquiries to R. E. Landrum, Ph.D., Department of Psychology, Boise State University, 1910 University Drive, Boise, ID 83725.

bipolar scales by citing studies by Conklin (1923) and Symonds (1924). Even more recent work by Neumann and Neumann (1981) acknowledged that the optimal number of categories issue has not been settled. However, more empirical examination of this issue is warranted, as they concluded "The short-scales (2- and 3-point scales) appeared too crude and usually resulted in more favorable conclusions, while the 7- and 10-point scales were too ambiguous as they involved some difficulty in discrimination between adjacent choices. This is only an impression though as no such data were collected" (p. 404). The present study was undertaken to use empirical data to assess the discriminative powers of different scale lengths using a unique and innovative exemplar approach.

Method

One hundred and forty-eight students enrolled in a general psychology course participated for course credit. Different scales were used to ask students to generate examples of specific scale points. One variable was type of scale tested (Strongly Disagree to Strong Agree, Poor to Excellent), and another variable was scale labeling (on some scales only the anchors were labeled; on other scales, each individual point was labeled). Participants were asked to generate an example of someone (faculty) that would fit each evaluation point, i.e., exemplars. Participants were then asked to differentiate between pairs of scale points ('can you tell me the difference between a 4 and a 3'). Participants were then asked to describe the characteristics of a particular scale point ('what would lead you to evaluate a teacher as a 2'). Finally, participants were asked to rate their own confidence in using a particular scale. Students were given about 30 minutes to complete this task.

Results

Five scales were tested: a 5-point scale with each point labeled strongly disagree to strongly agree, a 5-point scale with endpoints or anchors only labeled, a 10-point scale with endpoints labeled, a 4-point poor to excellent scale with each point labeled, and a 4-point scale with anchors only labeled. Preliminary testing indicated no significant differences between the two versions of the 5-point scales, so these data were combined. A similar result with the 4-point scales was found, so these data were also combined. The proportion of instances named for each scale point was the measure of interest; using the proportion adjusted for the potentially different outcomes due to scale lengths (4 vs 5 vs 10). The proportion of scale points completed did vary significantly with scale used ($F_{2,73}=3.23$, $p<.05$). The mean percentage completed for 4-point scales was 96.7%, for 5-point scales 87.8%, and for 10-point scales 79.3%. *Post hoc* tests indicate that the source of this effect lies in the difference in performance between a 4-point scale and a 10-point scale.

Discussion

When asked to generate exemplars of teachers on a 4-, 5-, or 10-point continuum, students reported fewer examples proportionally on the 10-point scale than on the 4- and 5-point scales. One explanation may be due to the effort required to generate 10 examples vs 4 examples. In a related vein, however, if students cannot think of an example of particular scale points, how effective will they be in using that scale to evaluate any particular faculty member? Labeling anchors did not differ significantly from labeling each scale point. Also, there were no significant differences in students' self-reported confidence in using the scales.

This study examined how well particular scale points are understood and examined the role of number of options, labeling, and anchors. Important personnel decisions are made based on the evaluation of faculty, and it would be beneficial to all involved to have some confidence in the validity of the evaluation and scale used. The results of this research are directly applicable to the design of evaluative measurements. If a typical respondent cannot generate an exemplar for each scale point used in the rating scheme, then perhaps that rating scale contains too many points for meaningful conclusions to be drawn.

REFERENCES

Andrich, D., & Masters, G. N. (1988) Rating scale analysis. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: an international handbook*. Oxford, Eng.: Pergamon.

Cliff, N. (1959) Adverbs as multipliers. *Psychological Review*, 66, 27-44.

Conklin, E. S. (1923) The scale of values method for studies in genetic psychology. *University of Oregon Publications*, 2, No. 1.

Guilford, J. P. (1954) *Psychometric methods*. (2nd ed.) New York: McGraw-Hill.

Meredith, G. M. (1975) Structure of student-based evaluation ratings. *Journal of Psychology*, 91, 3-9.

Neumann, L., & Neumann, Y. (1981) Comparison of six lengths of rating scales: students' attitudes toward instruction. *Psychological Reports*, 48, 399-404.

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

Tollefson, N., Wigington, H., & McKnight, P. (1983) Course ratings as measures of instructional effectiveness. *Instructional Science*, 12, 389-395.

Accepted December 28, 1998.