

Uses and Misuses of Student Evaluations of Teaching: The Interpretation of Differences in Teaching Evaluation Means Irrespective of Statistical Information

Teaching of Psychology
2015, Vol. 42(2) 109-118
© The Author(s) 2015
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0098628315569922
top.sagepub.com


Guy A. Boysen¹

Abstract

Student evaluations of teaching are among the most accepted and important indicators of college teachers' performance. However, faculty and administrators can overinterpret small variations in mean teaching evaluations. The current research examined the effect of including statistical information on the interpretation of teaching evaluations. Study 1 ($N = 121$) showed that faculty members interpreted small differences between mean course evaluations even when confidence intervals and statistical tests indicated the absence of meaningful differences. Study 2 ($N = 183$) showed that differences labeled as nonsignificant still influenced perceptions of teaching qualifications and teaching ability. The results suggest the need for increased emphasis on the use of statistics when presenting and interpreting teaching evaluation data.

Keywords

teaching evaluations, assessment, statistical significance

Psychologists have a long-standing respect for statistics. Teachers of psychology—especially in the areas of statistics and research methods—frequently chide students for interpreting small trends in data as if they were automatically meaningful. Statistical significance is typically covered across multiple semesters in psychology programs with the hope that students will exhibit appropriate statistical caution when interpreting numbers. Along the same lines, psychologists would never consider submitting quantitative research for publication without using statistics to back up claims about the results. Considering these characteristics, it is ironic that psychologists seem to forego their statistical standards when it comes to one specific type of data they likely collect and evaluate multiple times a year, that is, student evaluations of teaching. Consider the following: Does your department require statistical tests in tenure and promotion materials to determine whether student evaluation means are significantly higher or lower than average? Have you made a change to a course based on a dip in teaching evaluations without determining if the change was significant? Have you ever argued that you were a “good teacher” using only raw teaching evaluation means? It is likely that most teachers of psychology have engaged in these or similar uses of teaching evaluation data, uses that they would not approve of in other contexts. Model teachers of psychology collect and utilize teaching evaluations (STP Presidential Taskforce, 2013). In order to improve their thinking about and use of teaching evaluations, the current article will outline best practices in the use of

teaching evaluation data and demonstrate how easy it is to be misled by teaching evaluations.

Student evaluations of teaching are, arguably, the most influential single metric in the careers of college teachers. Teaching evaluations influence decisions about teachers' classroom abilities and about their general job performance (Beran, Violato, Kline, & Frideres, 2005; Gravestock & Gregor-Greenleaf, 2008; Shao, Anderson, & Newsome, 2007). Several sources of evidence support the validity of student evaluations for use in making judgments about teaching (for reviews of the following information see d'Apollonia & Abrami, 1997; Marsh & Roche, 1997). Multisection validity studies correlate student evaluations with learning outcomes across several sections of the same course, and the average correlation in these studies appears to be at least medium in size. Student evaluations also correlate with teachers' self-evaluations and with ratings made by trained observers about specific teaching skills. In addition, experimental manipulations of teaching quality lead to predictable differences in student evaluations. Overall, the consensus in the research literature appears to be that teaching evaluations are at least moderately valid and useful (Greenwald, 1997).

¹ Department of Psychology, McKendree University, Lebanon, IL, USA

Corresponding Author:

Guy A. Boysen, Department of Psychology, McKendree University, 701 College Rd., Lebanon, IL 62254, USA.
Email: gaboysen@mckendree.edu

Despite some consensus on basic validity, the controversy surrounding teaching evaluations is ongoing (Galbraith, Merrill, & Kline, 2012; Gravestock, Greenleaf, & Boggs, 2009). One source of controversy is the existence of known biases in teaching evaluations such as class size and grades (d'Apollonia & Abrami, 1997; Franklin, 2001; Greenwald & Gillmore, 1997; D. L. Smith, Cook, & Buskist, 2011). However, the source of controversy central to the current research is faculty's concern that teaching evaluations—no matter what their inherent validity—are misinterpreted and misused (Algozzine et al., 2004; Beran & Rokosh, 2009; Gravestock et al., 2009). Such concern is legitimate because even results from the very best measures can be rendered meaningless if not used or interpreted correctly.

There are a number of ways that teaching evaluations can be misused and misinterpreted. At the most basic level, the assumption of validity cannot be made for all student evaluations of teaching. Measures that have not gone through the validation process have unknown validity, and it is incorrect to assume that their outcomes are automatically meaningful (Algozzine et al., 2004). Even validated measures can be misinterpreted, however. Overreliance on single outcome measures is a common problem (Abrami, 2001; Algozzine et al., 2004; Theall & Franklin, 2001). Teaching is multifaceted, and it is not possible to quantify all that goes into teaching excellence with a single number. Similarly, there is often an overreliance on means from individual classes (Abrami, 2001). There are numerous reasons why evaluations from a single class might be skewed, and combining evaluations from multiple sections and courses increases the reliability of results. In general, the use of raw means without statistics to help guide interpretation is problematic (Abrami, 2001; Franklin, 2001). Basic measurement theory states that means are only estimates of true scores (Abrami, 2001; Cohen, Swerdlik, & Sturman, 2014). A teacher's true evaluation score falls within a range of scores determined by the number of observations and the standard error represented in the measurement. Furthermore, differences between two means may or may not be statistically meaningful, and interpretation of a difference as meaningful should not occur without the appropriate statistical tests of significance. It is quite easy, and often very tempting, to interpret raw means from student evaluations of teaching with a precision they simply do not possess (Algozzine et al., 2004; Boysen, Kelly, Raesly, & Casner, 2013; Franklin, 2001; Theall & Franklin, 2001).

Because the misinterpretation of teaching evaluations is so easy, experts have provided guidelines for their presentation and use. These recommended best practices should intuitively make sense to psychologists with statistical and research training. The first guideline is to combine means across items and courses (Abrami, 2001). Means are more reliable in aggregate, and combining means reduces the potential for interpretation of fluctuations caused by random error. Given the error inherent in all measurements, confidence intervals should be provided to represent the possible range of scores in which the true score falls (Abrami, 2001; Franklin, 2001). Representation of the

possible range of scores should help prevent the reification of teachers' means into representations of their inherent teaching ability. In order to provide a context for interpretation, comparison means, such as the mean of the teacher's department, should be provided as a reference (Abrami, 2001). Furthermore, statistical tests should be conducted to determine if differences between means are significant, and the results should be presented in tables or figures to aid in interpretation (Abrami, 2001; Franklin, 2001; Theall & Franklin, 2001). These suggestions are not particularly complicated or particularly conservative—especially considering the statistical standards typical in psychology—but there is reason to believe that faculty and administrators do not follow them.

Boysen, Kelly, Raesly, and Casner (2013) conducted three studies to document faculty members' and administrators' tendencies to overinterpret small differences in raw teaching evaluation means. Their research required participants to make judgments about fictional scenarios containing teaching evaluation data. The first study asked faculty members to assign merit-based awards to two teachers. Variations between the teachers' means did not exceed 0.30 on a five-point scale, but those differences still led to higher awards being assigned to teachers with higher means. A second study showed that heads of academic departments interpreted differences of the same small size when making recommendations to teachers about course revisions. A third and final study asked faculty members to rate teachers' need for improvement across several specific teaching skills such as "sets clear learning objectives." The skills corresponded to student evaluation means that varied by no more than 0.15 of a point. Nonetheless, faculty members' believed that these exceptionally small differences had meaningful implications for the relative need to improve skills. Overall, these studies illustrated that faculty and administrators are influenced by teaching evaluations in ways that go beyond their accepted reliability and validity.

Although there is evidence that faculty and administrators will overinterpret small differences in raw teaching evaluation means (Boysen et al., 2013), the previous research did not follow best practices for the presentation of teaching evaluations, practices specifically designed to attenuate misinterpretation. Experts suggest taking three steps when summarizing teaching evaluations (Abrami, 2001; Franklin, 2001). One, provide comparison means. Two, provide confidence intervals to indicate the potential range of true scores. Three, provide summaries of statistical tests comparing relevant means. The purpose of these suggestions is to reduce the guesswork in interpreting teaching evaluation means. Comparison means offer a standard for judgment, and confidence intervals and statistical significance provide a basis on which to determine if differences from the standard are meaningful. Boysen and colleague's (2013) research presented participants with multiple means in order to determine if they would interpret extant differences, but comparison means, confidence intervals, and tests of significance were not available. It is possible that providing these interpretive aids may have reduced misinterpretation. On the other hand, there is also reason to believe that statistical

information may fail to prevent interpretation of small numerical differences.

The results of Boysen and colleagues' (2013) study suggest that interpretation of teaching evaluations is guided by heuristic thinking. Because there are not typically set rules and criteria for the interpretation of teaching evaluations, judgments about them have a certain level of uncertainty. People use heuristics—mental shortcuts or rules of thumb—when making judgments under uncertain conditions (Tversky & Kahneman, 1974). One very simple heuristic people might use is that higher teaching evaluations are better. Using this heuristic, any difference between means might have an effect on interpretation. Anchoring is a more specific heuristic that is known to affect mathematical judgments (A. R. Smith & Windschitl, 2011; Tversky & Kahneman, 1974). People use numbers to anchor mathematical estimates, and the anchor number can affect estimates even if it is random and completely unrelated to the number being estimated (Tversky & Kahneman, 1974; Wilson, Houston, Etling, & Brekke, 1996). To illustrate, when starting with a randomly chosen anchor of 10, participants in one study estimated the percentage of countries in Africa that are part of the United Nations at 25%, but the estimate climbed to 65% when the anchor was 45 (Tversky & Kahneman, 1974). What is especially important to the current research is that these effects occur without conscious intention and happen even when people are warned against being biased by anchor numbers (Mussweiler & Englich, 2005; Wilson et al., 1996); this is similar to the interpretation of teaching evaluations differences irrespective of statistical information.

Considering the unintentional and difficult to control nature of heuristic thinking, it seems possible that small differences in teaching means will still have significant effects even if labeled as nonsignificant. Imagine a teacher with a mean overall student evaluation of 4.0. The average in the teacher's department is 4.3, but the difference is not statistically significant and should be disregarded. Will faculty members evaluating the teacher be able to ignore that higher mean and avoid its anchoring effect? Or, will the difference still affect impressions of the teacher despite statistical warnings against its interpretation? The purpose of the current research is to explore these questions.

Study 1

Study 1 presented faculty members with student evaluations from several different courses taught by one teacher. Based on these evaluations, participants provided ratings of the need to improve the quality of the specific courses and ratings of the teacher's quality overall. Means for the specific courses varied by small amounts shown to have a significant effect on interpretations of previous research (Boysen et al., 2013). Access to statistical information varied across three conditions; depending on condition, participants could utilize (a) a comparison mean; (b) a comparison mean and confidence intervals; or (c) a comparison mean, confidence intervals, and tests of statistical significance. The research question for Study 1 was as

follows: Will interpretation of teaching evaluations vary based on the statistical information provided?

Method

Participants

Participants consisted of faculty members ($N = 121$) listed on the websites of US colleges and universities. Participants were mostly male (59%) and White (86%), and the average age was 50 ($SD = 12$). Participants described their institutions as private baccalaureate colleges (40%), public baccalaureate colleges (16%), public research universities (16%), public master's college/universities (13%), private master's college/universities (13%), and private research universities (2%). The faculty members came from 29 different academic disciplines, none of which constituted more than 10% of the sample. Sampling occurred randomly. Researchers selected institutions offering 4-year degrees from the Carnegie Classification list. Next, researchers used the universities' websites to locate departments corresponding to 18 popular areas of study identified using the National Association for Educational Statistics website. Due to size and mission differences, not all of the colleges had departments representing each of the areas of study. For each available department, researchers randomly selected two faculty members and recorded their email addresses. These procedures yielded 484 valid faculty email addresses (25% response rate) from 20 different colleges. Each faculty member received an email invitation to participate in a brief survey in exchange for a chance to win a small monetary reward.

Material and Procedure

Participants completed the materials online in the form of a brief survey modeled on previous research (Boysen et al., 2013). The survey contained three scenarios, and the instructions asked participants to consider the scenarios as if they were on faculty committees and were in charge of making personnel and financial decisions at their college. There was one filler scenario and two experimental scenarios. The purpose of presenting multiple experimental scenarios was to collapse them into a single measure so as to increase the generalizability of the results. The order of the experimental scenarios was counterbalanced with the filler scenario placed between them so as to obscure the purpose of the research. The filler scenario described two faculty members' research and asked participants to allocate a \$2,000 reward for excellence in scholarship. The two experimental scenarios asked participants to imagine that they were "on a committee charged with evaluating the annual reappointment of faculty members to their current positions." The scenarios provided brief descriptions of faculty members, their typical teaching practices, and teaching evaluation means in table format (see Figure 1).

The teaching evaluation table included means for three of the faculty members' courses. Tables also included a teaching evaluation mean labeled "Comparison mean: Overall effectiveness rating for my department." As outlined in previous

| A | | | | | |
|---------|----------------------------------|--|---|--|--|
| Course | My overall effectiveness ratings | | Comparison mean: Overall effectiveness rating for my department | | |
| | Mean | | Mean | | |
| PSY 101 | 4.39 | | 4.25 | | |
| PSY 350 | 4.53 | | | | |
| PSY 450 | 4.27 | | | | |

| B | | | | | |
|---------|----------------------------------|-------------------------|---|-------------------------|--|
| Course | My overall effectiveness ratings | | Comparison mean: Overall effectiveness rating for my department | | |
| | Mean | 95% Confidence Interval | Mean | 95% Confidence Interval | |
| PSY 101 | 4.39 | 4.18–4.58 | 4.25 | 4.11–4.39 | |
| PSY 350 | 4.53 | 4.33–4.73 | | | |
| PSY 450 | 4.27 | 4.10–4.41 | | | |

| C | | | | | |
|---------|----------------------------------|-------------------------|---|-------------------------|--|
| Course | My overall effectiveness ratings | | Comparison mean: Overall effectiveness rating for my department | | Contrast: My mean vs. the comparison mean at 95% probability |
| | Mean | 95% Confidence Interval | Mean | 95% Confidence Interval | |
| PSY 101 | 4.39 | 4.18–4.58 | 4.25 | 4.11–4.39 | Not significantly different |
| PSY 350 | 4.53 | 4.33–4.73 | | | Not significantly different |
| PSY 450 | 4.27 | 4.10–4.41 | | | Not significantly different |

Figure 1. Format for the teaching evaluations in Study 1 included the (A) baseline, (B) confidence interval, and (C) statistical significance conditions. The format in panel C follows Franklin's (2001) recommended best practice for the presentation of teaching evaluation means. The means in the figure represent the high mean scenario. In the low mean scenario, the means and confidence intervals were 4.02 (3.89–4.23), 4.17 (4.03–4.31), and 4.28 (4.12–4.50); and the comparison mean was 4.31 (4.16–4.45). The analyses combined participants' ratings from the high and low mean scenarios.

research (Boysen et al., 2013), assuming a standard deviation of 0.50 and a reliability of 0.90, the confidence interval for a five-point teaching evaluation scale includes a total spread 0.32. In general, differences between means within this range should not be interpreted without further statistical information. Considering this statistical rule, faculty members' teaching evaluation means never differed from the comparison means by more than 0.32. Each faculty member's teaching evaluations included high, medium, and low means. In the high mean scenario, the faculty member's overall means for three courses were 4.27, 4.39, and 4.53; in the low mean scenario the means were 4.02, 4.17, and 4.28. Comparison means in the high and low scenarios were, respectively, 4.25 and 4.31. Thus, each faculty member had course evaluation means that were within 0.03, 0.14, and 0.28 of their corresponding comparison

mean. Following the procedure established in previous research (Boysen et al., 2013), the purpose of these variations was to average the scenarios together to increase generalizability and ensure that results were not biased by the fact that the teachers' ratings were systematically above or below average.

The experimental manipulation consisted of three variations in statistical information provided in relation to the teaching evaluation means (see Figure 1). In the baseline condition participants only saw the comparison mean. In the confidence interval condition, a 95% confidence interval accompanied all means. The intervals of all three course means overlapped with the confidence interval of the comparison mean. The confidence intervals for the all of the course means also overlapped. Confidence interval overlap suggests a difference between means that is not statistically reliable (Cumming & Finch,

2005). Statistical tests provide clear evidence for the reliability of differences, and in the statistical significance condition the means and confidence intervals were accompanied by explicit statements that the faculty member's means were "not significantly different" from the comparison mean. This final condition follows Franklin's (2001) recommended best practice for the presentation of teaching evaluation means.

After reading each scenario, participants rated 7 items using a seven-point scale ranging from *strongly disagree* to *strongly agree*. The three "need for improvement" items corresponded to the three specific courses with teaching evaluations. Each item stated "The instructor should work on improving _____" with the name of a course filling in the blank. Four items measured participants' general perceptions of the faculty member's teaching ability. The items asked participants to rate their agreement that the professor "is an above average teacher," "needs to work on developing teaching skills," and "is likely to exceed the teaching requirements for promotion and tenure." They also rated their agreement with the statement "I would highly recommend [the teacher] for reappointment." Factor analysis of these 4 items with varimax rotation suggested a one-factor solution explaining 53% of the variance. However, the item "needs to work on developing teaching skills" did not consistently load above 0.30 on the factor. Thus, the quality of teaching scale included only the other 3 items. Coefficient α for the combined items was .76 or higher, and this meets the conventional standards for research (Streiner, 2003), especially considering the small number of items (Cortina, 1993).

Results

The purpose of the analysis was to determine if varying statistical information would influence participants' tendency to interpret differences between teaching evaluations for specific courses. The dependent variables of interest were participants' ratings of need for improvement for the courses with high, medium, and low teaching evaluations. In order to conduct the analyses, total scores for each condition were necessary. The total scores consisted of the average need for improvement rating across the two scenarios for courses with high, medium, and low evaluations. A mixed factorial analysis of variance (ANOVA) examined the variables of statistical information (baseline, confidence interval, and statistical significance) and course evaluation (high, medium, and low) with statistical information serving as the between subjects factor and course evaluation serving as the within subjects factor. Results showed that the within subjects analysis was significant, $F(2, 220) = 41.87, p < .001, \eta_p^2 = .28$. Neither the effect of statistical information nor the interaction approached significance, all F s $< .68$, all p s $> .508$. Examination of the means indicated that need for improvement ratings was highest for the course with the low evaluation ($M = 4.09, SD = 1.56$) and lowest for the course with the high evaluation ($M = 3.60, SD = 1.54$), with means for the medium course falling in the middle ($M = 3.84, SD = 1.48$). Post hoc paired samples t tests showed that all three

means were significantly different, all t s > 5.12 , all p s $< .001$. These results indicate that differences in teaching evaluation means had a significant effect on interpretations regardless of the statistical information provided.

The second set of analyses examined for an effect of statistical information on ratings of overall teaching quality. Once again, the total score averaged ratings from the two scenarios. A one-way ANOVA examined for differences in overall teaching quality between the three experimental conditions (baseline, confidence interval, and statistical significance). The ANOVA was not significant, $F(1, 108) = 1.57, p = .213$, indicating that statistical information did not influence perceptions of overall teaching quality. Exploratory analyses replicated the first and second set of analyses separately for high and low teaching evaluation scenarios and repeated the analyses using only faculty members from the social sciences. Results in these exploratory analyses were the same as in the primary analysis. Overall, the results produced no evidence for the influence of statistical information on teaching evaluation interpretation.

Discussion

The purpose of Study 1 was to determine how statistical information affects the interpretation of small differences in teaching evaluations. Participants considered teaching evaluation means for three courses that differed in small amounts from a mean explicitly provided for the purpose of comparison. The results indicated that participants' ratings of the need to improve the courses varied regardless of the statistical information provided about the difference between the means. Even with overlapping confidence intervals and explicit statements about lack of statistical significance, participants still rated courses as needing significantly different levels of improvement based on how high or low they were in relation to the comparison mean. It is important to note that the difference between course means was never larger than 0.15. These results suggest that the inclusion of in-depth statistical information does not automatically eliminate the tendency to overinterpret small differences in teaching evaluations.

Although Study 1 provided strong evidence for the interpretation of teaching evaluation differences in the face of contradictory statistical evidence, there are some unanswered questions about the generalizability of the results. One major question is how far the influence of teaching evaluation differences extends. The dependent variables in Study 1 and past research (Boysen et al., 2013) were directly related to teaching itself. It is not clear if small differences in evaluations can also affect perceptions of broader variables such as a teacher's qualifications or general suitability for an instructional position. In addition, there has been no examination of variations in teachers' professional qualifications. The possibility exists for interaction between the quality of a teacher's credentials and the quality of their teaching evaluations. For example, it may be that teaching evaluations are only overinterpreted in cases when a teacher's evaluations and credentials are of low quality.

The purpose of Study 2 was to address these specific gaps in previous research.

Study 2

Study 2 offered a replication and extension of previous research. The methods included presentation of scenarios describing candidates for instructional positions whose teaching evaluations were not significantly different from comparison means, as indicated by confidence intervals and explicit statements of nonsignificance, and this replicates Study 1. In order to extend Study 1 and previous research, the teaching evaluations varied to include means that were high or low in relation to a comparison mean, and the quality of the candidates' qualifications also varied from high to low. The dependent variables represented another extension of previous research, and they went beyond perceptions of teaching ability to include perceptions of teaching experience and general suitability for an instructional position. These methods allowed the study to address two research questions. Will differences in teaching evaluations labeled as nonsignificant interact with differences in teaching qualifications? Will differences in teaching evaluations labeled as nonsignificant affect perception of characteristics other than teaching ability?

Method

Participants

Participants consisted of faculty members ($N = 183$) listed on the websites of U.S. colleges and universities. Participants were mostly male (67%) and White (82%), and the average age was 52 ($SD = 11$). Participants described their institutions as public master's college/universities (25%), private master's college/universities (24%), private baccalaureate colleges (23%), public research universities (16%), public baccalaureate colleges (7%), and private research universities (4%). The faculty members came from 26 different academic disciplines, none of which constituted more than 11% of the sample. Sampling and recruitment occurred using the same procedures as in Study 1. These procedures yielded 765 valid faculty email addresses (24% response rate) from 51 different colleges.

Material and Procedure

After providing their informed consent, participants completed a brief survey. The survey instructions for the experimental task asked participants to imagine that they were "on a committee evaluating part-time (adjunct) instructors" and that "the committee's task is to decide if each instructor should be placed on a 1-year temporary contract." Participants then read scenarios describing two part-time faculty members who had previously taught one course at their college and who were candidates for ongoing part-time positions. One of the fictional faculty members worked in the campus health center and taught psychology, and the other was a graduate student at a local university and taught communications. The order of the descriptions was randomized and separated by the same filler scenario described in Study 1.

The experiment was a 2×2 factorial design with the manipulations consisting of variations in candidates' qualifications and teaching evaluations. Qualification for the teaching position was either high or low based on the candidate's level of education and experience. To illustrate, in the high-qualification condition, the psychology teacher had a doctoral degree, completed a college teaching course, and worked for several semesters as a teaching assistant; in the low qualification condition the teacher had a master's degree and worked as a teaching assistant for one semester. Teaching evaluations were also either high or low in relation to the comparison mean. In the high condition, the overall teaching evaluation mean for the one course previously taught by the candidates was either 4.40 or 4.60, and the comparison mean from their departments was 4.02 or 4.11, respectively. In the low condition, the means were the same but reversed. Means in both conditions included confidence intervals that overlapped for the candidate's mean and the comparison mean. In addition, information labeled "Contrast: My mean vs. the comparison mean at 95% probability" indicated that the means were "not significantly different." Thus, the information provided about the means mimicked the full statistical information condition from Study 1 (see Figure 1C).

After reading the scenarios, participants completed a 6-item survey. Participants rated their agreement with statements related to the candidates' experience, teaching ability, and suitability for the position. The seven-point rating scale ranged from *strongly disagree* to *strongly agree*. The experience items stated that the candidate "possesses sufficient professional qualifications to be an instructor" and "has sufficient teaching experience to be an instructor." The teaching items stated that the candidate "is an above average teacher" and "needs to work on improving teaching" (reverse scored). One suitability item stated that the candidate "is an excellent candidate for the teaching position," and the other item stated "I would have reservations about hiring [this candidate] for a teaching position" (reverse scored). The 2 items related to experience, teaching ability, and suitability each combined to form subscales. Coefficient α for the experience and suitability subscales across the two candidates ranged from low to high by research standards (.67 to .87; Streiner, 2003), but teaching ability subscale was consistently low ($< .40$). It is possible that the lack of internal consistency emerged because participants believed that even "above average teachers" still should "work on improving teaching," which would lead to inconsistent responding on the scale. Nonetheless, retention of the subscale was justifiable because of the items' theoretical relation and the fact that both items showed the exact same pattern of results in terms of significance and the direction of the means.

Results

The analyses consisted of three separate factorial ANOVAs with the experience, teaching ability, and suitability subscales serving as dependent variables. Each subscale consisted of a

total score averaged across the two experimental scenarios. The independent variables for all three ANOVAs were qualification (high, low) and teaching evaluation (high, low). The first ANOVA examined ratings of experience. The main effect of qualification was significant, $F(1, 172) = 13.51$, $p < .001$, $\eta_p^2 = .07$, but the main effect of teaching evaluation was not significant, $F(1, 172) = 0.21$, $p = .646$, $\eta_p^2 < .01$. However, these results were qualified by a significant interaction, $F(1, 172) = 4.62$, $p < .033$, $\eta_p^2 = .03$. Post hoc Tukey's comparisons of experience ratings in the four conditions indicated that the difference between the high- and low-qualification conditions was not significant when the teaching evaluations were high (high qualification: $M = 5.47$, $SD = 1.37$; low qualification: $M = 5.19$, $SD = 1.17$), but the difference was significant when the teaching evaluations were low (high qualification: $M = 5.93$, $SD = 0.73$; low qualification: $M = 4.89$, $SD = 1.30$). In addition, the rating of experience in the high-qualification/low-evaluation condition was significantly higher than the rating in the low-qualification/high-evaluation condition. The means indicate that qualifications had relatively little impact when teaching evaluations were high but that qualifications were especially impactful when teaching evaluations were low. It must be reemphasized that even in the low teaching evaluation conditions, the difference between the teachers' means and comparison means was labeled as not significant and, thus, should not have been interpreted as meaningful.

The next analysis examined ratings of teaching quality. The main effect of teaching evaluation was significant, $F(1, 169) = 14.47$, $p < .001$, $\eta_p^2 = .08$. Examination of the means indicated that the main effect occurred because ratings of quality were higher in the high teaching evaluation condition ($M = 4.34$, $SD = 0.96$) than in the low teaching evaluation condition ($M = 3.74$, $SD = 1.06$). Neither the main effect of qualification (high qualification: $M = 4.06$, $SD = 0.93$; low qualification: $M = 3.99$, $SD = 1.16$) nor the interaction were significant, all $F_s < 1.16$, all $p_s > .283$. Exploratory analysis using only social science faculty demonstrated the same main effect of teaching evaluation. These results indicate that the difference between candidates' teaching evaluation means and the comparison means had a significant effect on interpretations even though they were clearly labeled as lacking statistical significance.

The final analysis examined ratings of candidates' suitability for an instructional position. The main effect of qualification was significant, $F(1, 174) = 14.52$, $p < .001$, $\eta_p^2 = .08$. Examination of the means indicated that the main effect occurred because ratings of suitability were higher in the high-qualification condition ($M = 5.16$, $SD = 1.03$) than in the low-qualification condition ($M = 4.48$, $SD = 1.24$). Neither the main effect of teaching evaluation (high evaluation: $M = 4.90$, $SD = 1.21$; low evaluation: $M = 4.70$, $SD = 1.19$) nor the interaction was significant, all $F_s < 1.36$, all $p_s > .245$. These results indicate that overall evaluations of suitability for a teaching position were not affected by the differences in teaching evaluations.

Discussion

Study 2 presented faculty members with descriptions of candidates for instructional positions whose teaching evaluation means differed nonsignificantly from comparison means. These methods allowed the study to address two research questions. Research Question 1 asked, will differences in teaching evaluations labeled as nonsignificant interact with teaching qualifications? The answer to this question was yes but only in relation to perceptions of teaching experience. Differences in qualifications did not affect perceptions of a teaching experience when a candidate had high teaching evaluations. However, if teaching evaluations were low, then perceptions of the teaching experience were significantly different between the high- and low-qualification conditions. These results suggest that high teaching evaluations mitigate the effects of teaching qualifications, and low teaching evaluations amplify the effects of teaching qualifications. Research Question 2 asked, will differences in teaching evaluations labeled as nonsignificant affect perception of characteristics other than teaching ability? The answer to this question was yes, but the effect was limited. As outlined above, teaching evaluations did have an effect on perceptions of teaching experience. Evaluations did not influence general perception of a candidate as suitable for an instructional position, only teaching qualifications affected that variable. However, it should be noted that, as in Study 1, differences in teaching evaluations that were clearly identified as not being statistically significant did lead to significant differences in perceptions of teaching ability.

Study 2 offered a conceptual replication of Study 1's finding that faculty members are influenced by nonsignificant differences in teaching evaluations. There are several possible explanations for these results. One potential reason for the overinterpretations is simple ignorance of statistics. Many participants may not have training that allows them to understand the meaning of "not significantly different." Although the results must be considered tentative and exploratory, analysis of the responses of social scientists in the sample indicated that they showed the same trends despite their, presumably, increased statistical knowledge. Such a finding suggests the alternative explanation that teaching evaluations may be processed heuristically (Tversky & Kahneman, 1974; Wilson et al., 1996). Research on heuristics indicates that people's numerical judgments can be influenced by anchor numbers provided to them even when they know the anchor is not meaningful. This is similar to participants in the current study being influenced by differences in teaching evaluations despite the statistical meaninglessness of the differences. A final explanation is that participants simply did not attend to the materials closely and failed to notice the statistical cues that should have led them to be wary of interpretation.

Another finding from Study 2 was that teaching evaluations can influence the perceived importance of a teacher's qualifications, and this fits with previously published concerns about the misuse of teaching evaluations. Franklin (2001) made the intriguing argument that failure to use statistics to interpret teaching

evaluations can lead them to be a sort of projective test in that people selectively interpret the numbers to fit their beliefs about a teacher; such use is especially pernicious because it masks subjective judgments behind a façade of seemingly objective numbers. Exactly this type of subjective interpretation appeared to be going on in this study. Faculty members tended to ignore differences in qualifications as long as a teacher had high student evaluations; however, when evaluations were low, differences in teaching qualifications had a significant influence on perceptions of experience. In other words, teachers' qualifications may be judged differently based on their student evaluations.

A final finding from Study 2 provides some good news in the otherwise dismal message of this research. It appears that there is a limit to the influence of small, nonsignificant differences in teaching evaluations. Participants rated candidates' overall suitability for a teaching position, and this variable was only influenced by the candidates' qualifications. Overall suitability was the variable in the current research most closely related to a final decision of whether or not to hire a candidate, and it is reassuring to find it unaffected by meaningless variations in teaching evaluations.

General Discussion

Experts emphasize the importance of utilizing statistical information to aid in the interpretation of teaching evaluation means (Abrami, 2001; Franklin, 2001). Specifically, student evaluations should be accompanied by a comparison mean, confidence intervals, and tests of significance. Across two studies in the current research, faculty members considered teaching evaluations that included these statistical details. Nonetheless, Study 1 showed that extremely small differences between student evaluations for individual courses led to significant differences in the perceived need to improve those courses and that variations in statistical information had no effect on these interpretations. Furthermore, Study 2 showed that differences clearly identified as nonsignificant still influenced perceptions of teachers' qualifications and teaching ability. These results suggest that even following the best practices in presenting teaching evaluations may still not stop faculty members from being influenced by small, nonsignificant differences in teaching evaluation means.

The results of the current studies are consistent with previous research. Boysen and colleagues (2013) examined the tendency of faculty members and administrators to interpret small differences in teaching evaluations. Differences in student evaluations that were too small to be meaningful affected participants' assignment of merit-based awards, judgment of course revisions, and evaluation of specific teaching skills. The current research went even further by showing that similar effects occur even when mean differences are labeled as being not statistically significant. Findings from the current research are also consistent with research on heuristics (Chen & Kemp, 2012; Tversky & Kahneman, 1974; Wilson et al., 1996). Participants' tendency to judge differences as meaningful in direct

contradiction to statistical information is particularly reminiscent of the research of Wilson et al. (1996). They had participants guess the number of doctors listed in the telephone directory after having thought about a large, unrelated number in a previous task. Despite explicit warning to avoid being influenced by the previously seen number, participants' estimates were significantly increased—more than doubled in fact—by the anchoring effect of the large number. Similarly, it appears that faculty members in the current study could not help but be influenced by having seen the comparison means. It should be noted, however, that the current research cannot eliminate more mundane explanations for the results such as ignorance of statistics or failure to attend to the experimental materials.

Teachers who want to effectively present their student evaluations can take away several suggestions from the current research. The first suggestion is to assume that statistics will be interpreted incorrectly. It is likely that some faculty and administrators will not know how to interpret statistics, and others will be influenced by differences even if they do understand statistics. As such, the second suggestion is to provide detailed qualitative explanations of trends in teaching evaluations (Franklin, 2001). Outline likely explanations for meaningful trends, discuss mitigating circumstances, and propose plans for improvement if means are legitimately low. Teachers might consider providing explicit verbal warnings to avoid interpretation of nonsignificant differences, but the current research indicates that such efforts may not be completely successful. Another suggestion is to emphasize teaching qualifications if student evaluations are lower than desired. There are two major reasons for this suggestion. One, the results of Study 2 illustrated that overall evaluations of suitability for an instructional position were most influenced by teaching qualifications, and arguably, that is the most important perception to influence. Two, the study also suggested that perceptions of experience are actually quite positive when a teacher has relatively high teaching qualifications and relatively low teaching evaluations.

Faculty members and administrators who are responsible for the evaluation of teachers can also take away lessons from the current research. The first suggestion is to require statistical information whenever teaching evaluations are presented. Although this suggestion may seem counterintuitive, given the results of the current research, requiring statistics could have several positive effects. This study illustrated that there is a tendency to interpret nonsignificant differences in teaching evaluation means when forming initial judgments, but it did not show that statistics are ignored when making intentional, well-reasoned arguments based on teaching evaluations. Requiring statistics should obviate both teachers' and administrators' ability to use meaningless differences to support claims about teaching; and their initial impressions of teaching quality may still be influenced by small differences, but they will be unable to intentionally reason that those impressions are backed by meaningful numbers. In fact, requiring statistics could offer protection for all interested parties; administrators could not use subjective impressions based on small

differences to dismiss the abilities of teachers, and teachers could not exploit small differences to inflate perceptions of their teaching ability. Another advantage of requiring statistics would be the effect of forcing large numbers of faculty and administrators to learn about their interpretation. Knowledge about statistics should only increase the precision with which teaching evaluations are interpreted. To that end, the final suggestion is to offer professional development to assist faculty and administrators in the interpretation of teaching evaluation statistics. This final point clearly illustrates the important role that psychology faculty can play in the improved use of teaching evaluations. Not only can psychologists utilize their statistical knowledge to improve their own presentation and interpretation of teaching evaluations, they can be leaders at their institutions by imparting their knowledge to others.

Although psychologists have greater statistical training than most other faculty members, they are not immune to failures in statistical reasoning. For example, one study sampled authors of articles in empirically focused American Psychological Association (APA) journals, individuals who should have sound statistical knowledge, and the group received an average score of 59% on a quiz of basic statistical knowledge (Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). Thus, psychologists may misinterpret teaching evaluations due to absent or incorrect knowledge about statistics. However, even people with expert knowledge can make mistakes when they rely on heuristic thinking for statistical problems. Tversky and Kahneman (1971) famously showed that reliance on (incorrect) intuition rather than statistical algorithms led members of the APA's Mathematical Psychology Group to greatly overestimate the probability of replicating a statically significant finding with a small sample size. Slow, effortful thinking by most psychology teachers would probably lead to the conclusion that small differences in teaching evaluations are unlikely to be statistically or practically meaningful, but fast, heuristic thinking would likely result in the interpretation of small differences. Be it from increased statistical knowledge or increased effort, the ideal outcome of the current research is for psychology teachers to avoid interpreting small differences in teaching evaluations without the appropriate statistical information.

Despite providing clear evidence for the influence of small, nonsignificant differences in teaching evaluations, the current research had several limitations. A primary limitation is the inability to determine exactly why participants were influenced by differences that were clearly identified as not being statistically significant. Heuristic thinking, ignorance of statistics, or inattention could have affected the results in isolation or in combination. Another limitation was the sample. Participants self-selected into the study and may be systematically different than faculty who chose not to participate. A final limitation was the artificial nature of the research task assigned to participants. The experimental materials were intentionally brief, and it is not clear that teaching evaluations would have the same effect when presented alongside a full professional dossier. It is also not clear if these immediate effects have more long-standing effects on perceptions of teachers.

Student evaluations of teaching are consequential, controversial, and inescapable. As such, teachers should use the same statistical caution in the presentation and interpretation of data from teaching evaluations as they do with research data. The current research illustrated that significant interpretation of teaching evaluation means can occur in the face of clear messages of statistical nonsignificance. Teachers of psychology, due to their training and professional identity, should lead the way in the intentional use of statistics to prevent overinterpretation of student evaluations. Although changing standards may lead to more work for both summarizers and interpreters of teaching evaluations, increasing accuracy and fairness in how teachers are evaluated is a worthwhile goal.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

References

- Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. *New Directions for Institutional Research, 109*, 59–87. doi:10.1002/ir.4
- Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., . . . Spooner, F. B. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching, 52*, 134–141.
- Beran, T., Violato, C., Kline, D., & Frideres, J. (2005). The utility of student ratings of instruction for students, faculty, and administrators: A 'consequential validity' study. *Canadian Journal of Higher Education, 35*, 49–70.
- Beran, T. N., & Rokosh, J. L. (2009). The consequential validity of student ratings: What do instructors really think? *Alberta Journal of Educational Research, 55*, 497–511.
- Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2013). The (mis) interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education, 39*, 641–656. doi:10.1080/02602938.2013.860950
- Chen, Z., & Kemp, S. (2012). Lie hard: The effect of self-assessments on academic promotion decisions. *Journal of Economic Psychology, 33*, 578–589. doi:10.1016/j.joep.2011.11.004
- Cohen, R. J., Swerdlik, M., & Sturman, E. (2014). *Psychological testing and assessment: An introduction to tests and measurement*. (8th ed.). New York, NY: McGraw-Hill.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104. doi:10.1037/0021-9010.78.1.98
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170–180. doi:10.1037/0003-066X.60.2.170
- d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*, 1198–1208. doi:10.1037/0003-066X.52.11.1198

- Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. *New Directions for Teaching and Learning*, 87, 59–87. doi:10.1002/tl.10001
- Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education*, 53, 353–374. doi:10.1007/s11162-011-9229-0
- Gravestock, P., Greenleaf, E., & Boggs, A. M. (2009). The validity of student course evaluations: An eternal debate? *Collected Essays on Learning and Teaching*, 2, 152–158.
- Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends*. Toronto, Canada: Higher Education Quality Council of Ontario.
- Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, 52, 1182–1186. doi:10.1037/0003-066X.52.11.1182
- Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, 52, 1209–1217. doi:10.1037/0003-066X.52.11.1209
- Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, 52, 1187–1197. doi:10.1037/0003-066X.52.11.1187
- Mussweiler, T., & Englich, B. (2005). Subliminal anchoring: Judgmental consequences and underlying mechanisms. *Organizational Behavior and Human Decision Processes*, 98, 133–143. doi:10.1016/j.obhdp.2004.12.002
- Shao, L. P., Anderson, L. P., & Newsome, M. (2007). Evaluating teaching effectiveness: Where we are and where we should be. *Assessment & Evaluation in Higher Education* 32, 355–371. doi:10.1080/02602930600801886
- Smith, A. R., & Windschitl, P. D. (2011). Biased calculations: Numeric anchors influence answers to math equations. *Judgment and Decision Making*, 6, 139–146.
- Smith, D. L., Cook, P., & Buskist, W. (2011). An experimental analysis of the relation between assigned grades and instructor evaluations. *Teaching of Psychology*, 38, 225–228. doi:10.1177/0098628311421317
- STP Presidential Taskforce. (2013). *Society for the teaching of psychology model teaching competencies*. Retrieved from <http://teachpsych.org/Resources/Documents/publications/2013%20Model%20Teaching%20Competencies.pdf>
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80, 99–103. doi:10.1207/S15327752JP A8001_18
- Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research*, 109, 45–56. doi:10.1002/ir.3
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76, 105–110. doi:10.1037/h0031322
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, 125, 387–402. doi:10.1037/0096-3445.125.4.387
- Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science*, 4, 49–53. doi:10.1111/j.1467-9280.1993.tb00556.x