

An Investigation Of 'Honesty Check' Items In Higher Education Course Evaluations

Kelly D. Bradley, University of Kentucky, USA

Kenneth D. Royal, University of Kentucky, USA

James W. Bradley, Bluegrass Community and Technical College, USA

ABSTRACT

The reliability and validity of course evaluations in higher education is often assumed. The typical Likert-type surveys utilized when students' evaluate the course and instructor often overlook measurement issues, or deal with them in an ineffective manner. Given the importance that is placed on higher education course evaluations, with results impacting such events as merit raises and promotion, the proper construction and use of evaluation tools is a critical issue. In an effort to assure 'honesty' in student responses, many institutions include items written positively and negatively, which are intended to measure the same construct. Using 537 course evaluations for a mathematics faculty member at a Midwest college, an item analysis is conducted with attention given to means and standard deviations, frequency counts, nonparametric correlations and tests of significant differences between questions that should, in theory, produce a similar measure or exactly opposite. A contention is made that the way the item is asked does matter, at least in some instances, and it should not be assumed that an item written in the positive and negative should directly correlate. The survey research community and institutions utilizing similar rating scale instruments will benefit from the results of this study, as well as the education community in general.

Keywords: Measurement, Course Evaluations, Reliability and Validity, Higher Education

INTRODUCTION

The practice of using reverse or negatively worded items on higher education course evaluations, and surveys in general, has been utilized for decades. The primary purpose of including this type of item is to ensure valid measures by safeguarding against acquiescence; basically, it is an “honesty-check” with the goal of identifying respondents who appear to select items haphazardly. Presumably, with proper identification of these respondents, institutions could remove these individuals, or make necessary adjustments prior to data analysis that would lead to valid results. This process, in of itself, can have a significant effect upon the reliability and validity of the evaluation process. Interestingly, the search for best practices in regards to the use of positively versus negatively worded items in the survey research literature is often muddled and contradictory.

THEORETICAL FRAMEWORK

Reliability And Validity

Falzhik and Jolson's 1974 study lays a foundation for discussion. The authors created two surveys on different topics. Each survey contained six positive statements and six negative statements, which were constructed as the inverse of the other. Results suggested the intensity of responses depended on whether the wording was positive or negative. The authors also found when a personalized direction of a question is changed to a non-personalized direction, suggestibility is decreased. Falzhik and Jolson went on to argue the historical predominance of positive wording on Likert-type surveys may potentially lead respondents to answer differently than they would if questions were worded with the occasional negative statement, all else being the same.

Conversely, Leenaars, Bringmann, and Balance (1978) found there was no clear evidence that positively worded statements were more accurate than negatively worded ones. Results did suggest that negative statements were more difficult to interpret, possibly leading to other problems. One such problem was identified by Noelle-Neumann's (1970) research which found negative or incomplete wording of items to have a discouraging effect upon respondents and their response behaviors.

Specific to reliability, Barnette (1997) conducted a study which examined the effects of positively versus negatively worded items by using a 2 x 3 study design. The six versions of the survey instrument included versions that were completely positively worded and worded with an equal mix of positive and negative statements. The different forms of the surveys also contained a mixture of rating scale structures as some ranged from Strongly Disagree (SD) to Strongly Agree (SA), SA to SD, and with an equal mix of SD to SA and SA to SD. Results revealed the surveys with all positive statements produced the highest reliability estimates. The survey with the lowest reliability was the survey that contained an equal mix of positively and negatively worded items and an equal mix of SD to SA and SA to SD scale structures. Schriesheim and Eisenbach's (1995) research echoed these findings, also reporting regularly worded items were the most reliable.

Other Issues Regarding Bias

There are also debates surrounding the issue of bias. In 1981, Schriesheim and Hill found mixing positive and negative items did not control for acquiescence. In 1998, Bergstrom and Lunz found it acceptable to provide both positive and negatively worded items on a questionnaire, particularly when applying the Rasch Rating Scale model, because there is sufficient evidence to suggest both types of statements are measuring the same construct. Garg presented evidence to the contrary. In a 1996 study, subjects' responses were greatly affected by the positively and negatively worded statements in attitude questionnaires, as subjects tended to respond positively to positively worded items and negatively towards negatively worded items. Reiser, Wallace and Schuessler (1986) found negatively worded items tended to draw more keyed responses than positively worded items. Generally, respondents were less inclined to reject a positive statement than to accept a negative one. The authors claimed, "all else being equal, it is less threatening and therefore more practical to agree than to disagree with statements on self and society, no matter what the specific content of those statements may be" (p. 23).

Similar to Reiser et al., Mook, Kleijn and Van der Ploeg (1991) found differences between responses when using positive versus negative worded items and attributed those differences to issues of item-intensity and socially desirable responses. Hazlett-Stevens, Ullman and Craske (2004) found response patterns differed between positively worded and reverse-scored items. The authors claimed it could be due to an inherent difficulty of answering reverse-coded items due to the double-negative nature of the items.

Survey Respondent Behavior And Associated Theories

There are numerous theories for specific response tendencies. Bishop, Oldendick and Tuchfarber (1982) theorized respondents answered attitude survey questions without an exhaustive search, as they typically respond to the first thing that comes to mind, or whatever is most accessible in memory. Schuman and Presser (1977) found less educated persons are affected more by agree/disagree statements than others. Offering a possible explanation for the education gap/response behavior phenomena, Narayan and Krosnick (1996) theorized people with higher education levels might be more sensitive to wording and various distinctions, whereas people with less education might be more influenced by the inclusion or omission of particular response options, wording, and formats.

Zaller and Felman (1992) argue most people do not possess preformed attitudes at the level of specificity demanded in surveys. When questioned, individuals call to mind a sample of ideas and use them to choose among the options offered. In turn, the resulting response reflects the thoughts that are most accessible to one's memory at the moment of response. Dillman and Christian's (2005) research suggested that identical surveys can reveal different responses depending on the mode through which it was administered. The authors contended that the visual layout chosen for the survey may have a direct effect upon respondents' answers.

Research Pertaining To Teaching Evaluations

In 1982, John Ory created six evaluation forms which measured items relating to both the course and the instructor. Each instrument consisted of 30 items with a varying number of negatively worded items (0, 10 or 20) that were placed either before or after a question that solicited an overall rating. Two experiments were conducted using this methodology with 75 undergraduate students. Results revealed overall ratings were not affected for both the course and the instructor. However, placement of certain items, particularly items placed after the overall rating question, produced lower ratings for the course in one of the experiments.

Two years earlier, Ory and Valois (1980) conducted two studies which investigated essentially the same questions of positively versus negatively worded items and placement of items. The study's instrument contained 30 items, half of which pertained to the instructor and half of which pertained to the course. Twenty of the 30 items contained a positively worded statement and its assumed inverse. In these studies, 455 undergraduate students completed one of six randomly assigned evaluation forms. ANOVA results found instructor ratings were significantly higher than the ratings of the course. Overall ratings for both the instructor and the course were not affected by either positively or negatively worded statements in either of the studies.

METHOD

Research Questions

The study was driven by the need to compare the positive and negative items on the evaluation that were intended to measure the same construct. The assumption is that the positively and negatively worded items would directly correlate and have no significant difference between the average items scores. This research was directed at exploring this assumption through the specific questions:

- * Are the response patterns consistent for the positive and negative paired items?
- * Do the positive and negative paired items directly correlate?
- * Does a significant difference exist between the positive and negative paired items?

Instrumentation And Data Collection

A census sample of all course evaluations over a four-year period for a mathematics professor at a private college in a large Midwest City was utilized. The instructor employed consistent procedures across the timeframe in distributing and collecting evaluation data. Here, the instructor was responsible for all classes and the courses were all mathematics based but the content was not consistent. Even though a variety of student majors were represented, all students shared the same core of mathematics curriculum. Class sizes were not equal; still, no class would be classified as small or large, resulting in comparable situations.

The instrument was constructed by the institution and used as an end of semester evaluation for the course/instructor for all classes offered. It was comprised of 35 closed-ended statements that students were asked to rate their agreement using a 4-point scale, where 1 = Never / Almost Never, 2 = Some of the Time, 3 = Usually and 4 = Always / Almost Always. Statements are available in the appendix. Space was also provided for additional comments; although, those comments are not included in this analysis. Instead, this study focuses on the pairs of positively and negatively coded items, five sets in total, which were constructed to serve as 'honesty checks' in the evaluation. The items included in the study are presented below in Table 1. The negatively coded items are indicated with a 'N'.

Table 1. Evaluation Items Utilized in the Study

Pair 110: Grading was unfair or biased.^N

24: Grading was fair and impartial.

Pair 219: The professor presented material in a disorganized, illogical, and unclear manner.^N

13: The professor presented the material in an organized, logical, and clear manner.

Pair 326: The professor was unprepared for class.^N

6: The professor was prepared for class.

Pair 428: The professor was unable to answer students' questions, or find ways to help students answer their own questions.^N

3: The professor was able to answer students' questions, or find ways to help students answer their own questions.

Pair 533: The overall objectives of the course were vague.^N

4: The overall objectives of the course were clearly presented.

Procedures

Prior to completing the evaluation, students were given assurance their identity would not be revealed in regards to their responses and were informed that participation was voluntary. The collected data were entered into Excel and analyses were conducted using Minitab and SPSS. No response was found to be out of range, meaning all data supplied were credible. Since students were informed that they were not required to respond to all items and it was reasonable to believe that a student may not have had an opinion on every item, missing data were treated as missing. Thus, there was not a need to impute means or other values. Valid responses were coded as 1, 2, 3 or 4, as described above. Negatively written items were reverse coded so that a 1 = 4, 2 = 3, 3 = 2, and 4 = 1.

In this study, the scale was thought of as a measurement continuum, largely due to the institution's approach. Cardinal numbers were used to define the scale; thus, variables are measured at an interval level, meaning that the difference between a score of 1 and 2 is assumed to equal the difference between a score of 3 and 4. Even though it is argued that the measurement of affective variables is a combination of ordinal and interval scales, it is commonly accepted to use interval-scale statistics due to the concept sociological variables are fairly crude predictors (Aiken, 1996; Boser, Palmer, & Daugherty, 1998; Lamon, 1997; Summers, 1977).

The analyses began with a descriptive summary of the 10 items, which were the five pairs of positively and negatively worded statements. Means, standard deviations and number of responses per item were computed. As well, paired differences were produced by subtracting the rating of the positive item from the reverse coded, negative item. Theoretically, the institution would assume this difference to be zero if the instrument is reliable and valid. Correlation estimates were then calculated for each pair of items. These items were assumed to be direct opposites of one another, which in theory would result in a perfect correlation. Given the rank order nature of the data, it was decided that Spearman's Rho would be the best correlation estimate. Next, paired t-test and confidence intervals were conducted. Alpha was set at 0.05; however, due to the family wise error, a Bonferroni correction was applied. Because there were five tests, the final p-value is compared against $\alpha_{\text{corrected}} = \alpha/\# \text{ of tests} = .05/5 = .01$. Frequencies for the paired differences are also presented. When significant differences were found, a graphical investigation of response patterns followed.

RESULTS/DISCUSSION

The analyses began with a descriptive overview of the items. The rating scale was a 4-point scale, where ratings 1, 2 were viewed as a negative endorsement and ratings 3, 4 were viewed as a positive endorsement. For that reason, the calculated mid-point of the scale, 2.5, is treated as the change of direction point, between a positive and negative endorsement. Thus, the perfect evaluation would have an average score of 4 with a standard deviation of 0. While in theory this is an achievable rating, for all practical purposes, a perfect rating is impossible.

Table 2 illustrates the means, standard deviations and paired differences for the five sets of positive and negative paired items. At first glance, all items indicate a positive rating, consistently above the midpoint, leaning toward the highest end of the rating scale. Considering the comparison of positively and negatively worded items, the means are not consistently higher for either type of presentation. Three means are higher for the positively worded items and two are higher for the negatively worded items, presented in the difference in means column. There is, with only one exception, more consistency in responses for the positively worded items, as indicated by smaller standard deviations. Generally speaking, at the descriptive level, it does not appear that these pairs of items are indicating a reliable and valid ‘honesty check’. The means are not equal. It does appear that students are reading the items and not flat lining, as indicated by the inconsistent pattern of mean differences.

Table 2. Means, Standard Deviations and Mean Differences for Paired Items

N	Negatively worded item (Reverse Coded)			Positively worded item			Δ M
	Item	M	SD	Item	M	SD	
528	10R	3.826	0.631	24	3.716	0.746	0.110
531	19R	3.847	0.577	13	3.815	0.447	0.032
530	26R	3.908	0.473	6	3.938	0.285	-0.030
532	28R	3.716	0.819	3	3.791	0.475	-0.075
527	33R	3.691	0.798	4	3.816	0.480	-0.125

If the items were exact opposites, once reverse coded, the positively and negatively worded items would directly correlate, meaning a score of 1. As presented in Table 3, we can see the correlations are at best, weakly correlated. This may not be a clear indication of the pattern responses, as the lower correlation estimates may be attributed to the lack of variability in responses. More specifically, the clustering of data at certain points makes it statistically impractical to present any type of linear relationship. Spearman Rho values range from 0.22 to 0.34, with the largest estimate occurring for items 33 and 4, related to the presentation of course objectives.

Table 3. Spearman’s Rho (correlation estimates) for Paired Items

	24	13	6	3	4
10R	0.29				
19R		0.22			
26R			0.25		
28R				0.26	
33R					0.34

After a Bonferroni correction, a p-value of less than .01 is said to be significant. When viewing Table 4, there are two significant differences out of the five comparisons at this level. Differences exist between the pairs of:

- * Item 10 (reverse coded) and Item 24
- * Item 33 (reverse coded) and Item 4

Item 10 and Item 24 reflect perceptions of grading, and as discussed above Item 33 and Item 4 reflect perceptions related to presentation of course objectives. The significant differences again indicate that lack of a stable ‘honesty check’ as the differences are not in a consistent direction. In the first case of 10 versus 24, the negatively worded item is endorsed at a higher rating; while in the second case of 33 versus 4, the positively worded item is endorsed at a higher rating. Again, this offers support for students not ‘flat lining’, meaning that students just respond positively or negatively to the items in general. It also suggests that students do not interpret the positively and negatively worded items exactly the same.

Table 4. Paired t-test and Confidence Interval for Paired Items

Pair	N	Confidence Interval	p-value
10R – 24	528	(0.032, 0.187)	0.005
19R – 13	531	(-0.025, 0.089)	0.274
26R – 6	530	(-0.071, 0.010)	0.144
28R – 3	532	(-0.150, -0.000)	0.050
33R – 4	527	(-0.195, -0.055)	0.000

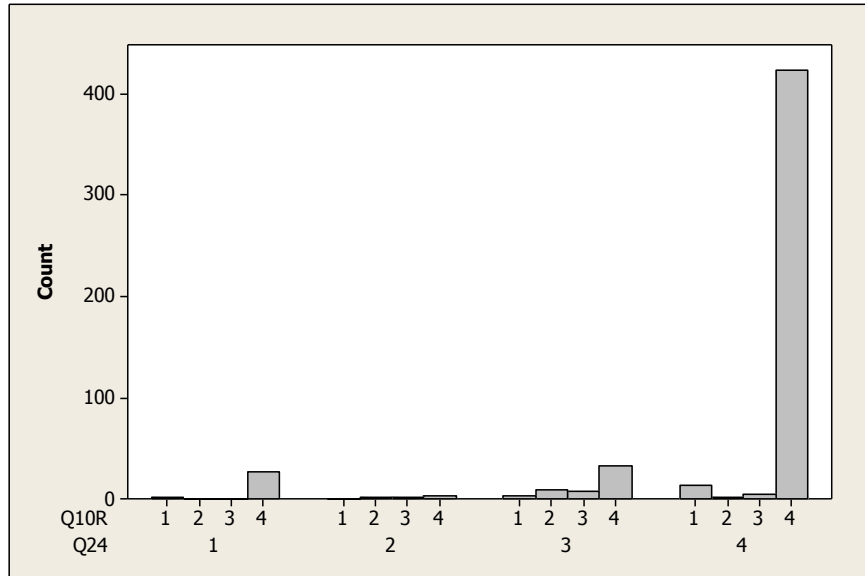
To further investigate the paired differences, a frequency table was constructed for the paired differences. The table represents the positively worded item subtracted from the negatively worded item (reverse coded). So the lowest possible score would be $1 - 4 = -3$ and the highest possible score would be $4 - 1 = 3$. These scores would also indicate the least amount of consistency between responses. The expectation would be a score of 0, indicating consistent responses on both items, i.e. $4 - 4$, $3 - 3$, $2 - 2$ and $1 - 1$. As seen in the table, the score of 0 has the largest percentages for all items. However, no pair received a 100% agreement; instead it lingered around 80%. Similar discrepancies, as discussed previously, are also viewed. It is interesting to see, with only the exception of the paired items 10 and 24, that there is a noticeable percentage of scores at -3 and -2 as compared to 3 and 2. On the other hand, there is a higher likelihood of a score of 1 than -1 in all cases.

Table 5. Frequencies for the Paired Differences

	10R – 24		19R – 13		26R – 6		28R – 3		33R – 4	
-3	13	2.5%	14	2.6%	8	1.5%	30	5.6%	25	4.7%
-2	5	1.0%	3	0.6%	5	0.9%	8	1.5%	12	2.3%
-1	13	2.5%	14	2.6%	4	0.8%	13	2.4%	24	4.6%
0	434	82.2%	426	80.2%	492	92.8%	410	77.1%	413	78.4%
1	33	6.3%	69	13.0%	20	3.8%	66	12.4%	49	9.3%
2	3	0.6%	5	0.9%	1	0.2%	2	0.4%	4	0.8%
3	27	5.1%	0	0.0%	0	0.0%	3	0.6%	0	0.0%
Total	528		531		530		532		527	

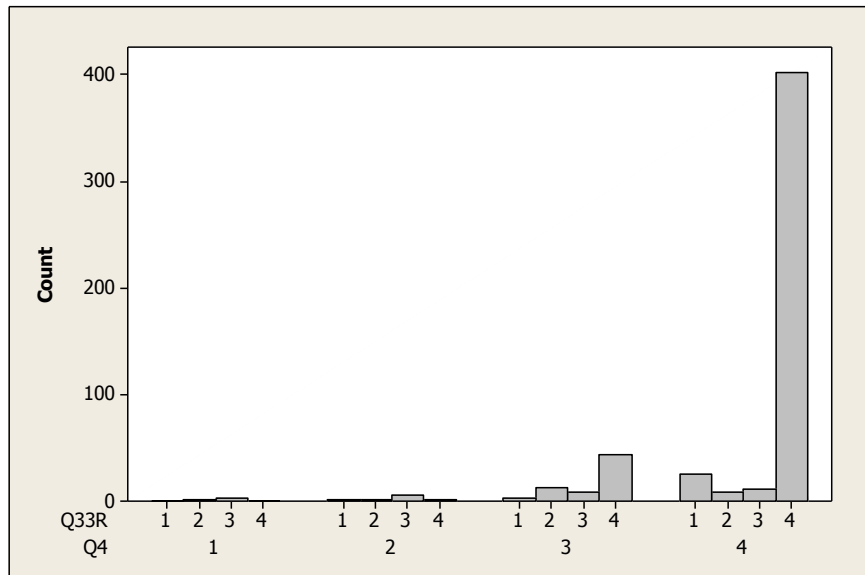
For a further review of the significant paired differences, graphs of response patterns were produced. In Figure 1, a comparison of Item 10 and Item 24 is presented. If there was a direct response pattern, the most frequent outcome would be 1 for 1, 2 for 2, and so on. Instead, for each response category of item 24, the most frequent outcome for 10R is 4.

Figure 1. Item 10 Reverse Coded Responses by Item 24 Response Patterns



The comparison of Item 33 and 4 also present an atypical response patterns, as compared to the theoretical assumption of 1-1 relationship. For item 4, rating 1 and 2 indicate the most frequent response for 33R is 3 and for ratings 3 and 4 it is 4. Again, it should be consistent across the board and it is not, providing further evidence that the ‘honesty checks’ are not functioning as intended.

Figure 2. Item 33 Reverse Coded Responses by Item 4 Response Patterns



CONCLUSION

Given the importance that many higher education institutions are giving to instructor evaluation results, it becomes of critical importance to have reliable and valid practices associated with the data collection process. When dealing with instructor and/or course evaluation data (as seen here and in most surveys), it is presumed the respondents have an accurate perception of the construct, rate items according to reproducible criteria, and accurately record their ratings within uniformly spaced levels. Wright (1997) would argue that ratings are simply responses based on fluctuating personal criteria. Even more, the responses are not always interpreted as intended. These issues seem to surface in this investigation.

The literature indicates that individuals respond differently to positively and negatively worded items. Here, it is clear that there is not a 1 to 1 relationship between the positively and negatively worded items. This is a real issue considering institutions are building this type of item comparisons into evaluations to serve as 'honesty checks'. In fact, these paired 'opposite' items may just be introducing noise, better known as measurement error.

Data-driven decision making is only reliable and valid when an evaluation plan is properly designed and carried out, with a careful consideration of measurement properties and the subsequent application. Here, a contention is made that while 'honesty' is important to capture, the best approach may not be reverse code items – i.e., trying to trick the respondent. It could be that items phrased in the positive and negative are just innately different, leading to inconsistent responses. Students are asked to use an increasing scale to rate the items, one which indicates the higher score the better the performance. Trying to capture valid responses through a tricky measure might be best explained as a systematic bias, leading to measurement error.

EDUCATIONAL IMPORTANCE

The survey research community and institutions analyzing similar rating scale data will benefit from the results of this study as it provides a data-driven example that the 'honesty checks' do not typically meet the assumptions proposed. The careful construction and proper analyses of evaluation instruments is often overlooked or underemphasized. A contention is made that the way the question is asked does matter, at least in some instances, and it should not be assumed that an item written in the positive and negative should directly correlate. The survey research community and institutions utilizing similar rating scale instruments will benefit from the results of this study, as well as the education community in general.

REFERENCES

1. Barnette, J. J. (1997). Effects of item and response set reversals on survey statistics. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL. March 24-28.
2. Bergstrom, B. A., & Lunz, M. E. (1998). Rating scale analysis: Gauging the impact of positively and negatively worded items. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA. April 13-17.
3. Bishop, G. F., Oldendick, R. W., & Tuchfarber, A. J. (1982). Effects of presenting one versus two sides of an issue in survey questions. *The Public Opinion Quarterly*, 46, 1, 69-85.
4. Dillman, D. A., & Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods*, 17, 1, 30-52.
5. Falzthzik, A. M., & Jolson, M. A. (1974). Statement polarity in attitude studies. *Journal of Marketing Research*, 11, 1, 102-105.
6. Garg, R. K. (1996). The influence of positive and negative wording and issue involvement on responses to Likert scales in marketing research. *Journal of the Market Research Society*, 38(3), 235-246.
7. Hazlett-Stevens, H., Ullman, J. B., & Craske, M. G. (2004). Factor structure of the Penn State Worry Questionnaire: Examination of a method factor. *Assessment*, 11(4), 361-370.
8. Krosnick, J. A. (1999). Maximizing questionnaire quality. In the book: *Measures of political attitudes*. John P. Robinson, Phillip R. Shaver, Lawrence S. Wrightsman (Eds), pp. 37-57). San Diego, CA. Academic Press, Inc.

9. Leenaars, A. A., Bringmann, W. G., & Balance, W. D. G. (1978). The effects of positive vs. negative wording on subjects' validity ratings of "true" and "false" feedback statements. *Journal of Clinical Psychology, 34*, 2, 369-370.
10. Mook, J., Kleijn, W. C., & Van der Ploeg, H. M. (1991). Symptom-positively and negatively worded items in two popular self-report inventories of anxiety and depression. *Psychological Reports, 69*(2), 551-560.
11. Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60*, 58-88.
12. Noelle-Neumann, E. (1970). Wanted: Rules for wording structured questionnaires. *The Public Opinion Quarterly, 34*(2), 191-201.
13. Ory, J. C. (1982). Item placement and wording effects on overall ratings. *Educational and Psychological Measurement, 42*(3), 767-775.
14. Ory, J. C., & Valois, R. F. (1980). The influence of negatively worded scale items on overall ratings. (ERIC Document Reproduction Service No. ED 199 293).
15. Reiser, M., Wallace, M., & Schuessler, K. (1986). Direction-of-wording effects in dichotomous social life feeling items. *Sociological Methodology, 16*, 1-25.
16. Schriesheim, C. A., & Eisenbach, R. J. (1995). An exploratory and confirmatory factor-analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *Journal of Management, 21*(6), 1177-1193.
17. Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement, 41*(4), 1101-1114.
18. Schuman, H., & Presser, S. (1977). The open and closed question. *American Sociological Review, 44*(5), 692-712.
19. Zaller, J., & Felman, S. (1992). A simple theory of the survey response: Answering questions versus revealing preferences. *American Journal of Political Science, 36*(3), 579-616.

APPENDIX

Student Evaluation of Faculty

Students were asked to use the following rating scale:

4 = Always / Almost Always

3 = Usually

2 = Some of the Time

1 = Never / Almost Never

If a statement does not apply, please leave it blank.

Complete List Evaluation Items

Item Statement

1. The professor demonstrated knowledge of the subject matter.
2. The professor established a connection between course work, other subjects and/or practical applications in the work world.
3. The professor was able to answer students' questions, or find ways to help students answer their own questions.
4. The overall objectives of the course were clearly presented.
5. The professor covered the material and objectives that were stated in the course syllabus.
6. The professor was prepared for class.
7. The professor was enthusiastic in teaching.
8. The professor made clear and appropriate use of the board/AV materials/handouts, etc.
9. The professor made good use of class time by teaching material relevant to course objectives.
10. Grading was unfair or biased.
11. The professor spoke clearly.
12. The professor clearly presented the lesson objective(s).
13. The professor presented the material in an organized, logical, and clear manner.
14. The professor was open to constructive feedback from students.
15. The professor used examples to reinforce understanding of the material.
16. The professor used instructional techniques that helped me learn.
17. The professor used examples to reinforce understanding of the material.
18. Course assignments reinforced the concepts presented in class.
19. The professor presented material in a disorganized, illogical, and unclear manner.
20. The grading system was clearly explained.
21. In this class, I believe students felt free to ask questions and express ideas.
22. The exams were clearly worded.
23. Student work was returned as announced.
24. Grading was fair and impartial.
25. The professor was available for appointments.
26. The professor was unprepared for class.
27. The professor demonstrated professionalism.
28. The professor was unable to answer students' questions, or find ways to help students answer their own questions.
29. The professor established an atmosphere of mutual respect.
30. The professor started and ended class on time.
31. Dates of major exams were conveyed in advance.
32. Course objectives were reflected in the exams.
33. The overall objectives of the course were vague.
34. I would recommend this course to a friend.
35. I would recommend this professor to a friend.