HOW DO STUDENTS AND EDUCATORS INTERPRET STUDENT EVALUATIONS OF TEACHING?

by

Katherine Neely

Submitted in partial fulfillment of the
Requirements for Departmental Honors in
the Department of Psychology
Texas Christian University
Fort Worth, Texas

May 6, 2019

HOW DO STUDENTS AND EDUCATORS INTERPRET STUDENT EVALUATIONS OF TEACHING?

Project Approved:

Supervising Professor: Uma Tauber, Ph.D.
Department of Psychology

Gary Boehm Ph.D.
Department of Psychology

Shauna McGillivray, Ph.D.
Department of Biology

ABSTRACT

Student evaluations of teaching (SETs) are a tool commonly employed at universities for assessing faculty members' teaching performance and eligibility for promotions. Survey items often ask students to make judgments about the professor's knowledgeability, teaching style, and class difficulty. Fair and consistent review of SETs is critical for faculty members as they seek to improve their teaching and gain professional recognition. The present study investigates the novel question of how SETs are interpreted. Undergraduate students and faculty participants were shown and asked to make judgments about a SET for a fictional professor. The four conditions varied in whether the fictional professor was rated lower or higher than the departmental average, and whether or not the professor gave daily in-class quizzes. There were no differences between student and faculty member responses overall, and the magnitude of the fictional professor's ratings had a significant impact on participants' judgments. Quizzing did not cause a significant difference in participants' ratings of the professor despite students and faculty both rating quizzing as beneficial for learning and increasing class difficulty.

How do Students and Educators Interpret Student Evaluations of Teaching?

Student evaluations of teaching (SETs) are a tool commonly employed at universities for gathering students' opinions in order to assess faculty members' teaching performance. SETs provide students with an outlet for their opinions by asking them to make judgments about professors' level of knowledge, teaching style, and class difficulty. Typically, ratings are made on a Likert-type scale ranging from 1 (very poor) to 5 (very good), and students can also leave open-ended comments to describe and explain their responses to the close-ended questions. My interest is in evaluating how SETs are interpreted with a particular focus and determine if students and faculty interpret SETs differently. As important, given the paucity of research on interpretations of SETs, this project was designed to explore the impact of the magnitude of a professor's ratings and the presence of quizzing on participants' interpretations.

SETs have become increasingly popular in recent decades, both in their frequency of use and in the significance of their impact on administrative decisions. A survey of 598 deans of four-year liberal arts colleges reported that they gave more weight to SETs when evaluating faculty members' teaching performance than exam scores or classroom observation (Seldin, 1998). A survey of a similar population of deans found that reliance on student ratings of teaching increased from 88.1% to 94.2% over a ten-year period (Miller & Seldin, 2014). According to Spooren, Brockx, and Mortelmans (2013), SETs provide two primary functions: (a) to help improve teaching quality and (b) to gather student input for faculty appraisal, such as considerations for promotions and tenure decisions. For this process to be functional, students must make fair and accurate judgments of their professors and administrators must in turn make accurate judgments about professors based on their student evaluations.

Research has shown that a number of factors unrelated to teaching quality can bias students' evaluations of teaching. A common criticism of SETs, particularly from faculty members themselves, is that they may assess professors' popularity and personality more so than teaching effectiveness. There is evidence to support this claim. Clayson and Sheffet (2006) found that students' perceptions of a teacher's personality are established very early on in a term, and first impressions of personality could be used to strongly predict course evaluation ratings 16 weeks later. As a consequence, SETs fail to reflect good teaching practices. As well, grading leniency, associated with "grade inflation," is a serious concern in higher education. Students who hold the belief that a professor is a lenient grader will typically give that professor globally higher ratings, so perceived grading leniency has been found to be positively correlated with teaching ratings (Olivares, 2001). Thus, students may give higher evaluations to educators who adopt teaching practices that are unrelated to actual learning, relative to those who do not.

Gender and physical appearance have also been demonstrated to influence SETs. While male professors' evaluations tend to be unaffected by the gender of students, female professors can receive significantly lower ratings from their male students compared to their female students (Basow, 1995). Furthermore, Rinolo, Johnson, Sherman, and Misso (2006) found that, across many studies and time periods, physically attractive instructors tend to receive higher ratings than those who are judged to be less attractive. This finding is especially pertinent to female instructors, whose looks may be more heavily scrutinized than their male counterparts.

Variables unrelated to the instructors themselves can also play a significant role in students' evaluations. Heckert (2006) found that classes that students were interested in and took place in the afternoon received the highest ratings. A professor's department can also influence their SETs. Certain departments and classes, such as statistics, finance, and business, tend to

culminate in lower evaluation ratings that may be due to the challenging nature of the content instead of to the professor per se (Stapleton & Murkison, 2001). These discrepancies can result in the unintended consequences for promotions based on reasons that are outside of faculty members' control and that can be unrelated to their teaching quality.

The interpretation of completed SETs is another challenge for administrators and faculty tasked with reviewing students' ratings and comments. For instance, reviewers struggle to accurately interpret the significance of small mean differences between instructors' average ratings (Boysen, Kelly, Raesly, & Casner 2014). Teaching evaluations are not precise measurements, so it can be challenging to compare raw averages and make decisions based on this information alone. Furthermore, reviewers can be ineffective at considering all the statistical information necessary for making accurate judgments. Even faculty members with prior statistical knowledge and training are prone to over-interpret small mean differences (Boysen, 2017). Overall, reviewers tend to fall into the heuristic that "bigger is better;" a professor who receives a higher score is better than the professor who receives a lower score, even though the difference between the two scores may not be statistically significant. This research suggests that participants in the present study will be strongly influenced by the magnitude of a professor's ratings.

Specific teaching practices may have an impact on SETs, with quizzing being of particular interest to the present study. There is significant evidence to show that regular quizzing leads to increased retention of information (Roediger & Karpicke, 2006). In practice, students who engage in self-testing methods in which they practice recalling information will remember studied information better than students who simply re-read the information. This phenomenon is collectively referred to as the testing effect (Rowland, 2014). Although quizzing

has substantial support as an effective method to promote learning, its impact on SETs has not been researched. One possibility is that students will view quizzing as an additional challenge and workload to a class that will cause them to hold a more negative perception of a professor. This hypothesis would be in line with Olivares (2001), in which educators were more often rewarded with high SET ratings when they incorporated lenient, less challenging policies in the classroom. Alternatively, SET reviewers (particularly faculty members) may recognize the cognitive benefits of quizzing and consequently reward professors for quizzing by rating them more favorably in terms of teaching efficacy. The present study aims to identify any differences between student and faculty groups when interpreting SETs, particularly with regard to their attitudes toward quizzing in the classroom.

Method

Participants

Student participants were 161 undergraduates from Texas Christian University who received course credit. Students were recruited from the psychology department participant pool. Faculty participants were 120 current professors at American universities who were compensated with a $15 gift card upon completion of the study. To recruit faculty, participating departments at Texas Christian University and Tarrant County College circulated an email requesting volunteers. Each faculty participant was instructed to email the researchers using a university-provided email address to verify their job status as a university faculty member. Survey submissions were screened to ensure participants did not complete the study multiple times.

Participants were randomly assigned to one of four conditions. Specifically, participants reviewed a SET for a professor who (a) had high ratings and gave quizzes (n = 41 students, n = 29 faculty), (b) had high ratings and did not give quizzes (n = 41 students, n = 30 faculty), (c)

had low ratings and gave quizzes (n = 40 students, n = 33 faculty), or (d) had low ratings and did not give quizzes (n = 39 students, n = 28 faculty). The magnitude of Professor W's ratings (either higher or lower than departmental averages) and the presence of quizzing (either daily quizzes or no daily quizzes) were manipulated between-participants.

Procedure

Students completed the survey in a computer lab independently, and faculty were emailed a link to the survey, which they completed on any computer at their own pace. Participants were given the following instructions,

> "In the following task, you will be asked to review the student evaluations for a professor. At the end of the semester, students were asked to rate several of the professor's qualities on a scale from 1 (strongly disagree) to 5 (strongly agree). Students were also encouraged to leave anonymous comments about the professor and the class. You should consider both the professor's average ratings and the students' comments when making your judgments."

Participants then viewed the SET for the fictional Professor W. Participants saw a bar graph that showed how Prof W compared to their department's average ranking in terms of teaching the material effectively, staying organized, and providing help to students outside of class. Some of the participants saw that the professor was rated modestly higher than average on these three qualities, and others saw that the professor was rated lower than average (see Appendix A).

Below that, participants saw five student comments. These comments served to inform the participants of whether or not the professor gave daily quizzes. They were intentionally written to provide no additional information about the quality of the professor's teaching or the class.

For example, a comment in the quizzing-present condition read, “I always attended lecture and this helped me make lots of notes to study off of for the quizzes.” The analogous comment in the quizzing-absent condition was, “Even though there weren’t daily quizzes, I always attended lecture and this helped me make lots of notes to study” (see Appendix A).

Next, participants responded to questions by reporting the degree to which their ratings of Professor W were influenced by the magnitude of his/her student evaluations, student comments, and quizzing. These statements were also rated from 1-5 (see Appendix A).

To ensure that each participant recognized the essential information about the professor, we included two manipulation check questions. Specifically, participants recalled if the professor they rated had higher or lower ratings relative to the departmental average and if the professor gave quizzes (see Appendix B). Finally, participants answered demographic questions and questions about their attitudes toward SETs and quizzing. The entire task was self-paced and questions were presented in a fixed order.

This research was approved by the IRB at Texas Christian University and all participants were treated ethically.

Results

Ratings of Professor W

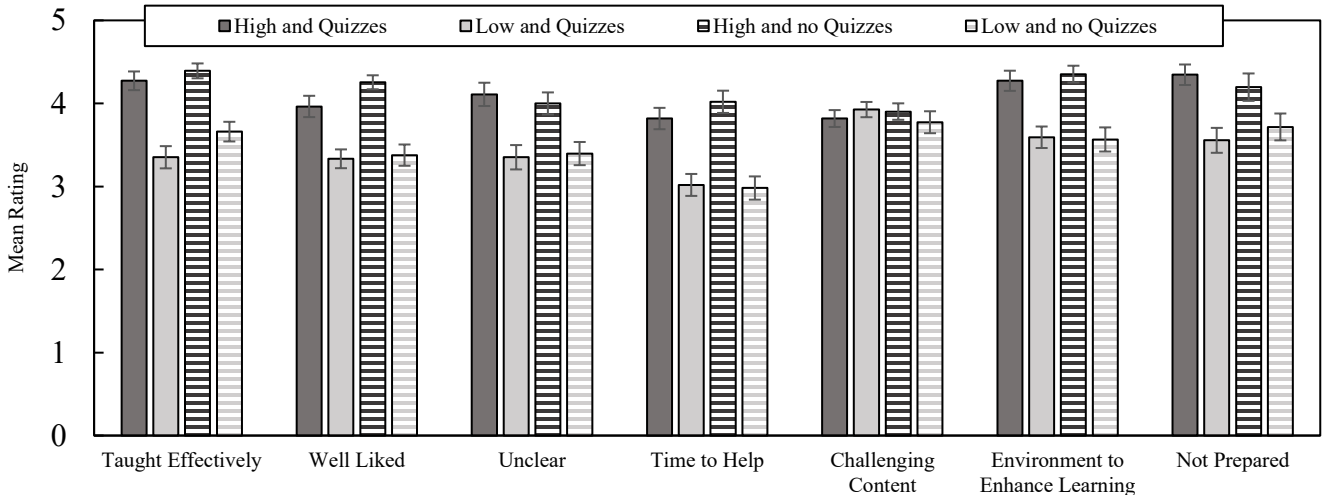


Figure 1. Combined student and faculty responses for Ratings of Professor W. Error bars represent standard error.

Two statements about Professor W were reverse scored: “Often, Professor W was unclear about what was expected from students.” and “Professor W was not usually prepared for class.” This ensured that a high score on these two questions, like the other questions in “Ratings of Professor W,” represented a positive quality in Professor W (see Appendix B).

A 2 (Participant status: Student, Faculty) x 2 (Magnitude of Professor W’s ratings: above departmental average, below departmental average) x 2 (Quizzes: present, absent) between-participants analysis of variance (ANOVA) was conducted for each of the seven questions represented in Figure 1. The *p*-value was adjusted to correct for the number of questions (Bonferroni correction), making the cut-off $p = .007$. For all questions, there was no significant main effect of group. Faculty and student responses did not significantly differ on any of the questions shared by both groups (Q1-Q7, see Appendix B). For example, faculty and student responses were statistically equivalent for Q1 (Students: $M = 3.91$, $SE = .08$; Faculty: $M = 3.93$, $SE = .09$; $F < 1$, $p = .86$). As such, data are collapsed across participant status (i.e., student or faculty) in all figures.

The magnitude of Professor W’s ratings had a consistent impact on participants’ responses. Specifically, both students and faculty gave higher ratings for six out of the seven questions when presented with high ratings of Professor W compared with those given low ratings of Professor W. For Q1, for example, participants who were given high faculty ratings for Professor W gave significantly higher responses ($M = 4.33$, $SE = .07$) than did participants who were given low faculty ratings for Professor W, ($M = 3.50$, $SE = .09$), $t(211) = 7.19$, $p < .001$. For Q5, “Professor W’s course content was sometimes challenging,” the magnitude of ratings did not impact responses ($F < 1$, $p = .79$).

Presence of quizzes did not impact responses for any item for either group. To illustrate, participants' responses on Q1 were not significantly affected by the presence or absence of quizzing, $F(1, 205) = 2.74$, $p = .10$, $\eta^2_p = .01$. This was true for all questions in this section, Q1-Q11 (see Appendix B).

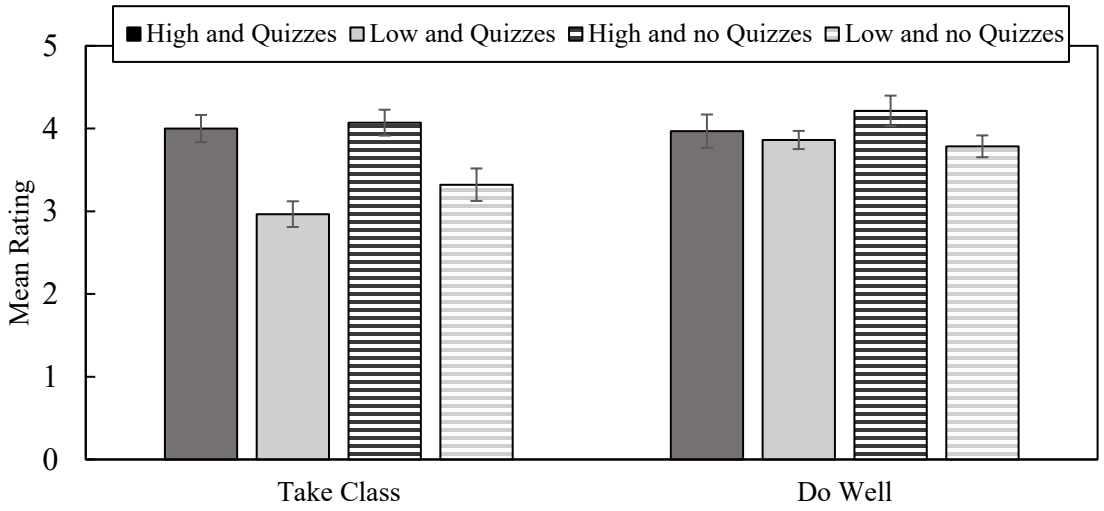


Figure 2. Combined student and faculty responses for Ratings of Professor W. Error bars represent standard error.

Questions shown only to students (Q8 and Q9) were analyzed with two 2 (Magnitude of Professor W's ratings: above departmental average, below departmental average) x 2 (Quizzes: present, absent) between-participants ANOVAs, with the p value adjusted to a cutoff of $p = .025$ (Bonferroni correction). For Q8, a significant main effect of the magnitude of Professor W's ratings arose, $F(1, 113) = 28.85$, $p < .001$, $\eta^2_p = .20$, and there was no main effect of quizzing, $F(1, 113) = 1.66$, $p = .20$, $\eta^2_p = .01$. For Q9, there was no main effect of the magnitude of Professor W's ratings, $F(1, 113) = 2.75$, $p = .10$, $\eta^2_p = .02$, and no main effect of quizzing, $F < 1$, $p = .60$. Thus, students who were given high ratings for Professor W were significantly more likely to report wanting to take a class with the professor and the expectation of doing well in said class, relative to students who were given low ratings for Professor W.

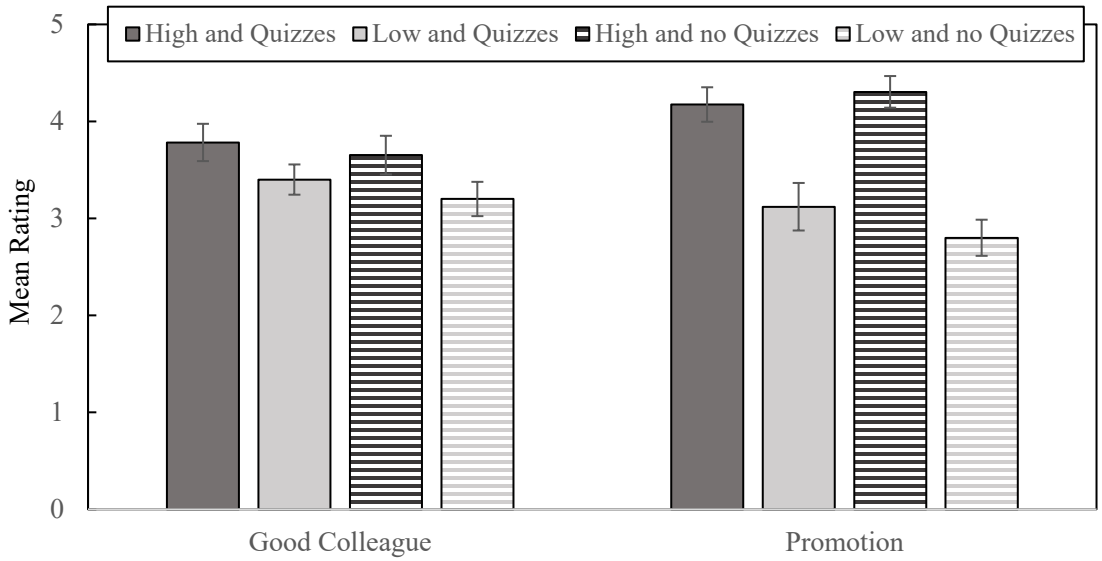


Figure 3. Combined student and faculty responses for Ratings of Professor W. Error bars represent standard error.

Questions shown only to faculty (Q10 and Q11) were also analyzed with two 2 (Magnitude of Professor W's ratings: above departmental average, below departmental average) x 2 (Quizzes: present, absent) between-participants ANOVAs, with the *p* value adjusted to a cutoff $p = .025$ (Bonferroni correction). For Q10, a significant main effect of the magnitude of Professor W's ratings was found, $F(1, 92) = 5.56$, I = .02, $\eta^2_p = .06$, and there was no main effect of quizzing, $F < 1$, $p > .35$. For Q11, a significant main effect of the magnitude of Professor W's ratings was found, $F(1, 92) = 43.55$, $p < .001$, $\eta^2_p = .32$, and there was no main effect of quizzing, $F < 1$, $p = .63$. Thus, faculty who were given high ratings for Professor W were significantly more likely to report that Professor W would be a good colleague and would be likely to promoted relative to faculty who were given low ratings for Professor W.

Factors influencing participants' ratings

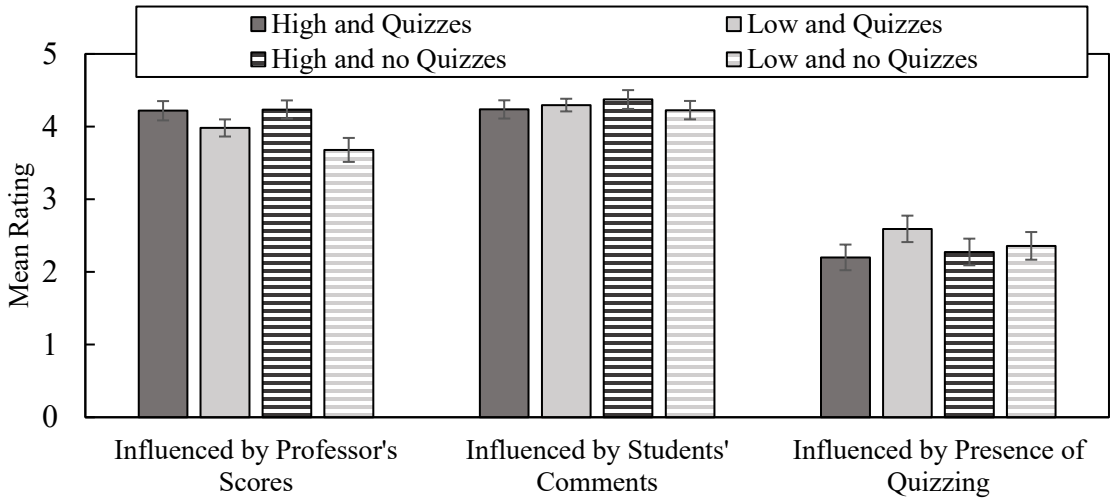


Figure 4. Combined student and faculty responses for Factors influencing participants' ratings. Error bars represent standard error.

Participants reported that they were influenced by the magnitude of Professor W's ratings and by student comments. Participants agreed with questions 12 and 13, and disagreed with question 14 (see Figure 4). There was a significant difference in responses between the three questions, $F(2, 410) = 183.81$, $p < .001$, $\eta^2_p = .47$ No other effects were significant, $p > .140$.

Manipulation check

A total of 68 participants answered at least one manipulation check question incorrectly ($n = 44$ students and 24 faculty). This suggests that these participants did not carefully read the information provided about Professor W. As such, their data were excluded from all analyses (resulting $N = 213$, $n = 117$ students, $n = 96$ faculty). Groups reviewed a professor who had: high ratings and gave quizzes ($n = 32$ students, $n = 23$ faculty), (b) high ratings and did not give quizzes ($n = 28$ students, $n = 23$ faculty), (c) low ratings and gave quizzes ($n = 29$ students, $n = 25$ faculty), or (d) low ratings and did not give quizzes ($n = 28$ students, $n = 25$ faculty).

Demographics

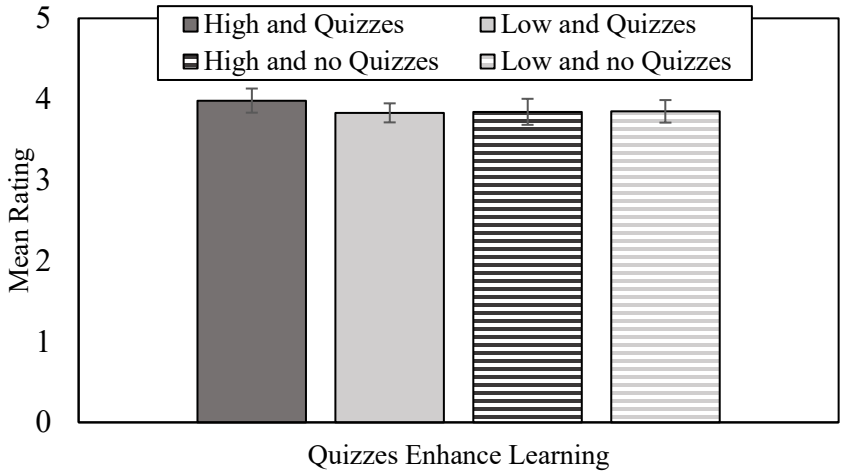


Figure 5. Combined student and faculty responses for Q28, "Taking quizzes in class enhances my/my students' learning." Error bars represent standard error.

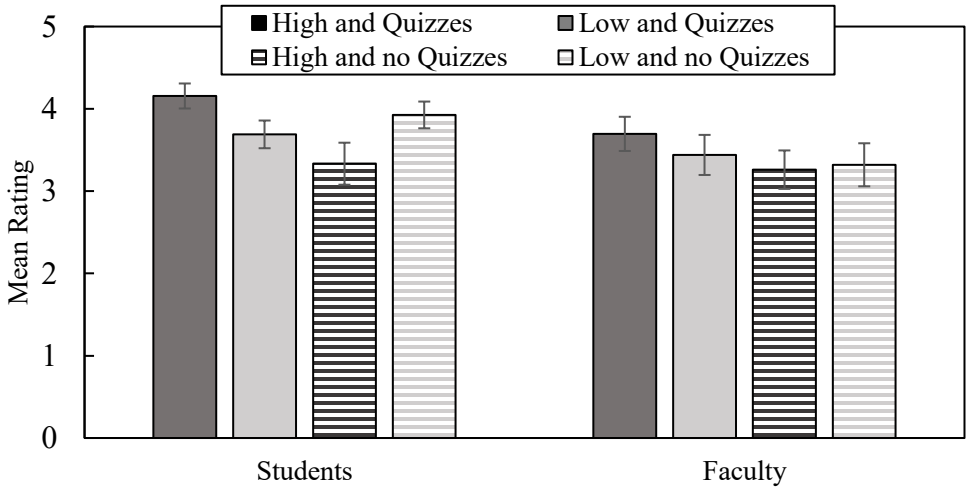


Figure 6. Divided student and faculty responses to Q31, "Quizzes make a class more challenging." Error bars represent standard error.

Several demographic questions were asked about the participants' attitudes toward quizzing in the classroom. Notably, participants in both groups agreed with statement Q28 regarding the beneficial effect of quizzing on learning (F <1) (see Figure 5). Students more strongly agreed with Q31 (M = 3.78, SE = .10) than did faculty (M = 3.43, SE = .11) (see Figure 6).

Discussion

Faculty and student groups, despite their different roles in the classroom setting, did not interpret the SETs differently. Although faculty have different experience with SETs, typically reviewing them instead of completing them, there was no effect of experience on their interpretations. Previous research by Cain et al. (2018) showed that students and faculty do tend to converge in their agreement about a professor's teaching effectiveness on individual class days. Our results suggest that these two groups may continue to agree when interpreting global assessments of teaching effectiveness seen on SETs. Cain et al. (2018) also reported that students and faculty can diverge in their opinions on the level of challenge within a course, particularly that students rated a course more challenging than did faculty near the end of a semester. Although both our student and faculty groups agreed with the statement that taking quizzes make a class more challenging (see Appendix B), students more strongly agreed with the statement than did faculty (see Figure 6). As Cain et al. (2018) suggests, an increased perceived level of challenge may be due to students' fatigue at the end of a semester, and our student sample may have also felt that quizzing was a strenuous practice that challenges them.

Although students reported that quizzing was more challenging than did faculty, both groups were in equal agreement with the statement that taking quizzes enhances learning (see Figure 5). This suggests both groups have an awareness of testing effect. Testing and quizzing, including at-home self-testing practices such as flashcards and practice tests, have robust evidence for enhancing long-term learning (Schwieren, Barenberg, & Dutke, 2017) Our student comments specifically mentioned that Professor W gave "daily quizzes" or did not (see Appendix A). Although daily quizzes come with the valid concerns that they take up class time and cause student fatigue, this practice typically has positive effects on students' classroom

performance and long-term retention of material (Batsell, Perry, Hanley, & Hostetter, 2017). However, our results showed no main effect of quizzing on any question, meaning that our participants did not use quizzing as an influential factor when interpreting Professor W's SETs. Our manipulation check ensured that participants attended to the student comments carefully and remembered whether or not Professor W gave quizzes, so this neglect was not a matter of simple inattention. The participants, instead, assigned much more importance to the magnitude of Professor W's ratings.

The magnitude of ratings had strong effect on our participants' interpretations of Professor W's SETs including questions about teaching effectiveness, clarity, and preparedness. This suggests that, for example, participants who were given information that Professor W was rated higher than average used this information to rate Professor W higher globally relative to those who were given information that Professor W was rated lower than average. Conclusions drawn by Boysen et al. (2014) provide a framework by which to interpret this finding:

> Despite the importance of teaching evaluations and the simplicity of the principles for their interpretation, the current studies illustrate the relative ease with which faculty members and department heads can be led to make inappropriate generalisations from limited data. Despite the absence of information needed for accurate statistical interpretation, small differences in teaching evaluation means significantly impacted faculty and administrators' judgements about the skills and merits of teachers portrayed in fictional vignettes. (p. 653)

Similar to our study, Boysen et al. (2014) found that faculty members were prone to overinterpret the significance of small mean differences between faculty members' ratings on SETs. We did not provide any numerical statistical data (such as means, confidence intervals, or

standard deviation) to accompany the graphs in the SET materials because we did not know our participants' level of statistical knowledge and did not want to complicate the graphs. As such, participants were given no indication of whether Professor W's rating was significantly different from the departmental average, yet our participants consistently rated the higher-rated Professor W significantly higher than the lower-rated Professor W. The absence of statistical information and guidelines is a potential limitation in comparing our study to Boysen et al. (2014), and the effects of including such additional information warrants further investigation.

Our participants may have been more swayed by the magnitude of ratings than quizzing because the ratings data was presented in a bar graph, whereas the quizzing information was embedded within student comments. The visual information may have been easy for participants to interpret and assign value to than the information about quizzing, particularly because Professor W was compared against a standard in the bar graph. This difference in modality was maintained because this is often how results are summarized in real SETs. Although participants showed a general awareness of the testing effect, they did not use this knowledge to assign either positive or negative value to the presence or absence of quizzing. Should this discrepancy translate into real-world settings, this poses the problem that SETs are not being used to accurately represent educators' teaching practices, and administrators may not be paying adequate attention to qualitative information within SETs. Boysen et al. (2014) tested the effectiveness of a warning to be read by administrators prior to reviewing SETs, and found that reviewing explanations about proper statistical interpretation practices could significantly increase the accuracy of administrators' interpretations of data on SETs. Future directions for this research could implement a similar warning aimed at increasing SET reviewers' attention to aspects of a SET beyond professors' average ratings. Specifically, reminding reviewers of the

beneficial effects of quizzing on student learning may prompt reviewers to more favorably assess professors who regularly quiz their students.

A limitation of our research was the self-selection of faculty participants. While students signed up to participate in our study without any knowledge of its topic, faculty participants were recruited with an email that briefly described the study. Those who chose to participate likely held some special interest in the topic. As well, student participants were primarily female first-year undergraduates. This convenience sample was by necessity, but both of these methods of participant selection may have limited the generalizability of our findings. In the future, a larger and more diverse pool of administrators who regularly review SETs should be recruited for a follow-up study. This would increase the ecological validity of the SET interpretations and provide more insight into what factors most strongly influence these interpretations. A better understanding of these factors will contribute to the improvement of SET design, the administration of SETs to students, and the interpretations of SETs.

References

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*(4), 656–665.

Batsell, W. R., Perry, J. L., Hanley, E., & Hostetter, A. B. (2017). Ecological validity of the testing effect: The use of daily quizzes in introductory psychology. *Teaching of Psychology 2017, 44*(1) 18– 23.

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*(6), 641–656.

Boysen, G. A. (2017). Statistical knowledge and the over-interpretation of student evaluations of teaching. *Assessment & Evaluation in Higher Education, 42*(7), 1095–1102.

Cain, K. M., Wilkowski, B. M., Barlett, C. P., Boyle, C. D., & Meier, B. P. (2018). Do we see eye to eye? Moderators of correspondence between student and faculty evaluations of day-to-day teaching. *Teaching of Psychology, 45*(2), 107–114.

Clayson, D. E., & M. J. Sheffet (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education, 28,* 149–60.

Heckert, T. M., Latier, A., Ringwald, A., & Silvey, B. (2006). Relation of course, instructor, and student characteristics to dimensions of student ratings of teaching effectiveness. *College Student Journal, 40*(1), 195–203.

Miller, J. E., & Seldin, P. (2014). Changing practices in faculty evaluation: Can better evaluation make a difference? *American Association of University Professors.* Retrieved from https://www.aaup.org/article/changing-practices-faculty-evaluation

Olivares, O. J. (2001). Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology, 26*, 382–399.

Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006) Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *The Journal of General Psychology, 133*(1), 19–35.

Roediger, H. L. & Karpicke, J. D. (2006). Test-enhanced learning. Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463.

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching, 16*(2), 179–196.

Seldin, P. (1998, February). The teaching portfolio. Paper presented for the American Council on Education, Department Chairs Seminar, San Diego, CA.

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*(4), 598–642. http://dx.doi.org/10.3102/0034654313496870.

Stapleton, R. J., & Murkison, G. (2001). Optimizing the fairness of student evaluations: A study of correlations between instructor excellence, study production, learning production, and expected grades. *Journal of Management Education, 25*, 269–291.

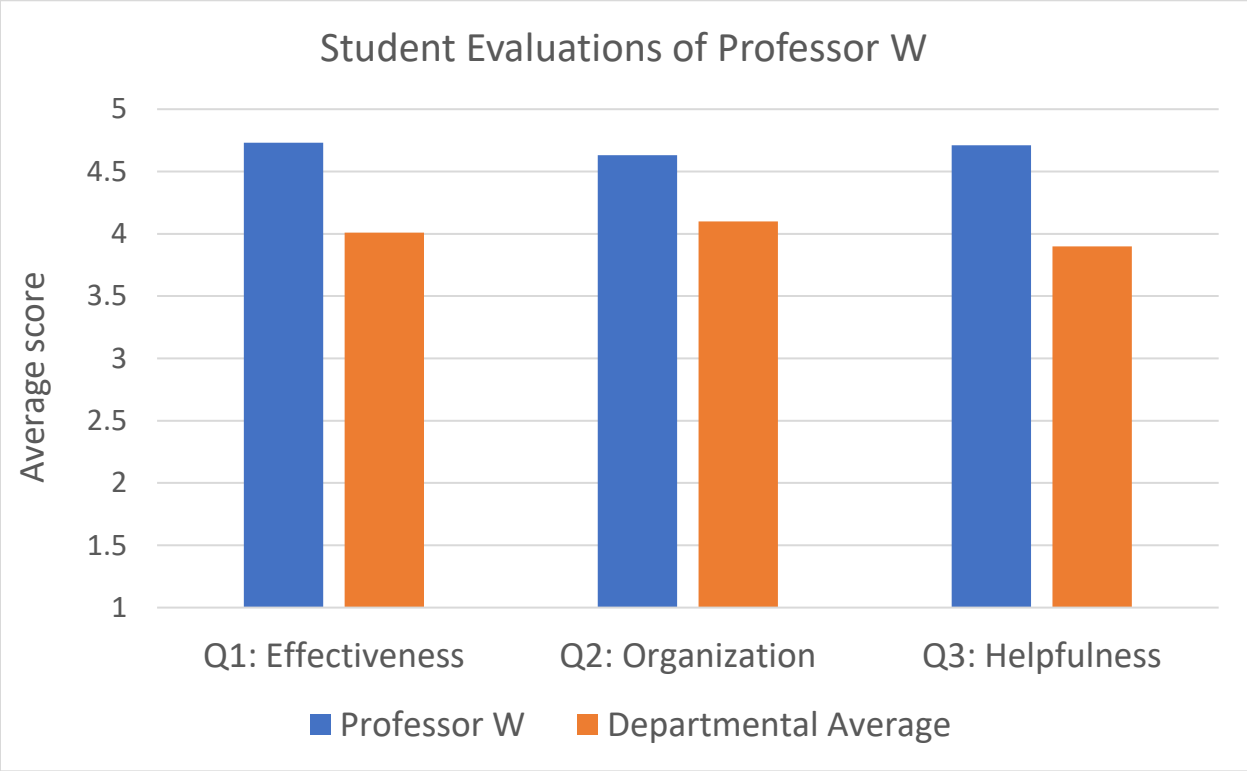
Appendix A

SET of Professor W

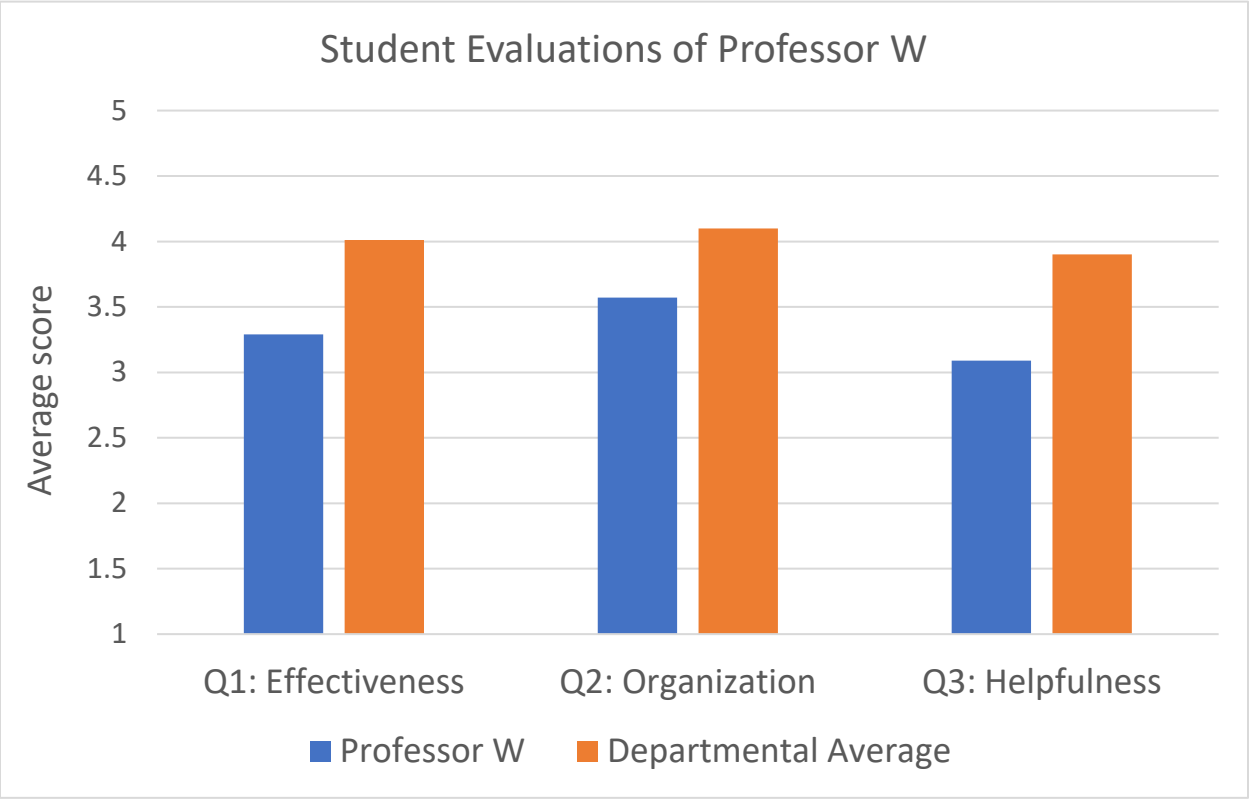
In the following task, you will be asked to review the student evaluations for a professor. At the end of the semester, students were asked to rate several of the professor's qualities on a scale from 1 (strongly disagree) to 5 (strongly agree). Students were also encouraged to leave anonymous comments about the professor and the class. Based on this information, we would like to you to make several ratings about the professor.

Students used this five-point scale to respond to the following questions:
(1) Strongly disagree (2) Disagree (3) Neither agree nor disagree (4) Agree (5) Strongly agree
Q1: Overall, my professor taught the material effectively.
Q2: My professor organized materials and class time well.
Q3: I felt welcome to seek my professor's help outside of class.

Below you will find Professor W's average ratings for each question as well as average ratings for those in Professor W's department.



Note: this graph was shown only to participants in the high-rated conditions



Note: this graph was shown only to participants in the low-rated conditions

Students wrote the following comments about Professor W:

- We got a lot done in class, and I formed a study group with some of my classmates. This helped me prepare for our quizzes.
- I didn't think I'd like the material because this class isn't for my major, but it was actually kind of interesting stuff.
- The daily quizzes were manageable. But it was kind of annoying that there was so much content to study.
- Just another class I'm taking to graduate and I'm happy with a B.
- I always attended lecture and this helped me make lots of notes to study off of for the quizzes.

Note: these comments were shown only to participants in the quizzing-present conditions

Students wrote the following comments about Professor W:

· We got a lot done in class, and I formed a study group with some of my classmates. This helped me stay up on the material.
· I didn't think I'd like the material because this class isn't for my major, but it was actually kind of interesting stuff.
· There were no quizzes. But, it was kind of annoying that there was so much content to study.
· Just another class I'm taking to graduate and I'm happy with a B.
· Even though there weren't daily quizzes, I always attended lecture and this helped me make lots of notes to study.

Note: these comments were shown only to participants in the quizzing-absent conditions

Appendix B
Survey Questions

Ratings of Professor W

Taking this information into consideration, please rate the extent to which you agree or disagree with the following statements:

Q1. Professor W taught the class effectively.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

Q2. Professor W is well-liked by the students.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

Q3. Often, Professor W was unclear about what was expected from students.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

Q4. When students needed help, Professor W would make extra time to help them.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

Q5. Professor W's course content was sometimes challenging.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

Q6. Professor W created a classroom environment that helped students learn.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree

Strongly agree

Q7. Professor W was not usually prepared for class.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Q8. I would elect to take this class taught by Professor W.*

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Q9. I would probably do well in this class taught by Professor W.*

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Q10. Professor W would be a good colleague.**

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Q11. Professor W would be looked on favorably for a promotion.**

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

** indicates questions shown only to students, ** indicates questions shown only to faculty*

Factors influencing participants' ratings

Think back to the questions you just answered. Consider which aspects of Professor W's evaluation affected your opinions about Professor W.

Q12. My ratings were influenced by Professor W's average scores in the graph.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree

Strongly agree

Q13. My ratings were influenced by the students' written comments about Professor W.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Q14. My ratings were influenced by whether or not Professor W gave quizzes.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Manipulation check

Q15. How did Professor W's average ratings on the bar graph compare with the departmental averages?

Professor W's ratings were lower
Professor W's ratings were the same
Professor W's ratings were higher
Unsure

Q16. Did Professor W give quizzes?

Yes
No
Unsure

Demographics

These final questions are demographic in nature. We will use this information for research purposes only, and it will be kept strictly confidential. Your name will not be linked with any of this information. Please answer each question to the best of your ability.

Q17. What is your age? _____

Q18. Which year are you in your undergraduate studies? Choose the answer that most closely describes you.*

First year
Second year
Third year
Fourth year
Fifth year or more
Prefer not to respond

Q19. What is your gender?

Male
Female
Other _____
Prefer not to respond

Q20. How do you identify your race/ethnicity? Choose one or more that apply to you.

Asian
Black/African
Caucasian/White
Hispanic/Latinx
Native American
Pacific Islander
Other _____
Prefer not to respond

Q21. Is English your first language?

Yes
No
Prefer not to respond

Q22. What is your best estimate of your current GPA?* _____

Q23. What is your major?* _____

Q24. At which institution(s) do you currently teach?** _____

Q25. For which department(s) do you teach?** _____

Q26. How often do you complete evaluations for your professors at the end of the semester?*

Never
Sometimes
About half the time
Most of the time
Always

Q27. About how often do your professors give quizzes?* / About how often do you give your students quizzes?**

Never
About once or twice per semester
About once per month
Every other week
At least once per week

Q28. Taking quizzes in class enhances my learning.* / Taking quizzes in class helps students learn.*

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree

Strongly agree

Q29. Having regular quizzes decreases the likelihood that I will attend class.* / Having regular quizzes decreases the likelihood that my students will attend class.**

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Q30. I enjoy taking quizzes in my classes.*

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Q31. Quizzes make a class more challenging.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

*Note: * indicates questions shown only to students, ** indicates questions shown only to faculty*