Reducing the Workload in Your Class Won't "Buy" You Better Teaching Evaluation Scores: Re-Refutation of a Persistent Myth

Kay C Dee

Department of Biomedical Engineering, Tulane University, New Orleans, LA 70118

Abstract

Although much of the educational literature characterizes the relationship between course workload and teaching evaluation scores as small (with higher workload/difficulty courses rated slightly more favorably), many faculty believe and some reports claim that students "reward" instructors of low-workload courses with good teaching evaluation scores. This study therefore examined whether engineering students' perceptions of course workload were related to perceived instructor performance. Course-averaged student evaluations of teaching for each class offered through the Tulane University School of Engineering from Fall 1997 to Fall 2002 (878 courses) were collected. Evaluations contained seventeen items rated on five-point scales similar to Likert scales, including an item regarding the instructor's overall performance and an item regarding the amount of work required for the course. Pearson's and Spearman's correlation coefficients and two-tailed significance levels were calculated for evaluation items, the number of respondents, and the general course level. Subsequent model adequacy checking revealed that the course evaluation data violated major assumptions inherent in the use of parametric statistical methods (*e.g.*, Pearson's correlations, t tests). Therefore, statistical outliers were removed, data transformations (natural log, standardized scores) equalized variance and normalized the distributions of scores within items, and correlation coefficients were re-calculated and compared for each data transformation.

Over all analyses performed, the largest correlation between evaluation items regarding the amount of coursework and instructor performance was only 0.15, such that courses requiring more work received poorer evaluations. However, curve fitting revealed essentially no easily-generalized relationship between these two items (R^2 = 0.01 and 0.04 for linear and quadratic curve fits). In contrast, scores on other items, such as "the instructor gave organized lectures," were highly correlated (coefficients between 0.80 and 0.92) with overall instructor performance scores. A number of strong inter-item correlations suggested and factor analysis confirmed that the evaluation form used at Tulane fundamentally assessed two distinct factors, apparently "instructor performance" and "amount of work," which accounted for 70% of the variance across items. Although correlation does not imply causation, this study re-confirms that engineering faculty seeking improved teaching evaluations should focus on improving instructional practices associated with content organization and delivery, or with instructor-student interactions, instead of worrying about the potential effects of perceived course workloads.

Introduction

Student evaluations of teaching and various factors believed to affect how students rate instructor performance are the subjects of numerous literature reports[1-6]. Many faculty believe that high levels of course workload and/or intellectual rigor may cause students to give poor overall course/instructor evaluations; in one study[7], 54% of faculty (but only 26% of students) agreed or strongly agreed with the statement "To get favorable evaluations, professors demand less from students." A number of literature reports characterize the relationship between perceived course workload and student evaluations of instructor quality to be small[2, 3, 5, 6]. Unfortunately engineering faculty tend to discount such reports, especially when the evaluations considered were from non-engineering courses, since engineering faculty (and students) often view engineering as a uniquely demanding division of academic culture. The overall goal of this study was therefore to determine whether reducing course workload would be likely to "buy" an engineering professor better evaluation scores, by searching for relationships between perceived course workload and student evaluations of teaching for recent engineering courses.

Methods

Evaluation Form and Data Collection

Teaching evaluation scores for all classes offered through the Tulane School of Engineering from the Fall of 1997 to the Fall of 2002 were collected and used for this study. The evaluation form consisted of a 17-item "bubble sheet" questionnaire followed by space for written comments (Appendix, Figures A1 and A2). Items one through fourteen of the questionnaire asked students to use a scale ranging from 1.0 ("strongly agree") to 5.0 ("strongly disagree") to indicate their level of agreement with a given statement (essentially an inverted Likert scale); items fifteen through seventeen of the form used slightly different five-point scales. The evaluation form did not change between Fall 1997 and Fall 2002, and was the only formal, institutionalized teaching evaluation conducted in the School of Engineering during this time period. Completed bubble sheet forms were optically scanned. Course-averaged scores for items one through seventeen of the questionnaire were computed and the resulting seventeen scores for all courses (listed by course number) were distributed by the Dean's office to all faculty in the School of Engineering, each semester.

Initial Analysis

Course-averaged scores for questionnaire items one through seventeen were entered into SPSS (SPSS Inc., Chicago, IL), along with the number of respondents and the level (100, 200, 300, *etc.*) of each course (878 courses). Course-averaged item scores less than 1.0, which occurred when only a portion of the total respondents answered a particular evaluation question, were omitted. Course evaluations which had been completed by fewer than five respondents were removed, reducing the data to 823 sets of teaching evaluation scores. Pearson's and Spearman's correlation coefficients and two-tailed significance levels (with missing values excluded pairwise) were calculated, using SPSS, for evaluation items, the number of respondents, and the general course level.

Model Adequacy Checking

Use of the Pearson correlation (a parametric method) implies four major assumptions: 1) equal variances within the groups of data to be correlated[8]; 2) normal distributions of data within the variables to be correlated[9]; 3) a linear relationship between the variables to be correlated[9], and 4) homoscedasticity, or approximately the same spread of data about a best-fit line at all levels of the variables[10]. Additionally, statistical outliers (which could disproportionately alter correlations and curve fits) should be identified and removed from data prior to analysis. Cochran's and Hartley's statistics[11] were used to test for homogeneity of variance. To determine whether the data followed a normal distribution, normal probability plots and histograms were created, and numerical values of skewness and kurtosis were checked. Scatterplots were created to visually check for homoscedasticity and potentially linear relationships between variables; linear regressions were performed to confirm linear relationships. Boxplots were created to check for the presence of statistical outliers within the data.

Model Revisions and Re-Analysis

Outliers, defined here as data points which fell outside the range of ± three standard deviations away from the mean score for a given item, were detected in the data for questions 1, 5, 7, 8, 10, 11, and 17 on the evaluation form. After confirming data entry accuracy, these outlier data points were flagged for further examination and removed from the data set. Course evaluations with more than 54 respondents were outliers in regards to the number of respondents, and were therefore removed from the data set. The reduced data set was then subjected to a variety of transformations with the hope of yielding more normally-distributed, less skewed data distributions. Ultimately, a natural log transformation was used ($y' = ln(y)$, where y = original data value and y' = transformed data value) and the resulting data were standardized into z scores (Figure A2). Pearson's and Spearman's correlations were again computed for the transformed/standardized data (with missing values excluded pairwise). Mean evaluation scores from courses which received at least one outlier evaluation value were compared to the mean scores from all courses, using Wilcoxon signed ranks tests. Finally, factor analysis[12] for data reduction was performed using principal component analysis (with Varimax rotation[13]).

Results

Initial Analysis

Selected characteristics of the original data set are presented in Table A1. Correlation coefficients between teaching evaluation items calculated during this initial analysis are not presented here due to space considerations. Scores from a few evaluation items were nearly linearly related to scores from the "instructor's performance" item (Figure 1). However, scores from the "amount of work" item were neither linearly (Figure 1) nor quadratically well-related to the "instructor's performance" scores.

Model Adequacy Checking

The data passed neither Hartley's nor Cochran's tests for homogeneity of variance. Normal probability plots and histograms showed that the data deviated from normal

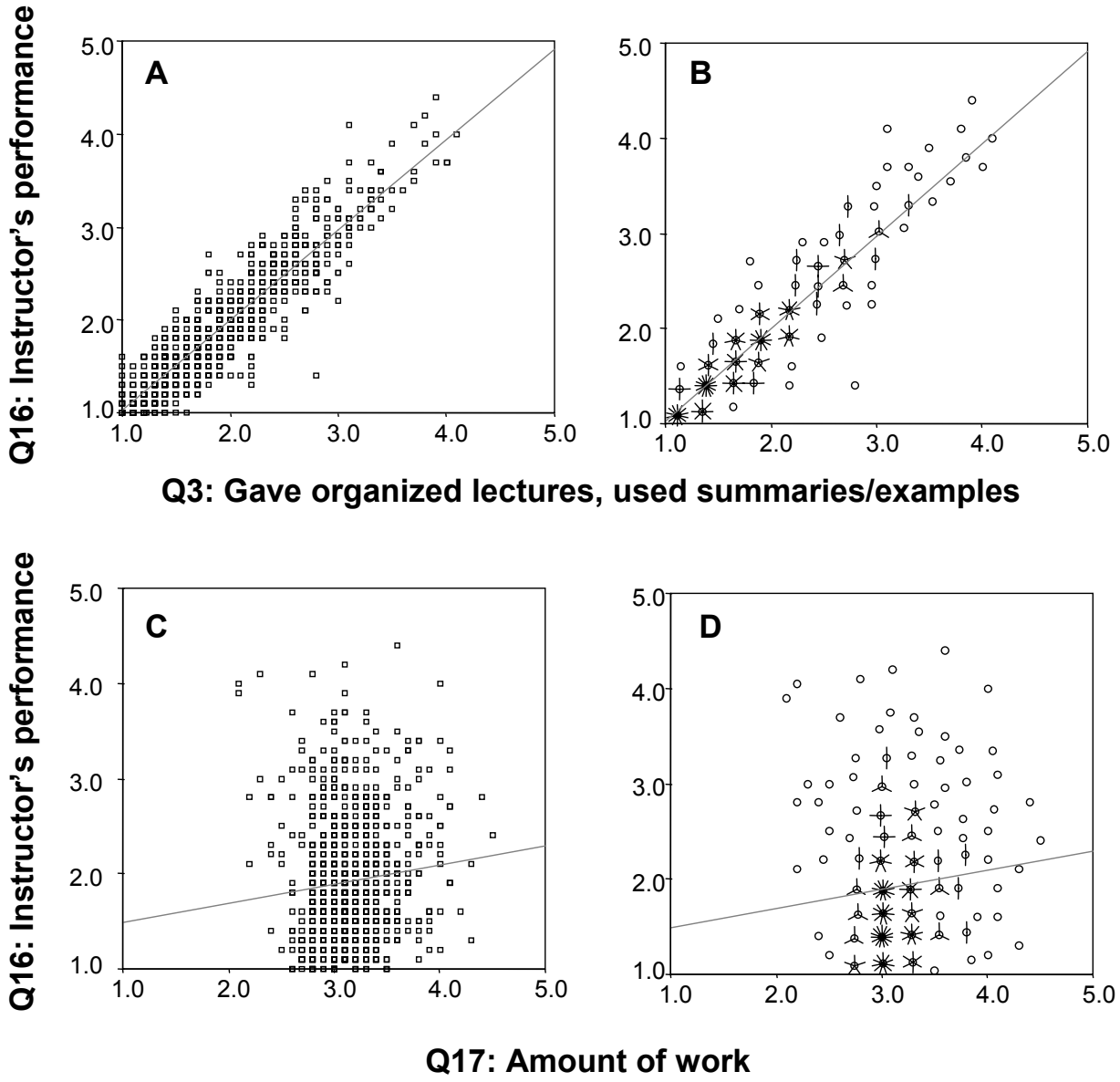


Figure 1. Example Scatterplots. Examples of a fairly linear relationship (Q16 plotted versus Q3; $R^2 = 0.842$) are given in **A** and **B**. Examples of a nearly non-existent linear relationship (Q16 plotted versus Q17; $R^2 = 0.010$) are given in **C** and **D**. Frames **A** and **C** are standard scatterplots; frames **B** and **D** display "sunflowers," where each 'petal' on a point represents five cases.

distributions. As can be seen in Figure 2, data were not spread over the full range of scores, but instead were skewed toward the minimum possible value of 1.0 (the best possible score). Scatterplots revealed that the assumption of homoscedasticity was probably acceptable for some instances (*e.g.,* Figure 1, frame B) but much less appropriate for other instances (*e.g.*, Figure 1, frame D). Scatterplots indicated and linear regressions confirmed that while some teaching evaluation items were linearly associated, other items weren't. Finally, the boxplot shown in Figure 2 indicated the potential presence of statistical outliers within the data set. The data were therefore transformed and re-analyzed.

Model Revisions and Re-Analysis

After removal of outliers, transformations were applied to the data set (Figure A2). It was difficult to tell from visual displays which data transformation produced the best results. The natural log transform generally reduced skew and inequalities in variance, but generally increased kurtosis. Standardizing the data by calculating z scores equalized the variances but did not modify skew or kurtosis. Since none of the transforms yielded a data set which clearly met all of the statistical method assumptions, correlation coefficients were re-calculated (Table A2) for all transformed data sets. In general, many of the weak original correlations were slightly stronger upon re-calculation, and many of the strong original correlations were slightly weaker upon recalculation. Re-calculated correlations differed from original correlations by up to as much as 0.08. Re-calculated Pearson's correlations were similar to the re-calculated Spearman's correlations, usually differing only by 0.01 to 0.02. Table 1 highlights the highest and lowest inter-item correlations identified in both the original and the transformed data sets.

The scale on the teaching evaluations was ordered such that the minimum score of 1.0 was the best rating possible, and the maximum score of 5.0 was the worst rating possible. Negative correlations between course level and teaching evaluation items were observed, implying that lower-level courses tended to receive higher (and thus poorer) scores on evaluations. The small magnitudes of these negative correlations, however, reduce the importance of these results. Very high (>0.90) correlations were observed between questions regarding whether "the instructor tries to find out if the material is being understood" and "the instructor is genuinely interested in teaching and student progress", and between questions regarding whether "the instructor gave organized lectures, summarized major points and used enough examples to clarify points" and overall instructor performance expressed as "The instructor's performance is:" (scoring options for this item ranged from 1.0, excellent, to 5.0, poor). The three evaluation items which tended not to correlate strongly with other items were question 1, "The instructor was present and on time for classes during the semester,"; question 15, "The text is:" (scoring options for this item ranged from 1.0, excellent, to 5.0, poor); and question 17, "The amount of work was:" (scoring options for this item ranged from 1.0, definitely too little, to 5.0, definitely too much).

The highest correlation obtained between questions 16 (instructor's performance) and 17 (amount of work) was only 0.15. The data transformations had essentially no effect on the variance accounted for by linear ($R^2 = 0.017$) or quadratic ($R^2 = 0.040$ to 0.043) curve fits after plotting instructor's performance scores versus amount of work scores.

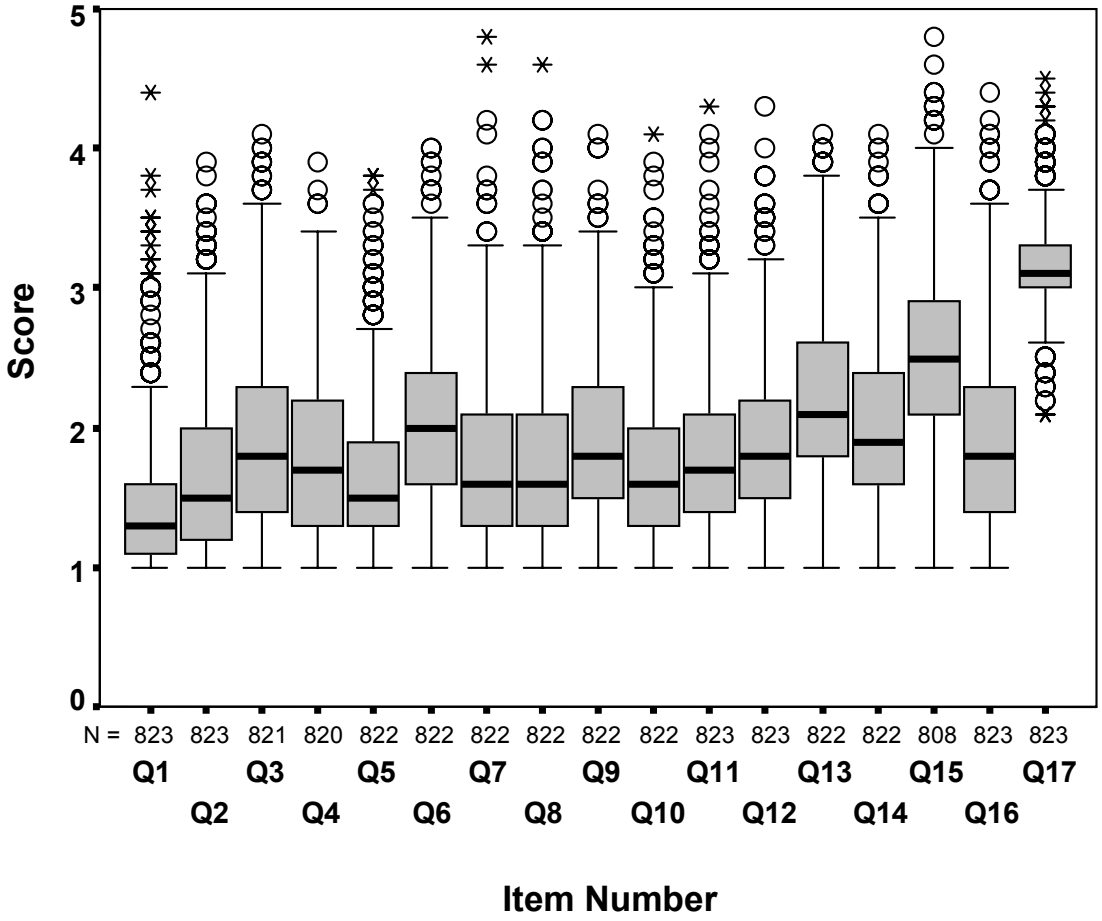


Figure 2. Boxplots of Teaching Evaluation Scores. Open circles indicate outliers, defined for this figure as values within 1.5 and 3 interquartile ranges from the edges of the box. Asterisks indicate extreme outliers, defined as values more than 3 interquartile ranges from the edges of the box. Although the minimum score possible was 1.0, the vertical axis starts at 0.0 for clarity of view. Recall that the box encloses the interquartile range (essentially the range of the middle 50% of the data); the line across the box indicates the median; the "whiskers" extend from the box to the data points closest to but still within a limit of 1.5 interquartile ranges away from the edges of the box (*i.e.,* the first or third quartiles).

Table 1. Visual Summary: Highest and Lowest Correlation Coefficients Between Teaching Evaluation Items, Regardless of Data Manipulation. *Table continues onto next page.*

	Level	Responses	Q1 Present, on-time	Q2 Well-prepared, Didn't waste time	Q3 Organized lectures, summaries and examples	Q4 Effective delivery	Q5 Clarified objectives, announced assignments in advance	Q6 Stimulated interest in the field	Q7 Answered questions carefully / patiently
Level									
Responses									
Q1									
Q2									
Q3									
Q4									
Q5									
Q6									
Q7									
Q8									
Q9									
Q10									
Q11									
Q12									
Q13									
Q14									
Q15									
Q16									
Q17									

Legend:

- $r \geq 0.80$
- $r \leq 0.20$
- $0.21 \leq r \leq 0.79$

Table 1, Continued.

	Q8 Helpful outside the classroom	**Q9** Tried to find out if material was understood	**Q10** Interested in teaching and students	**Q11** Evaluated performance accurately / fairly	**Q12** The course increased knowledge, competence	**Q13** Provided frequent feedback	**Q14** The homework was carefully planned / appropriate	**Q15** The text was:	**Q16** The instructor's peformance was:	**Q17** The amount of work was:
Level										
Responses										
Q1										
Q2										
Q3										
Q4										
Q5										
Q6										
Q7										
Q8										
Q9										
Q10										
Q11										
Q12										
Q13										
Q14										
Q15										
Q16										
Q17										

Legend:

	$r \geq 0.80$
	$r \leq 0.20$
	$0.21 \leq r \leq 0.79$

Dark red cells identify locations where Pearson's and/or Spearman's correlation coefficients were calculated to be greater than or equal to 0.80 across all data analyses. Light blue cells identify locations where the magnitude (absolute value) of either Pearson's or Spearman's correlation coefficents were calculated to be equal to or less than 0.20 across all data analyses. "Level" refers to the level of a course (100 or freshman-level, 200 or sophomore level, *etc.*, up to 700 or graduate-level courses). "Responses" refers to the number of individual student ratings that were used to calculate the course-average rating. All the Q1- Q17 items refer to instructor practices/characteristics except where specifically noted otherwise (*e.g.*, "The course increased knowledge / competence").

Outliers

Although outlying data points were removed prior to re-analysis, data from ostensibly extreme courses (*i.e.,* requiring the absolute most or least work) were relevant to the overall goals of this study. Question 1 on the evaluation ("The instructor was present and on-time for classes during the semester") generated the largest number of outlier (> 3 standard deviations away from the mean) values, and data from question 1 were not highly correlated with data from many other evaluation items. Figure 3 shows that compared to scores for all courses, courses which received outlier (higher, and therefore poorer) scores on question 1 also received significantly ($p < 0.05$) poorer scores on questions 2, 3, 13, and 16; this implies a link between instructors not being present/on time for classes and: wasting class time or not being well-prepared, not giving organized lectures/examples, not providing frequent feedback on student performance, and not receiving a good overall performance rating. Figure 3 also shows that compared to scores for all courses, courses which received outlier (higher, and therefore poorer) scores on questions other than one also received significantly ($p < 0.01$) poorer scores on all evaluation items except question seventeen (which assessed students' perceptions of the amount of work required for the course). In other words, although outlier, extremely poor ratings of specific instructor characteristics were associated with poor ratings of other instructor/course characteristics, these poorly-reviewed courses were *not* perceived by students as having significantly higher workloads.

Instrument Analysis

The number of high inter-item correlations calculated during the completion of this project sparked a question: how many distinct underlying factors did the teaching evaluation form actually assess? Factor analysis for data reduction revealed that only two components were needed to explain 70% of the total variance across all evaluation items. The first component appeared to represent teaching/course practices (*i.e.*, questions 16, 3, 12, 14, 8 and 4 loaded with this factor); the second component, with an eigenvalue nearly ten times smaller than the first component, appeared to represent the amount of work required in the course (*i.e.*, question 17).

Discussion

Analytical Methods

Parametric statistical methods (*e.g.,* analyses of variance, F and t tests, Pearson's correlations, *etc.*) are commonly-used and powerful, but can yield misleading results[8] when the populations under study do not meet the assumptions inherent in such methods. Nonparametric methods, such as Spearman's rho correlation, reach conclusions based on the relative ranks of the observed data rather than on the observed values of the data. Seeking to be as rigorous as possible, this study used a logarithmic transform to de-skew data, and a z score transform to force equality of variances, and compared results from both parametric and nonparametric methods. Although in this case no data transform clearly emerged as the most appropriate, the fact that assumptions were checked and alternate data transforms considered should lend credence to the overall results of the present study. The variation in the coefficients calculated in this study is, perhaps, a good reminder to not focus intently on specific, calculated numerical correlation values but instead to look for general ranges or patterns (as displayed in Table 1).

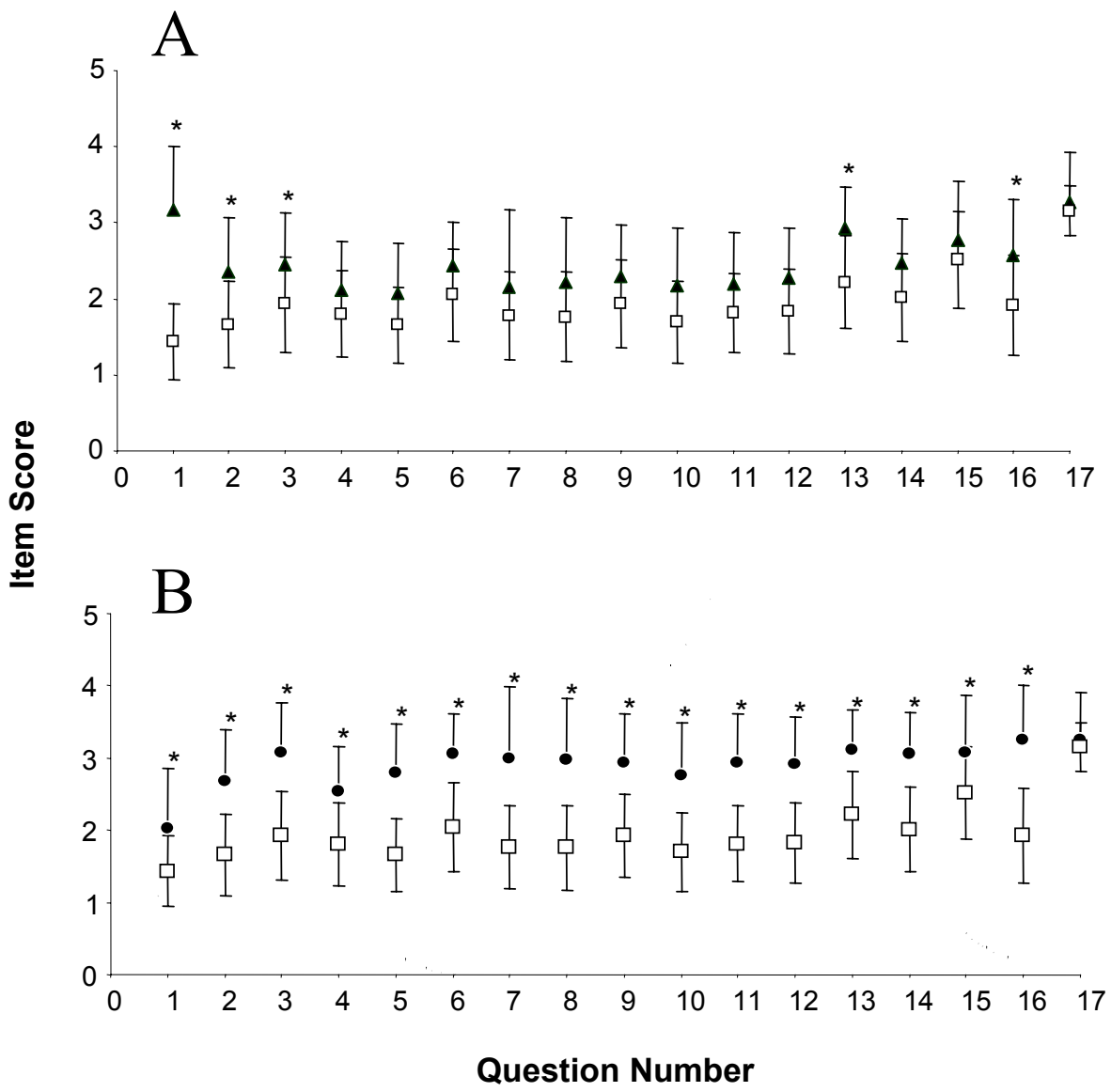


Figure 3. Teaching Evaluation Scores for Courses with Outlying Values, Compared to Scores for All Courses. Data shown are means ± standard deviations (S.D.). Open squares in frames A and B: the overall data summarized in Table A1.

A) ***Courses with outlying values for question one.*** Filled triangles: courses with data ≥ 3 S.D. larger than the overall mean for Q1; only positive vertical error bars plotted for visual clarity; n = 14. *$p < 0.05$ (Wilcoxon signed ranks test).

B) ***Courses with outlying values for questions other than one.*** Filled circles: courses with data ≥ 3 S.D. away from the overall mean for Q5, Q7, Q8, Q10, Q11, or Q17; only positive vertical error bars plotted for visual clarity; n = 34. *$p < 0.01$ (Wilcoxon signed ranks test).

Inter-item Correlations

Some teaching evaluation items were highly, and seemingly logically, correlated with each other. For instance, it appeared that instructors who tried to find out if material was being understood tended to be perceived as interested in teaching and students. Other high correlations indicated that favorable ratings of the instructor's performance were associated with: being well-prepared and not wasting time; effectively delivering organized lectures, summaries, and examples; answering questions carefully and patiently; trying to find out if material was being understood; appearing interested in teaching and students. Favorable ratings of instructor performance were also associated with courses that increased interest in the field and which students perceived as increasing their knowledge and competence.

Workload and Instructor Performance

The largest correlation calculated in this project between the evaluation item regarding the instructor's performance and the item regarding the amount of work required for the course was only 0.15. The scoring options for the instructor performance ranged from 1.0 (excellent) to 5.0 (poor), and the scoring options for the amount of work required ranged from 1.0 (definitely too little) to 5.0 (definitely too much). Therefore the non-negative correlation in ***this study implies a potential (but small) association between courses which required more work and poorer reviews of instructor performance in those courses.*** In contrast, Marsh[5, 6] has reported the correlation of higher levels of course workload/difficulty with higher student ratings of instructional quality, citing, for example, a correlation coefficient of 0.16 between student ratings of course workload/difficulty and of the instructor overall[6]. Similarly, Chau and Hocevar[16] calculated a mean correlation of 0.13 between the workload/difficulty factor and all other factors of the Students' Evaluations of Educational Quality (SEEQ) rating instrument. Analysis of the Individual Development & Educational Assessment (IDEA) system revealed low (including many near-zero) correlations between IDEA items thirty-one ("Amount of reading") or thirty-two ("Amount of work in other (non-reading) assignments") and items regarding the instructor's teaching procedures[14]; some of the largest correlations were reported between the item "Stimulated students to intellectual effort beyond that required by most courses" and the items regarding the amount of reading and amount of non-reading work (0.23 and 0.32, respectively)[14].

The small correlation between instructor performance and amount of work calculated in this project was similar in magnitude to the studies cited above, but opposite in direction. This may be due to the different student populations and/or courses used for analyses. This study used data only from engineering courses, while Marsh's investigations have utilized data from a variety of undergraduate course types, with samples ranging from 183[6] to many hundreds (*e.g.*, [15]) of evaluations. The SEEQ and IDEA analyses cited above were based on data from 6,322 classes and 104,237 classes, respectively[14, 16], again from a wide variety of course types (including, for example, social science, business, and engineering[16]). An alternative interpretation might be that the correlation is so small that for all practical purposes it (both direction and magnitude) is essentially meaningless – especially for predictive purposes.

Upon first glance, the correlations between workload and instructor performance calculated in the present study might appear to be very meaningful, since statistically significant ($p < 0.01$) results were obtained. But the meaning of the statistical test should be carefully

considered. Even after conservatively setting the probability of committing either a Type I or Type II statistical error, large sample sizes allow small effect sizes to be detected as statistically significant[17, 18]. For the present study, with n = 820 and setting a small probability of committing either a Type I or Type II error (alpha = beta = 0.01), Cohen's *d* measure of effect size is calculated[11] to be 0.16. This small[17] Cohen's *d* can be related to the data variances (Table A1) in the present study to estimate that a difference ($\mu - \mu_o$) in teaching evaluation scores of about 0.1 could be detected as statistically different. While a small effect may still be an important effect, in this case the approximate detectable difference (0.1) is the smallest division reported at the Tulane School of Engineering on the teaching evaluation scale of 1.0 to 5.0. ***The small effect size, combined with the low correlation coefficient between workload and instructor performance, reduces the ultimate real-world importance of the correlation.***

The "practical significance"[17] of the correlation between perceived course workload and instructor performance is further reduced after considering that linear and quadratic relationships poorly fit the data from these evaluation items. Plots of quadratic relationships between workload[19] (or aspects thereof[15]) and overall teacher ratings have been published, with standardized beta weights (essentially, regression coefficients when variables have been expressed in standardized form) reported rather than R^2 values, ranging from –0.17 to 0.06[15]. Qualitatively, the data from this study (with or without outliers included) seem similar in terms of scatter and heteroscedasticity to the main portions of the previously-published plots[15, 19], but lack the extreme values shown in the published plots. This is important to note, since extreme values can greatly influence the shape of a curve fit, and the published plots[15, 19] have been cited as evidence of "nonlinear" relationships between perceived workload and instructor effectiveness. ***This study therefore provides additional evidence of the lack of any linear relationship between perceived workload and instructor effectiveness, and demonstrates that quadratic relationships may also poorly fit these data.***

In other words, even though we could confidently detect a very small difference in teaching evaluation scores, we calculated a small correlation at best between students' perceptions of engineering course workload and instructor performance, and were unable to determine an even marginally-acceptable descriptive relationship (either linear or quadratic) between workload and instructor performance evaluation items.

There are literature reports that link lighter-workload classes with better teaching evaluations. For example, a study by Ryan *et al.*[20] summarized faculty-reported changes in instructional practices in response to student evaluations of teaching as "mainly reduced course work demands on students"[21]. Indeed, nearly 38% of the faculty respondents in the study by Ryan *et al.* reported they had 'greatly or somewhat decreased' the difficulty level of their instructional activities. However, the other most commonly-reported changes were 'greatly or somewhat increased' explicit specification of course objectives, provision of handouts and other course aids, and attention to organization of course content. These changes were interpreted by Ryan *et al.* to cause "reduced coursework demands" on students "indirectly by providing increased course content structure (such as specification of course objectives and attention to course organization) and supporting information that could facilitate learning (such as the provision of handouts and use of A/V aids)."[22]. An alternate and equally viable interpretation is that instructional practices (specifying objectives, facilitating learning) were improved.

Course workload figured prominently in Greenwald and Gillmore's relational models of student evaluations of courses, workload, and expected grades; the final models presented expected grade as a direct mediator and workload as an indirect mediator of course evaluations [23]. An extended and detailed critique by Marsh and Roche[19] of Greenwald and Gillmore's work concluded that the relationship between expected grade and course evaluations was eliminated by including perceived learning as a factor. Marsh and Roche also noted that workload not perceived by students to be associated with learning had a slightly negative effect on student evaluations of teaching[19]. In a later paper, Marsh[15] constructed models which separated course workload into "Good" (perceived by students to be valuable to their learning) and "Bad" (unnecessary, excessive, *etc.*) hours – as originally proposed but not published[15] by Gillmore and Greenwald – with Good workload positively and Bad workload negatively associated with overall teaching rating [15]. Note that the teaching evaluation data used in the present study did not discriminate between Good and Bad workloads. It is possible that the slight negative relationship observed in the present study between course workload and student evaluation of teaching could be due to the confounding of Good and Bad workloads, with students perceiving slightly more Bad than Good workload in engineering courses. Future work with revised and validated teaching evaluation forms could help to determine more detailed relationships between engineering student perceptions of course workload and overall instructor performance ratings.

Conclusions

The overall goal of this study was to search for relationships between student evaluations of teaching and perceived course workload in engineering courses, to determine whether reducing course workload would be likely to "buy" an engineering professor better evaluation scores. The largest correlation observed between the evaluation item regarding the instructor's overall performance and the item regarding the amount of work required was only 0.15. The small detectable effect size reduced the real-world importance of this correlation, and perceived workload and instructor performance scores were neither linearly nor quadratically related. ***Lower course workloads did not appear to "buy" engineering professors better evaluation scores.*** In contrast, other teaching evaluation items were highly (r = 0.80 – 0.92), linearly, and logically correlated with each other. Engineering faculty interested in improving their teaching evaluation scores may therefore want to consider being well-prepared/not wasting time; effectively delivering organized lectures, summaries, and examples; answering questions carefully and patiently; trying to find out if material is being understood; and appearing interested in teaching and students rather than focusing on course workload as a possible evaluation bias.

Acknowledgements

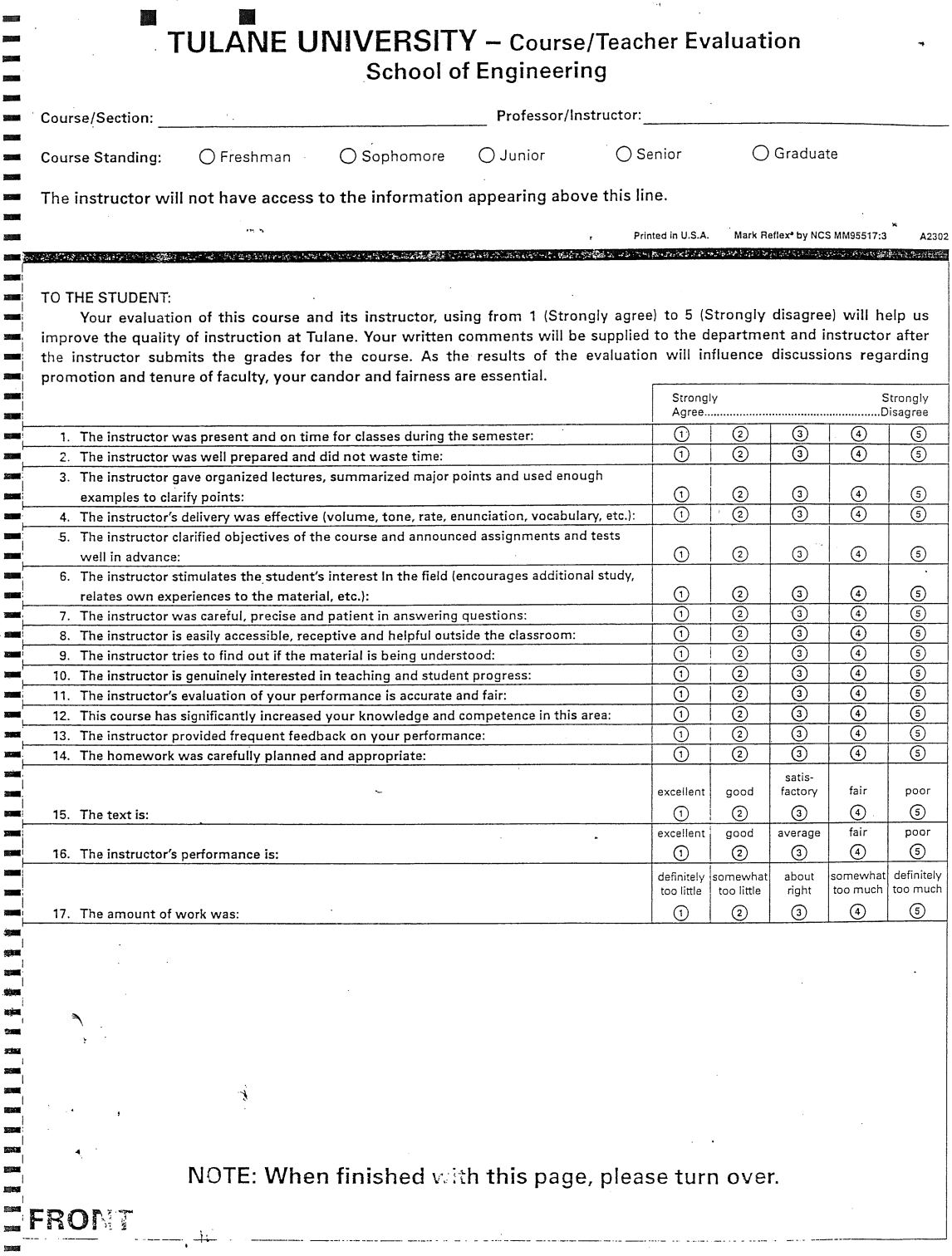
Discussions with colleagues inspired me to complete this project and submit it for publication; I am grateful for the motivation. I thank Dr. Glen A. Livesay for helpful comments on preliminary statistical analyses and drafts of this paper.

References

1. McKeachie, W.J. Student ratings of faculty: a reprise. *Academe*. October: 384-397, 1979.
2. Cashin, W.E., *IDEA Paper no. 32: Student ratings of teaching: the research revisited*. 1995, Center for Faculty Evaluation & Development, Division of Continuing Education, Kansas State University: Manhattan, Kansas.
3. Wachtel, H.K. Student evaluation of college teaching effectiveness: a brief review. *Assessment & Evaluation in Higher Education*. 23: 191-211, 1998.
4. Linsky, A.S. and Straus, M.A. Student evaluations, research productivity, and eminence of college faculty. *Journal of Higher Education*. 46: 89-102, 1975.
5. Marsh, H.W. Students' evaluations of university teaching: research findings, methodological issues, and directions for future research. *International Journal of Educational Research*. 11: 253-388, 1987.
6. Marsh, H.W. Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*. 76: 707-754, 1984.
7. Sojka, J., Gupta, A.K., and Deeter-Schmelz, D.R. Student and faculty perceptions of student evaluations of teaching: A study of similarities and differences. *College Teaching*. 50: 44-49, 2002.
8. Glantz, S.A., *Primer of Biostatistics*, New York: McGraw-Hill Health Professions Division, 1997, pg. 324.
9. See Glantz, op. cit. (reference 8), pg. 257.
10. Nunnally, J.C., *Psychometric Theory*, New York: McGraw-Hill Book Company, 1967, pg. 125.
11. Kirk, R.E., *Experimental Design: Procedures for the Behaviorial Sciences (second edition)*, Monterey, CA: Brooks/Cole Publishing Company, 1982, pg. 78.
12. See Nunnally, op. cit. (reference 10), pgs. 288-371.
13. SPSS software Tutorial: "Factor Analysis", SPSS for Windows, Standard version, 2001
14. Sixbury, G.R. and Cashin, W.E., *IDEA Technical Report No. 9: Description of database for the IDEA diagnostic form*. 1995, Center for Faculty Evaluation & Development, Kansas State University.
15. Marsh, H.W. Distinguishing between good (useful) and bad workloads on students' evaluations of teaching. *American Educational Research Journal*. 38: 183-212, 2001.
16. Chau, H. and Hocevar, D. Higher-order factor analysis of multidimensional students' evaluations of teaching effectiveness. Proceedings of the *The Annual Conference of the American Educational Research Association (AERA)*. New Orleans, LA: Educational Resources Information Center (ERIC). 1994.
17. Kirk, R.E. Practical significance: a concept whose time has come. *Educational and Psychological Measurement*. 56: 746-759, 1996.
18. Nickerson, R.S. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*. 5: 241-301, 2000.
19. Marsh, H.W. Effects of grading leniency and low workload on students' evaluations of teaching: popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*. 92: 202-228, 2000.
20. Ryan, J.J., Anderson, J.A., and Birchler, A.B. Student evaluation: the faculty responds. *Research in Higher Education*. 12: 317-333, 1980.
21. See Ryan, et al., op. cit. (reference 21), pg. 317.
22. See Ryan, et al., op. cit. (reference 21), pg. 322.
23. Greenwald, A.G. and Gillmore, G.M. No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology*. 89: 743-751, 1997.

Author Biography

KAY C DEE is an Associate Professor of Biomedical Engineering and a member of the Educational Research and Methods division of ASEE. She teaches undergraduate and graduate courses on the topics of cell and tissue engineering, cell/tissue biomechanics, science fiction and bioethics, and teaching engineering.



TULANE UNIVERSITY – Course/Teacher Evaluation
School of Engineering

Course/Section: _____ Professor/Instructor: _____

Course Standing: ○ Freshman ○ Sophomore ○ Junior ○ Senior ○ Graduate

The instructor will not have access to the information appearing above this line.

Printed in U.S.A. Mark Reflex® by NCS MM95517:3 A2302

TO THE STUDENT:

Your evaluation of this course and its instructor, using from 1 (Strongly agree) to 5 (Strongly disagree) will help us improve the quality of instruction at Tulane. Your written comments will be supplied to the department and instructor after the instructor submits the grades for the course. As the results of the evaluation will influence discussions regarding promotion and tenure of faculty, your candor and fairness are essential.

	Strongly Agree				Strongly Disagree
1. The instructor was present and on time for classes during the semester:	①	②	③	④	⑤
2. The instructor was well prepared and did not waste time:	①	②	③	④	⑤
3. The instructor gave organized lectures, summarized major points and used enough examples to clarify points:	①	②	③	④	⑤
4. The instructor's delivery was effective (volume, tone, rate, enunciation, vocabulary, etc.):	①	②	③	④	⑤
5. The instructor clarified objectives of the course and announced assignments and tests well in advance:	①	②	③	④	⑤
6. The instructor stimulates the student's interest In the field (encourages additional study, relates own experiences to the material, etc.):	①	②	③	④	⑤
7. The instructor was careful, precise and patient in answering questions:	①	②	③	④	⑤
8. The instructor is easily accessible, receptive and helpful outside the classroom:	①	②	③	④	⑤
9. The instructor tries to find out if the material is being understood:	①	②	③	④	⑤
10. The instructor is genuinely interested in teaching and student progress:	①	②	③	④	⑤
11. The instructor's evaluation of your performance is accurate and fair:	①	②	③	④	⑤
12. This course has significantly increased your knowledge and competence in this area:	①	②	③	④	⑤
13. The instructor provided frequent feedback on your performance:	①	②	③	④	⑤
14. The homework was carefully planned and appropriate:	①	②	③	④	⑤
15. The text is:	excellent ①	good ②	satis-factory ③	fair ④	poor ⑤
16. The instructor's performance is:	excellent ①	good ②	average ③	fair ④	poor ⑤
17. The amount of work was:	definitely too little ①	somewhat too little ②	about right ③	somewhat too much ④	definitely too much ⑤

NOTE: When finished with this page, please turn over.

FRONT

Figure A1. Tulane University School of Engineering teaching evaluation, page 1.

DO NOT WRITE IN THIS AREA

In the space provided below, please make any comments which you wish to direct to the instructor. The instructor will have the opportunity to review these after the final grades have been submitted. Your comments will help the instructor improve the course and enhance teaching effectiveness.

COURSE (What do you think should be changed to improve this course? What from this course was the most valuable to you? What was the least valuable?):

INSTRUCTOR (Please give examples of where the instructor was particularly effective or ineffective. What could the instructor do to be more effective?):

TEXT, TEACHING MATERIALS, AND LAB (if any): (Were they relevant to the rest of course? Were they helpful in understanding the material?):

Figure A2. Tulane University School of Engineering teaching evaluation, page 2.

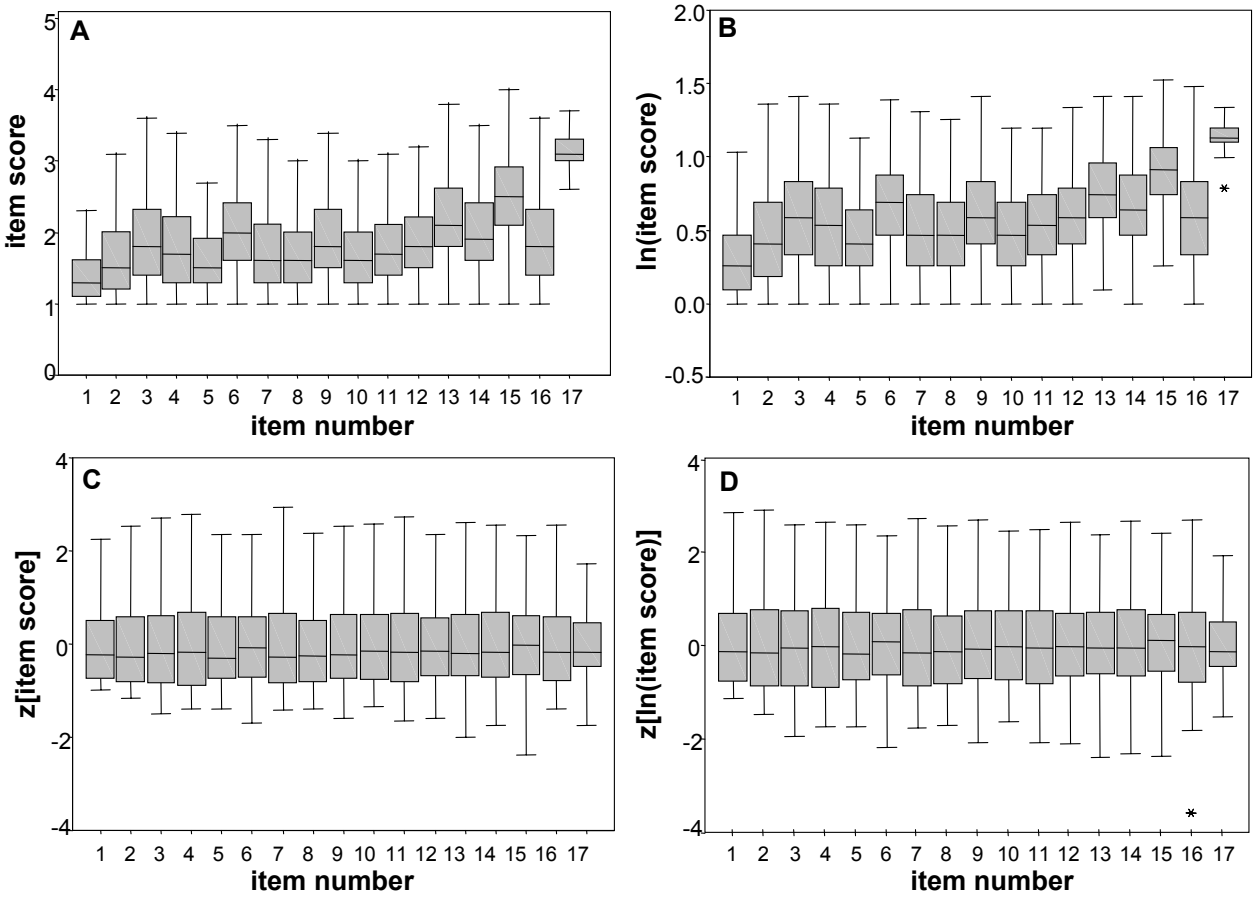


Figure A3. Boxplots for Various Data Transforms. The original data set with outliers removed is represented in frame **A**. Frame **B** shows the reduced data set after a natural log transformation. Frame **C** displays z scores of the reduced data set and frame **D** shows z scores calculated from the natural log-transformed data set. Vertical axes in all frames are scaled for convenient viewing rather than by maximum or minimum possible values of the data displayed.

Table A1. Data Descriptives, All Course Evaluations With Five or More Respondents.

	Course Level	*Number of Respondents*	*Q1*	*Q2*	*Q3*
N	823	823	823	823	821
Mean ± S.E.M.	3.52 ± 0.06	19.35 ± 0.40	1.436 ± 0.017	1.662 ± 0.020	1.923 ± 0.022
Median	3.0	17.0	1.3	1.5	1.8
Variance	2.75	133.94	0.241	0.322	0.384
Range	1 – 7	5 – 119	1.0 – 4.4	1.0 – 3.9	1.0 – 4.1
Skewness	0.397	2.123	1.922	1.231	0.823
Kurtosis	-0.857	11.710	4.665	1.274	0.320

	Q4	*Q5*	*Q6*	*Q7*	*Q8*
N	820	822	822	822	822
Mean ± S.E.M.	1.802 ±0.020	1.661 ± 0.018	2.042 ± 0.022	1.770 ± 0.020	1.760 ± 0.020
Median	1.7	1.5	2.0	1.6	1.6
Variance	0.326	0.250	0.381	0.331	0.338
Range	1.0 – 3.9	1.0 – 3.8	1.0 – 4.0	1.0 – 4.8	1.0 – 4.6
Skewness	0.644	1.244	0.487	1.162	1.25
Kurtosis	-0.141	1.842	-0.208	2.015	2.13

	Q9	*Q10*	*Q11*	*Q12*	*Q13*
N	822	822	823	823	822
Mean ± S.E.M.	1.929 ± 0.020	1.699 ± 0.019	1.811 ± 0.018	1.883 ± 0.019	2.216 ± 0.021
Median	1.8	1.6	1.7	1.8	2.1
Variance	0.334	0.292	0.272	0.310	0.366
Range	1.0 – 4.1	1.0 – 4.0	1.0 – 4.3	1.0 – 4.3	1.0 – 4.1
Skewness	0.701	1.002	1.114	0.978	0.506
Kurtosis	0.311	1.081	1.856	0.992	-0.092

	Q14	*Q15*	*Q16*	*Q17*
N	822	808	823	823
Mean ± S.E.M.	2.010 ± 0.020	2.517 ± 0.022	1.921 ± 0.023	3.157 ± 0.012
Median	1.9	2.5	1.8	3.1
Variance	0.337	0.404	0.433	0.110
Range	1.0 – 4.1	1.0 – 4.8	1.0 – 4.4	2.1 – 4.5
Skewness	0.755	0.300	0.874	0.634
Kurtosis	0.354	0.354	0.409	1.203

S.E.M.= Standard Error of the Mean. Q1 through Q17 refer to the seventeen items on the teaching evaluation questionnaire.

Table A2. Pearson's "*r*" and Spearman's "ρ" Correlation Coefficients Between Teaching Evaluation Items, For Data Sets "A" Through "D" Depicted in Figure A3. *Table continues over three more pages.*

		Level r, ρ	Responses r, ρ	Q1 r, ρ	Q2 r, ρ	Q3 r, ρ	Q4 r, ρ	Q5 r, ρ	Q6 r, ρ	Q7 r, ρ	Q8 r, ρ
Level		*(816)*									
Responses	A	-0.47*, -0.48*									
	B	-0.47*, -0.48*	*(816)*								
	C	-0.47*, -0.48*									
	D	-0.47*, -0.48*									
Q1	A	-0.17*, -0.15*	0.10*, 0.15*	*(797)*							
	B	-0.18*, -0.15*	0.11*, 0.15*								
	C	-0.17*, -0.15*	0.10*, 0.15*								
	D	-0.18*, -0.15*	0.11*, 0.15*								
Q2	A	-0.12*, -0.12*	0.15*, 0.18*	**0.59*, 0.67***	*(816)*						
	B	-0.14*, -0.12*	0.16*, 0.18*	**0.64*, 0.67***							
	C	-0.12*, -0.12*	0.15*, 0.18*	**0.59*, 0.67***							
	D	-0.14*, -0.12*	0.16*, 0.18*	**0.64*, 0.67***							
Q3	A	-0.12*, -0.08*	0.16*, 0.18*	0.46*, **0.52***	**0.84*, 0.81***	*(814)*					
	B	-0.13*, -0.08*	0.16*, 0.18*	0.49*, **0.52***	**0.83*, 0.81***						
	C	-0.12*, -0.08*	0.16*, 0.18*	0.46*, **0.52***	**0.84*, 0.81***						
	D	-0.13*, -0.08*	0.16*, 0.18*	0.49*, **0.52***	**0.83*, 0.81***						
Q4	A	-0.27*, -0.25*	0.16*, 0.20*	0.34*, 0.39*	**0.61*, 0.64***	**0.74*, 0.75***	*(813)*				
	B	-0.28*, -0.25*	0.17*, 0.20*	0.37*, 0.39*	**0.63*, 0.64***	**0.75*, 0.75***					
	C	-0.27*, -0.25*	0.16*, 0.20*	0.34*, 0.39*	**0.61*, 0.64***	**0.74*, 0.75***					
	D	-0.28*, -0.25*	0.17*, 0.20*	0.37*, 0.39*	**0.63*, 0.64***	**0.75*, 0.75***					
Q5	A	-0.03, 0.00	0.04, 0.07^	0.42*, 0.46*	**0.70*, 0.70***	**0.73*, 0.74***	**0.53*, 0.55***	*(803)*			
	B	-0.04, 0.00	0.05, 0.07^	0.44*, 0.46*	**0.71*, 0.70***	**0.74*, 0.74***	**0.55*, 0.55***				
	C	-0.03, 0.00	0.04, 0.07^	0.42*, 0.46*	**0.70*, 0.70***	**0.73*, 0.74***	**0.53*, 0.55***				
	D	-0.04, 0.00	0.05, 0.07^	0.44*, 0.46*	**0.71*, 0.70***	**0.74*, 0.74***	**0.55*, 0.55***				
Q6	A	-0.33*, -0.33*	0.30*, 0.34*	0.38*, 0.44*	**0.63*, 0.62***	**0.76*, 0.73***	**0.78*, 0.78***	**0.59*, 0.58***	*(815)*		
	B	-0.35*, -0.33*	0.30*, 0.34*	0.41*, 0.44*	**0.63*, 0.62***	**0.74*, 0.73***	**0.78*, 0.78***	**0.59*, 0.58***			
	C	-0.33*, -0.33*	0.30*, 0.34*	0.38*, 0.44*	**0.63*, 0.62***	**0.76*, 0.73***	**0.78*, 0.78***	**0.59*, 0.58***			
	D	-0.35*, -0.33*	0.30*, 0.34*	0.41*, 0.44*	**0.63*, 0.62***	**0.74*, 0.73***	**0.78*, 0.78***	**0.59*, 0.58***			
Q7	A	-0.25*, -0.23*	0.25*, 0.28*	0.37*, 0.42*	**0.62*, 0.62***	**0.78*, 0.77***	**0.69*, 0.72***	**0.59*, 0.60***	**0.75*, 0.75***	*(807)*	
	B	-0.26*, -0.23*	0.25*, 0.28*	0.39*, 0.42*	**0.63*, 0.62***	**0.77*, 0.77***	**0.72*, 0.72***	**0.60*, 0.60***	**0.75*, 0.75***		
	C	-0.25*, -0.23*	0.25*, 0.28*	0.37*, 0.42*	**0.62*, 0.62***	**0.78*, 0.77***	**0.69*, 0.72***	**0.59*, 0.60***	**0.75*, 0.75***		
	D	-0.26*, -0.23*	0.25*, 0.28*	0.39*, 0.42*	**0.63*, 0.62***	**0.77*, 0.77***	**0.72*, 0.72***	**0.60*, 0.60***	**0.75*, 0.75***		
Q8	A	-0.20*, -0.19*	0.21*, 0.24*	0.38*, 0.42*	**0.55*, 0.53***	**0.67*, 0.64***	**0.54*, 0.54***	**0.55*, 0.56***	**0.63*, 0.62***	**0.76*, 0.76***	*(805)*
	B	-0.22*, -0.19*	0.23*, 0.24*	0.41*, 0.42*	**0.55*, 0.53***	**0.66*, 0.64***	**0.56*, 0.54***	**0.56*, 0.56***	**0.64*, 0.62***	**0.76*, 0.76***	
	C	-0.20*, -0.19*	0.21*, 0.24*	0.38*, 0.42*	**0.55*, 0.53***	**0.67*, 0.64***	**0.54*, 0.54***	**0.55*, 0.56***	**0.63*, 0.62***	**0.76*, 0.76***	
	D	-0.22*, -0.19*	0.23*, 0.24*	0.41*, 0.42*	**0.55*, 0.53***	**0.66*, 0.64***	**0.56*, 0.54***	**0.56*, 0.56***	**0.64*, 0.62***	**0.76*, 0.76***	
Q9	A	-0.22*, -0.19*	0.23*, 0.26*	0.37*, 0.41*	**0.63*, 0.62***	**0.80*, 0.78***	**0.72*, 0.73***	**0.65*, 0.66***	**0.82*, 0.81***	**0.77*, 0.78***	**0.74*, 0.73***
	B	-0.23*, -019*	0.24*, 0.26*	0.39*, 0.41*	**0.64*, 0.62***	**0.79*, 0.78***	**0.74*, 0.73***	**0.66*, 0.66***	**0.81*, 0.81***	**0.77*, 0.78***	**0.74*, 0.73***
	C	-0.22*, -0.19*	0.23*, 0.26*	0.37*, 0.41*	**0.63*, 0.62***	**0.80*, 0.78***	**0.72*. 0.73***	**0.65*, 0.66***	**0.82*, 0.81***	**0.77*, 0.78***	**0.74*, 0.73***
	D	-0.23*, -0.19*	0.24*, 0.26*	0.39*, 0.41*	**0.64*, 0.62***	**0.79*, 0.78***	**0.74*, 0.73***	**0.66*, 0.66***	**0.81*, 0.81***	**0.77*, 0.78***	**0.74*, 0.73***

Table A2, Continued.

		Level r, ρ	Responses r, ρ	Q1 r, ρ	Q2 r, ρ	Q3 r, ρ	Q4 r, ρ	Q5 r, ρ	Q6 r, ρ	Q7 r, ρ	Q8 r, ρ
Q10	A	-0.20*, -0.19*	0.20*, 0.24*	0.40*, 0.45*	**0.62*, 0.62***	**0.78*, 0.77***	**0.72*, 0.74***	**0.62*, 0.63***	**0.82*, 0.82***	**0.83*, 0.84***	**0.79*, 0.78***
	B	-0.22*, -0.19*	0.21*, 0.24*	0.42*, 0.45*	**0.63*, 0.62***	**0.78*, 0.77***	**0.74*, 0.74***	**0.63*, 0.63***	**0.82*, 0.82***	**0.84*, 0.84***	**0.79*, 0.78***
	C	-0.20*, -0.19*	0.20*, 0.24*	0.41*, 0.45*	**0.62*, 0.62***	**0.78*, 0.77***	**0.72*, 0.74***	**0.62*, 0.63***	**0.82*, 0.82***	**0.83*, 0.84***	**0.79*, 0.78***
	D	-0.22*, -0.19*	0.21*, 0.24*	0.42*, 0.45*	**0.63*, 0.62***	**0.78*, 0.77***	**0.74*, 0.74***	**0.63*, 0.63***	**0.82*, 0.82***	**0.84*, 0.84***	**0.79*, 0.78***
Q11	A	-0.14*, -0.11*	0.28*, 0.30*	0.34*, 0.39*	**0.59*, 0.59***	**0.73*, 0.72***	**0.54*, 0.55***	**0.64*, 0.65***	**0.68*, 0.67***	**0.73*, 0.74***	**0.67*, 0.67***
	B	-0.15*, -0.11*	0.29*, 0.30*	0.37*, 0.39*	**0.60*, 0.59***	**0.72*, 0.72***	**0.55*, 0.55***	**0.65*, 0.65***	**0.67*, 0.67***	**0.73*, 0.74***	**0.68*, 0.67***
	C	-0.14*, -0.11*	0.28*, 0.30*	0.34*, 0.39*	**0.59*, 0.59***	**0.73*, 0.72***	**0.54*, 0.55***	**0.64*, 0.65***	**0.68*, 0.67***	**0.73*, 0.74***	**0.67*, 0.67***
	D	-0.15*, -0.11*	0.29*, 0.30*	0.37*, 0.39*	**0.60*, 0.59***	**0.72*, 0.72***	**0.55*, 0.55***	**0.65*, 0.65***	**0.67*, 0.67***	**0.73*, 0.74***	**0.68*, 0.67***
Q12	A	-0.13*, -0.12*	0.21*, 0.23*	0.37*, 0.44*	**0.75*, 0.71***	**0.82*, 0.78***	**0.64*, 0.64***	**0.70*, 0.69***	**0.76*, 0.74***	**0.63*, 0.63***	**0.59*, 0.57***
	B	-0.16*, -0.12*	0.22*, 0.23*	0.40*, 0.44*	**0.73*, 0.71***	**0.80*, 0.78***	**0.65*, 0.64***	**0.70*, 0.69***	**0.75*, 0.74***	**0.64*, 0.63***	**0.59*, 0.57***
	C	-0.13*, -0.12*	0.21*, 0.23*	0.37*, 0.44*	**0.75*, 0.71***	**0.82*, 0.78***	**0.64*, 0.64***	**0.70*, 0.69***	**0.76*, 0.74***	**0.63*, 0.63***	**0.59*, 0.57***
	D	-0.16*, -0.12*	0.22*, 0.23*	0.40*, 0.44*	**0.73*, 0.71***	**0.80*, 0.78***	**0.65*, 0.64***	**0.70*, 0.69***	**0.75*, 0.74***	**0.64*, 0.63***	**0.59*, 0.57***
Q13	A	-0.18*, -0.14*	0.15*, 0.16*	0.41*, 0.42*	**0.64*, 0.62***	**0.67*, 0.66***	**0.53*, 0.55***	**0.71*, 0.72***	**0.64*, 0.63***	**0.58*, 0.59***	**0.58*, 0.60***
	B	-0.18*, -0.14*	0.16*, 0.16*	0.43*, 0.42*	**0.64*, 0.62***	**0.68*, 0.66***	**0.55*, 0.55***	**0.71*, 0.72***	**0.65*, 0.63***	**0.59*, 0.59***	**0.60*, 0.60***
	C	-0.18*, -0.14*	0.15*, 0.16*	0.41*, 0.42*	**0.64*, 0.62***	**0.67*, 0.66***	**0.53*, 0.55***	**0.71*, 0.72***	**0.64*, 0.63***	**0.58*, 0.59***	**0.58*, 0.60***
	D	-0.18*, -0.14*	0.16*, 0.16*	0.43*, 0.42*	**0.64*, 0.62***	**0.68*, 0.66***	**0.55*, 0.55***	**0.71*, 0.72***	**0.65*, 0.63***	**0.59*, 0.59***	**0.60*, 0.60***
Q14	A	-0.15*, -0.12*	0.19*, 0.21*	0.33*, 0.41*	**0.70*, 0.69***	**0.78*, 0.77***	**0.56*, 0.58***	**0.72*, 0.73***	**0.67*, 0.64***	**0.66*, 0.66***	**0.60*, 0.61***
	B	-0.16*, -0.12*	0.19*, 0.21*	0.37*, 0.41*	**0.70*, 0.69***	**0.78*, 0.77***	**0.58*, 0.58***	**0.73*, 0.73***	**0.66*, 0.64***	**0.66*, 0.66***	**0.61*, 0.61***
	C	-0.15*, -0.12*	0.19*, 0.21*	0.33*, 0.41*	**0.70*, 0.69***	**0.78*, 0.77***	**0.56*, 0.58***	**0.72*, 0.73***	**0.67*, 0.64***	**0.66*, 0.66***	**0.60*, 0.61***
	D	-0.16*, -0.12*	0.19*, 0.21*	0.37*, 0.41*	**0.70*, 0.69***	**0.78*, 0.77***	**0.58*, 0.58***	**0.73*, 0.73***	**0.66*, 0.64***	**0.66*, 0.66***	**0.61*, 0.61***
Q15	A	-0.17*, -0.17*	0.17*, 0.20*	0.22*, 0.25*	0.33*, 0.32*	0.40*, 0.39*	0.29*, 0.31*	0.32*, 0.34*	0.38*, 0.38*	0.33*, 0.32*	0.31*, 0.30*
	B	-0.19*, -0.17*	0.18*, 0.20*	0.21*, 0.25*	0.30*, 0.32*	0.38*, 0.39*	0.30*, 0.31*	0.31*, 0.34*	0.38*, 0.38*	0.32*, 0.32*	0.31*, 0.30*
	C	-0.17*, -0.17*	0.17*, 0.20*	0.22*, 0.25*	0.33*, 0.32*	0.40*, 0.39*	0.29*, 0.31*	0.32*, 0.34*	0.38*, 0.38*	0.33*, 0.32*	0.31*, 0.30*
	D	-0.19*, -0.17*	0.18*, 0.20*	0.21*, 0.25*	0.30*, 0.32*	0.38*, 0.39*	0.30*, 0.31*	0.31*, 0.34*	0.38*, 0.38*	0.32*, 0.32*	0.31*, 0.30*
Q16	A	-0.21*, -0.20*	0.20*, 0.23*	0.47*, **0.53***	**0.81*, 0.79***	**0.92*, 0.90***	**0.80*, 0.81***	**0.72*, 0.71***	**0.84*, 0.83***	**0.82*, 0.83***	**0.74*, 0.71***
	B	-0.24*, -0.20*	0.20*, 0.23*	0.50*, **0.53***	**0.80*, 0.79***	**0.90*, 0.90***	**0.81*, 0.81***	**0.72*, 0.71***	**0.84*, 0.83***	**0.83*, 0.83***	**0.73*, 0.71***
	C	-0.21*, -0.20*	0.20*, 0.23*	0.47*, **0.53***	**0.81*, 0.79***	**0.92*, 0.90***	**0.80*, 0.81***	**0.72*, 0.71***	**0.84*, 0.83***	**0.82*, 0.83***	**0.74*, 0.71***
	D	-0.24*, -0.20*	0.20*, 0.23*	0.50*, **0.53***	**0.80*, 0.79***	**0.90*, 0.90***	**0.81*, 0.81***	**0.72*, 0.71***	**0.84*, 0.83***	**0.83*, 0.83***	**0.73*, 0.71***
Q17	A	-0.21*, -0.21*	0.23*, 0.25*	0.01, 0.05	0.05, 0.07	0.10*, 0.10*	0.11*, 0.11*	0.09*, 0.08^	0.21*, 0.20*	0.19*, 0.16*	0.13*, 0.14*
	B	-0.21*, -0.21*	0.23*, 0.25*	0.01, 0.05	0.05, 0.07	0.10*, 0.10*	0.12*, 0.11*	0.09^, 0.08^	0.21*, 0.20*	0.17*, 0.16*	0.13*, 0.14*
	C	-0.21*, -0.21*	0.23*, 0.25*	0.01, 0.05	0.05, 0.07	0.10*, 0.10*	0.11*, 0.11*	0.09*, 0.08^	0.21*, 0.20*	0.19*, 0.16*	0.13*, 0.14*
	D	-0.21*, -0.21*	0.23*, 0.25*	0.01, 0.05	0.05, 0.07	0.10*, 0.10*	0.12*, 0.11*	0.09^, 0.08^	0.21*, 0.20*	0.17*, 0.16*	0.13*, 0.14*

Table A2, Continued.

		Q9 r, ρ	Q10 r, ρ	Q11 r, ρ	Q12 r, ρ	Q13 r, ρ	Q14 r, ρ	Q15 r, ρ	Q16 r, ρ	Q17 r, ρ
Level										
Responses	A									
	B									
	C									
	D									
Q1	A									
	B									
	C									
	D									
Q2	A									
	B									
	C									
	D									
Q3	A									
	B									
	C									
	D									
Q4	A									
	B									
	C									
	D									
Q5	A									
	B									
	C									
	D									
Q6	A									
	B									
	C									
	D									
Q7	A									
	B									
	C									
	D									
Q8	A									
	B									
	C									
	D									
Q9	A									
	B	*(815)*								
	C									
	D									

Table A2, Continued.

		Q9 r, ρ	Q10 r, ρ	Q11 r, ρ	Q12 r, ρ	Q13 r, ρ	Q14 r, ρ	Q15 r, ρ	Q16 r, ρ	Q17 r, ρ
Q10	A	**0.90*, 0.90***								
	B	**0.90*, 0.90***	*(808)*							
	C	**0.90*, 0.90***								
	D	**0.90*, 0.90***								
Q11	A	**0.75*, 0.73***	**0.76*, 0.75***							
	B	**0.74*, 0.73***	**0.76*, 0.75***	*(807)*						
	C	**0.75*, 0.73***	**0.76*, 0.75***							
	D	**0.74*, 0.75***	**0.76*, 0.75***							
Q12	A	**0.77*, 0.75***	**0.71*, 0.70***	**0.70*, 0.69***						
	B	**0.76*, 0.75***	**0.71*, 0.70***	**0.70*, 0.69***	*(816)*					
	C	**0.77*, 0.75***	**0.71*, 0.70***	**0.70*, 0.69***						
	D	**0.76*, 0.75***	**0.71*, 0.70***	**0.70*, 0.69***						
Q13	A	**0.72*, 0.72***	**0.65*, 0.65***	**0.65*, 0.66***	**0.71*, 0.70***					
	B	**0.73*, 0.72***	**0.66*, 0.65***	**0.67*, 0.66***	**0.71*, 0.70***	*(815)*				
	C	**0.72*, 0.72***	**0.65*, 0.65***	**0.65*, 0.66***	**0.71*, 0.70***					
	D	**0.73*, 0.72***	**0.66*, 0.65***	**0.67*, 0.66***	**0.71*, 0.70***					
Q14	A	**0.73*, 0.70***	**0.66*, 0.66***	**0.73*, 0.73***	**0.76*, 0.73***	**0.74*, 0.75***				
	B	**0.72*, 0.70***	**0.67*, 0.66***	**0.73*, 0.73***	**0.75*, 0.73***	**0.75*, 0.75***	*(815)*			
	C	**0.73*, 0.70***	**0.66*, 0.66***	**0.73*, 0.73***	**0.76*, 0.73***	**0.74*, 0.75***				
	D	**0.72*, 0.70***	**0.67*, 0.66***	**0.73*, 0.73***	**0.75*, 0.73***	**0.75*, 0.75***				
Q15	A	0.39*, 0.39*	0.34*, 0.34*	0.34*, 0.33*	0.45*, 0.44*	0.37*, 0.38*	0.41*, 0.40*			
	B	0.39*, 0.39*	0.34*, 0.34*	0.35*, 0.33*	0.43*, 0.44*	0.37*, 0.38*	0.40*, 0.40*	*(801)*		
	C	0.39*, 0.39*	0.34*, 0.34*	0.34*, 0.33*	0.45*, 0.44*	0.37*, 0.38*	0.41*, 0.40*			
	D	0.39*, 0.39*	0.34*, 0.34*	0.35*, 0.33*	0.43*, 0.44*	0.37*, 0.38*	0.40*, 0.40*			
Q16	A	**0.86*, 0.85***	**0.87*, 0.86***	**0.77*, 0.76***	**0.84*, 0.81***	**0.71*, 0.71***	**0.77*, 0.76***	0.39*, 0.39*		
	B	**0.85*, 0.85***	**0.87*, 0.86***	**0.76*, 0.76***	**0.82*, 0.81***	**0.72*, 0.71***	**0.77*, 0.76***	0.39*, 0.39*	*(816)*	
	C	**0.86*, 0.85***	**0.87*, 0.86***	**0.77*, 0.76***	**0.84*, 0.81***	**0.71*, 0.71***	**0.77*, 0.76***	0.39*, 0.39*		
	D	**0.85*, 0.85***	**0.87*, 0.86***	**0.76*, 0.76***	**0.82*, 0.81***	**0.72*, 0.71***	**0.77*, 0.76***	0.39*, 0.39*		
Q17	A	0.16*, 0.15*	0.13*, 0.13*	0.22*, 0.22*	0.09*, 0.08^	0.08^, 0.08^	0.24*, 0.20*	0.08^, 0.09^	0.13*, 0.15*	
	B	0.15*, 0.15*	0.13*, 0.13*	0.21*, 0.22*	0.09^, 0.08^	0.07^, 0.08^	0.21*, 0.20*	0.08^, 0.09^	0.14*, 0.15*	*(809)*
	C	0.16*, 0.15*	0.13*, 0.13*	0.22*, 0.22*	0.09*, 0.08^	0.08^, 0.08^	0.24*, 0.20*	0.08^, 0.09^	0.13*, 0.15*	
	D	0.15*, 0.15*	0.13*, 0.13*	0.21*, 0.22*	0.09^, 0.08^	0.07^, 0.08^	0.21*, 0.20*	0.08^, 0.09^	0.14*, 0.15*	

Data sets A, B, C, and D are the same data sets portrayed in Figure A3. $*p < 0.01$; $\hat{}p < 0.05$ (two-tailed). The number of valid cases for each item is presented in parentheses along the diagonal of the table. Correlation coefficients greater than 0.5 are highlighted in bold; correlation coefficients greater than or equal to 0.9 are additionally highlighted in red and underlined. "Level" refers to the level of a course (100 or freshman-level, 200 or sophomore level, *etc.*, up to 700 or graduate-level courses); "Responses" refers to the number of individual student ratings that were used to calculate the course-average rating. Q1 through Q16 refer to items one through sixteen on the evaluation form, for which lower numerical scores implied better performance. For Q17/item seventeen, lower numerical scores implied too little work, while higher numerical scores implied too much work.