

THE EFFECT OF DECREASING RESPONSE OPTIONS ON STUDENTS' EVALUATION OF INSTRUCTION

R. Eric Landrum and Keli A. Braitman

Abstract. This study examined the statistical effect of changing from a 10-point to a 5-point response scale on students' evaluation of instruction. Participants were 5,616 students enrolled in classes offered by the College of Social Sciences and Public Affairs at a large Western university, who completed both the old evaluation (10-point response) and the new evaluation (5-point response). On average, students used a greater range of points on the scale for the 5-point scale (32 percent) than on the 10-point scale (19 percent). The effects of changing the response scale are discussed.

Keywords: *rating scales, student evaluation of instruction*

Society's investment in higher education is substantial. According to recent statistics, 17.3 million students enrolled in an institution of higher education in fall 2004 (National Center for Education Statistics 2006). Additionally, the federal government provides \$50.9 billion in financial aid support (grants and loans) for those attending a college or university (Morgan 2002). Given the

massive investment of various resources (e.g., time, money, effort), it seems reasonable that some mechanism be available for the evaluation of the delivery of academic instruction, and the most visible mechanism on the college campus is students' evaluation of instruction.

Although researchers agree that student evaluations of teaching inventories can be designed to elicit scores with validity and reliability (e.g., d'Apollonia and Abrami 1997; Marsh 1987; Marsh and Roche 1997), the current debate seems to center on measuring multidimensional aspects of teaching (i.e., multiple factors) or measuring teaching effectiveness unidimen-

sionally (i.e., global evaluation). Those in the unidimensionality camp believe that "student ratings should be used to make only crude judgments of instructional effectiveness (exceptional, adequate, and unacceptable)" (d'Apollonia and Abrami 1997, 1205). Of more interest to this study, however, is how the structure of the evaluation (more specifically, the response scale used) affects the respondent's use of that evaluation.

How individuals use response scales in survey research is not a new area of inquiry. Several researchers have addressed response options issues, but not within the context of student evaluation of teaching. For example, Sidick, Barrett, and Doverspike (1994) studied the transition from 5-option to 3-option multiple-choice items (much like this article addresses the transition from 10-option to 5-option choices) and found the 3-option items to be preferable. Neumann and Neumann's (1981) study is most directly related to this one, however. First, they posited that raters do have a limited amount of discriminative ability and suggested that rating scales beyond 7 or 9 categories does not make much sense. Second, Neumann and Neumann observed in their own data that respondents tended not to use the extremes of the case, especially with 7- and 10-point scales. They found that all the distributions were negatively skewed, with respondents preferring a 5-point scale. This finding raises some

R. Eric Landrum is in the department of psychology at Boise State University. Keli A. Braitman is with the Insurance Institute for Highway Safety, Arlington, Virginia.
Copyright © 2008 Heldref Publications

interesting questions. If one was to shorten a rating scale, would the overall range of values used increase? How would the ratings change on comparable evaluation items?

This study takes a different approach than Neumann and Neumann (1981). We modified 10-option scale evaluation items for use on a 5-option scale. This process occurred within a college (with multiple departments) at a large western university. One unique aspect of this study is that students enrolled in classes filled out both the old 10-option evaluation form and the new 5-option evaluation form, and we could link the responses. Hence, this procedure allowed for direct testing of the effects of scale change on teaching evaluations with a large, multi-department sample.

Method

Participants

Students enrolled in classes offered by departments in the College of Social Sciences and Public Affairs participated. Students from anthropology ($n = 724$), communications ($n = 814$), criminal justice ($n = 86$), history ($n = 1,143$), political science ($n = 446$), psychology ($n = 1,962$), and sociology ($n = 441$) participated, for a total of 5,616 participants.

Materials and Procedure

Students completed two different course evaluation instruments, the one that had been in use for some time, using a 10-point rating system and the newly developed evaluation using a 5-point rating scale (see appendix for the "old" and "new" rating scale).

Students completed each evaluation using Scantron-type bubble sheets and pencils. We informed students that a new evaluation instrument was being tested, and to help validate the new instrument, we instructed students to put the last four digits of their social security numbers on both of the Scantron-type bubble sheets (old rating scale and new rating scale). We assured students that no effort would be made to match their ratings to their identities. On average, it took students 10 minutes to complete each of the rating forms. The university data center scanned all of the evaluations, and after depart-

ments completed their analyses, the rating sheets were recoded and rescanned for this study.

Results and Discussion

To examine whether a greater range of evaluative values is used when a greater range is available (e.g., a 10-point scale versus a 5-point scale), a range value was obtained for each participant on each instrument. We calculated the difference between the maximum and minimum evaluation score for each participant (on each instructor) separately for the two evaluation instruments across all questions. An alpha of .05 was used for all significance tests. There was a statistically significant difference between ranges for the two instruments $t(3,401) = 9.75, p < .05$.

In this study, we found that students, in absolute terms, used a larger range of values on a 10-point scale than on a 5-point scale. This finding should be interpreted with caution, however. The effect size of this difference was quite small ($d = .17$), and the difference between the means was relatively small ($M = 1.93, SD = 1.92$ for the 10-point scale versus $M = 1.62, SD = 1.09$ for the 5-point scale). Students using a 5-point response scale tended to use about 32 percent of the possible range of values, whereas the same students using a 10-point response scale tended to use only about 19 percent of the range of possible values. Given that the mean range for both types of scales was less than two points, the 5-point response scale seems to be the more appropriate of the two scales. With the 10-point scale, given that the mean range for respondents is less than two points, most of the response options were unused.

To further examine the statistical effect of different response scales, we compared nearly identical items from the two scales. For the 10-point response scale, one item was, "The instructor was well prepared for each class." For the 5-point response scale, a nearly identical item was, "The instructor seemed well prepared for class." To examine whether the response scale affected mean ratings for these items, raw scores were converted to proportions by dividing the raw evaluative rating by the number of response options. A paired samples t test was performed to compare the mean prop-

portion ratings between the two items. A statistically significant difference was found, $t(3,382) = 29.41, p < .05, d = .51$, indicating that the use of a 5-point scale resulted in lower mean ratings ($M = 0.68, SD = 0.16$) than a 10-point scale ($M = 0.77, SD = 0.18$). This finding suggests that changing from 10 response options to 5 may influence student ratings of teachers. One possible explanation for this finding is that 10 response options provide students with more evaluative options for differentiating among their teachers. However, we report that students used a limited range of this scale. Thus, a more plausible explanation is that it becomes increasingly difficult to differentiate between the meaning of adjacent response values with 10 options. For example, students may have a harder time differentiating between a 6 and 7 on the 10-point scale than differentiating between a 4 and a 5 on a 5-point scale. In fact, such a suggestion already exists in the literature (Landrum 1999). Accordingly, student ratings may be less accurate with 10 response options compared with 5 response options.

Conclusion

The purpose of this study was to examine whether students use a greater range of scores when available (e.g., a 10-point scale versus a 5-point scale) and to examine the statistical effect of changing from a 10-point to a 5-point scale. Previous researchers have provided conflicting recommendations about the length of response scales. Whereas T. R. Hinkin (1998) recommends a 5-point Likert-type scale, D. H. Wedell, A. Parducci, and M. Lane (1990) suggest that increasing response options increases reliability.

One limitation of our study is that we only compared a 5-point scale and a 10-point scale. It would also be informative to evaluate the comparative differences of a 7-point scale (cf. Preston and Colman 2000). Although this comparison was not conducted in this study, it would have been preferable to keep the questions constant and only change the response scales from the old to the new evaluation form. Additionally, we would have liked to counterbalance the presentation of old and new evaluation instruments, but given the nature of the data collection process

and the large data set, this was not plausible. A necessary next step would be to examine the effect of different response formats on the reliability and validity of teacher evaluations. In the current study, we found that teacher ratings were higher for the 10-point scale than for the 5-point scale (when scores are converted to a common metric). Future research should explore the question about whether the use of a 5-point scale leads to a superior evaluative outcome.

REFERENCES

- d'Apollonia, S., and P. C. Abrami. 1997. Navigating student ratings of instruction. *American Psychologist* 52:1198–208.
- Hinkin, T. R. 1998. A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods* 1:104–21.
- Landrum, R. E. 1999. Scaling issues in faculty evaluations. *Psychological Reports* 84:178–80.
- Marsh, H. W. 1987. Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Education Research* 11:253–388.
- Marsh, H. W., and L. A. Roche. 1997. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist* 52:1187–197.
- Morgan, R. 2002. Are federal grants taking a back seat to student loans? *Chronicle of Higher Education* 49 (11): A23–A26.
- National Center for Education Statistics. 2006. *Digest of education statistics 2005*. Washington, DC: U.S. Government Printing Office.
- Neumann, L., and Y. Neumann. 1981. Comparison of six lengths of rating scales: Students' attitudes toward instruction. *Psychological Reports* 48:399–404.
- Preston, C. C., and A. M. Colman. 2000. Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* 104:1–15.
- Sidick, J. T., G. V. Barrett, and D. Doverspike. 1994. Three-alternative multiple choice tests: An attractive option. *Personnel Psychology* 47:829–35.
- Wedell, D. H., A. Parducci, and M. Lane. 1990. Reducing the dependence of clinical judgment on the immediate context: Effects of number of categories and type of anchors. *Journal of Personality and Social Psychology* 58:319–29.

APPENDIX

Student Evaluation of Teaching “Old” Rating Form with 0–9 Scale Used

Evaluation item

1. Instructor demonstrated a thorough knowledge of the subject matter.
2. Course appears to have been carefully planned.
3. Instructor was well prepared for each class.
4. Instructor presented course material in an understandable fashion.
5. Methods used to evaluate my learning were clearly stated.
6. Overall, I would rate this instructor as a good teacher.

Note. For these questions, the 0 anchor was labeled “strongly disagree” and the 9 anchor was labeled “strongly agree.” None of the intermediate rating points were labeled.

Student Evaluation of Teaching “New” Rating Form with 1–5 Scale Used

Evaluation item

1. Instructor's presentations increased my knowledge of the subject.
2. Instructor's methods of evaluation were fair.
3. Instructor was available during office hours.
4. I would recommend this instructor to another student.
5. I felt free to participate (e.g., ask questions) in this class.
6. Instructor seemed well prepared for class.
7. Instructor expressed ideas clearly.
8. Objectives of the course were met.
9. Assignments and exam results were returned in a timely fashion.
10. Assignments were of value to my learning.

Note. For these questions, a 5-point Likert-type scale was used, with each individual rating point labeled (1 = “strongly disagree”; 2 = “disagree”; 3 = “uncertain”; 4 = “agree”; 5 = “strongly agree”).