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

Claudia J. Stanny1,\* and James E. Arruda2

1 Center for University Teaching, Learning, and Assessment, University of West Florida

2 Department of Psychology, University of West Florida

\* Corresponding author: Claudia J. Stanny1 , , cstanny@institution.edu.

**Abstract (Maximum 250 words)**

When institutions administer student evaluations of teaching (SETs) online, response rates are lower relative to paper-based administration. We 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 3rd year changed the response rate or the average SET score. To control for instructor characteristics, we based the data analysis 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 2 years, the magnitude of this change was minimal (0.11 on a five-item Likert-like scale). We discuss practical implications of these findings for interpretation of SETs and the role of SETs in the evaluation of teaching quality.

***Keywords****:* college teaching; student evaluations of teaching; online administration; response rate; assessment

1. **Introduction**

Student ratings and evaluations of instruction have a long history as sources of information about teaching quality (Berk, 2013). Student evaluations of teaching (SETs) 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 et al., 2014; Nulty, 2008; Spooren et al., 2013). They have also examined how decision makers evaluate SET scores (Boysen, 2015a, 2015b; Boysen et al., 2014; Dewar, 2011). In the last 20 years, considerable attention has been directed toward the consequences of administering SETs online (Morrison, 2011; Stowell et al., 2012) because low response rates may have implications for how decision makers should interpret SETs.

1. **Text body (Background& Literature Survey, Research Methodology, Method & Materials, Results, Discussions, Conclusions, References, etc.)**

In the present study we examined patterns of responses for online and paper-based SET scores at a midsized, regional, comprehensive university in the United States. 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? Whereas much earlier research relied on small samples often limited to a single academic department, we gathered SET data on a large sample of courses (N = 364) that included instructors from all colleges and all course levels over 3 years. We controlled for individual differences in instructors by limiting the sample to courses taught by the same instructor in all 3 years. The university offers nearly 30% of course sections online in any given term, and these courses have always administered online SETs. This allowed us to examine 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 and online for online classes in the first 2 years of data gathered) to uniform use of online forms for all classes in the final year of data collection.

1. **Results**
	1. **Response Rates**

 Response rates are presented in Table 1. The findings indicate that response rates for face-to-face courses were much higher than for online courses, but only when face-to-face course evaluations were administered in the classroom. 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]](#footnote-1) The strength of the overall interaction effect was .22 (ηp2). Simple main-effects tests revealed statistically significant differences in the response rates for face-to-face courses and online courses for each of the 3 observation years.[[2]](#footnote-2) The greatest differences occurred during Year 1 (*p* < .001) and Year 2 (*p* < .001), when evaluations were administered on paper in the classroom for all face-to-face courses and online for all online courses. 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 small (ηp2 = .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.

* 1. **Evaluation Ratings**

 The same 2 × 3 × 3 analysis of variance model 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 = 0.18, p = .03 (ηp2 = .01; see Footnote 1). Evaluation ratings associated with the Year 3 administration (M = 3.26, SD = 0.60) were significantly lower than the evaluation ratings associated with both the Year 1 (M = 3.35, SD = 0.53) and Year 2 (M = 3.38, SD = 0.54) administrations. Thus, all courses received lower SET scores in Year 3, regardless of course delivery method and course level. However, the size of this effect was small (the largest difference in mean rating was 0.11 on a five-item scale). The second statistically significant main effect involved delivery mode, *F*(1, 358) = 23.51, *MSE* = 0.52, *p* = .01 (ηp2 = .06; see Footnote 2). Face-to-face courses (*M* = 3.41, *SD* = 0.50) received significantly higher mean ratings than did online courses (*M* = 3.13, *SD* = 0.63), regardless of evaluation year and course level. No other factors or interactions included in the analysis were statistically reliable.

* + 1. Stability of Ratings

 The scatterplot presented in Figure 1 illustrates the relation between SET scores and response rate. Although the correlation between SET scores and response rate 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 to evaluate the variability of SET scores above and below the 60% response rate, which several researchers have recommended as an acceptable threshold for response rates (Berk, 2012, 2013; Nulty, 2008). The variability of scores above and below the 60% threshold was not statistically reliable, *F*(1, 362) = 1.53, *p* = .22.

**Table 1:** Means and Standard Deviations for Response Rates (Year)

|  |  |  |
| --- | --- | --- |
| Administration year | Face-to-face course | Online course |
| *M* | *SD* | *M* | *SD* |
| Year 1: 2012 | 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 |

*Note. Student evaluations of teaching (SETs) were administered 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.*

1. **Discussion**

 Online administration of SETs in this study was associated with lower response rates, yet it is curious that 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 in response rates for face-to-face and online courses might be attributed to characteristics of the students who enrolled or 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 2 years, the magnitude of the numeric difference was small (differences ranged from 0.08 to 0.11, based on a 0–4 Likert-like scale). This difference is similar to the differences Risquez et al. (2015) reported for SET scores after statistically adjusting for the influence of several potential confounding variables. A substantial literature has discussed the appropriate and inappropriate interpretation of SET ratings (Berk, 2013; Boysen, 2015a, 2015b; Boysen et al., 2014; Dewar, 2011; Stark & Freishtat, 2014).



**Figure 1:** Scatterplot Depicting the Correlation Between Response Rates and Evaluation Ratings

Note. Evaluation ratings were made during the 2014 fall academic term.

1. **Conclusion**

 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 et al., 1966). In addition, psychologists have documented various heuristics and biases that contribute to the misinterpretation of quantitative data (Gilovich et al., 2002), including SET scores (Boysen, 2015a, 2015b; 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.

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.

**Acknowledgment**

The research leading to these results has received Research Project Grant Funding from the xxxxxx, Research Grant Agreement No. GRANT AA E2 /01010. The authors would like to acknowledge support from the xxxxxx. Or The research leading to these results has received no Research Grant Funding.

**Author contribution:** All authors have contributed, read, and agreed to the published version of the manuscript results.

 **Conflict of interest:** The authors declare no conflict of interest.

**References (APA style 7th Edition- font size is 10. 1 space).**

1. Al-Hatmi, M. O., & Yousif, J. H. (2017). A review of Image Enhancement Systems and a case study of Salt &pepper noise removing. International Journal of Computation and Applied Sciences (IJOCAAS), 2(3), 171-176.
2. Alighaleh, P., Khosravi, H., Rohani, A., Saeidirad, M. H., & Einafshar, S. (2022). The detection of saffron adulterants using a deep neural network approach based on RGB images taken under uncontrolled conditions. Expert Systems with Applications, 198, 116890.
3. Alkishri, W., Widyarto, S., Yousif, J. H., & Al-Bahri, M. (2023). Fake Face Detection Based on Colour Textual Analysis Using Deep Convolutional Neural Network. Journal of Internet Services and Information Security (JISIS), 13(3), DOI: 10.58346/JISIS.2023.I3.009.
4. Atomi, W. H. (2012). The effect of data preprocessing on the performance of artificial neural networks techniques for classification problems (Doctoral dissertation, Universiti Tun Hussein Onn Malaysia).
5. Calderon, S., Fallas, F., Zumbado, M., Tyrrell, P. N., Stark, H., Emersic, Z., ... & Solis, M. (2018, October). Assessing the impact of the deceived non local means filter as a preprocessing stage in a convolutional neural network based approach for age estimation using digital hand x-ray images. In 2018 25th IEEE International Conference on Image Processing (ICIP) (pp. 1752-1756). IEEE.



1. A Greenhouse–Geisser adjustment of the degrees of freedom was performed in anticipation of a sphericity assumption violation. [↑](#footnote-ref-1)
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 1st, 2nd, and 3rd years. [↑](#footnote-ref-2)