Perceived Grading and Student Evaluation of Instruction

Erin M. Buchanan[1], Becca N. Huber[1], Arden Miller[1], David W. Stockburger[2], & Marshall Beauchamp[3]

[1] Missouri State University

[2] US Air Force Academy

[3] University of Missouri - Kansas City

Author Note

A portion of this research was presented at the meeting of the Southwestern Psychological Association, April, 2009, San Antonio, TX. The authors would like to thank Melissa Fallone for comments on earlier drafts and Stephen Martin for his help with restructuring the data.

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, perceived grading, reliability

Perceived Grading and Student Evaluation of Instruction

Student evaluations of professors have been disputed over time with regard to validity and reliability. The impact of student evaluations on professor advancement can be great and often acts as a deciding factor in professor promotion or demotion, along with the access to certain funding opportunities, coursework choice, and tenureship. There are certain variables researched that result in improving evaluations, such as giving higher grades (**???**; **???**; **???**). Student evaluations have also been found to be influenced by likability, attractiveness, and dress (**???**; **???**; **???**). Further, 20 years ago, (Neath 1996) suggests 20 tips in which professors may bolster their evaluations from students that have no relationship with proven instructional methods or further learning retention among the student body, such as being a male professor and only teaching only male students. In more recent research, (**???**) confirms that student evaluations of teaching are biased against female instructors, and the authors conclude student evaluations are more representative of the students' grading expectations and biases rather than an evaluation of objective instructional methods. All together, these findings elicit the argument that student evaluations are not necessarily measuring whether the instructional methods of professors are sound, rather student evaluations of instruction are measuring whether or not the professor is likeable among the students with regard to their expectation of their performance in the classroom, in addition to the instructor meeting pre-existing biases.

However, some authors (**???**; **???**; **???**) have discovered professors are not able to increase their positive evaluations by only providing their students with higher grades. We believe this is due to what we consider the effect of "perceived grading". We operationally define perceived grading as the students' perceptions of assignment appropriateness, grading fairness, and the expected course grade at the time the evaluations are being completed. We believe social psychology theory would support that students with low perceived grading may reduce cognitive dissonance and engage in ego defense by giving low evaluations of professors who give them lower grades (**???**), resulting in decreased validity and reliability of

the proposed construct, professor instruction. We argue that both social psychology theory and the evidence from student evaluations supports that higher perceived grading can lead to better student evaluations of instruction. This theory and evidence from student evaluation leads us to further posit student evaluations of professors as biased methods of data collection and irrelevant to the quality of the instructor and the instructional methods used over the course of a semester.

Much of the literature on student evaluations involves diverse and complex analyses (e.g., (**???**)) 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. As we know, failure threatens the ego [(**???**); 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 (**???**). One common strategy involves diminishing the value of the activity (**???**), which would result in lowered perceived value of a course.

Similarly, Cognitive Dissonance Theory (**???**) predicts that people who experience poor performance but perceive themselves as competent will experience dissonance, of which they can reduce through negative evaluations of the instruction (**???**). Attribution research (**???**) 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 (**???**).

Considerable research has been conducted in support of widely distributed evaluation systems. (**???**) 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 professor attributes. This restriction in range suppresses correlation.

However, (**???**) provided causal evidence of lowered student evaluations due to expected grades. In her study of 444 students completing faculty evaluations at two separate points in a semester, students who expected to get Fs significantly lowered their evaluations while students who expected to receive As and Bs significantly raised their evaluations.

(**???**) argued that the individual is also 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 (do we have a source for this claim???). 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 (**???**; **???**; **???**).

Additionally, (**???**) 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 student evaluation of instructional quality. We must either accept that the least trained and qualified instructors are actually better teachers, or we must believe this result suggests that student evaluations have given us false information on the quality of instruction via their perceptions of grading.

(**???**) provided further 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, (**???**) 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, which was one aim of the current study.

It is pertinent to note that 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. Hence, analysis of these correlations within the same discipline and course would be expected to strengthen the relationship between expected grades and quality measures, offering more valid results.

Further, 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(??? do we have a source for halo effects from internal consistency reliability???). 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. As a result, we should be able to deduce if student ratings can be considered a valid measure of an instructor's teaching skills if they are or are not able to reliably differentiate instructors within the same course across different semesters.

If ratings are, in fact, valid measures of instructor attributes, it should be expected that ratings would have some stability across semester and specific 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, in the current study, we first sought to establish reliability of ratings for the instructors across courses and semesters.

The current study 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 at the same university. After examining reliability, we sought to show that items on instructor evaluations were positively correlated for undergraduate and graduate students (???didn't we

want to eliminate graduate students from the analyses because they're a special population???), demonstrating that overall course evaluations are related to the perceived grading of the students. We also expected correlations to be substantially higher than those obtained by previous researchers who used individual students as their unit of analysis, since we used the course as the 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 (**???**). 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 4313 undergraduate, 397 mixed-level undergraduate, and 687 graduate psychology classes taught from 1987 to 2016 that were evaluated by students using the same 15-item instrument. The graduate courses were excluded from analyses due to the ceiling effects on expected grades. 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.

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 scale from 1 (*strongly disagree*) to 5 (*strongly agree*):

```
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
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 as graduate assistant, per-course faculty, instructors, and tenure-track faculty. This data was considered structured by instructor; therefore, all analyses below were coded in *R* using the *nlme* package (**???**) to control for correlated error of instructor as a random intercept in a multilevel model. The overall dataset was screened for normality, linearity, homogeneity, and homoscedasticity using procedures from (**???**). Data generally met assumptions with a slight skew and some heterogeneity. This data was not screened for outliers because it was assumed that each score was entered correctly from student evaluations. The complete set of all statistics can be found online at http://osf.io/jdpfs. This page also includes the manuscript written inline with the statistical analysis with the *papaja* package (**???**) for interested researchers/reviewers.

Reliability of Instructor Scores DONE

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 every instructor, semester, and course combination. Therefore, this created eight possible combinations of matching v. no match for instructor by semester by course. Multilevel models were used to calculate correlations on each fo the eight combinations controlling for response size for both courses (i.e., course 1 number of ratings and course 2 number of ratings) and random

intercepts for instructor(s). Correlations were calculated separately for each question, however, the overall pattern of the data was the same for each of the eight combinations, and these were averaged for Table @ref:(tab:rel-table). The complete set of all correlations can be found online. Because the large sample size can bias statistical significance, we focused on the size of the correlations. The correlations were largest for the same instructor in the same semester and course, followed by the same instructor in the same semester with a different course and the same instructor in a different semester with the same course. The first shows that scores are somewhat reliable (i.e., rs $\sim$ .45) for instructors teaching two or more of the same class at the same time. The correlations within instructor then drop to rs $\sim$ .09 for the same semester or same course. All other correlations are nearly zero, with the same semester, same course, and different instructor as the next largest at rs $\sim$ .05. Given these values are still low for traditional reliability standards, this results may indicate that student demand characteristics or course changes impact instructor ratings.

Correlations of Evaluation Questions DONE

Table 2 presents the inter-correlations for the five relevant evaluation questions using instructor as a random intercept in a multilevel model with evaluation sample size as an adjustor variable. The partial correlation (pr) is the standardized coefficient from the multilevel model analysis between items while adjusting for sample size and random effects of instructor. The raw coefficient b, standard error, and significance statistics are also provided. We found class expected grade was related to class overall rating, exams reflecting the material, grading fairness, and appropriateness of assignments; however, these partial correlations were approximately half of all other pairwise correlations. The correlations between grading related items were high, representing some consistency in evaluation, as well as the overall course evaluation to grading questions.

Instructor Status and Ratings DO WE WANT THIS

We compared teaching assistants, per-course faculty, instructors, and ranked faculty in undergraduate courses that included evaluations for all four types of teacher, usually general education classes (i.e., Introductory Psychology), required major courses (i.e., Statistics, Research Methods), and popular electives (i.e., Abnormal Psychology). This anlaysis included 179 teachers: 49 teaching assistants, 54 per-course instructors, 17 instructors, and 59 full-time faculty who taught 2744 courses: 266 teaching assistants, 400 per-course instructors, 354 instructors, and 1724 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). 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 version of the Sobel test (Baron & Kenny, 1986), as well as bootstrapped samples to determine the 95% confidence interval of the mediation (Preacher & Hayes, 2008; **???**) 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, (**???**) 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 (**???**; **???**; **???**) in addition to leniency and low demands (**???**). One must question whether a factor like instructor warmth, which relates to student evaluation (**???**), 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, (**???**) 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 (**???**) 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 (**???**; **???**) 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 (**???**). (**???**) also examine the difficulty in adjusting faculty evaluation for bias and determined that the complex nature of ratings makes unbiased evaluation nearly impossible. (**???**) 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 (**???**). 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 (**???**) or (**???**) in implementing teaching portfolios and syllabus review, particularly because a recent meta-analysis of student evaluations showed they are unrelated to student learning (**???**).

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

Table 1

Correlations for Instructor, Semester, and Course Combinations

Instructor	Semester	Course	*b*	*SE*	*t*	*df*	*p*
Different Instructor	Different Semester	Different Course	-.001	.000	10144295	-3.58	.013
Different Instructor	Same Semester	Different Course	.006	.002	152801	2.91	.048
Different Instructor	Different Semester	Same Course	.008	.001	517353	6.24	.027
Different Instructor	Same Semester	Same Course	.054	.010	6265	5.40	$< .001$
Same Instructor	Different Semester	Different Course	-.038	.003	108849	-13.13	$< .001$
Same Instructor	Same Semester	Different Course	.095	.020	1872	4.66	$< .001$
Same Instructor	Different Semester	Same Course	.090	.004	55057	21.77	$< .001$
Same Instructor	Same Semester	Same Course	.446	.023	1401	19.63	$< .001$

Table 2

t Statistics for Undergraduate Correlations

Coefficient	*pr*	*b*	*SE*	*df*	*t*	*p*
Overall to Exams	.637	.828	.014	4447	60.813	$< .001$
Overall to Fair	.606	.903	.016	4447	57.837	$< .001$
Overall to Assignments	.675	.999	.016	4447	63.251	$< .001$
Overall to Grade	.344	.597	.022	4447	27.167	$< .001$
Exams to Fair	.655	.751	.012	4447	61.387	$< .001$
Exams to Assignments	.615	.700	.014	4447	50.425	$< .001$
Exams to Grade	.311	.416	.018	4447	23.066	$< .001$
Fair to Assignments	.720	.715	.011	4447	63.912	$< .001$
Fair to Grade	.375	.438	.016	4447	27.865	$< .001$
Assignments to Grade	.344	.404	.015	4447	26.913	$< .001$