Abstract

We analyzed student evaluations for 3,585 classes collected over 20 years to determine stability and evaluate the relationship of perceived grading to global evaluations, perceived fairness, and appropriateness of assignments. Using class as the unit of analysis, we found small evaluation reliability when professors taught the same course in the same semester, with much weaker correlations for differing courses. Expected grade and grading related questions correlated with overall evaluations of courses. Differences in course evaluations on expected grades, grading questions, and overall grades were found between full-time faculty and other types of instructors. These findings are expanded to a model of grading type questions mediating the relationship between expected grade and overall course evaluations with a moderating effect of type of instructor.

*Keywords*: Student evaluation, teacher evaluation, teacher effectiveness, perceived grading

Perceived Grading and Student Evaluation of Instruction

The controversy over perceived grading and student evaluations continues, with arguments that professors can (Greenwald & Gilmore, 1997; Isley & Singh, 2005; Krautmann & Sander, 1999) or cannot (Centra, 2003; Marsh, 1987; Marsh & Roche, 2000) improve evaluations by giving higher grades. Social psychology theory would support that students with low grades can reduce cognitive dissonance and engage in ego defense by giving low evaluations to teachers who give them low grades (Maurer, 2006). We argue that both social psychology theory and the evidence from student evaluations supports that higher grades can lead to better student evaluations of teaching.

Much of the research on student evaluations involves diverse and complex analyses (e.g., Marsh, 1987) and lacks social-psychological theoretical guidance on human judgment. To expect that student evaluations would not be influenced by expected grade would contradict a long-standing history of social psychology research on cognitive dissonance, attribution, and ego threat. Failure threatens the ego (Miller, 1985; Snyder, Stephan & Rosenfield, 1978) and motivates us to find rationales to defend the ego. Failing students, or those performing below personal expectations, would be expected to defend their ego by attributing low grades to poor teaching or unfair evaluation practices (Maurer, 2006). One common strategy involves diminishing the value of the activity (Miller & Klein, 1989), which would result in lowered value of a course.

Similarly, Cognitive Dissonance Theory (Festinger, 1957) predicts that people who experience poor performance but perceive themselves as competent will experience dissonance, by which they can reduce through negative evaluations of the instruction (Maurer, 2006). Attribution research (Weiner, 1992) also supports the argument that among low achievement motivation students, failure is associated with external attributions for cause, and the most plausible external attribution is the quality of instruction and grading practices. Although arguments regarding degree of influence are reasonable, the position that they are not affected is inconsistent with existing and established theory. Thus, it is not surprising that the majority of faculty perceive student evaluations to be biased by perceived grading and course choice (Marsh, 1987).

Considerable research has been conducted in support of widely distributed evaluation systems. Centra (2003) reported that in a study of 9,194 class averages using the Student Instructional Support, the relationship between expected grades and global ratings was only .20. He further argued that when variance due to perceived learning outcomes was regressed from the global evaluation, the effect of expected grades was eliminated. However, a student’s best assessment of “perceived learning outcome” is their expected grade, and thus, these should be highly correlated. When perceived learning is regressed from the global evaluations, it is not surprising that suppression effects would eliminate or could even reverse the correlation between expected grade and global evaluation. In general, there are several reasons why the relationship of expected grade to global evaluations is suppressed. For example, faculty ratings are generally very high on average (i.e. quality instructors are hired), which restricts variation; thus, weakening their reliability as a measure of teacher attributes. This restriction in range suppresses correlation.

Perhaps a more significant reason the grade/evaluation effect is often minimized in research is due to the confusion between leniency in perceived grading and expected grade. We describe this concept as leniency because the inflated grades are higher than grade descriptors would imply. The grading system in the university where data were collected describes a C as “Average” and a B as “Above Average”. Isley and Singh (2005) demonstrated that student evaluations are influenced by the difference between expected grade and the student’s GPA. Expected grade is experienced by the student as leniency to the extent it reflects differences from course expectations based upon course effort and typical grades for the student. The authors recognize that many professors would argue that they give the majority of their students As because they are extraordinary teachers or because they use a mastery system designed to allow all students to receive an A. This ideology reflects an abandonment of grades as a means to differentiate students, a departure from “C = average”, and a lack of matching levels of challenge to the achievement potential of students. For our purposes, we refer to it as leniency to the extent that grading is above the university standard of “C = average”.

Marsh (1987) argued that the individual is not the proper unit of analysis because such analyses could suggest false findings related to individual differences in students. Therefore, he argued the use of class as the suggested unit of analysis. We agree, both for his reasoning and because analyses with individual ratings can mask significant relationships as well. Individual differences in expectancy will attenuate the correlation less when class average is used as the unit of analysis. To the extent that the same class average would be expected across all courses, an assumption we will challenge, the class average for expected grade is a good measure of perceived grading as an instructor attribute. Course quality, not individual attributes of students, is what we are attempting to assess when we are using student evaluations of courses. Several studies provide support that when class is the unit of analysis, expected grade is a more significant biasing factor in student evaluations (Blackhart, Peruche, DeWall & Joiner, 2006; DuCette & Kenney, 1982; Ellis, Burke, Lomire, & McCormick, 2003).

Blackhart et al., (2006) analyzed 167 psychology classes in a multiple regression analysis with class as the unit of analysis and found that the two most significant predictors of instructor ratings were average grade given by the instructor and instructor status (TA or rank of faculty). Because of the limited number of classes, the power of the analysis was limited. However, in addition to the concern regarding the relationship between grades and global course evaluations, it was found that TAs were rated more highly than ranked faculty. This finding raises additional questions on validity. We must either accept that the least trained and qualified teachers are actually better teachers, or we must believe this result suggests that student evaluations have given us false information on the quality of teaching via their perceptions of grading.

DuCette and Kenney (1982) provided evidence that using course as a unit of analysis increased the correlation between expected grade and other course ratings. Within specific groupings of classes, these correlations ranged from .23 to .53. Two factors limited the level of their relationships. First, the classes used were all upper division courses and graduate courses. Secondly, over 90% of the students in these classes expected an A or a B. Consequently, the correlations between expected grade and global course ratings would be reduced due to the absence of variation in expected grades. Similarly, Ellis et al. (2003) found a correlation of .35 between average course grade and average rating of the instructor in 165 classes during a two-year period. However, these studies did not consider the predictive relationship for instructors across different courses and semesters.

Different disciplines and subject areas have diverse GPA standards, and students have differing grade and workload expectations in different courses as well. For example, an instructor in Anatomy giving a 3.00 GPA might be considered lenient while an Education instructor giving a 3.25 GPA might be considered hard (examples for illustration only). To have a valid measure of workload and leniency factors, correlations should be conducted with varied teachers of the same course. Further, different populations take courses in different disciplines, resulting in potential population differences between anatomy classes and education classes, which could create or mask findings as well. Hence, analysis of these correlations within the same discipline and course was expected to strengthen the relationship between expected grades and quality measures, offering more valid results.

In most studies of student evaluations, reliability is established through internal consistency reliability. However, this form of reliability is confounded with halo effects (i.e. a cognitive bias that influences ratings based on an overall perception of the person teaching, rather than the individual components of the course), and tells only whether the individual responding to the questions is consistent and reliable. By having many different classes for the same instructor, we can establish the reliability of ratings across the same and different courses during the same and different semesters. Student ratings cannot be considered a valid measure of an instructor’s teaching skills if they are unable to reliably differentiate instructors.

If ratings are valid measures of instructor attributes, it should be expected that ratings would have some stability across semester and course taught. If variation were due to instructor attributes and not the course they are assigned, we would expect ratings to be most stable across two different courses during the same semester. We would expect these correlations to decline somewhat for the same course in a different semester, since faculty members may improve or decline with experience. But if they are reliable and stable enough to use in making choices about retention, their stability should be demonstrated across different semesters as well. Therefore, we first sought to establish reliability of ratings for the instructors across courses and semesters.

In the current study, we used data collected over a 20-year period to allow for more powerful analyses, with such analyses occurring within many sections of the same course. After examining reliability, we sought to show that items on instructor evaluations were positively correlated for undergraduate and graduate students, demonstrating that overall course evaluations are related to perceptions of grading and expected grade. We also expected correlations to be substantially higher than those obtained by researchers who used individual students as their unit of analysis. Next, we examined if rating differences across these questions were found between types of instructors compared to full-time faculty, such as teaching-assistants and per-course faculty. The presumption of university hiring requirements that include a terminal degree for regular faculty is that better-trained faculty will be more effective teachers. Therefore, if student evaluations are a valid measure, better-trained full-time faculty should receive higher ratings than per-course instructors and teaching assistants. However, existing literature appears to contradict this expectation (Blackhart et al., 2006). Given these differences, we proposed and examined a moderated mediation analysis to portray the expected relationship of the variables across instructor type.

**Method**

The archival study was conducted using data from the psychology department at a large Midwestern public university. We used data from 2,846 undergraduate, 244 mixed graduate and undergraduate, and 429 graduate psychology classes taught from 1987 to 2007 that were evaluated by students using the same 15-item instrument. Data was collected for course evaluation rather than for this study. Faculty followed set procedures in distributing scan forms no more than two weeks before the conclusion of the semester. A student was assigned to collect the forms and deliver them to the departmental secretary. The instructor was required to leave the room while students completed the forms. Classes with fewer than five student responses were not retained in the data pool.

We focused upon the five items, which seemed most pertinent to the issues of perceived grading and evaluation. We were most interested in how grades related to global course evaluation and grading/assignment evaluations. These items were presented with a five-point strongly agree/disagree scale:

1. The overall quality of this course was among the top 20% of those I have taken.

2. The examinations were representative of the material covered in the assigned readings and

class lectures.

3. The instructor used fair and appropriate methods in the determination of grades.

4. The assignments and required activities in this class were appropriate.

5. What grade do you expect to receive in this course? (A = 5, B, C, D, F = 1).

**Results**

All data were checked for course coding errors, and type of instructor was coded. This data was considered structured by instructor; therefore, all analyses below were coded in *R* using the *nlme()* package (Pinheiro et al., 2017) to control for instructor as a random intercept in a multilevel model. Each dataset was screened separately normality, linearity, homogeneity, and homoscedasticity using procedures from Tabachnick and Fidell (2012). This data was not screened for outliers because it was assumed that each score was a correct representation of instructor rating.

**Reliability of instructor scores across courses**

Reliability of ratings of instructors can be inferred by the consistency of ratings across courses and semester, assuming that we infer there is a stable good/poor instructor attribute and that these multiple administrations of the same question are multiple assessments of that attribute. A file was created with all possible course pairings for each instructor, such as each pairwise combination could only be represented once. For example, professor A’s course pairings might include the following pairs: Fall 2000 Statistics - Fall 2000 Statistics (complete match), Fall 2001 Statistics - Fall 2000 Statistics (course match), Fall 2000 Statistics - Fall 2000 Research Methods (semester match) and Fall 2000 Statistics - Fall 2001 Research Methods (no match). The minimum number of combinations for an instructor was one (especially for our teaching assistants) and the maximum number of pairings was 8301 (for long term faculty). The average number of pairings was *M* = 1013.34 (*SD* = 1992.58).

Multilevel models were used to calculate correlations on each combination of same and different courses and same and different semesters after controlling for nested instructor data. Data appeared linear and normal with problems indicated on homoscedasticity, as residuals were not evenly spread. Correlations are presented in Table 1 split by semester and course, and a table of all *t*, *p*, and *df* values for these correlations can be found at http://osf.io/jdpfs. Reliability was found with medium significant correlations for same semester and course pairing across questions. Interestingly, the same course across different semesters did not correlate for scores, which may indicate that student demand characteristics or course changes impact instructor ratings. Different course ratings for same and different semesters were significantly correlated across questions but these correlations were small. Interestingly, grading was significantly negatively correlated for different semesters for both same and different course pairings.

**Correlations between relevant evaluation items**

**Undergraduate courses***.* Table 2 presents the inter-correlations of the five relevant items in two forms, again controlling for instructor as a random intercept in a multilevel model. Using class average on each item as the unit of analysis, we found class expected grade was related to class overall rating, exams reflecting the material, grading fairness, and appropriateness of assignments. We found high correlations with grade expectancy, overall quality, exams reflecting material, grading fairness, and appropriateness of assignments when correlations were computed within the same course (Table 2, upper right-hand side).

**Graduate courses***.* Of concern here was that the students in the Psychology graduate classes had a grade average exceeding 3.9 on a 4.0 scale, yielding little variation. We saw substantial correlations between grade expectancy and overall quality, exams reflecting material, grading fairness, and appropriateness of assignments (Table 2, lower left-hand side). However, these correlations are lower than the comparison undergraduate correlations.

**Instructor Status and student ratings**

We compared teaching assistants, per-course faculty, instructors, and ranked faculty in undergraduate general education classes, since these courses have the largest proportion of TAs and per-course faculty. This analysis included 68 teachers: 18 teaching assistants, 14 per-course faculty, 10 instructors, and 26 full-time faculty, who taught 535 courses: 69 teaching assistants, 94 per-course faculty, 85 instructors, and 287 full-time faculty. All comparisons were made against full-time faculty to control for Type 1 error using a multilevel model with a dummy coded instructor variable, and dummy coded *t* values were used to determine which comparison groups were different from full-time faculty. Overall means and standard deviations are presented in Table 3, and the complete set of *t* value comparisons for these analyses can be found online. As shown in the Table 3, the ratings of all groups were fairly high, hovering around 4.00 on a 5.00 point scale, and the expected grade for courses was approximately a B.

For overall ratings, faculty were found to be rated less highly than teaching assistants, *p* = .027, but not significantly different than per-course faculty (*p* = .181) or instructors (*p* = .814). When rating if exams were representative of course material, full-time faculty were rated lower than both teaching assistants (*p* < .001) and per-course faculty (*p* = .047), but were not significantly different than instructors (*p* = .740). Full-time faculty were rated as less fair and appropriate in their grades than teaching assistants (*p* = .003), while per-course faculty (*p* = .128) and instructors (*p* = .657) had similar scores to faculty. Teaching assistants were designated to have more appropriate assignments than faculty (*p* < .001), while per-course (*p* = .060) and instructors (*p* = .073) had the same ratings as faculty on assignments. Finally, faculty showed significantly lower expected grades than teaching assistants (*p* < .001) and per-course faculty (*p* = .044), while having similar grades to instructors (*p* = .705).

**Moderated Mediation**

Given the correlations between items and differences between items and ranked faculty, we proposed a mediation relationship between expected grade, perceived grading, and overall course grades that varies by instructor type. Figure 1 demonstrates the predicted relationship between these variables. We hypothesized that expected course grade would impact the overall course rating, but this relationship would be mediated by the perceived grading in the course, which was calculated by averaging questions about exams, fairness of grading, and assignments. Therefore, as students expected to earned higher grades (leniency), their perception and ratings of the grading would increase, thus, leading to higher overall course scores. This relationship was tested using traditional and newer approaches to mediation (Baron & Kenny, 1986; Hayes, 2013). All categorical interactions were compared to ranked faculty. Each step of the model is described below. Because significant interactions were found, we calculated each group separately (Figure 1) to portray these differences in path coefficients. Tables of *t* values for the overall and separated analyses are available at http://osf.io/jdpfs.

**C Path*.*** First, expected grade was used to predict the overall rating of the course, along with the interaction of type of instructor and expected grade. The expected grade positively predicted overall course rating, *p* < .001, wherein higher expected grades was related to higher overall ratings for the course (*b* = 0.39). A significant interaction between type and expected grade rating was found for instructors versus faculty. In looking at Figure 1, we find that instructors (*b* = 0.56) have a stronger relationship between expected grade and overall course rating than faculty (*b* = 0.39, interaction *p* = .020), while per-course (*b* = 0.41, interaction *p* = .621) and teaching assistants (*b* = 0.71, interaction *p* = .068) were not significantly different than faculty on the *c* path coefficient.

**A Path*.*** Expected grade was then used to predict the average of the grading related questions, along with the interaction of type of instructor. Higher expected grades were related to higher ratings of appropriating grading (*b* = 0.21, *p* < .001), and a significant interaction of faculty and all three other instructor types emerged: teaching assistants (*p* = .001), per-course faculty (*p* = .001), and instructors (*p* < .001). As seen in Figure 1, faculty (*b* = 0.21) have a much weaker relationship between expected grade and average ratings of grading than teaching assistants (*b* = 0.55), per-course (*b* = 0.41), and instructors (*b* = 0.45).

**B and C’ Paths*.*** In the final model, expected grade, average ratings of grading, and the two-way interactions of these two variables with type were used to predict overall course evaluation. Average rating of grading was a strong significant predictor of overall course rating (*b* = 1.10, *p* < .001), indicating that a perception of fair grading was related positively to overall course ratings. An interaction between per-course faculty and fair grading emerged, *p* < .001, wherein faculty (*b* = 1.10) had a less positive relationship than per-course (*b* = 1.28), while teaching assistants (*b* = 1.37, interaction *p* = .071) and instructors (*b* = 1.16, interaction *p* = .187) were not significantly different coefficients.

The relationship between expected grade and overall course rating decreased from the original model (*b* = 0.16, *p* < .001). However, the interaction between this path and per-course (*p* < .001) and instructors (*p* = .041) versus faculty was significant, while faculty versus teaching assistants’ paths were not significantly different (*p* = .133). Faculty relationship between expected grade and overall course scoring, while accounting for ratings of grading was stronger (*b* = 0.16) than instructors (*b* = 0.04) and per-course (*b* = -0.10), but not that of teaching assistants (*b* = -0.04).

**Mediation Strength*.*** We then analyzed the indirect effects (i.e. the amount of mediation) for each type of instructor separately, using both the Aroian versionof the Sobel test (Baron & Kenny, 1986), as well as bootstrapped samples to determine the 95% confidence interval of the mediation (Preacher & Hayes, 2008; Hayes, 2013) because of the criticisms on Sobel. For confidence interval testing, we ran 5,000 bootstrapped samples examining the mediation effect and interpreted that the mediation was different from zero if the confidence interval did not include zero. For teaching assistants, we found mediation significantly greater than zero, indirect = 0.74 (*SE* = 0.14), *Z* = 5.15, *p* < .001, 95% CI[0.48, 1.02]. Per-course faculty showed mediation between expected grade and overall course rating, indirect = 0.52 (*SE* = 0.09), *Z* = 6.06, *p* < .001, 95% CI[0.36, 0.73]. Instructors showed a similar indirect mediation effect, indirect = 0.53 (*SE* = 0.07), *Z* = 7.31, *p* < .001, 95% CI[0.40, 0.66]. Last, faculty showed the smallest mediation effect, indirect = 0.23 (*SE* = 0.02), *Z* = 8.71, *p* < .001, 95% CI[0.19, 0.28], wherein the confidence interval did not include zero, but also did not overlap with any other instructor group.

**Discussion**

The findings support the model that student evaluations of Psychology faculty are related to what one might consider leniency (i.e., overall average scores of B) in grading through perceptions of assignment appropriateness, grading fairness, and the expected course grade. This position is supported both in the strong relationships between expected grade and global ratings by the evidence that greater training and experience is related to poorer evaluations, lower expected grades, and lower relationships between grading and evaluations. Faculty received lower scores than teaching assistants in every category and often lower scores than per-course faculty, but not instructors. Mediation analyses showed that expected grade is positively related to overall course ratings, although this relationship is mediated by the perceived grading in the course. Therefore, as students have higher expected grades, the perceived grading scores increase, and the overall course score also increases. Moderation of this mediation effect indicated differences in the strength of the relationships between expected grade, grading questions, and overall course rating, wherein faculty generally had weaker relationships between these variables.

Because the study was not experimental, causal conclusions from this study alone need to be limited. However, Salmons (1993) provides some evidence of the causal direction of student ratings of instructors and expected grades. She had 444 students complete faculty evaluations after 3-4 weeks of classes, and again after 13 weeks. Students who expected to get Fs significantly lowered their evaluations while students who expected to receive As and Bs significantly raised their evaluations.

It is compelling that the correlations suggest that we can do a better job of understanding global ratings, perception of exams, fairness, and appropriateness of assignments based upon what grade students expected as compared to relating these ratings using ratings for the same course in a different semester or ratings for a different course in the same semester for instructor (i.e., correlations between items in the same semester are higher than reliability estimates across the board). It is very likely that these correlations with expected grade are suppressed by the loading of scores at the high end of the scale for course ratings and expected grade. Generally, evaluation items reflect scores at the high end of the 1-5 scale (see Table 3) even when items are intentionally constructed to move evaluators from the ends. The item, “The overall quality of this course was among the top 20% of those I have taken,” is conspicuously designed to move subjects away from the top rating. Yet average global ratings remain about a 4.00. The grade expectation average was approximately 4.00, which relates to a B average or 3.00 GPA.

One way of establishing convergent validity would be a finding that better trained and more experienced teachers get higher ratings than less well trained instructors. If the measure were valid, we would expect that regular faculty and full time instructors would get higher ratings than per course faculty and teaching assistants. To argue otherwise is to challenge the merits of higher education units with a faculty of professors with doctoral status. If the university were a researcher powerhouse where faculty research is emphasized over teaching and graduate assistants are admitted from the highest ranks of undergraduates, the finding that teaching assistants and per course faculty get higher ratings might be less of a challenge to the validity of these ratings. However, the university at which the data were collected is a non-doctoral program with greater emphasis on teaching and moderate emphasis on research, and teaching assistants are master’s candidates with less substantial admission expectations than doctoral programs. Hence, these findings challenge the convergent validity of the teaching evaluations.

Like most studies in this area, a major limitation is the absence of an independent measure of learning. Of course, this limitation is based upon the belief that the goal is to create educated persons, not just satisfied consumers. Even when common tests are used, these are invalid if the instructors are aware of the course content. Teachers seeking high evaluations are able to improve their ratings and scores by directly addressing the content of the specific test items. ETS now allows faculty who administer Major Field Tests to access the specific items which thereby invalidates it as a measure for these purposes. Ultimately, answering questions about the validity of student evaluations is a daunting task without such measures.

Evidence suggests that student evaluations are influenced by likability, attractiveness, and dress (Buck & Tiene, 1989; Feely, 2002; Gurung & Vespia, 2007) in addition to leniency and low demands (Greenwald & Gilmore, 1997). One must question whether a factor like instructor warmth, which relates to student evaluation (Best & Addison, 2000), is really fitting to the ultimate purposes of a college education. In a unique setting where student assignments to courses were random and common tests were used, Carrell and West (2010) demonstrated that teaching strategies that enhanced student evaluations led to poorer performance in subsequent classes. With the sum of invalid variance from numerous factors being potentially high, establishment of a high positive relationship to independent measures of achievement is essential to the acceptance of student evaluations as a measure of teaching quality.

Perception of the influence of leniency on teacher evaluations is far more detrimental to the quality of education than the biased evaluations themselves. It is unlikely that good teachers, even if more challenging, will get bad evaluations (i.e. evaluations where the majority of students rate the course poorly). Good teachers are rarely losing their positions due to low quality evaluations. But Marsh (1987) found that faculty perceives evaluations to be biased based upon course difficulty (72%), expected grade (68%), and course workload (60%). If one’s goal is high merit ratings and teaching awards, and the most significant factor is student evaluations of teaching, then putting easier and low-level questions on the test, adding more extra credit, cutting the project expectations, letting students off the hook for missing deadlines, and boosting borderline grades would all be likely strategies for boosting evaluations.

Effective teachers will get positive student ratings even when they have high expectations and do not inflate grades. But, many excellent teachers will score below average. It is maladaptive to try to increase a 3.90 global rating to a 4.10, because it often requires that the instructor try to emphasize avoidance of the lowest rating (1.00) because these low ratings in a skewed distribution have in inordinate influence on the mean. This effort of competing against the norms is likely to lead to grade inflation and permissiveness for the least motivated and most negligent students. Some researchers (Ellis et al., 2003; Greenwald et al., 1999) argue that student evaluations of instruction should be adjusted on the basis of grades assigned. However, there are problems with such an approach. The regression Betas are likely to differ based upon course and many other factors. In our research and in research by DuCettte and Kenney (1982), substantial variation in correlations was found across different course sets. Establishing valid adjustments would be problematic at best. Further, such an approach would punish instructors when they happen to get an unusually intelligent and motivated class (or teach an honors class) and give students the grades they deserve. Student evaluations are not a proper motivational factor for instructors in grade assignment, whether it is to inflate or deflate grades.

It would seem nearly impossible to eliminate invalid bias in student ratings of instruction. Yet, they may tell us a teacher is ineffective when the majority give poor ratings. It is the normative, competitive use that makes student evaluations of teaching subject to problematic interpretation. This finding is especially critical in light of recent research that portrays that student evaluations are largely biased against female teachers, and that student bias in evaluation is related to course discipline and student gender (Boring, Ottoboni, & Stark, 2016). Boring et al. also examine the difficulty in adjusting faculty evaluation for bias and determined that the complex nature of ratings makes unbiased evaluation nearly impossible. Stark and Freishtat (2014) further explain that evaluations are often negatively related to more objective measures of teaching effectiveness, and biased additionally by perceived attractiveness and ethnicity. In line with the current paper, he suggests dropping overall teaching effectiveness or value of the course type questions because they are influenced by many variables unrelated to actual teaching. Last, they suggest the distribution and response rate of the data are critical information, and this point becomes particularly important when recent research shows that online evaluations of teaching experience a large drop in response rates (Stanny & Arruda, 2017). Our study contributes to the literature of how student evaluations are a misleading and unsuccessful measure of teaching effectiveness, especially focusing on reliability and the impact of grading on overall questions. We conclude that it may be possible to manipulate these values by lowering teaching standards, which implies that high stakes hiring and tenure decisions should probably follow the advice of Palmer, Bach, and Streifer (2014) or Stanny, Gonzalez, and McGowan (2015) in implementing teaching portfolios and syllabus review, particularly because a recent meta-analysis of student evaluations showed they are unrelated to student learning (Uttl, White, & Gonzalez, 2017).

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social

psychological research: Conceptual, strategic, and statistical considerations. *Journal of*

*Personality and Social Psychology*, *51*(6), 1173-1182. doi: 10.1037//0022-3514.51.6.1173

Best, J.B., & Addison, W.E., (2000). A preliminary study of perceived warmth of professor and student evaluations. *Teaching of Psychology, 27,* 60-62.

Blackhart, G. C., Peruche, B.M., DeWall, C.N, & Joiner, T.E. Jr. (2006). Factors influencing teaching evaluations in higher education. *Teaching of Psychology, 33,* 37-39. doi:10.1207/s15328023top3301\_9

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not

measure teaching effectiveness. *ScienceOpen Research*. doi:10.14293/s2199-

1006.1.sor-edu.aetbzc.v1

Buck, S., & Tiene, D. (1989). The impact of physical attractiveness, gender, and teaching

philosophy on teacher evaluations. *Journal of Educational Research, 82,* 172-177.

doi:10.1080/00220671.1989.10885887

Carrell, S.E., & West, J.E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118,* 409-432.

Centra, J.A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44,* 495-518. doi:10.1023/A:1025492407752

DuCette, J., & Kenney, J. (1982). Do grading standards affect student evaluations of teaching? Some new evidence on an old question. *Journal of Educational Psychology, 74*, 308-314. doi:10.1037/0022-0663.74.3.308

Ellis, L., Burke, D.M., Lomire, P., & McCormick, D.R. (2003). Student grades and average ratings of instructional quality: The need for adjustment. *The Journal of Educational Research, 97,* 35-40. doi:10.1080/00220670309596626

Feeley, T.H. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education, 51,* 225-236. doi:10.1080/03634520216519

Festinger, L. (1957). *A theory of cognitive dissonance*, Stanford, CA: Stanford University Press.

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

Gurung, R.A.R., & Vespia, K.M. (2007). Looking good, teaching well? Linking liking, looks, and learning. *Teaching of Psychology, 34,* 5-10. doi:10.1080/00986280709336641

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis*.

New York: The Guilford Press.

Isley, P., & Singh, H. (2005). Do higher grades lead to favorable evaluations? *The Journal of Economic Education, 36,* 29-42. doi:10.3200/JECE.36.1.29-42

Krautman, A. C., & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review, 18,* 59-63. doi:10.1016/S0272-7757(98)00004-1

Miller, A. (1985). A developmental study of the cognitive basis of performance impairment after failure. *Journal of Personality and Social Psychology, 49*, 529-538. doi:10.1037/0022-3514.49.2.529

Miller, A., & Klein, J. (1989). Individual differences in ego value of academic performance and persistence after failure. *Contemporary Educational Psychology, 14*, 124-132. doi:10.1016/0361-476X(89)90030-1

Marsh, H.W. (1987). Students’ evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11,* 253-288. doi:10.1016/0883-0355(87)90001-2

Marsh, H.W., & Roche, L.A. (2000). Effects of grading leniency and low workload on students’ evaluations of Teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92,* 202-228. doi:10.1037/0022-0663.92.1.202

Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. *Teaching of Psychology, 33, 176-179.* doi:10.1207/s15328023top3303\_4

Palmer, M. S., Bach, D. J., & Streifer, A. C. (2014). Measuring the promise: A learning-focused

syllabus rubric. *To Improve the Academy: A Journal of Educational Development, 33*,

14–36. doi:10.1002/tia2.20004

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2017). nlme: Linear and

Nonlinear Mixed Effects Models. R package version 3.1-131, https://CRAN.R-

project.org/package=nlme

Salmons, S.D. (1993). The relationship between students’ grades and their evaluation of instructor performance. *Applied H.R.M. Research, 4,* 102-114.

Stanny, C. J., & Arruda, J. E. (2017). A Comparison of student evaluations of teaching with online and paper-based administration. *Scholarship of Teaching and Learning in Psychology*, 1-10. doi:10.1037/stl0000087

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

Stark, P., & Freishtat, R. (2014). An Evaluation of course evaluations. *ScienceOpen Research*. doi:10.14293/s2199-1006.1.sor-edu.aofrqa.v1

Snyder, M.L., Stephan, W.G., & Rosenfield, D. (1978). Attributional egotism. In J.H. Harvey, W.J. Ickes, and R.R. Kidd (Eds), *New directions in attribution research (Vol. 2)* Hillsdale, N.J.: Erlbaum.

Tabachnick, B. G., and Fidell, L. S. (2012). *Using multivariate statistics*, 6th ed. Boston:

Pearson.

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty’s teaching

effectiveness: Student evaluation of teaching ratings and student learning are not related.

*Studies in Educational Evaluation, 54*, 22–42. doi:10.1016/j.stueduc.2016.08.007

Weiner, B. (1992). *Human Motivation: Metaphors, theories, and research.* Newbury Park, CA: Sage.

Table 1.

*Partial Correlations Between Class Ratings of Instructors Across Variations in Course and Semester Pairings*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Evaluation Item | | | | |
|  | Overall | Exams | Fair | Assignments | Grade |
| Semester pair type | Same Course pairing | | | | |
| Same semester  1806 cases | **.32** | **.39** | **.31** | **.27** | **.43** |
| Different semester  41048 cases | <-.01 | -.01 | **.06** | <.01 | **-.18** |
|  | Different course | | | | |
| Same semester  1582 cases | **.05** | **.02** | **.03** | **.02** | **.08** |
| Different semester  75,138 cases | **-.08** | **-.08** | **-.08** | **-.08** | **-.17** |

*Note*. Bolded correlations are significant at *p* < .05

Table 2.

*Item Correlations Using All Classes*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Undergraduate Courses | | | | | |
|  | | Overall | Exams | Fair | Assignments | Grade |
| Graduate Courses | Overall | - | .81 | .87 | .93 | .59 |
| Exams | .65 | - | .73 | .68 | .44 |
| Fair | .86 | .61 | - | .68 | .41 |
| Assignments | >.99 | .48 | .66 | - | .44 |
| Grade | .44 | .13\* | .34 | .35 | - |

*Note*. All correlations are significant at *p* < .001, except \* *p* = .092.

Table 3.

*Student Evaluation in General Education Classes as a Function of the Status of the Course Instructor*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Status of the Course Instructor | | | |
|  | TA  M SD | Per Course  M SD | Instructor  M SD | Ranked  M SD |
| Overall | 4.03 0.52 | 3.95 0.75 | 3.71 0.50 | 3.74 0.47 |
| Exams | 4.48 0.34 | 4.27 0.57 | 4.06 0.36 | 4.11 0.35 |
| Fair | 4.57 0.25 | 4.43 0.45 | 4.35 0.26 | 4.40 0.26 |
| Assignments | 4.31 0.31 | 4.18 0.39 | 4.17 0.30 | 4.03 0.31 |
| Grade | 4.30 0.32 | 4.19 0.33 | 4.02 0.30 | 4.00 0.30 |
|  | | | | |

|  |  |
| --- | --- |
| Expected  Grade  Perceived  Grading  Overall Rating  *a* 0.55  *b* 1.37  *c* 0.71  *c'* -0.04  Teaching Assistants | Expected  Grade  Perceived  Grading  Overall  Rating  *a* 0.41  *b* 1.28  *c* 0.41  *c'* -0.10  Per-Course Faculty |
| Expected  Grade  Perceived  Grading  Overall  Rating  *a* 0.45  *b* 1.16  *c* 0.56  *c'* 0.04  Full-Time Instructors | Expected Grade  Perceived Grading  Overall  Rating  *a* 0.21  *b* 1.10  *c* 0.39  *c'* 0.16  Full-Time Faculty |
| *Figure 1.* Mediation models for moderated mediation analysis indicating mediation effects for each type of teacher. Expected grading indicates student entered grade expected in the course, perceived grading is an average score of fairness, appropriateness, and exam grading questions, and overall rating indicates the omnibus rating for a course. | |