Getting More Out of Educational Workshop Evaluations: Positively Packing the Rating Scale

Joni M. Lakin *Shankharupa Chaudhuri*

Auburn University

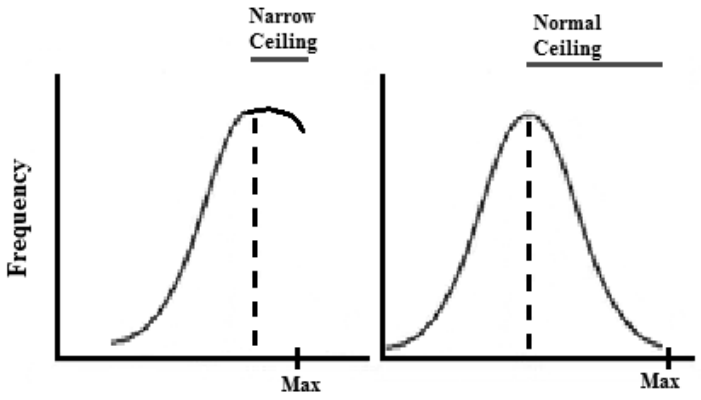
Collecting evaluations following a professional development workshop and similar events has become common practice for assessing workshop quality. However, these evaluation forms often do not reflect best practices in survey development and result in average ratings that are uniformly high and uninformative. The purpose of this study was to evaluate the effects of rating scales on workshop evaluations using an experimental design where participants were randomly assigned one of four evaluation forms with different rating scales following workshops. We found that using a typical "poor to excellent" scale yielded the highest average (most homogeneous) ratings while scales that were off-centered, with more positive than negative scale points (positively packed), yielded ratings with more variability. Recommendations for designing workshop evaluations are provided.

Getting More Out of Educational Workshop Evaluations: Positively Packing the Rating Scale

It has become standard practice in the U.S. to end workshops and professional development events with an evaluation form, intended to provide feedback to the workshop provider and assess perceived benefits of the event. However, the typical design of these evaluation forms may lead to undesirable effects on the quality of feedback received. Specifically, from our work in providing such workshops, we were concerned about the limited variability we saw in workshop ratings that we attributed to a *leniency bias* causing *ceiling effects*. A ceiling effect is the observation that the ratings

on a particular item or scale have average scores and score ranges that are very close to the maximum rating point with little room for variation in scores above the mean. See Figure 1 for an example. The purpose of this experimental study was to explore how the design of the rating scale impacted the ceiling statistics observed on workshop evaluation forms.

Figure 1. Example of distribution with normal and narrow ceiling at the upper end of the distribution



The trouble with ceiling effects

In workshop evaluations, ceiling effects (in our experience) seem to be the norm, with universally high ratings from every workshop. This may reflect truly strong performance, but we suspect there is an upward bias of respondents who do not feel the need to disparage presenters or a workshop. In other words, workshop evaluations seem to be vulnerable to a leniency bias (Anastasi, 1992) that results in a strong ceiling effect.

A ceiling effect is the observation that a variable has typical scores that are very close to the maximum score of the variable. On Likert-type scales, this means that the typical

rating is near the maximum scale point (or anchor point). This introduces psychometric problems in using the scale and trusting the utility of the scores given. Ceiling effects reflect reduced variability or restriction of range, which attenuates correlations with other variables (Cronbach, 1990). It can also compress or obscure mean differences, limiting the opportunity to find improvements over time, relative effects of different interventions, or differences between groups.

In psychometric thinking, ceiling effects are clearly a problem because of these statistical challenges. For a coordinator of workshops, however, universally high ratings may initially seem to be a good thing. Anyone organizing a workshop will make an effort to select presenters they believe will be effective and engaging and pick topics they believe will offer value to participants. To receive many low ratings of their workshops would indicate they were not choosing presenters carefully. On the other hand, anyone who organizes more than one workshop will recognize that workshops *do* vary in quality. Researchers we work with are often puzzled that these clear variations are often not reflected by the evaluations.

If the goal of workshop evaluations is to gather *formative* feedback to lead to improvements and identify the most valuable workshops, then being unable to detect differences in workshop quality makes the feedback from attendees worthless in improving future workshops. To summarize, if a workshop organizer's goal is to validate their workshops, advocate for their effectiveness, and/or gather summary evidence that the events were appreciated by attendees, then ceiling effects and universally high ratings are acceptable. However, if organizers want to differentiate between the generally good workshops and the truly exceptional workshops (those worth replicating, investing in, and disseminating), then more differentiation in evaluation ratings is essential. In this study, we explored how to obtain

greater rating variations in the context of an organization that provides professional development workshops on educational topics.

Prior work on scales

Some basic features of rating scales have been extensively studied and show that rating scales (including the number of scale points and anchor labels for those points) are critical to the quality of ratings obtained by a survey instrument (Weng, 2005). One of the most commonly used rating scales in behavioral sciences, especially for evaluation tools, is anchored as Above Average, Average, Below Average (French-Lazovik & Gibson, 1984). However, French-Lazovik and Gibson showed that "average" may not be a useful midpoint on a scale, because their subjects perceived rating human behaviors with the term "average" as actually being a negative evaluation. Thus, using average as a midpoint label may push respondents towards the higher categories and contribute to reduced variability in responses.

In order to ensure that the raters use the entire scale instead of solely focusing on the extreme negative or positive points, it may be necessary to change the scale labels to reflect less extreme points on the scale or move the center point of the scale. According to findings from several studies, adding more positive labels to the scale, also referred to as *positively packing*, leads the respondent to use more of the scale range (English et. al., 2009; Hancock & Klockars, 1991; Lam and Klockars, 1982; Klockars & Yamagishi, 1988; Schwarz, Knauper, Hippler, Neumann, & Clark, 1991; Wyatt & Meyers, 1987).

Although Guilford suggested moving the center point of a rating scale over 70 years ago, the practice is uncommon in current program evaluation practice and has limited research supporting its utility in decreasing rating leniency specifically for program and workshop evaluations. One relevant study

was conducted by Vita et al. (2013), who explored rating scales and anchor labels for workshop evaluations at professional/academic meetings with the goal of reducing ceiling effects. They combined several manipulations into one comparison, moving the center point of the rating scale ("average") from point 3 on a 5 point scale to point 2, so there were more scale points above than below "average". They also associated the anchor points with normative information so that their experimental scale had these five points: 1 (below expectations), 2 (average), 3 (truly above average), 4 (outstanding), 5 (top 5%). Their goal was to create a scale that "better differentiates above-average ratings from what is truly outstanding" (p. 48). They found that their scale changes did improve the ceiling, although the study somewhat confounded format (paper vs. online) and changes across two years of the professional meeting. This study expands on and replicates their study, applying the off-center average and normative information in separate manipulations to determine the source of their effects. The current study addressed the following research questions:

1. Do the anchor labels affect the average rating given to workshops?
2. Do the anchor labels affect the scale ceiling of ratings?
3. Do the anchor labels create more variation in scores across evaluation questions?

Methods

Participants were instructional faculty, staff, and graduate students who attended one or more of 11 professional development events organized by a Center that provides professional development in teaching and learning at a southeastern university. Over the course of a year, 378

participants completed session evaluation forms at the end of these workshops.

Instruments

At the center, the same evaluation form is used for all workshops. It asks a short series of questions with Likert-type rating scales as well as open-ended questions about what was effective or ineffective in the session. To increase the utility of this research for the center, we maintained their question wording, but modified the ratings scale. For the rating scales, four general questions were asked about each workshop:

1. "Did you find the topic useful?"
2. "Were the presenter(s) effective?"
3. "Opportunities for interaction & discussion?"
4. "Overall quality of the session?"

For this study, the only variation across the different evaluation forms was the scale given to respond to the questions. Four formats of the scale were used with the anchor point labels shown in Table 1. First, the control condition was the generic scale previously used by the Center for their evaluation questions: 1 (Low) to 5 (High). On the basis of survey development guidelines, the first experimental condition changed the anchor descriptors to be more relevant to the target construct (quality of session): 1 (Poor) to 5 (Excellent). On the basis of prior work (Vita et al., 2013) and our own experiences with evaluations, we knew that participants would rarely indicate that a presenter was poor quality both because the Center chooses relatively effective presenters and because attendees at workshops are generally kind to volunteer presenters. To address this upward bias, we developed a third scale which added labels for each scale point and moved the "average" anchor from the middle to the left end of the scale. This is a *positively packed* scale

compared to scale 1 and 2. Finally, a fourth scale provided norm-referenced anchor points (4 = Well above average, top 25%, 5 = Excellent, top 5%), consistent with the Vita et al. (2013) and to further encourage respondents to only use the top scale point for truly exceptional presentations.

Table 1: Rating scales

1–Low to High	2–Poor to Excellent	3–Positively Packed	4–Positively Packed, with Norms
1 (Low)	1 (Poor)	1 (Below average)	1 (Below average)
2	2	2 (Average)	2 (Average)
3	3	3 (Above average)	3 (Above average)
4	4	4 (Well above average)	4 (Well above average, top 25%)
5 (High)	5 (Excellent)	5 (Excellent)	5 (Excellent, top 5%)

Procedure

After each workshop, participants completed a workshop evaluation form. At the Center, the evaluation forms are always placed on the tables before the session begins so that participants can complete one before they leave. For the purposes of this experiment, four different variations of the evaluation form were spiraled together, so that attendees sitting at the same table would get different forms of the evaluation sheet. At the end of all workshops at the Center, attendees are strongly encouraged to complete an evaluation form. There was no special mention of variations in the forms to the attendees during the research study. All evaluations were anonymous.

Results

Eleven workshops were given with the evaluation form manipulation from August 2013-February 2014. The average ratings varied from 3.0 to 5.0 and participation ranged from 14 to 91 attendees per event (Graduate Teaching Assistant orientation sessions were the most highly attended).

To address research question 1, the average ratings across scale formats were compared. A one-way independent ANOVA showed that average ratings differed across scale formats ($F(3,377) = 3.616$, $p = .013$). Tukey post-hoc tests indicated that scale 2 ("poor to excellent") had significantly higher ratings than scales 3 and 4 (the two off-center scales—"positively packed" and "positively packed with norms"), consistent with more leniency effect on this scale. See Table 2. We were concerned that variations across workshops might confound formatting effects, because the number of each rating scale format administered was not exactly proportional across workshops (which did vary in average ratings). When ratings were standardized across workshops, there was still a significant effect of format ($F(3,374) = 2.848$, $p = .037$). Therefore, confounding of evaluation forms and workshops was not an issue in the analyses.

The ANOVA results confirmed that the anchor labels could affect average ratings of participants. Another indication that the scale has reduced the upward bias was the percent of participants giving ratings of 4 or 5 (on each of the individual questions, but aggregated here). Scales 3 and 4 ("positively packed" and "positively packed with norms") showed the fewest high ratings (4/5), indicating that respondents were more selective in using these ratings and that there is more room to differentiate between good and truly excellent workshops with these scales. Comparing the proportions of ratings across all four scales, we found significant chi-square tests indicating that the rating distributions of all scales, except 3 and 4, differed from each

other significantly ($p < .001$). Scales 3 and 4 had the highest proportion of ratings at the low end of the scale (ratings 1-3).

Table 2: Descriptive Statistics for Average Ratings across Four Rating Questions

		Low to High	Poor to Excellent	Off-center with norms	Off-center no norms
N		105	93	90	90
Rating	1	2%	1%	1%	2%
	2	5%	4%	8%	10%
	3	18%	12%	22%	17%
	4	32%	30%	32%	33%
	5	42%	54%	36%	38%
M		4.04	4.32	3.95	3.96
SD		0.86	0.78	0.86	0.95
Skew		-0.89	-1	-0.43	-0.96
Kurt.		0.54	0.58	-0.72	0.15
Ceiling statistic		1.11	0.87	1.22	1.09

Note. **Results for the four rating questions were similar to the results for the average ratings. Therefore, we just report average ratings here.**

Ceiling effects

Research question 2 addressed the presence of a ceiling effect with the four different scales. Skew and kurtosis statistics indicated that these four rating scales resulted in different degrees of non-normality in the distributions. Given the standards errors for kurtosis, there was a non-significant degree of kurtosis for each rating scale. There was significant negative skew for each scale, although scale 3 ("positively packed") had the least skew, only just exceeding the 2 standard error cutoff for significant skew.

In addition to skew and kurtosis statistics, a ceiling statistic (Bracken, 2007) was calculated to frame the distribution differences in practical terms. A ceiling statistic is calculated as the maximum score on the scale (in this case, 5) minus mean score, divided by the standard deviation (Bracken, 2007). In mathematical terms:

$$C.E. = \frac{Max - \bar{X}}{\sigma}$$

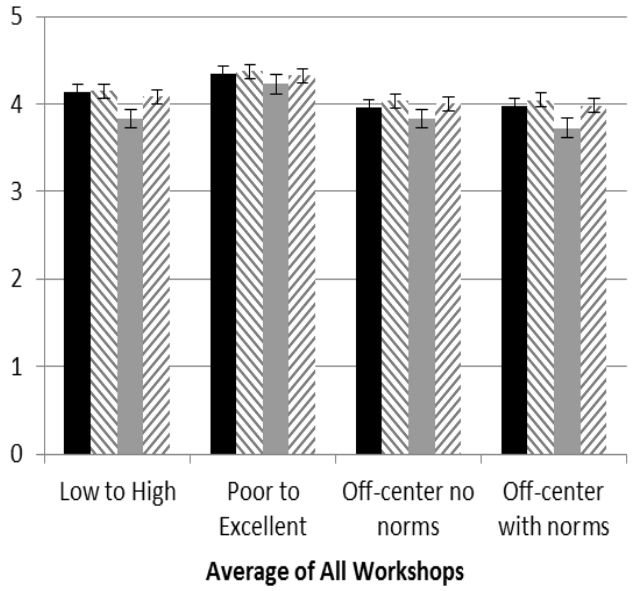
Larger values of this statistic indicate more ceiling (i.e., more room for variability in scores around the mean). This is a useful statistic for quickly determining the degree to which scales have a negative skew due to a score ceiling. Table 2 shows the ceiling statistics by scale format. All of the formats yielded ceilings that were more restricted than the 2SD criteria suggested by Bracken (2007). However, format 3 ("positively packed") yielded the most ceiling, followed by format 1 ("low to high") and 4 ("positively packed with norms").

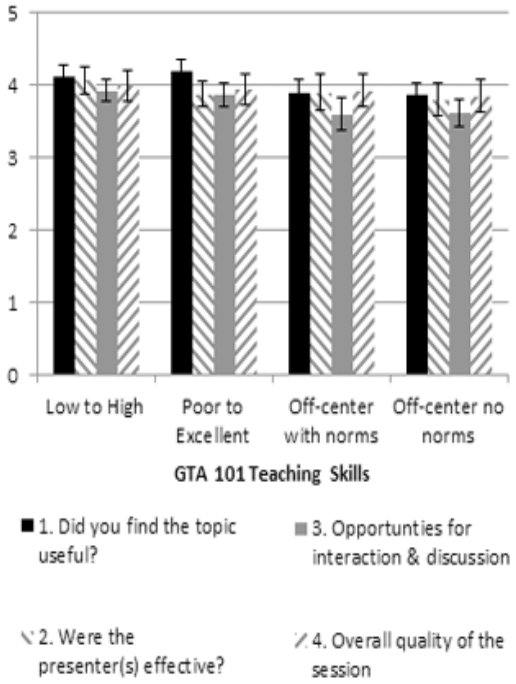
The superiority of format 3 was surprising, as we expected that normative judgments (format 4) would make the highest ratings less common.

Individual Evaluation Questions

Question 3 addressed the effects of the anchor scales across the four different evaluation questions posed to respondents. Differentiation across the questions was of particular interest, because that information *should* help workshop organizers pinpoint specific areas for improvement. Considering the four evaluation questions independently in Figure 2 (for all workshops combined as well as three larger workshops separately), it appears that

scale 2 ("poor to excellent") yielded the least differentiation across evaluation questions (in


5
4
3
2
1
0
Low to High
Poor to Excellent
Off-center no norms
Off-center with norms
Average of All Workshops


5
4
3
2
1
0
Low to High
Poor to Excellent
Off-center with norms
Off-center no norms
GTA 101 Teaching Skills
1. Did you find the topic useful?
2. Were the presenter(s) effective?
3. Opportunties for interaction & discussion
4. Overall quality of the session

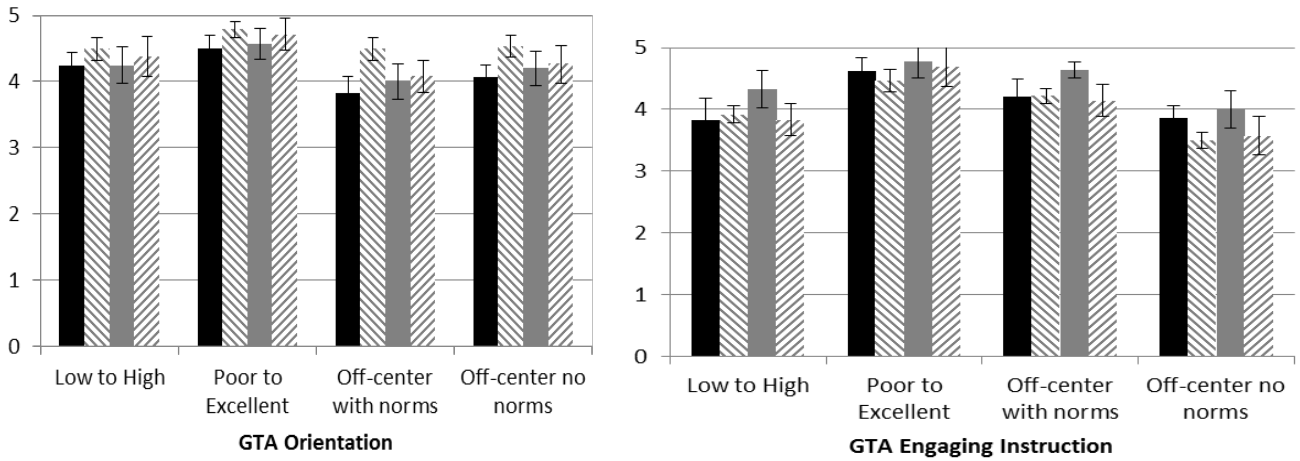


Figure 2. **Mean rating by evaluation question and format, with standard error bars (a) combined workshops data, (b-d) individual workshop results with over 50 responses.**

other words, consistently high ratings on each dimension). With the other scales, it is clear that the attendees were somewhat less satisfied with the opportunity for interaction (in the averaged workshop results) and discussion than they were with the presenter, topic, and overall quality. This would be useful information to workshop presenters that is obscured on scale 2 ("poor to excellent"). For the three individual workshops, sample sizes are too small for many significant results, but more variation is observed for scales 3 and 4 ("positively packed" and "positively packed with norms"). For "Engaging Instruction" there was significant variation across scale formats and significantly stronger ratings for opportunity for interaction when scales 3 and 4 are used.

Figure 3 shows the average response to Question 4 ("Overall quality of the session") for all of the workshops. This allows us to look at variation across workshops where the previous analysis focused on differences across dimensions of each workshop. Figure 3 differs from the previous result. Now scale 1 ("low to high") appears to show the least variation.

Scales 2 and 4 ("poor to excellent" and "positively packed with norms") show markedly more differentiation across workshop than scale 1.

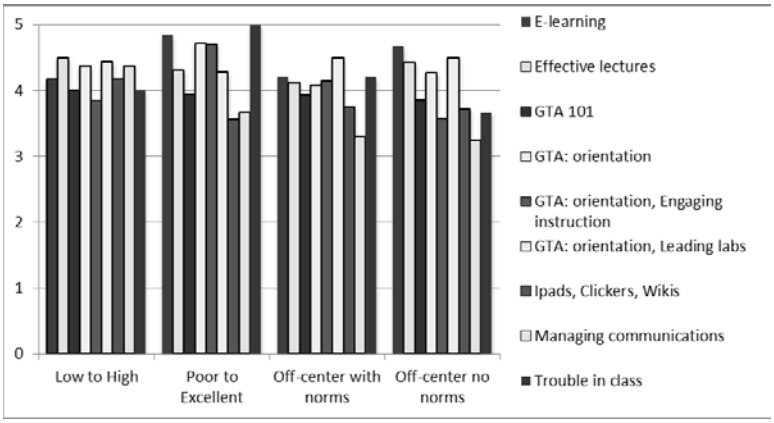


Figure 3. Average overall ratings (Question 4) for each workshop (excluding means based on less than 3 responses).

The ceiling statistics varied somewhat by question, as can be seen in Figure 4. Scale 2 ("poor to excellent") consistently had the least variability and ceiling. The two off-center scales (3 and 4) showed larger ceilings and overall more variability across questions. We looked at similar results by workshop, but the sample sizes by workshop were too small in most cases to make substantive interpretations.

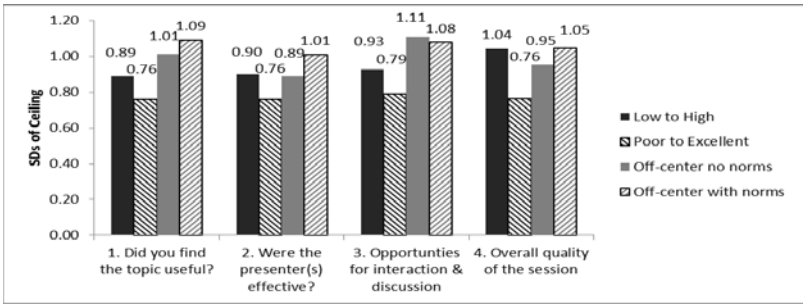


Figure 4. **Ceiling statistics by question and format (2SD is ideal)**

Discussion and Conclusions

As we expected, the different anchor labels used in the four different scales had significant impacts on the average rating, the amount of ceiling a scale showed, and the differentiation across evaluation questions. We found that using a poor to excellent scale without mid-point labels (scale 2) yielded the highest average rating, reflecting less differentiation across the four evaluation questions by respondents and the least amount of ceiling. Scale 1 ("low to high") and scale 3 ("positively packed"), which used an off-centered scale with all five anchor points labeled, yielded the best results. We would recommend the use of scale 3 in the future because it provides the most useful scale psychometrically. It is also best practice, based on survey research literature, to label all anchor points and to make the scale specific to the construct being studied, rather than generic as scale 1 was. Scale 1 leaves too much ambiguity for the respondents to interpret.

The superiority of format 3 ("positively packed") over 4 ("positively packed with norms") was surprising, as we expected that normative judgments that were implied by format 4 would make the highest ratings less common. This adds a new perspective on the findings of Vita et al. (2013) who combined the positively packing with normative information. It appears from our results that only the positively packing was effective in modifying respondent ratings.

Future research could use cognitive interviews to explore how respondents interpret anchor labels with normative information. Our study as well as Vita et al. (2013) assumed normative information would encourage more conservative use of the top scale point. We may find, as French-Lazovik and Gibson (1984) did with the term "average", that respondents are interpreting anchor labels in unexpected ways.

The narrow ceilings observed, even on the best scale in our study, make it clear that more work is needed. For example, future research should explore methods of providing directions or guidelines to respondents to encourage honest and constructive criticism through evaluation forms. Perhaps explaining the benefit of honest feedback could further reduce the ceiling effects observed. In other contexts, it may also be feasible to look at the ratings of the same individuals across different workshop events. Being able to explore scale effects for the same individuals could increase the power to see trends in ratings.

Given the widespread use of session evaluation forms for a myriad of programs and workshops, the use of ratings scales deserves more careful research. Workshop evaluations with no variability do not help to differentiate excellent programs or to identify areas for improvement. When realistic and detailed evaluation results are desired (as opposed to merely gathering high ratings for marketing purposes), the anchor labels and scale are essential to gathering high-quality and useful evaluation data.

References

Anastasi, A. (1992). *Psychological testing.* New York: McMillan.

Bracken, B. A. (2007). Creating the optimal preschool testing situation. In B. A. Bracken, & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 137-154). Mahwah, NJ: Lawrence Erlbaum Associates.

Cronbach, L.J. (1990). *Essentials of psychological testing.* New York: Harper & Row.

Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. *Journal of Applied Psychology, 65*, 147-154.

English, A., Rose, D., & McLellan, J. (2009). Rating scale label effects on leniency bias in 360- degree feedback.

The 24th Annual Meeting of the Society for Industrial Organizational Psychologists.

French-Lazovik, G., & Gibson, C. L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. *Applied Psychological Measurement, 8*(1), 49-57.

Klockars, A. J., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement, 25*(2), 85-96.

Lam, T. C. M., & Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement, 19*, 317-322.

Landrum, R. E. (1999). Scaling issues in faculty evaluations. *Psychological Reports, 84*, 178-180.

Likert, R., Roslow, S., & Murphy, G. (1934). A simplified and reliable method of scoring the Thurstone attitude scales. *Journal of Social Psychology, 5*, 228-238.

Sangster, R. L., Willits, F. K., Saltiel, J., Lorenz, F.O., & Rockwood, T. H. (2001). *The effects of numerical labels on response scales*, U.S. Bureau of Labor Statistics.

Schwarz, N., B. Knauper, H. J. Hippler, E. Noelle Neumann, & Clark, L., (1991) Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*(4), 570-582.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, UK: Cambridge University Press.

Wang, C.-N., & Weng, L.-J. (2002). Evaluating the use of exploratory factor analysis in Taiwan: 1993-1999 *Chinese Journal of Psychology, 44*, 239-251.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956-972.