CITATION

Stanny, C. J., & Arruda, J. E. (2017, June 19). A Comparison of Student Evaluations of Teaching With Online and Paper-Based Administration. *Scholarship of Teaching and Learning in Psychology.* Advance online publication. http://dx.doi.org/10.1037/stl0000087

A Comparison of Student Evaluations of Teaching with Online and Paper-Based Administration

Claudia J. Stanny and James E. Arruda

University of West Florida

Author Note

Claudia J. Stanny, Center for University Teaching, Learning, and Assessment; James E. Arruda, Department of Psychology.

Data collection and preliminary analysis were sponsored by the Office of the Provost and the Student Assessment of Instruction Task Force.

Portions of these findings were presented as a poster at the 2016 National Institute on the Teaching of Psychology, St. Pete Beach, FL.

Correspondence concerning this article should be addressed to Claudia J. Stanny (cstanny@uwf.edu) at the Center for University Teaching, Learning, and Assessment, BLDG 53, University of West Florida, 11000 University Parkway, Pensacola, FL, 32514

A Comparison of Student Evaluations of Teaching with Online and Paper-Based Administration

Date Submitted: May 20, 2016

Revision: January 13, 2017

Revision: March 14, 2017

Revision: May 15, 2017

Abstract

When institutions administer student evaluations of teaching (SET) online, response rates are lower relative to paper-based administration. Researchers analyzed average SET scores from 364 courses taught during the fall term in 3 consecutive years to determine whether administering SET forms online for all courses in the third year changed the response rate or the average SET score. To control for instructor characteristics, the data analysis was based on courses for which the same instructor taught the course in each of three successive fall terms. Response rates for face-to-face classes declined when SET administration occurred only online. Although average SET scores were reliably lower in Year 3 than in the previous two years, the magnitude of this change was minimal (.11 on a five-item Likert-like scale). The authors discuss practical implications of these findings for interpretation of SETs and the role of SETs for the evaluation of teaching quality.

KEYWORDS: college teaching, student evaluations of teaching, online SET administration, response rate, assessment

A Comparison of Student Evaluations of Teaching with Online and Paper-Based Administration

Student ratings and evaluations of instruction have a long history as a source of information about teaching quality (Berk, 2013). Student evaluations of teaching (SET) often play a significant role in high-stakes decisions about hiring, promotion, tenure, and teaching awards. As a result, researchers have examined the psychometric properties of SETs and the possible impact of variables such as race, gender, age, course difficulty, and grading practices on average student ratings (Griffin, Hilton, Plummer, & Barret, 2014; Hativa, 2013; Marsh, 2007; Nulty, 2008; Spooren, Brockx, & Morteelmans, 2013). They have also examined how decision-makers evaluate SET scores (Dewar, 2011, Boysen, 2015a, 2015b; Boysen, Kelly, Raesly, & Casner, 2014). In the last 20 years, considerable attention has been directed toward the consequences of administering SETs online (Cates, 1993; Layne, DeChristoforo, McGinty, 1999; Morrison, 2011; Stowell, Addison, & Smith, 2012; Venette, Sellnow, & McIntyre, 2010) because low response rates may have implications for how decision-makers should interpret SETs.

Administering SETs online creates multiple benefits. Online administration saves the cost of paper, printing, and staff time devoted to scanning paper forms and typing written comments (Layne, DeChristoforo, & McGinty, 1999; Miller, 1987). Online administration enables instructors to devote more class time to instruction (versus administering paper-based forms) and can improve the integrity of the process. Students who are not pressed for time (as occurs with short, in-class SET administrations) are more likely to reflect on their answers and write detailed comments when they complete an SET online. For example, Layne et al. report that 76% of students wrote comments on electronic SETs whereas only 50% of students wrote comments on paper-based SETs. Although several researchers report that students write the same number of comments when they complete SETs online and on paper (Morrison, 2011; Stowell et al., 2012; Venette et al., 2010), these researchers also report that students write longer and more detailed comments in online SETs (based on word counts). Because electronic aggregation of responses bypasses the time-consuming task of transcribing comments (sometimes written in challenging handwriting), instructors can receive summary data and verbatim comments shortly after the close of the term instead of weeks or months into the following term.

In spite of the many benefits of online administration, instructors and students express concerns about online administration of SETs. Regardless of assurances of confidentiality, students express concern that their responses are not confidential when they must use their student identification to log into the system (Dommeyer, Baum, & Hanna, 2002; Layne et al., 1999). Students report they feel more confident that their responses will be anonymous when they complete paper SET forms (Layne et al., 1999). Unfortunately, breaches of confidentiality can occur with paper-based administration. For example, an instructor might recognize student handwriting (one reason some students do not write comments on paper-based forms) or an instructor might remain present during SET administration (Avery, Bryant, Mathios, Kang, & Bell, 2006).

In-class, paper-based administration creates social expectations that might motivate students to complete SETS. In contrast, students who are concerned about confidentiality or do not understand how instructors and institutions use SET findings to improve teaching might ignore requests to complete an online SET (Dommeyer, Baum, & Hanna, 2002). Instructors worry that low response rates will reduce the validity of the findings if students who do not complete an SET differ in significant ways from students who do (Stowell et al., 2011). For example, students who do not attend class regularly often miss class the day that SETs are administered. However, all students (including non-attending students) can complete the forms when they are administered online. Faculty fear that SET findings based on a low-response sample will be dominated by students in extreme categories, who may be particularly motivated to complete online SETS (e.g., students with grudges and students with extremely favorable attitudes), and that SET findings will inadequately represent the voice of average students (Reiner & Arnold, 2010).

The potential for biased SET findings associated with low response rates has been examined in the published literature. Contrary to faculty fears that online SETs might be dominated by low-performing students, Avery et al. (2006) found that students with *higher* GPAs were more likely to complete online evaluations. Although Griffin et al. (2014) observed a moderate correlation between a global measure of student GPA and average ratings on SETs, the strength of this relationship varied widely across individual instructors and courses. Moreover, the correlation was negative for 21% of the instructors in their sample. Thus, Griffin's findings suggest that adopting a strategy of attempting to manipulate SET scores by grading leniently can backfire. More recently, Jaquett, Van Maaren, and Williams (2017) report that students who had positive experiences in their classes (including the grade they expected to earn) reported that they were more likely to submit course evaluations.

Institutions can expect lower response rates when they administer SETs online (Avery et al., 2006; Dommeyer, Baum, Hanna, & Chapman, 2004; Johnson, 2002, Layne et al., 1999; Morrison, 2011; Nulty, 2008; Reiner & Arnold, 2010; Stowell et al., 2012; Venette et al., 2010). However, most researchers find that the mean SET rating does not change significantly when they compare SETs administered on paper with those completed online. These findings have been replicated in multiple settings using a variety of research methods. Several researchers randomly assigned sections of the same course to complete forms in each method. Some of these researchers evaluated SETs for courses taught in a single program or simply selected small samples of instructors and courses (Avery et al., 2006; Dommeyer et al., 2004; Morrison, 2011; Stowell et al., 2012; Venette et al., 2010). Two of these studies included the strong design feature of randomly assigning course sections to either paper-based or online SET administration (Dommeyer et al., 2004; Stowell et al., 2012). Other researchers obtained larger samples of courses across the university, holding instructor and course constant (Layne et al., 1999; Reiner & Arnold, 2010; Risquez, Vaughan, & Murphy, 2015). Although response rates were significantly lower when SETs were administered online, the average rating did not differ significantly for online and paper-based SETs. Unlike previous researchers, Risquez et al. used a regression model to control the influence of possible confounding factors (academic discipline, class size, number of years teaching, student self-reports of interest and preparation, and time of data collection). They found that method of administration (paper or online) had a minimal impact on average SET scores (a change in .08 on a 5-point Likert scale). Unfortunately, they did not report summary data for unadjusted SET ratings.

Exceptions to the pattern of minimal or non-significant differences in average SET scores appear in Morrison (2011) and Nowell, Gale, and Handley (2010), who examined small samples of business course SETs (29 courses in each sample). Both studies reported lower average scores when SETs were administered online. However, they also found that SET scores for individual items varied more within an instructor when SETs were administered online than for scores based on paper forms. Students who completed SETS on paper tended to record the same response for all questions whereas students who completed forms online tended to respond differently to different questions. Both research groups argue that scores obtained online might not be directly comparable to scores obtained through paper-based forms. They advocate that institutions administer SETs entirely online or entirely on paper to ensure consistent, comparable evaluations across faculty. As noted earlier (Layne et al., 1999), online administration provides students with more time for reflection before selecting their response to individual questions.

Each university presents a unique environment and culture that could influence how seriously students take SETs and how they respond to decisions to administer SETs online. Although a few large-scale studies of the impact of online administration exist (Reiner & Arnold, 2010; Risquez et al., 2015), a local replication answers questions about characteristics unique to that institution and generates evidence about the generalizability of existing findings. Reiner and Arnold gathered data at a large, research-intensive university in the United States (Purdue University); Risquez et al. gathered data at a large university in Ireland (University of Limerick).

The present study examines patterns of responses for online and paper-based SET scores at a mid-sized regional comprehensive university in the United States. Like previous researchers, we posed two questions. First, does the response rate or the average SET score change when an institution administers SET forms online instead of on paper? Second, what is the minimal response rate required to produce stable average SET scores for an instructor? Unlike much of the earlier research, which relied on small samples (often limited to a single academic department), we gathered SET data on a large sample of courses (*n* = 364 courses) that represented instructors from all colleges (representing a wide variety of disciplines) and all course levels (introductory and general education courses through graduate seminars) across three years. Moreover, the sample controlled individual differences in instructors by limiting the sample to courses taught by the same instructor in each of the three years of data collection. The university offers a significant proportion of courses online (nearly 30% of course sections in any given term) and these courses have always administered SET forms online. As a result, the sample provided an opportunity to examine both the combined effects of changing the method of delivery for SETs (paper-based to online) for traditional classes and changing from a mixed method of administering SETs (paper for traditional classes, online for online classes in the first two years of data gathered) to uniform use of online forms for all classes in the final year of data collection.

Method

Sample

Response rates and evaluation ratings were retrieved from archived course evaluation data. The archive of SET ratings did not include information about personal characteristics of the instructor (sex, age, or years of teaching experience) and students were not provided with any systematic incentive to complete the paper or online versions of the SET. We extracted data on response rates and evaluation ratings for 364 courses that had been taught by the same instructor during each of three consecutive fall terms (2012, 2013, 2014). The sample included faculty who teach in each of the five colleges at the university: 109 instructors (30% of the sample) taught in the College of Social Science and Humanities, 82 (23%) taught in the College of Science and Engineering, 75 (21%) taught in the College of Education and Professional Studies, 40 (11%) taught in the College of Business, and 58 (16%) taught in the College of Health. Each instructor provided data on one course. Approximately 259 instructors (71% of the sample) provided ratings for face-to-face courses and 105 (29%) provided ratings for online courses, which accurately reflects the proportion of face-to-face and online courses offered at the university. The sample included 107 courses (29%) that were 1000- and 2000-numbered courses (beginning undergraduate, taken mainly by first and second year students), 205 courses (56%) that were

3000- and 4000-numbered courses (advanced undergraduate, taken mainly by third and fourth year students), and 52 (14%) graduate-level courses (5000- and 6000-numbered courses).

Instrument

The course evaluation instrument is a set of 18 items developed by the State University System. The first eight items were designed to measure the quality of the instructor, concluding with a global rating of instructor quality (Item 8: *Overall assessment of instructor*). The remaining items ask students to evaluate components of the course, concluding with a global rating of course organization (Item 18: O*verall, I would rate the course organization*). No formal data on the psychometric properties of the items are available, although all items have obvious face validity.

Students were asked to rate each instructor as *Poor* (0), *Fair* (1), *Good* (2), *Very Good* (3), or *Excellent* (4) in response to each item. Evaluation ratings, which ranged from 0 to 4, were subsequently calculated for each course and instructor. A median rating was computed when an instructor taught more than one section of a course during a term.

The institution limited our access to SET data for the three years of data requested. Researchers obtained scores for Item 8 (*Overall assessment of instructor*) for all three years but could only obtain scores for Item 18 (*Overall, I would rate the course organization*) for the third year. Researchers computed the correlation between scores on Item 8 and Item 18 (from course data recorded in the third year only) to estimate the internal consistency of the evaluation instrument. These two items, which serve as composite summaries of preceding items (1-7 for Item 8 and 9-17 for Item 18), were strongly related (*r*(362) = .92). Feistauer and Richter (2016) also report strong correlations between global items in a large analysis of SET responses.

Design

This study took advantage of a natural experiment created when the university decided to administer all course evaluations online. The authors requested SET data for the fall semesters for two years preceding the change, when students completed paper-based SET forms for face-to-face courses and completed an online SET form for online courses, and data for the fall semester of the implementation year, when students completed online SET forms for all courses. Data analysis employed a 2 x 3 x 3 factorial design in which Course Delivery Method (Face-to-Face, Online) and Course Level (Beginning Undergraduate, Advanced Undergraduate, and Graduate) were between subjects factors and Evaluation Year (Year 1 - 2012, Year 2 - 2013, and Year 3 - 2014) was a repeated measure factor. The dependent measures were response rate (measured as a percentage of class enrollment) and rating for Item 8 (*Overall evaluation of instructor*).

Data analysis was limited to scores on Item 8 because the institution agreed to release data on this one item only. Data for scores on Item 18 were made available for SET forms administered in Year 3 to address questions about variation in responses across items. The strong correlation between scores on Item 8 and scores on Item 18 suggested that Item 8 could be used as a surrogate for all of the items. These two items were of particular interest because faculty, department chairs, and review committees frequently rely on these two items as stand-alone indicators of teaching quality for annual evaluations and tenure and promotion reviews.

Results

Response rate

The findings for response rates (presented in Table 1) indicate that response rates for face-to-face courses were much higher than for online courses, but only when course evaluations were administered in the classroom (i.e., for face-to-face courses). In the Year 3 administration, when all course evaluations were administered online, response rates for face-to-face courses declined (*M* = 47.18, *SD* = 20.11), but were still slightly higher than for online courses (*M* = 41.60, *SD* = 18.23). These findings produced a statistically significant interaction between Course Delivery Method and Evaluation Year, *F*(1.78, 716) = 101.34, MSe = 210.61, *p* < .001.[1] The strength of the overall interaction effect was .22 (η^2_{partial}). Simple main effects tests revealed statistically significant differences in the response rates for face-to-face courses and online courses for each of the three observation years.[2] The greatest differences occurred during Years 1 (*p* < .001) and 2 (*p* < .001), when evaluations were administered on paper in the classroom for all face-to-face courses whereas online classes submitted online evaluation forms. More importantly, although the difference in response rate between face-to-face and online courses during the Year 3 administration was statistically reliable (when both face-to-to-face and online courses were evaluated with online surveys), the effect was quite small (η^2_{partial} = .02). Thus, there was minimal difference in response rate between face-to-face and online courses when evaluations were administered online for all courses. No other factors or interactions included in the analysis were statistically reliable.

Insert Table 1 about Here

Evaluation ratings

The same 2 x 3 x 3 ANOVA model described above was used to evaluate mean SET ratings. This analysis produced two statistically significant main effects. The first main effect involved Evaluation Year, *F*(1.86, 716) = 3.44, MSe = .18, *p* = .03 (η^2_{partial} = .01).[3] Evaluation ratings associated with the Year 3 administration (*M* = 3.26, *SD* = .60) were significantly lower than the evaluation ratings associated with both the Year 1 (*M* = 3.35, *SD* = .53) and Year 2 (*M* = 3.38, *SD* = .54) administrations. Thus, all courses received lower SET scores, regardless of course delivery method and course level. However, the size of this effect was quite small (the largest difference in mean rating was .11 on a five-item scale).

The second statistically significant main effect involved Delivery Mode, *F*(1, 358) = 23.51, MSe = .52, *p* = .01 (η^2_{partial} = .06).[4] Face-to-face courses (*M* = 3.41, *SD* = .50) received significantly higher mean ratings than did online courses (*M* = 3.13, *SD* = .63), regardless of evaluation year and course level. No other factors or interactions included in the analysis were statistically reliable.

Stability of ratings

The scatterplot presented in Figure 1 describes the relation between SET scores and response rate. Although the correlation between response rate and evaluation ratings was small and not statistically significant (*r*(362) = .07), visual inspection of the plot of SET scores suggests that SET ratings became less variable as response rate increased. We conducted Levene's Test (Brown & Forsythe, 2012) to evaluate the variability of SET scores above and below the 60% response rate, which several researchers recommend as an acceptable threshold for response rates (Berk, 2012, 2013; Nulty, 2008; Seldin and Associates, 2006). The variability of scores above and below the 60% threshold was not statistically reliable (*F*(1, 362) = 1.53, *p* = .22).

Insert Figure 1 about here

Discussion and Recommendations

As observed on multiple campuses, online administration of SETs produced lower response rates. Curiously, online courses experienced a 10% increase in response rate when all courses were evaluated with online forms in Year 3. Online courses had suffered from chronically low response rates in previous years, when face-to-face classes continued to use paper-based forms. The benefit to response rates observed for online courses when all SET forms were administered online might be attributed to increased communications that encouraged students to complete the online course evaluations. Despite this improvement, response rates for online courses continued to lag behind those for face-to-face courses. Differences between response rates for face-to-face and online courses might be attributed to characteristics of the students who enrolled or they might be attributed to differences in the quality of student engagement created in each learning modality. Avery et al. (2006) found that higher-performing students (defined as students with higher GPAs) were more likely to complete online SETs.

Although the average SET rating was significantly lower in Year 3 than in the previous two years, the magnitude of the numeric difference was quite small (differences ranged from .08 to .11, for scores based on a 0 – 4 Likert scale). This difference is similar to the differences Risquez et al. (2015) reported for SET scores after adjusting them statistically for the influence of several potential biasing variables. Moreover, the difference in average SET rating is comparable to non-significant differences in SET scores reported across courses taught by a single given instructor. For example, these differences are smaller than those Boysen (2015a, 2015b) used to evaluate the tendency of reviewers to over-interpret non-significant differences in average SET ratings. A substantial literature discusses the appropriate and inappropriate interpretation of SET ratings (Berk, 2013; Boysen, 2015a, 2015b; Boysen et al., 2014; Dewar, 2011; Stark & Freishtat, 2014).

The small sample sizes created by low response rates often raise concerns among faculty about the variability of SET scores. However, our analysis indicates that classes with high response rates produced equally variable SET scores as did classes with low response rates. Reviewers should take extra care when they interpret SET scores and recognize that SET scores are inherently variable when they make judgments about faculty expertise in teaching or compare SET scores for different faculty. Decision makers often ignore questions about whether means derived from small samples accurately represent the population mean (Tversky & Kahneman, 1971). Boysen (2015a, 2015b) reports that reviewers frequently treat all numeric differences as if they were equally meaningful as measures of true differences. That is, reviewers manifest the cognitive bias Tversky and Kahneman identified as the belief in the law of small numbers. Boysen (2015a, 2015b) discusses the difficulty of overcoming this bias and describes the persistence of reviewers, who continued to give credibility to numeric differences even after receiving explicit warnings that underlying variability clearly indicates that these differences are not meaningful.

Because low response rates produce small sample sizes, we expected that the SET scores based on small samples (courses with low response rates) would be more variable than those based on larger class samples (courses with high response rates). Although the published literature recommends that response rates should reach the criterion of 60-80% when SET data will be used for high-stakes decisions (Berk, 2012, 2013; Nulty, 2008; Seldin and Associates, 2006), our findings did not produce a significant reduction in SET score variability. Nulty (2008) argues that acceptable response rates depend in part on class enrollment. For example, Nulty argues that the average SET score for a large class (one enrolling 50 or more students) might achieve a sampling error as small as 10% with a sample as small as 17 (a 35% response rate). Nulty's calculations assume random selection and a representative sample of responses (i.e., students who complete the SET do not differ in important ways from students who do not submit an SET). Decision makers must also consider whether systematic differences between respondents and non-respondents produce response biases that might undermine the representativeness of the sample and the interpretability of the findings.

Implications for Practice

When decision makers use SET data to make high-stakes decisions (faculty hires, annual evaluations, tenure, promotion, teaching awards), institutions would be wise to take steps to ensure that SETs have acceptable response rates. Berk (2013) and others (Dommeyer et al., 2004; Jaquett, VanMaaren, & Williams, 2016; Nulty, 2008) discuss effective strategies to improve response rates for SETs. These strategies include offering empirically validated incentives, creating high-quality technical systems with good human factors characteristics, and promoting an institutional culture that clearly supports the use of SET data and other information to improve the quality of teaching and learning. Incentives might include early access to end-of-term grades, access to summary data for SETs when registering for classes, extra credit when a target percentage of students in the class complete the SET, and lotteries for prizes awarded to students who submit forms. Programs and instructors must discuss why information from SETs is important for decision-making and provide students with tangible evidence for how SET information guides decisions about curriculum improvement. Instructors can show that they value feedback from students when they conduct mid-course evaluations, discuss the findings with their students, and implement reasonable changes to course activities and teaching strategies based on student feedback. Technical support for the administration of online SETs should ensure that the software employed is user-friendly. Online systems should be convenient and easy for students to access, the instructions should be clear, the system should operate reliably, and the institution should provide students with compelling evidence that the administration system protects the confidentiality of their responses.

In addition to ensuring adequate response rates on SETs, decision makers should demand multiple sources of evidence about teaching quality. High-stakes decisions should never rely exclusively on numeric data from SETs. Reviewers often treat SET ratings as a surrogate for a measure of the impact an instructor has on student learning. However, a recent meta-analysis (Uttl, White, & Gonzalez, 2016/in press) questions whether SET scores have any relation to student learning. SETs can be useful because they provide insight into how students experience the teaching of an instructor. However, student feedback about teaching is limited by the student's expertise in evaluating the rigor of course content or the validity of instructor-created assessments of student learning. Therefore, SETs provide no direct evidence of student learning. Linse (2017) provides a useful analysis of misconceptions about SET scores and provides guidelines for how faculty and administrators should interpret SET scores. Reviewers need evidence in addition to SET ratings to evaluate teaching. This additional information entails evaluating disciplinary content expertise, skill with classroom management, the ability of the instructor to engage learners with lectures or other activities, impact on student learning, or evidence of success with efforts to modify and improve courses and teaching strategies (Berk, 2013; Hativa, 2013; Seldin & Associates, 2006; Stark & Freishtat, 2014). As with other forms of assessment, any one measure may be limited in terms of the quality of information it provides. Therefore, multiple measures are more informative than any single measure.

A portfolio of evidence can better inform high stakes decisions (Berk, 2013; Seldin and Associates, 2006). Portfolios might include summaries of class observations by senior faculty, the chair, and/or peers. Examples of assignments and exams can document the rigor of learning (especially if accompanied by redacted samples of student work). Course syllabi can identify intended learning outcomes, describe instructional strategies that reflect the rigor of the course (required assignments and grading practices), and provide other information about course content, design, instructional strategies, and how the instructor interacts with students (Palmer, Bach, & Streifer, 2014; Stanny, Gonzalez, & McGowan, 2015).

Psychology has a long history of devising creative strategies to measure the “unmeasurable,” whether the targeted variable is a mental process, an attitude, or the quality of teaching (e.g., Webb, Campbell, Schwartz, & Sechrest, 1966). In addition, psychologists have documented various heuristics and biases that contribute to the misinterpretation of quantitative data (Gilovich, Griffin, & Kahneman, 2002), including the misinterpretation of SET scores (Boysen, 2015a and b; Boysen et al., 2014). These skills enable psychologists to offer multiple solutions to the challenge posed by the need to objectively evaluate the quality of teaching and the impact of teaching on student learning (e.g., Seldin & Associates, 2006; Seldin, Miller, et al., 2010).

Online administration of SET forms presents multiple desirable features, including rapid feedback to instructors, economy, and support for environmental sustainability. However, institutions should adopt implementation procedures that do not undermine the usefulness of the data gathered. Moreover, institutions should be wary of emphasizing procedures that produce high response rates only to lull faculty into believing that SET data can be the primary (or only) metric used for high-stakes decisions about the quality of faculty teaching. Instead, decision makers should expect to use multiple measures to evaluate the quality of faculty teaching.

References

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: does an online delivery system influence student evaluations? *The Journal of Economic Education, 37,* 21-37. https://dx.doi.org/10.3200/JECE.37.1.21-37

Berk, R. A. (2012). Top 20 strategies to increase the online response rates of student rating scales. *International Journal of Technology in Teaching and Learning, 8,* 98-107.

Berk, R. A. (2013). *Top 10 flashpoints in student ratings and the evaluation of teaching.* Sterling, VA: Stylus.

Boysen, G. A. (2015a). Significant interpretation of small mean differences in student evaluations of teaching despite explicit warning to avoid overinterpretation. *Scholarship of Teaching and Learning in Psychology, 1,* 150-162. https://dx.doi.org/10.1037/stl0000017

Boysen, G. A. (2015b). Preventing the overinterpretation of small mean differences in student evaluations of teaching: an evaluation of warning effectiveness. *Scholarship of Teaching and Learning in Psychology, 1,* 269-282. https://dx.doi.org/10.1037/stl0000042

Boysen, G., A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education, 39*, 641-656. https://dx.doi.org/10.1080/02602938.2013.860950

Brown, M. B., & Forsythe, A. B. (2012). Robust tests for the equality of variances. *Journal of the American Statistical Association, 69,* 364-376. https://doi.org/10.1080/01621459.1974.10482955

Buller, J. L. (2012). *Best practices in faculty evaluation: A practical guide for academic leaders.* San Francisco, CA: Jossey-Bass.

Cates, W. M. (1993). A small-scale comparison of the equivalence of paper-and-pencil and computerized versions of student end-of-course evaluations. *Computers in Human Behavior, 9,* 401-409. https://dx.doi.org/10.1016/0747-5632(93)90031-M

Dewar, J. M. (2011). Helping stakeholders understand the limitations of SRT data: are we doing enough? *Journal of Faculty Development, 25,* 40-44.

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. 2004. Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education, 29,* 611-623. https://dx.doi.org/10.1080/02602930410001689171

Dommeyer, C. J., Bum, P., & Hanna, R. W. (2002). College students' attitudes toward methods of collecting teaching evaluations: in-class versus on-line. *Journal of Education for Business, 78,* 11-15. https://dx.doi.org/10.1080/08832320209599691

Feistauer, D. & Richter, T. (2016). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education.* https://dx.doi.org/10.1080/02602938.2016.1261083

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.) 2002. *Heuristics and biases: The psychology of intuitive judgment.* New York, NY: Cambridge University Press.

Griffin, T. J., Hilton, J., III., Plummer, K., & Barret, D. (2014). Correlation between grade point averages and student evaluation of teaching scores: taking a closer look. *Assessment & Evaluation in Higher Education, 39,* 339-348. https://dx.doi.org/10.1080/02602938.2013.831809

Hativa, N. (2013). *Student ratings of instruction: recognizing effective teaching (2nd ed).* [S.I.]: Oron Publications.

Jaquett, C. M., VanMaaren, V. B., & Williams, R. L. (2017). Course factors that motivate students to submit end-of-course evaluations. *Innovative Higher Education, 42*, 19-31. https://dx.doi.org/10.1007/s10755/016-9368-5

Jaquett, C. M., VanMaaren, V. G., & Williams, R. L. (2016). The effect of extra-credit incentives on student submission of end-of-course evaluations. *Scholarship of Teaching and Learning in Psychology, 2,* 49-61. https://dx.doi.org/10.1037/stl0000052

Johnson, T. (2002). Online student ratings: will students respond? Paper presented at the annual conference of the American Educational Research Association, New Orleans, LA. [ERIC Document Reproduction Service ED 465794]

Layne, B., H., DeChristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education*, *40,* 221-232. https://dx.doi.org/10.1023/A:1018738731032

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation.* https://dx.doi.org/10.1016/j.stueduc.2016.12.0004

Marsh, H. W. (2007). Students' evaluations of university teaching: a multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: an evidence-based perspective* (pp. 319-384). New York: Springer.

Miller, R. I. (1987). *Evaluating faculty for promotion and tenure.* San Francisco, CA: Jossey-Bass.

Morrison, R. (2011). A comparison of online versus traditional student end-of-course critiques in resident courses. *Assessment & Evaluation in Higher Education*, *36*, 627-641. https://dx.doi.org/10.1080/02602931003632399

Nowell, C., Gale, L. R., & Handley, B. (2010). Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. *Assessment & Evaluation in Higher Education, 35,* 463-475. https://dx.doi.org/10.1080/02602930902862875

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & Evaluation in Higher Education, 33,* 301-314. https://dx.doi.org/10.1080/02602930701293231

Palmer, M. S., Bach, D. J., & Streifer, A. C. (2014). Measuring the promise: A learning-focused syllabus rubric. *To improve the academy: A journal of educational development*, *33,* 14-36. https://dx.doi.org/10.1002/tia2.20004

Reiner, C. M., & Arnold, K. E. (2010). Online course evaluation: student and instructor perspectives and assessment potential. *Assessment Update, 22*, 8-10. https://dx.doi.org/10.1002/au

Risquez, A., Vaughan, E., & Murphy, M. (2015). Online student evaluations of teaching: what are we sacrificing for the affordances of technology? *Assessment & Evaluation in Higher Education*, *40,* 210-134. https://dx.doi.org/10.1080/02602938.2014.890695

Seldin, P., & Associates (Eds). (2006). *Evaluating faculty performance: A practical guide to assessing teaching, research, and service.* Bolton, MA: Anker.

Seldin, P., Miller, J. E., & Seldin, C. A. & Associates. (2010). *The teaching portfolio: A practical guide to improved performance and promotion/tenure decisions.* San Francisco: Jossey-Bass.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Review of Educational Research, 83,* 598-642. https://dx.doi.org/10.3102/0034654313496870

Stanny, C. J., Gonzalez, M., & McGowan, B. (2015) Assessing the culture of teaching and learning through a syllabus review, *Assessment and Evaluation in Higher Education, 40,* 898-913. https://dx.doi.org/10.1080/02602938.2014.956684

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research.* https://dx.doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Stowell, J. R., W. E. Addison, and J. L. Smith (2012). Comparison of online and classroom-based student evaluations of instruction. *Assessment & Evaluation in Higher Education, 37*, 465-473. https://dx.doi.org/10.1080/02602938.2010.545869

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76,* 105-110. https://doi.org/10.1037/h0031322

Uttl, B., White, C. A., & Gonzalez, D. W. (2016/*in press*). Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation.* https://dx.doi.org/10.1016/j.stueduc.2016.08.007

Venette, S., Sellnow, D., & McIntyre, K. (2010). Charting new territory: assessing the online frontier of student ratings of instruction. *Assessment & Evaluation in Higher Education, 35*, 101-115. https://dx.doi.org/10.1080/020602930802618336

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: nonreactive research in the social sciences.* Chicago, IL: Rand McNally.

Figure Title

Figure 1. A scatterplot depicting the correlation between response rate and evaluation ratings during the 2014 fall academic term.

Footnotes

[1] A Greenhouse-Geisser adjustment of the degrees of freedom was performed in anticipation of a sphericity assumption violation.

[2] A test of the homogeneity of variance assumption revealed no statistically significant difference in response rate variance between the two delivery modes for the first, second, and third years.

[3] A Greenhouse-Geisser adjustment of the degrees of freedom was performed in anticipation of a sphericity assumption violation.

[4] A test of the homogeneity of variance assumption revealed no statistically significant difference in evaluation rating variance between the two delivery modes, collapsed across years.

Table 1

Mean and Standard Deviations for Response Rates (Course Delivery Method by Evaluation Year)

Course Delivery Method	Face-to-Face		Online	
	M	*SD*	*M*	*SD*
Administration Year				
Year 1 (2012)[a]	71.72	16.42	32.93	15.73
Year 2 (2013)	72.31	14.93	32.55	15.96
Year 3 (2014)	47.18	20.11	41.60	18.23

[a] SETs were administrated in two modalities in Years 1 and 2 (paper-based for face-to-face courses and online for online courses). SETs were administered online for all courses in Year 3.

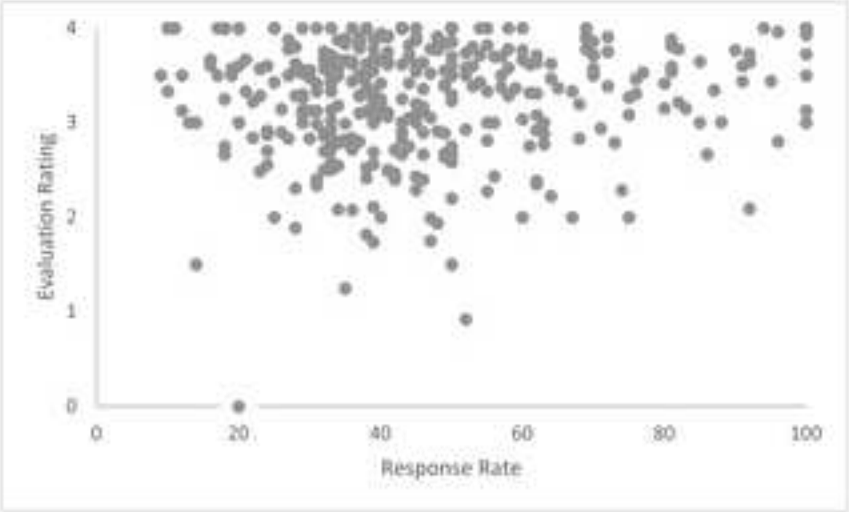

Figure 1. A scatterplot depicting the correlation between response rate and evaluation ratings during the 2014 fall academic term.