[1]   Perceived Grading and Student Evaluation of Instruction

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

[3]   Beauchamp[3]

[4]   [1] Missouri State University

[5]   [2] US Air Force Academy

[6]   [3] University of Missouri - Kansas City

[7]   Author Note

12   Abstract

13   We analyzed student evaluations for 3,585 classes collected over 20 years to determine

14   stability and evaluate the relationship of perceived grading to global evaluations, perceived

15   fairness, and appropriateness of assignments. Using class as the unit of analysis, we found

16   small evaluation reliability when professors taught the same course in the same semester,

17   with much weaker correlations for differing courses. Expected grade and grading related

18   questions correlated with overall evaluations of courses. Differences in course evaluations on

19   expected grades, grading questions, and overall grades were found between full-time faculty

20   and other types of instructors. These findings are expanded to a model of grading type

21   questions mediating the relationship between expected grade and overall course evaluations

22   with a moderating effect of type of instructor.

23   *Keywords:* Student evaluation, teacher evaluation, perceived grading, reliability

₂₄   Perceived Grading and Student Evaluation of Instruction

₂₅   Student evaluations of professors have been disputed over time with regard to validity

₂₆ and reliability. The impact of student evaluations on professor advancement can be great

₂₇ and often acts as a deciding factor in professor promotion, demotion, or inform access to

₂₈ certain funding opportunities, coursework choice, and tenureship. There are certain variables

₂₉ researched that result in improving evaluations, such as giving higher grades (Greenwald &

₃₀ Gillmore, 1997; Isely & Singh, 2005; Krautmann & Sander, 1999). Student evaluations have

₃₁ also been found to be influenced by likability, attractiveness, and dress (Buck & Tiene, 1989;

₃₂ Gurung & Vespia, 2007; Hugh Feeley, 2002). Further, 20 years ago, (Neath 1996) suggests 20

₃₃ tips in which professors may bolster their evaluations from students that have no

₃₄ relationship with proven instructional methods or further learning retention among the

₃₅ student body, such as being a male professor and only teaching only male students. In more

₃₆ recent research, Boring, Ottoboni, and Stark (2016) confirms that student evaluations of

₃₇ teaching are biased against female instructors, and the authors conclude student evaluations

₃₈ are more representative of the students' grading expectations and biases rather than an

₃₉ evaluation of objective instructional methods. All together, these findings elicit the argument

₄₀ that student evaluations are not necessarily measuring whether the instructional methods of

₄₁ professors are sound, rather student evaluations of instruction are measuring whether or not

₄₂ the instructor met the students' expectations of their performance in the classroom, in

₄₃ addition to the instructor meeting pre-existing biases.

₄₄   However, some authors (Centra, 2003; Marsh, 1987; Marsh & Roche, 2000) have

₄₅ discovered professors are not able to increase their positive evaluations by only providing

₄₆ their students with higher grades to meet expectations. We believe this is due to what we

₄₇ consider the effect of "perceived grading". We operationally define perceived grading as the

₄₈ students' perceptions of assignment appropriateness, grading fairness, and the expected

₄₉ course grade at the time the evaluations are being completed. Social psychology theory

₅₀ would support that students with low perceived grading may reduce cognitive dissonance

⁵¹ and engage in ego defense by giving low evaluations of professors who give them lower grades

⁵² (Maurer, 2006), resulting in decreased validity and reliability of the proposed construct,

⁵³ professor instruction. We argue both social psychology theory and the evidence from student

⁵⁴ evaluations supports that higher perceived grading can lead to better student evaluations of

⁵⁵ instruction. For example, Salmons (1993) provided causal evidence of lowered student

⁵⁶ evaluations due to expected grades. In her study of 444 students completing faculty

⁵⁷ evaluations at two separate points in a semester, students who expected to get Fs

⁵⁸ significantly lowered their evaluations while students who expected to receive As and Bs

⁵⁹ significantly raised their evaluations (Salmons, 1993). This theory and evidence from student

⁶⁰ evaluation leads us to further argue student evaluations of professors are biased methods of

⁶¹ data collection and irrelevant to the quality of the instructor and the instructional methods

⁶² used over the course of a semester.

⁶³ Much of the literature on student evaluations involves diverse and complex analyses

⁶⁴ (e.g., Marsh (1987)) and lacks social-psychological theoretical guidance on human judgment.

⁶⁵ To expect that student evaluations would not be influenced by expected grade would

⁶⁶ contradict a long-standing history of social psychology research on cognitive dissonance,

⁶⁷ attribution, and ego threat. As we know, failure threatens the ego [Miller (1985);] and

⁶⁸ motivates us to find rationales to defend the ego. Failing students, or those performing below

⁶⁹ personal expectations, would be expected to defend their ego by attributing low grades to

⁷⁰ poor teaching or unfair evaluation practices (Maurer, 2006). One common strategy involves

⁷¹ diminishing the value of the activity (Miller & Klein, 1989), which would result in lowered

⁷² perceived value of a course.

⁷³ Similarly, Cognitive Dissonance Theory (Festinger, 1957) predicts that people who

⁷⁴ experience poor performance but perceive themselves as competent will experience

⁷⁵ dissonance, of which they can reduce through negative evaluations of the instruction

⁷⁶ (Maurer, 2006). Attribution research (Weiner, 1992) also supports the argument that among

⁷⁷ low achievement motivation students, failure is associated with external attributions for

78 cause, and the most plausible external attribution for the student in the evaluation context

79 is the quality of instruction and grading practices. Although arguments regarding degree of

80 influence are reasonable, the position that they are not affected is inconsistent with existing

81 and established theory. Thus, it is not surprising that the majority of faculty perceive

82 student evaluations to be biased by perceived grading and course choice (Marsh, 1987).

83   <<<<<<< HEAD Considerable research has been conducted in support of widely

84 distributed evaluation systems. Centra (2003) reported in a study of 9,194 class averages

85 using the Student Instructional Support, the relationship between expected grades and

86 global ratings was only .20. He further argued that when variance due to perceived learning

87 outcomes was regressed from the global evaluation, the effect of expected grades was

88 eliminated (Centra, 2003). However, a student's best assessment of "perceived learning

89 outcome" is their expected grade, and thus, these should be highly correlated. When

90 perceived learning is regressed from the global evaluations, it is not surprising that

91 suppression effects would eliminate or could even reverse the correlation between expected

92 grade and global evaluation. In general, there are several reasons why the relationship of

93 expected grade to global evaluations is suppressed. For example, faculty ratings are generally

94 very high on average (i.e. quality instructors are hired), which restricts variation; thus,

95 weakening their reliability as a measure of professor attributes. This restriction in range

96 suppresses correlation. ======= Considerable research has been conducted in support of

97 widely distributed evaluation systems. Centra (2003) reported that in a study of 9,194 class

98 averages using the Student Instructional Support, the relationship between expected grades

99 and global ratings was only .20. He further argued that when variance due to perceived

100 learning outcomes was regressed from the global evaluation, the effect of expected grades was

101 eliminated. However, a student's best assessment of "perceived learning outcome" is their

102 expected grade, and thus, these should be highly correlated. When perceived learning is

103 regressed from the global evaluations, it is not surprising that suppression effects would

104 eliminate or could even reverse the correlation between expected grade and global evaluation.

105 In general, there are several reasons why the relationship of expected grade to global

106 evaluations is suppressed. For example, faculty ratings are generally very high on average

107 (i.e. quality instructors are hired), which restricts variation; thus, weakening their reliability

108 as a measure of professor attributes. This restriction in range suppresses correlation.

109 >>>>>>> a60e1b8c93a3086ea9ee92e7800f2e5fb1fd8360

110   Marsh (1987) argued that the individual is also not the proper unit of analysis because

111 such analyses could suggest false findings related to individual differences in students.

112 Therefore, he argued the use of class as the suggested unit of analysis (Marsh, 1987). We

113 agree, both for his reasoning and because analyses with individual ratings can mask

114 significant relationships as well (do we have a source for this claim???). Individual

115 differences in expectancy will attenuate the correlation less when class average is used as the

116 unit of analysis. To the extent that the same class average would be expected across all

117 courses, an assumption we challenged, the class average for expected grade is a good measure

118 of perceived grading as an instructor attribute. Course quality, not individual attributes of

119 students, is what we attempted to assess when we used student evaluations of courses.

120 Several studies provide support that when class is the unit of analysis, expected grade is a

121 more significant biasing factor in student evaluations (Blackhart, Peruche, DeWall, & Joiner,

122 2006; DuCette & Kenney, 1982; Ellis, Burke, Lomire, & McCormack, 2003).

123   Additionally, Blackhart et al. (2006) analyzed 167 psychology classes in a multiple

124 regression analysis with class as the unit of analysis and found that the two most significant

125 predictors of instructor ratings were average grade given by the instructor and instructor

126 status (TA or rank of faculty). Because of the limited number of classes, the power of the

127 analysis was limited. However, in addition to the concern regarding the relationship between

128 grades and global course evaluations, it was found that TAs were rated more highly than

129 ranked faculty. This finding raises additional questions on validity student evaluation of

130 instructional quality. We must either accept that the least trained and qualified instructors

131 are actually better teachers, or we must believe this result suggests that student evaluations

132 have given us false information on the quality of instruction via their perceptions of grading.

133   DuCette and Kenney (1982) provided further evidence that using course as a unit of

134 analysis increased the correlation between expected grade and other course ratings. Within

135 specific groupings of classes, these correlations ranged from .23 to .53. Two factors limited

136 the level of their relationships. First, the classes used were all upper division courses and

137 graduate courses. Secondly, over 90% of the students in these classes expected an A or a B.

138 Consequently, the correlations between expected grade and global course ratings would be

139 reduced due to the absence of variation in expected grades. Similarly, Ellis et al. (2003)

140 found a correlation of .35 between average course grade and average rating of the instructor

141 in 165 classes during a two-year period. However, these studies did not consider the

142 predictive relationship for instructors across different courses and semesters, which was one

143 aim of the current study.

144   It is pertinent to note that different disciplines and subject areas have diverse GPA

145 standards, and students have differing grade and workload expectations in different courses,

146 as well. For example, an instructor in Anatomy giving a 3.00 GPA might be considered

147 lenient while an Education instructor giving a 3.25 GPA might be considered hard (examples

148 for illustration only). To have a valid measure of workload and leniency factors, correlations

149 should be conducted with varied teachers of the same course. Further, different populations

150 take courses in different disciplines, resulting in potential population differences between

151 anatomy classes and education classes, which could create or mask findings. Hence, analysis

152 of these correlations within the same discipline and course would be expected to strengthen

153 the relationship between expected grades and quality measures, offering more valid results.

154   Further, in most studies of student evaluations, reliability is established through

155 internal consistency reliability. However, this form of reliability is confounded with halo

156 effects (i.e. a cognitive bias that influences ratings based on an overall perception of the

157 person teaching, rather than the individual components of the course), and tells only

158 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 (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 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

₂₁₂ considered structured by instructor; therefore, all analyses below were coded in R using the

₂₁₃ *nlme* package (Pinheiro, Bates, Debroy, Sarkar, & Team, 2017) to control for correlated error

₂₁₄ of instructor as a random intercept in a multilevel model. The overall dataset was screened

₂₁₅ for normality, linearity, homogeneity, and homoscedasticity using procedures from

₂₁₆ Tabachnick and Fidell (2012). Data generally met assumptions with a slight skew and some

₂₁₇ heterogeneity. This data was not screened for outliers because it was assumed that each

₂₁₈ score was entered correctly from student evaluations. The complete set of all statistics can

₂₁₉ be found online at http://osf.io/jdpfs. This page also includes the manuscript written inline

₂₂₀ with the statistical analysis with the *papaja* package (**???**) for interested

₂₂₁ researchers/reviewers.

Reliability of Instructor Scores DONE

₂₂₃   Reliability of ratings of instructors can be inferred by the consistency of ratings across

₂₂₄ courses and semester, assuming that we infer there is a stable good/poor instructor attribute

₂₂₅ and that these multiple administrations of the same question are multiple assessments of

₂₂₆ that attribute. A file was created with all possible course pairings for every instructor,

₂₂₇ semester, and course combination. Therefore, this created eight possible combinations of

₂₂₈ matching v. no match for instructor by semester by course. Multilevel models were used to

₂₂₉ calculate correlations on each fo the eight combinations controlling for response size for both

₂₃₀ courses (i.e., course 1 number of ratings and course 2 number of ratings) and random

₂₃₁ intercepts for instructor(s). Correlations were calculated separately for each question,

₂₃₂ however, the overall pattern of the data was the same for each of the eight combinations, and

₂₃₃ these were averaged for Table @ref:(tab:rel-table). The complete set of all correlations can be

₂₃₄ found online. Because the large sample size can bias statistical significance, we focused on

₂₃₅ the size of the correlations. The correlations were largest for the same instructor in the same

₂₃₆ semester and course, followed by the same instructor in the same semester with a different

₂₃₇ course and the same instructor in a different semester with the same course. The first shows

that scores are somewhat reliable (i.e., rs ~ .45) for instructors teaching two or more of the
same class at the same time. The correlations within instructor then drop to rs ~ .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 ~ .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.

263   . All comparisons were made against full-time faculty to control for Type 1 error using

264   a multilevel model with a dummy coded instructor variable, and dummy coded t values were

265   used to determine which comparison groups were different from full-time faculty. Overall

266   means and standard deviations are presented in Table 3, and the complete set of t value

267   comparisons for these analyses can be found online. As shown in the Table 3, the ratings of

268   all groups were fairly high, hovering around 4.00 on a 5.00 point scale, and the expected

269   grade for courses was approximately a B.

270   For overall ratings, faculty were found to be rated less highly than teaching assistants,

271   p = .027, but not significantly different than per-course faculty (p = .181) or instructors (p

272   = .814). When rating if exams were representative of course material, full-time faculty were

273   rated lower than both teaching assistants (p < .001) and per-course faculty (p = .047), but

274   were not significantly different than instructors (p = .740). Full-time faculty were rated as

275   less fair and appropriate in their grades than teaching assistants (p = .003), while per-course

276   faculty (p = .128) and instructors (p = .657) had similar scores to faculty. Teaching

277   assistants were designated to have more appropriate assignments than faculty (p < .001),

278   while per-course (p = .060) and instructors (p = .073) had the same ratings as faculty on

279   assignments. Finally, faculty showed significantly lower expected grades than teaching

280   assistants (p < .001) and per-course faculty (p = .044), while having similar grades to

281   instructors (p = .705).

Moderated Mediation

283   Given the correlations between items and differences between items and ranked faculty,

284   we proposed a mediation relationship between expected grade, perceived grading, and overall

285   course grades that varies by instructor type. Figure 1 demonstrates the predicted

286   relationship between these variables. We hypothesized that expected course grade would

287   impact the overall course rating, but this relationship would be mediated by the perceived

288   grading in the course, which was calculated by averaging questions about exams, fairness of

289 grading, and assignments. Therefore, as students expected to earned higher grades

290 (leniency), their perception and ratings of the grading would increase, thus, leading to higher

291 overall course scores. This relationship was tested using traditional and newer approaches to

292 mediation (Baron & Kenny, 1986; Hayes, 2017). All categorical interactions were compared

293 to ranked faculty. Each step of the model is described below. Because significant interactions

294 were found, we calculated each group separately (Figure 1) to portray these differences in

295 path coefficients. Tables of t values for the overall and separated analyses are available at

296 http://osf.io/jdpfs.

297   **C Path.**   First, expected grade was used to predict the overall rating of the course,

298 along with the interaction of type of instructor and expected grade. The expected grade

299 positively predicted overall course rating, $p < .001$, wherein higher expected grades was

300 related to higher overall ratings for the course ($b = 0.39$). A significant interaction between

301 type and expected grade rating was found for instructors versus faculty. In looking at Figure

302 1, we find that instructors ($b = 0.56$) have a stronger relationship between expected grade

303 and overall course rating than faculty ($b = 0.39$, interaction $p = .020$), while per-course ($b =$

304 0.41, interaction $p = .621$) and teaching assistants ($b = 0.71$, interaction $p = .068$) were not

305 significantly different than faculty on the c path coefficient.

306   **A Path.**   Expected grade was then used to predict the average of the grading related

307 questions, along with the interaction of type of instructor. Higher expected grades were

308 related to higher ratings of appropriating grading ($b = 0.21$, $p < .001$), and a significant

309 interaction of faculty and all three other instructor types emerged: teaching assistants ($p =$

310 $.001$), per-course faculty ($p = .001$), and instructors ($p < .001$). As seen in Figure 1, faculty

311 ($b = 0.21$) have a much weaker relationship between expected grade and average ratings of

312 grading than teaching assistants ($b = 0.55$), per-course ($b = 0.41$), and instructors ($b =$

313 0.45). B and C' Paths. In the final model, expected grade, average ratings of grading, and

314 the two-way interactions of these two variables with type were used to predict overall course

315 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; Hayes, 2017) 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

343   The findings support the model that student evaluations of Psychology faculty are

344 related to what one might consider leniency (i.e., overall average scores of B) in grading

345 through perceptions of assignment appropriateness, grading fairness, and the expected course

346 grade. This position is supported both in the strong relationships between expected grade

347 and global ratings by the evidence that greater training and experience is related to poorer

348 evaluations, lower expected grades, and lower relationships between grading and evaluations.

349 Faculty received lower scores than teaching assistants in every category and often lower

350 scores than per-course faculty, but not instructors. Mediation analyses showed that expected

351 grade is positively related to overall course ratings, although this relationship is mediated by

352 the perceived grading in the course. Therefore, as students have higher expected grades, the

353 perceived grading scores increase, and the overall course score also increases. Moderation of

354 this mediation effect indicated differences in the strength of the relationships between

355 expected grade, grading questions, and overall course rating, wherein faculty generally had

356 weaker relationships between these variables.

357   Because the study was not experimental, causal conclusions from this study alone need

358 to be limited. However, Salmons (1993) provides some evidence of the causal direction of

359 student ratings of instructors and expected grades. She had 444 students complete faculty

360 evaluations after 3-4 weeks of classes, and again after 13 weeks. Students who expected to

361 get Fs significantly lowered their evaluations while students who expected to receive As and

362 Bs significantly raised their evaluations.

363   It is compelling that the correlations suggest that we can do a better job of

364 understanding global ratings, perception of exams, fairness, and appropriateness of

365 assignments based upon what grade students expected as compared to relating these ratings

366 using ratings for the same course in a different semester or ratings for a different course in

367 the same semester for instructor (i.e., correlations between items in the same semester are

368 higher than reliability estimates across the board). It is very likely that these correlations

370 with expected grade are suppressed by the loading of scores at the high end of the scale for

371 course ratings and expected grade. Generally, evaluation items reflect scores at the high end

372 of the 1-5 scale (see Table 3) even when items are intentionally constructed to move

373 evaluators from the ends. The item, "The overall quality of this course was among the top

374 20% of those I have taken," is conspicuously designed to move subjects away from the top

375 rating. Yet average global ratings remain about a 4.00. The grade expectation average was

376 approximately 4.00, which relates to a B average or 3.00 GPA.

377   One way of establishing convergent validity would be a finding that better trained and

378 more experienced teachers get higher ratings than less well trained instructors. If the

379 measure were valid, we would expect that regular faculty and full time instructors would get

380 higher ratings than per course faculty and teaching assistants. To argue otherwise is to

381 challenge the merits of higher education units with a faculty of professors with doctoral

382 status. If the university were a researcher powerhouse where faculty research is emphasized

383 over teaching and graduate assistants are admitted from the highest ranks of undergraduates,

384 the finding that teaching assistants and per course faculty get higher ratings might be less of

385 a challenge to the validity of these ratings. However, the university at which the data were

386 collected is a non-doctoral program with greater emphasis on teaching and moderate

387 emphasis on research, and teaching assistants are master's candidates with less substantial

388 admission expectations than doctoral programs. Hence, these findings challenge the

389 convergent validity of the teaching evaluations.

390   Like most studies in this area, a major limitation is the absence of an independent

391 measure of learning. Of course, this limitation is based upon the belief that the goal is to

392 create educated persons, not just satisfied consumers. Even when common tests are used,

393 these are invalid if the instructors are aware of the course content. Teachers seeking high

394 evaluations are able to improve their ratings and scores by directly addressing the content of

395 the specific test items. ETS now allows faculty who administer Major Field Tests to access

396 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; Gurung & Vespia, 2007; Hugh Feeley, 2002) in addition to leniency and low demands (Greenwald & Gillmore, 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

424 requires that the instructor try to emphasize avoidance of the lowest rating (1.00) because

425 these low ratings in a skewed distribution have in inordinate influence on the mean. This

426 effort of competing against the norms is likely to lead to grade inflation and permissiveness

427 for the least motivated and most negligent students. Some researchers (Ellis et al., 2003;

428 Greenwald & Gillmore, 1997) argue that student evaluations of instruction should be

429 adjusted on the basis of grades assigned. However, there are problems with such an

430 approach. The regression Betas are likely to differ based upon course and many other factors.

431 In our research and in research by DuCettte and Kenney (1982), substantial variation in

432 correlations was found across different course sets. Establishing valid adjustments would be

433 problematic at best. Further, such an approach would punish instructors when they happen

434 to get an unusually intelligent and motivated class (or teach an honors class) and give

435 students the grades they deserve. Student evaluations are not a proper motivational factor

436 for instructors in grade assignment, whether it is to inflate or deflate grades.

437   It would seem nearly impossible to eliminate invalid bias in student ratings of

438 instruction. Yet, they may tell us a teacher is ineffective when the majority give poor ratings.

439 It is the normative, competitive use that makes student evaluations of teaching subject to

440 problematic interpretation. This finding is especially critical in light of recent research that

441 portrays that student evaluations are largely biased against female teachers, and that

442 student bias in evaluation is related to course discipline and student gender (Boring et al.,

443 2016). Boring et al. (2016) also examine the difficulty in adjusting faculty evaluation for bias

444 and determined that the complex nature of ratings makes unbiased evaluation nearly

445 impossible. Stark and Freishtat (2014) further explain that evaluations are often negatively

446 related to more objective measures of teaching effectiveness, and biased additionally by

447 perceived attractiveness and ethnicity. In line with the current paper, he suggests dropping

448 overall teaching effectiveness or value of the course type questions because they are

449 influenced by many variables unrelated to actual teaching. Last, they suggest the

450 distribution and response rate of the data are critical information, and this point becomes

451 particularly important when recent research shows that online evaluations of teaching

452 experience a large drop in response rates (Stanny & Arruda, 2017). Our study contributes to

453 the literature of how student evaluations are a misleading and unsuccessful measure of

454 teaching effectiveness, especially focusing on reliability and the impact of grading on overall

455 questions. We conclude that it may be possible to manipulate these values by lowering

456 teaching standards, which implies that high stakes hiring and tenure decisions should

457 probably follow the advice of Palmer, Bach, and Streifer (2014) or Stanny, Gonzalez, and

458 McGowan (2015) in implementing teaching portfolios and syllabus review, particularly

459 because a recent meta-analysis of student evaluations showed they are unrelated to student

460 learning (Uttl, White, & Gonzalez, 2017).

References

461

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

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

464   *of Personality and Social Psychology*, *51*(6), 1173–1182.

465   doi:10.1037//0022-3514.51.6.1173

466 Best, J. B., & Addison, W. E. (2000). A preliminary study of perceived warmth of professor

467   and student evaluations. *Teaching of Psychology*, *27*(1), 60–62. Retrieved from

468   http://psycnet.apa.org/record/2000-07173-018

469 Blackhart, G. C., Peruche, B. M., DeWall, C. N., & Joiner, T. E. (2006). Factors influencing

470   teaching evaluations in higher education. *Teaching of Psychology*, *33*(1), 37–39.

471   doi:10.1207/s15328023top3301_9

472 Boring, A., Ottoboni, K., & Stark, P. (2016). Student Evaluations of Teaching (Mostly) Do

473   Not Measure Teaching Effectiveness. *ScienceOpen Research*.

474   doi:10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

475 Buck, S., & Tiene, D. (1989). The Impact of Physical Attractiveness, Gender, and Teaching

476   Philosophy on Teacher Evaluations. *The Journal of Educational Research*, *82*(3),

477   172–177. doi:10.1080/00220671.1989.10885887

478 Carrell, S. E., & West, J. E. (2010). Does Professor Quality Matter? Evidence from Random

479   Assignment of Students to Professors. *Journal of Political Economy*, *118*(3), 409–432.

480   doi:10.1086/653808

481 Centra, J. A. (2003). Will teachers recieve higher student evaluations by giving higher

482   grades and less course work? *Research in Higher Education*, *44*(5), 495–518.

483   doi:10.1023/A:1025492407752

484 DuCette, J., & Kenney, J. (1982). Do grading standards affect student evaluations of

485   teaching? Some new evidence on an old question. *Journal of Educational Psychology*,

486   *74*(3), 308–314. doi:10.1037/0022-0663.74.3.308

487 Ellis, L., Burke, D. M., Lomire, P., & McCormack, D. R. (2003). Student Grades and

Average Ratings of Instructional Quality: The Need for Adjustment. *The Journal of Educational Research*, *97*(1), 35–40. doi:10.1080/00220670309596626

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press. Retrieved from https://books.google.com/books?hl=en{\&}lr={\&}id=voeQ-8CASacC{\&}oi=fnd{\&}pg=PA1{\&}dq=A+theory+of+cognitive+dissonance{\&}ots=9y58Kxq9 theory of cognitive dissonance{\&}f=false

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, *52*(11), 1209–1217. doi:10.1037/0003-066X.52.11.1209

Gurung, R. A., & Vespia, K. (2007). Looking Good, Teaching Well? Linking Liking, Looks, and Learning. *Teaching of Psychology*, *34*(1), 5–10. doi:10.1080/00986280709336641

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis : a regression-based approach.* (T. D. Little, Ed.) (2nd ed., p. 692). The Guilford Press. Retrieved from https://www.guilford.com/books/ Introduction-to-Mediation-Moderation-and-Conditional-Process-Analysis/ Andrew-Hayes/9781462534654

Hugh Feeley, T. (2002). Evidence of Halo Effects in Student Evaluations of Communication Instruction. *Communication Education*, *51*(3), 225–236. doi:10.1080/03634520216519

Isely, P., & Singh, H. (2005). Do Higher Grades Lead to Favorable Student Evaluations? *The Journal of Economic Education*, *36*(1), 29–42. doi:10.3200/JECE.36.1.29-42

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

Marsh, H. W. (1987). Students' evaluations of University teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*(3), 253–388. 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*(1), 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*(3), 176–179.
doi:10.1207/s15328023top3303_4

Miller, A. (1985). A developmental study of the cognitive basis of performance impairment
after failure. *Journal of Personality and Social Psychology*, *49*(2), 529–538.
doi:10.1037/0022-3514.49.2.529

Miller, A., & Klein, J. S. (1989). Individual differences in ego value of academic performance
and persistence after failure. *Contemporary Educational Psychology*, *14*(2), 124–132.
doi:10.1016/0361-476X(89)90030-1

Palmer, M. S., Bach, D. J., & Streifer, A. C. (2014). Measuring the Promise: A
Learning-Focused Syllabus Rubric. *To Improve the Academy*, *33*(1), 14–36.
doi:10.1002/tia2.20004

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., & Team, R. C. (2017). nlme: Linear and
nonlinear mixed effects models. Retrieved from
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*(2), 102–114. Retrieved from
http://psycnet.apa.org/record/2000-14222-002

Stanny, C., Gonzalez, M., & McGowan, B. (2015). Assessing the culture of teaching and
learning through a syllabus review. *Assessment & Evaluation in Higher Education*,
40(7), 898–913. doi:10.1080/02602938.2014.956684

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*, *3*(3), 198–207. doi:10.1037/stl0000087

Stark, P., & Freishtat, R. (2014). An Evaluation of Course Evaluations. *ScienceOpen*

542   *Research.* doi:10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

543 Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (Sixth.). Boston, MA:

544   Pearson.

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

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

547   related. *Studies in Educational Evaluation, 54*, 22–42.

548   doi:10.1016/j.stueduc.2016.08.007

549 Weiner, B. (1992). *Human motivation : metaphors, theories, and research* (p. 391). Sage.

Table 1

Correlations for Instructor, Semester, and Course Combinations

Instructor	Semester	Course	b	SE	df	t	p
Different Instructor	Different Semester	Different Course	-.001	.000	10144295	-3.580	.013
Different Instructor	Same Semester	Different Course	.006	.002	152801	2.906	.048
Different Instructor	Different Semester	Same Course	.008	.001	517353	6.236	.027
Different Instructor	Same Semester	Same Course	.054	.010	6265	5.402	$< .001$
Same Instructor	Different Semester	Different Course	-.038	.003	108849	-13.130	$< .001$
Same Instructor	Same Semester	Different Course	.095	.020	1872	4.659	$< .001$
Same Instructor	Different Semester	Same Course	.090	.004	55057	21.769	$< .001$
Same Instructor	Same Semester	Same Course	.446	.023	1401	19.631	$< .001$

Table 2

t Statistics for Inter-item Relationship

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 Expected 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 Expected Grade	.311	.416	.018	4447	23.066	< .001
Fair to Assignments	.720	.715	.011	4447	63.912	< .001
Fair to Expected Grade	.375	.438	.016	4447	27.865	< .001
Assignments to Expected Grade	.344	.404	.015	4447	26.913	< .001

Table 3

t Statistics for Moderated Mediation

DV	IV	b	SE	df	t	p
Overall Course	Expected Grade	0.542	0.026	4336	20.650	< .001
Overall Course	Teaching Assistant	-0.198	0.109	191	-1.814	.071
Overall Course	Per-Course	0.018	0.076	191	0.235	.815
Overall Course	Instructor	-0.096	0.081	191	-1.187	.237
Overall Course	EG X TA	0.255	0.060	4336	4.234	< .001
Overall Course	EG X PC	0.077	0.079	4336	0.976	.329
Overall Course	EG X IN	-0.049	0.105	4336	-0.464	.643
Average Grading	Expected Grade	0.360	0.017	4336	21.790	< .001
Average Grading	Teaching Assistant	-0.049	0.059	191	-0.839	.403
Average Grading	Per-Course	0.060	0.041	191	1.441	.151
Average Grading	Instructor	0.083	0.044	191	1.860	.064
Average Grading	EG X TA	0.173	0.038	4336	4.556	< .001
Average Grading	EG X PC	0.167	0.049	4336	3.403	.001
Average Grading	EG X IN	0.056	0.064	4336	0.870	.384
Overall Course	Expected Grade	0.148	0.020	4332	7.537	< .001
Overall Course	Teaching Assistant	-0.133	0.058	191	-2.286	.023
Overall Course	Per-Course	-0.056	0.041	191	-1.354	.177
Overall Course	Instructor	-0.198	0.045	191	-4.388	< .001
Overall Course	Average Grading	0.000	0.000	4332	1.768	.077
Overall Course	EG X TA	0.034	0.051	4332	0.654	.513
Overall Course	EG X PC	-0.299	0.063	4332	-4.714	< .001
Overall Course	EG X IN	-0.173	0.080	4332	-2.164	.031
Overall Course	AG X TA	0.061	0.054	4332	1.136	.256
Overall Course	AG X PC	0.359	0.061	4332	5.922	< .001
Overall Course	AG X IN	0.142	0.087	4332	1.634	.102

Table 4

t Statistics for Individual Mediations

Group	DV	IV	b	SE	df	t	p
Teaching Assistant	Overall Course	Expected Grade	0.510	0.092	219	5.534	< .001
Teaching Assistant	Average Grading	Expected Grade	0.407	0.049	219	8.326	< .001
Teaching Assistant	Overall Course	Expected Grade	-0.010	0.077	218	-0.126	.900
Teaching Assistant	Overall Course	Average Grading	1.265	0.084	218	15.017	< .001
Per-Course	Overall Course	Expected Grade	0.605	0.071	425	8.536	< .001
Per-Course	Average Grading	Expected Grade	0.505	0.040	425	12.640	< .001
Per-Course	Overall Course	Expected Grade	-0.109	0.051	424	-2.163	.031
Per-Course	Overall Course	Average Grading	1.426	0.049	424	28.991	< .001
Instructor	Overall Course	Expected Grade	0.836	0.054	504	15.511	< .001
Instructor	Average Grading	Expected Grade	0.562	0.035	504	15.967	< .001
Instructor	Overall Course	Expected Grade	0.194	0.044	503	4.375	< .001
Instructor	Overall Course	Average Grading	1.144	0.045	503	25.230	< .001
Tenure Track	Overall Course	Expected Grade	0.537	0.027	3185	19.817	< .001
Tenure Track	Average Grading	Expected Grade	0.359	0.017	3185	20.722	< .001
Tenure Track	Overall Course	Expected Grade	0.142	0.021	3184	6.891	< .001
Tenure Track	Overall Course	Average Grading	1.097	0.020	3184	56.152	< .001