



3-23-2018

Do We See Eye to Eye? Moderators of Correspondence Between Student and Faculty Evaluations of Day-to-Day Teaching

Kathleen M. Cain
Gettysburg College

Benjamin M. Wilkowski
University of Wyoming

Christopher P. Barlett
Gettysburg College


See next page for additional authors

Roles

Student Authors

Colleen D. Boyle '15, Gettysburg College

Follow this and additional works at: <https://cupola.gettysburg.edu/psyfac>

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), and the [School Psychology Commons](#)

Share feedback about the accessibility of this item.

Cain, K. M., Wilkowski, B. M., Barlett, C. P., Boyle, C. D., & Meier, B. P. (2018). Do We See Eye to Eye? Moderators of Correspondence Between Student and Faculty Evaluations of Day-to-Day Teaching. *Teaching of Psychology*, 45(2), 107–114.

This is the author's version of the work. This publication appears in Gettysburg College's institutional repository by permission of the copyright owner for personal use, not for redistribution. Cupola permanent link: <https://cupola.gettysburg.edu/psyfac/83>

This open access article is brought to you by The Cupola: Scholarship at Gettysburg College. It has been accepted for inclusion by an authorized administrator of The Cupola. For more information, please contact cupola@gettysburg.edu.

Do We See Eye to Eye? Moderators of Correspondence Between Student and Faculty Evaluations of Day-to-Day Teaching

Abstract

Students and instructors show moderate levels of agreement about the quality of day-to-day teaching. In the present study, we replicated and extended this finding by asking how correspondence between student and instructor ratings is moderated by time of semester and student demographic variables. Participants included 137 students and 5 instructors. On 10 separate days, students and instructors rated teaching effectiveness and challenge level of the material. Multilevel modeling indicated that student and instructor ratings of teaching effectiveness converged overall, but more advanced students and Caucasian students converged more closely with instructors. Student and instructor ratings of challenge converged early but diverged later in the semester. These results extend our knowledge about the connection between student and faculty judgments of teaching.

Keywords

student evaluations of teaching, student–instructor agreement, teaching effectiveness, S.E.T.


Disciplines

Educational Assessment, Evaluation, and Research | Psychology | School Psychology

Authors

Kathleen M. Cain, Benjamin M. Wilkowski, Christopher P. Barlett, Colleen D. Boyle, and Brian P. Meier

Do We See Eye to Eye? Moderators of Correspondence Between Student and Faculty Evaluations of Day-to-Day Teaching

Teaching of Psychology
2018, Vol. 45(2) 107–114
© The Author(s) 2018
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0098628318762862
journals.sagepub.com/home/top


Kathleen M. Cain¹, Benjamin M. Wilkowski²,
Christopher P. Barlett¹, Colleen D. Boyle¹, and Brian P. Meier¹

Abstract

Students and instructors show moderate levels of agreement about the quality of day-to-day teaching. In the present study, we replicated and extended this finding by asking how correspondence between student and instructor ratings is moderated by time of semester and student demographic variables. Participants included 137 students and 5 instructors. On 10 separate days, students and instructors rated teaching effectiveness and challenge level of the material. Multilevel modeling indicated that student and instructor ratings of teaching effectiveness converged overall, but more advanced students and Caucasian students converged more closely with instructors. Student and instructor ratings of challenge converged early but diverged later in the semester. These results extend our knowledge about the connection between student and faculty judgments of teaching.

Keywords

student evaluations of teaching, S.E.T., student–instructor agreement, teaching effectiveness

For decades, students, administrators, and especially faculty have debated the validity of student evaluations of faculty teaching. The available evidence offers some support for student evaluations as valid measures of teaching quality, but this evidence is not consistent (Spooren, Brockx, & Mortelmans, 2013). For example, students tend to emphasize quality of instruction over other instructor characteristics (e.g., faculty charisma and humor) in their evaluations (Barth, 2008; Pan et al., 2009; Remedios & Lieberman, 2008), and students' evaluations of teaching are correlated with actual teaching behavior (Renaud & Murray, 2005). However, students and faculty sometimes disagree on what constitutes high-quality instruction (Bosshardt & Watts, 2001). Students' expected grades influence their evaluations, although the importance and strength of this pattern with respect to validity of student evaluations has been debated (e.g., Brockx, Spooren, & Mortelmans, 2011; Greenwald & Gillmore, 1997; Marsh & Roche, 2000; Olivares, 2001). Variations in personal characteristics of the instructor, such as gender and race, account for small but statistically significant variation in student evaluations (e.g., Beran & Violato, 2005; Smith, Yoo, Farr, Salmon, & Miller, 2007), and variations in other characteristics such as the instructor's physical attractiveness also predict student evaluations (e.g., Gurung & Vespia, 2007; Hamermesh & Parker, 2005; see Spooren et al., 2013 for a thorough review of the literature).

Most studies of student evaluations of teaching (S.E.T.) have focused on global assessments of courses offered at the

end of the semester. Among these studies, a few have found that student evaluations tend to correlate with instructors' own evaluations of their teaching (e.g., Basow & Montgomery, 2005; Roche & Marsh, 2000), providing further support for the validity of S.E.T. However, it is also important to examine the extent to which students and instructors agree on the quality of specific classes on particular days, given that a typical course lasts 3–4 months and that end of the semester evaluations might therefore capture biases in recall of teaching effectiveness. Drews, Burroughs, and Nokovich (1987) suggested that an important yet overlooked component of the evaluation of teaching effectiveness is the extent to which student and faculty evaluations of teaching converge on a day-to-day basis. Convergence in ratings of teaching effectiveness over a number of individual class days minimizes the influence of factors such as instructor personal characteristics and expected grades and may reflect the “nitty gritty” of specific class events and exercises rather than global impressions. A lack of agreement between student and instructor ratings of individual class days could reflect a fundamental problem with the evaluation

¹ Department of Psychology, Gettysburg College, Gettysburg, PA, USA

² Department of Psychology, University of Wyoming, Laramie, WY, USA

Corresponding Author:

Kathleen M. Cain, Department of Psychology, Gettysburg College, 300 North Washington Street, Gettysburg, PA 17325, USA.

Email: kcain@gettysburg.edu

process or systematic bias on the part of the students and/or the instructor.

Drews et al. (1987) compared student and faculty evaluations of teaching using Likert-type scale rating forms filled out by students and faculty in four courses on 15 days across a single semester and found moderate positive correlations between student and faculty ratings of teaching effectiveness on individual days of class. For example, there were significant positive correlations (ranging from .24 to .53, $ps < .05$) between student and faculty ratings of whether the instructor used time well, communicated well, and was organized; whether it was a productive day; and whether the material was challenging. These correlations were determined for agreement over all days of teaching and student ratings, without examining any potential impact of time of semester. Drews et al. (1987, p. 25) interpreted the findings overall as supporting the “credibility of students as judges of teaching effectiveness.”

We are unaware of research that has replicated and extended the work of Drews et al. (1987), and we therefore sought to do so using multilevel modeling techniques that allowed us to examine moderator variables. Specifically, like Drews et al., we asked whether student and faculty evaluations of teaching were significantly related over individual class days throughout the duration of the courses. We expected to find significant relationships between student and instructor ratings on individual days of instruction. Importantly, we extended the Drews et al. study by asking how the relationship between student and faculty evaluations of teaching remains constant or changes over the course of a semester. Understanding how the relationship between student and faculty evaluations changes over the course of a semester helps us to predict when student evaluations are most and least likely to coincide with the interpretation of instructors. To the best of our knowledge, no research has examined this question, and therefore, we did not make a prediction about how time in the semester might affect agreement between students and teachers. Students and faculty may become better acquainted with each other over the semester, resulting in a greater convergence of ratings late in the semester. Alternatively, their ratings might diverge late in the semester, as stress levels and workload mount while the semester progresses. They may also be equally in agreement throughout the semester, perhaps by relying consistently on the same criteria for judgment.

We also extended the Drews et al. (1987) research by examining the role of student demographic variables that influence student evaluations of day-to-day teaching. Although student variables have been shown to predict overall S.E.T., to our knowledge, no studies have asked how these characteristics influence students' ratings of individual class periods or how they predict correspondence with instructor self-ratings. For general course evaluations, several studies have shown that students' overall grade point averages (GPA) correlate with ratings of instructor effectiveness and also that more advanced students tend to rate instructors more favorably (e.g., Al-Issa & Sulieman, 2007; Badri, Abdulla, Kamali, & Dodeen, 2006; Griffin, Hilton, Plummer, & Barret, 2014;

Marsh, 1980; Spooren, 2010). These variables may be seen as “cognitive” in the sense that they likely reflect individual differences in cognitive skill or maturity. In other words, students who have more “expertise” as students—by virtue of greater past academic success or more years of experience—tend to give more favorable ratings to instructors than do new students and those with low GPAs. In our study, we expected that students with higher GPAs and more advanced students would agree more with faculty in their evaluations of day-to-day teaching.

Other demographic variables are perhaps more “personal” in that they reflect students' different preferences, experiences, and unique personal characteristics. For example, students tend to evaluate courses in their own majors more positively (Ting, 2000). Several studies suggest that students' ratings differ by gender, too. Although results are somewhat inconsistent, in several studies, females rated teaching more favorably than did males (e.g., Kohn & Hatfield, 2006; Santhanam & Hicks, 2002). In some studies, student gender interacted with instructor gender, but findings have not pointed to a single consistent pattern (e.g., Basow & Montgomery, 2005; Basow, Phelan, & Capotosto, 2006).

Although numerous studies have asked how students' teaching evaluations vary by race of instructor (e.g., Basow, Codos, & Martin, 2013; Smith, 2007), to our knowledge few studies have asked how students' own race predicts their evaluations of teaching effectiveness, and no studies have asked how students' race is related to the correspondence between student evaluations and instructor self-evaluations. This omission is somewhat surprising in light of the many studies of other aspects of race and classroom learning. Ehrenberg, Goldhaber, and Brewer (1995) and Dee (2004) suggested that teachers may evaluate same-race students differently than other-race students, implying that same-race students may also respond differently to same-race instructors. Centra (1993) speculated that students may evaluate same-race faculty more favorably. Li (1993), however, reported that Asian and American students rate the same instructors similarly. The extent to which minority and majority students agree on quality of teaching and the extent to which they agree with instructors' own evaluations of teaching are important questions for understanding the role of diversity in student learning.

Given the lack of consistent evidence about the role of student “personal” variables in evaluating teaching, we did not advance hypotheses for these variables. Rather, we simply posed the question of how students' major, gender, and race predict correspondence between their ratings and their instructors' ratings of teaching effectiveness during single class periods. If students and instructors agree despite variations in student personal variables, this finding would support the validity of S.E.T.

Overall, our main purposes in this study were to (1) replicate the Drews et al. (1987) finding that student and instructor ratings of teaching quality on individual class days are correlated, (2) ask how agreement on teaching quality between professors and students varies with time of semester, and (3) examine the

role of student demographic variables in moderating the link between professors' and students' perceptions of day-to-day teaching quality.

Method

Participants

Participants were 137 students enrolled in one of five psychology courses at Gettysburg College as well as the five instructors for these courses. The students, 96 females and 41 males, were predominantly Caucasian (88.3%). The sample represented students from all years in college, including 29.9% first-year students, 27.7% sophomores, 21.2% juniors, and 21.2% seniors. Of the students in the study, 42.3% were psychology majors, 18.2% were undeclared, and 39.5% were majors in disciplines other than psychology. The average GPA of students enrolled in the study was 3.21 (on a 4.33 scale). The number of students in each course ranged from 25 to 30. Students who happened to be enrolled in more than one participating course were asked to only participate in the study in the first course in which they heard the study introduced, and no students participated in the study for more than one course.

The five courses, all taught in the Spring 2015 semester, included one section of Psychology 101 and four 200-level psychology elective courses (social psychology, sensation and perception, developmental psychology, and brain and behavior). These courses were offered in a liberal arts college with a traditional undergraduate enrollment and with a psychology curriculum that emphasizes scientific approaches to psychology. The courses were taught by five different instructors, three males and two females, with a wide range of teaching experience (3–25 years) at the institution. All of the instructors were Caucasian.

Measures

On the first day of the study, students completed a brief demographic questionnaire that inquired about gender, race, major, year in school, and GPA. On that day and all remaining days of the study, students completed a short questionnaire with 4 items in which they rated the quality of the class that day. The specific items, rated on a 9-point Likert-type scale from *strongly disagree* to *strongly agree*, were: "Today, the instructor was effective in conducting the class," "Today's class offered a good learning experience for students," "Today, the material covered in class was challenging," and "Overall, today's class was good." Faculty members received the same 4-item questionnaire to rate their own teaching for each class day. We used a small number of items given the repeated assessment, and we chose to use items that tapped a general evaluation of that day's teaching along with the challenge level of the material. Our items roughly coincide with those used by Drews et al. (1987).

Initial analyses indicated that 3 items (related to instructor effectiveness, learning experience, and general positivity) were all highly related to each other, all unstandardized b s > .57, all

Table 1. Overall Means (and SDs) for Instructor and Student Ratings.

	Student	Instructor
Overall evaluation*	7.69 (.47)	6.83 (.99)
Challenge	6.02 (.86)	6.25 (1.56)

* $p < .0001$.

p s < .0001. However, the item pertaining to challenging material was not as strongly related to the others, all b s < .10. Thus, the three strongly related items were averaged to form a single index of overall evaluation, but the item pertaining to challenging material was analyzed separately.

Procedure

In prior discussion with the participating course instructors, the researchers and instructors agreed to start the study in the fourth week of the semester and to continue for 10 weeks, ending 1 week before the course itself ended. Data were collected once per week but were not collected on days when there were exams, guest lectures, or other unusual events in class. In those weeks, data were collected intentionally on other days. Beyond that restriction, selection of dates for data collection in any given week was random.

On the first day of the study, a research assistant visited the class shortly before the end of the class period and described the study. She distributed informed consent forms, demographic questionnaires, and the first set of evaluations to students, and the faculty member completed the first evaluation right after class ended. After the first class, the instructor received a packet of student evaluations and one self-evaluation on the morning of each day when data collection was scheduled. The instructor distributed the student evaluations at the end of class, collected them in an envelope without viewing them, and returned them along with his or her self-evaluation to the research assistant. After the first day, the study took about 2 min. of class time per week. Students only completed the evaluation on days they were present in class. We received a grand total of 1,116 evaluations from the 137 students in our sample. Thus, the average student completed 8.15 of the 10 possible evaluations. Only 2 evaluations were missing from instructors of 50 (i.e., a 96% completion rate).

Students' responses were anonymous and tracked by unique identifiers. The procedure was approved by the Institutional Review Board at Gettysburg College.

Results

Initial Results

As shown in Table 1, instructors rated their own teaching more negatively than did their students, $b = .43$, standard error (SE) = .07, $p < .0001$, but instructors and students did not differ in their ratings of how challenging the material was, $b = -.11$, $SE = .10$, $p = .27$.

Analytic Strategy

The current data set exhibited a rather complex nested structure. Each individual student evaluation was simultaneously nested within the student who provided it and the class period for which it was provided. Students and class periods were, in turn, nested within the course. Because of this structure, the traditional assumption of independent observations is violated in three different fashions (i.e., evaluations provided by the same student on different days are not independent, evaluations provided by different students on the same day are not independent, and evaluations of the same course are not independent).

All analyses were thus conducted using multilevel modeling (Raudenbusch & Bryk, 2002; Snijders & Bosker, 1999). This analytic technique was specifically designed to examine nested data structures in which observations are not independent. Beyond this issue, this analytic technique provides a number of other advantages. It appropriately takes into account the sample size at each level of analysis (i.e., in the current study, individual evaluations, students, days, and courses) and does not inappropriately inflate statistical power. Furthermore, it can handle randomly missing data. Finally, it allows us to examine interactions across different levels of analysis.

For all analyses, we thus created a three-level model, with a cross-classification at Level 2. Individual evaluations were modeled at Level 1. Day and Student were modeled at Level 2, and Course was modeled at Level 3. Instructor's evaluations of each day and day-of-semester were treated as day-level variables, and student demographic variables were treated as student-level variables. To appropriately separate between-course from within-course effects, all predictors were centered around the mean value of the course (Enders & Tofghi, 2007; Raudenbusch & Bryk, 2002).

For each of the two outcome variables (overall evaluation and challenging material), we first estimated the unconditional model (Raudenbusch & Bryk, 2002). This step evaluated whether there was significant random variation in ratings across days, students, and courses. These analyses indicated that both variables significantly varied across days (Overall-Evaluation: $z = 4.15, p < .0001$; Challenging-Material: $z = 4.13, p < .0001$) and across students (Overall-Evaluation: $z = 7.10, p < .0001$; Challenging-Material: $z = 7.34, p < .0001$) but not across courses (Overall-Evaluation: $z = .70, p = .24$; Challenging-Material: $z = .89, p = .19$). The nonsignificant variation across courses likely reflects the small sample size at this level (i.e., $n = 5$ courses). This nonsignificant result is also relatively inconsequential in the current context, since none of the variables we examined were at this level of analysis.

Correspondence in Overall Evaluations

We next entered Overall Instructor Evaluations of individual class periods as a predictor of Overall Student Evaluations of the same class period to see whether these two converge. The results indicated that these two variables were significantly

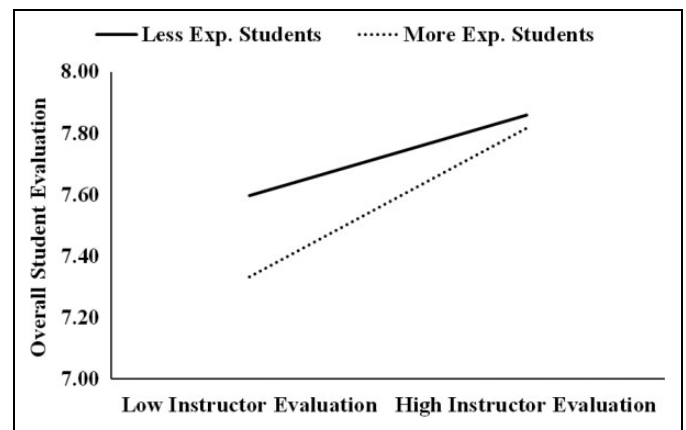


Figure 1. Overall student evaluations as a function of overall instructor evaluations and student academic year (experience).

related, $b = .22, SE = .07, p = .002$. This indicates that at the aggregate level, as hypothesized, student and instructor overall effectiveness ratings of individual class periods converged at an above-chance level.

We next sought to examine whether Time of Semester or student demographic variables (i.e., GPA, Academic Year [coded as 1, 2, 3, or 4 for first-year, sophomore, junior, or senior year, respectively], Major [contrast-coded as *Psychology* = 1; *Not Psychology* = -1], Gender [contrast-coded as *female* = 1; *male* = -1], or Race [contrast coded as *Minority* = 1; *Caucasian* = -1]) moderated this relationship. To do so, we simultaneously entered each of these variables as a predictor of Overall Student Evaluations to test their main effect. More importantly, we also entered their interaction with Overall Instructor Evaluations. The results reported below are virtually unchanged if each moderator is examined without controlling for other potential moderators (i.e., no change in the significance, direction, or pattern of effects).

In this analysis, the Overall Instructor Evaluation \times Student Academic Year interaction was significant, $b = .07, SE = .03, p = .02$. The interaction between Overall Instructor Evaluation and Student Race (minority vs. White) was also significant, $b = -.11, SE = .04, p = .01$. All other effects did not reach significance, all $ps > .12$. In a subsequent analysis, we also asked if the match between the student and the instructor gender exhibited a main effect or interaction with Instructor Evaluations, and we found no significant effects, $ps > .40$.

To understand the nature of the Overall Instructor Evaluation \times Student Academic Year interaction, we next estimated the mean overall student evaluation values at high ($M + 1 SD$) and low ($M - 1 SD$) levels of overall instructor evaluations and student academic year (Aiken & West, 1991; Preacher, Curran, & Bauer, 2006). The results are depicted in Figure 1. Finally, simple slope analyses (Aiken & West, 1991; Preacher et al., 2006) were conducted to test the convergence of student and instructor evaluations for more versus less experienced students. A steeper slope (i.e., b values) indicates stronger convergence between student and instructor evaluations. This

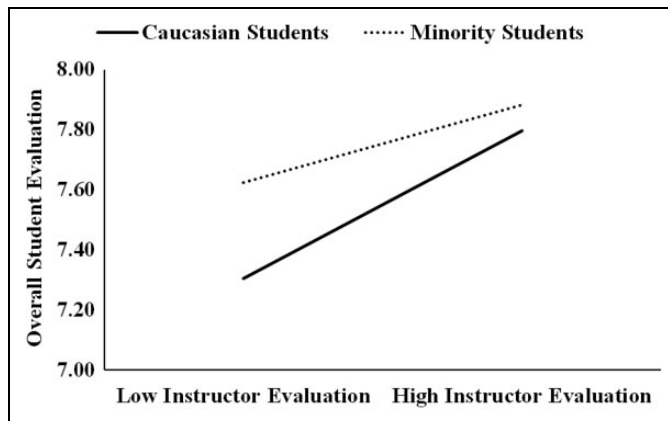


Figure 2. Overall student evaluations as a function of overall instructor evaluations and students' race.

analysis indicated that overall instructor evaluations significantly predicted overall student evaluations for less experienced students in the class, $b = .16$, $SE = .08$, $p = .03$. However, as hypothesized, this relationship was stronger for more experienced students, $b = .30$, $SE = .08$, $p = .0002$.

The estimated means for the Overall Instructor Evaluation \times Student Race interaction are depicted in Figure 2. Simple slope analyses indicated that Overall Instructor Evaluations were a significant predictor of Caucasian students' Overall Evaluations, $b = .26$, $SE = .07$, $p = .0005$. However, this relationship was weaker for minority students, $b = .17$, $SE = .10$, $p = .09$, and did not reach traditional levels of significance. This analysis quite clearly suggests that instructor ratings converged more weakly with minority student than with Caucasian student ratings. However, it is likely that convergence with minority students would become significant, if less strong, with a larger sample of minority students, as there were only 16 minority students in the current sample.

Correspondence in Challenging Material Ratings

We next assessed whether students and instructors converged in their evaluations of how challenging the material was. When Instructor Challenging Material ratings were entered as a predictor of Student Challenging Material ratings, a significant effect emerged, $b = .30$, $SE = .06$, $p < .0001$. Thus at the aggregate level, as hypothesized, student and instructor ratings of the challenge level of the material in individual class periods converged at an above-chance level.

We next assessed whether time of semester or student demographic variables (i.e., Student GPA, Student Academic Year, Student Major [Psychology vs. Not Psychology], Student Gender, or Student Race [Caucasian vs. Minority]) moderated this effect. These variables were coded the same way as in the prior analysis. As with overall evaluations, we simultaneously assessed the main effect of each of these variables as well as their interaction with Instructor Evaluations of Challenging Material. When this was done, the main effect of Student GPA approached but did not reach traditional levels of significance,

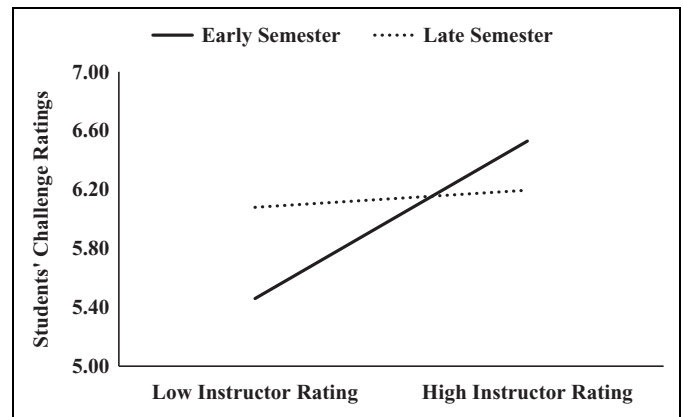


Figure 3. Students' challenge ratings as a function of instructors' challenge ratings and time of semester.

$b = -.64$, $SE = .34$, $p = .056$, such that students with higher GPAs overall rated the material as less challenging. The Student-Gender main effect was also significant, $b = .41$, $SE = .15$, $p = .005$, such that females overall rated class material as more challenging than males.

More central to current concerns was the fact that the Instructor Challenging Material Ratings \times Time of Semester interaction was significant, $b = -.06$, $SE = .02$, $p = .001$. The interaction between Instructor Challenge Ratings and Student GPA also approached but did not reach traditional levels of significance, $b = .15$, $SE = .08$, $p = .055$. All other effects were nonsignificant, all $ps > .14$. In a subsequent analysis, we also asked if the match between the student and instructor gender exhibited a main effect or interaction with instructor evaluations, and we found no significant effects, $ps > .55$. All interactive effects were virtually unchanged when each moderator was examined without controlling for other potential moderators (i.e., no change in direction or significance of effects).

The estimated means for the Instructor Challenge Ratings \times Time of Semester interaction are depicted in Figure 3. Subsequent simple slope analyses indicated that Instructor Challenge ratings significantly predicted Student Challenge ratings early in the semester, $b = .40$, $SE = .07$, $p < .0001$. However, this effect was not significant late in the semester, $b = .04$, $SE = .10$, $p = .65$. In order to understand this effect a bit better, we computed the predicted means for students and instructor ratings for early and later in the semester. The instructors (estimated $M = 5.83$) and students (estimated $M = 5.82$) converged early in the semester on their challenge ratings, but the instructors appeared to perceive the later part of the semester as more challenging (estimated $M = 6.70$) than students did (estimated $M = 6.19$).

The estimated means for the Instructor Challenge Ratings \times Student GPA interaction are depicted in Figure 4. Subsequent simple slope analyses indicated that Instructor Challenge ratings significantly converged with the Challenge ratings of students with a lower GPA, $b = .21$, $SE = .07$, $p = .001$. However, as hypothesized, this effect was somewhat stronger for students with a higher GPA, $b = .24$, $SE = .07$, $p = .0005$.

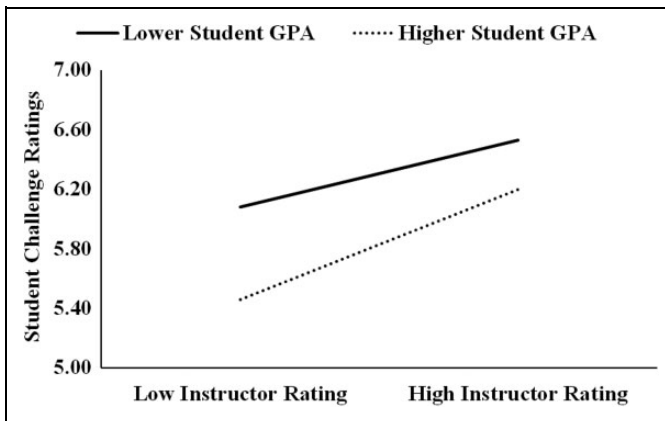


Figure 4. Students' challenge ratings as a function of instructors' challenge ratings and student overall grade point averages.

Discussion

In the present study, faculty rated their teaching effectiveness more negatively than did students, but their ratings of challenge did not differ. These results indicate that instructors and students have similar ideas about the extent of the challenge level of the material, but, on average, instructors' ratings of their performance were lower than student ratings by almost a full point on a 9-point scale.

Most importantly, similar to Drews et al. (1987), we found clear evidence of instructor and student convergence in agreement about teaching effectiveness on individual class days. In other words, students and faculty generally agreed that specific days in class had more or less effective teaching. This convergence did not vary with time of the semester. However, students who were less experienced and students who were from minority groups were less strongly in agreement with their instructors' self-evaluations. For ratings of the challenge level of the material, students and faculty again agreed with each other overall. For this variable, however, agreement decreased over the course of the semester such that there was no significant agreement at all by the end of the semester. In addition, students with higher GPAs agreed more closely with their instructors on the challenge level of the material, although we note that the interaction was not significant at the traditional level.

These findings largely support the idea that students can make valid judgments of teaching, at least when considering instructors' evaluations as a criterion. The agreement between students and instructors on overall teaching effectiveness was not influenced by time of semester, student GPA, student major, or gender, and the agreement between students and instructors on challenging material was not moderated by student academic year, student major, race, or gender (although females overall rated class material as more challenging than did males). The fact that agreement was unaffected by so many variables points to the general robustness of the correspondence between students and faculty.

Despite this general agreement, some important areas of disagreement emerged. Student academic year was a

significant moderator of the link between student and instructor ratings for overall evaluation. Although instructor ratings significantly predicted less experienced students' ratings, the relationship between student and instructor ratings became stronger for more advanced students. Perhaps with increasing experience in the college classroom, students come to share more of their instructors' views of high-quality teaching, and thus their judgment of their instructors is more similar to the instructors' judgments of themselves. These results suggest that students in upper level classes would be more likely to agree with their professors on teaching effectiveness compared to students in the lower level classes such as those assessed in the present study. Additionally, upper level classes generally have fewer students, which may enhance agreement between students and faculty on day-to-day teaching effectiveness.

Caucasian students were more likely to agree with their Caucasian instructors' self-evaluations of overall effectiveness than were minority students. This finding is consistent with the academic year effect in that both findings suggest that students whose backgrounds are more similar to the instructors' backgrounds agree with the instructor more. The finding should be interpreted with caution, however, as there were relatively few minority students in the sample (16). Also, no parallel examination of majority and minority students' agreement with instructors can be made for minority faculty, as all of the instructors in the present sample were Caucasian. Even so, the finding raises important questions about the experiences of minority students in classes taught by Caucasian instructors. For example, does the lower level of agreement stem from differences in the kinds of teaching styles preferred? Does it speak to majority–minority differences in students' performance or comfort levels in classroom settings? What strategies can majority faculty employ to ensure they are gauging their minority students' classroom experiences effectively? It will be important for future research to address these questions.

The overall agreement between students and faculty in ratings of challenge is not surprising; indeed, in the Drews et al. (1987) study, the highest correlation was found for student and faculty ratings of challenge. What is novel and surprising is the fact that this general pattern of agreement on challenge at the start of the semester falls apart at the end of the semester. The material at the start of the semester is often easier and more familiar to students than the material at the end of the course, and both they and their instructors may rate challenge similarly at the beginning due to experiences with this material in other courses. At the end of the semester, the material presented in many classes is more difficult and more unfamiliar for students. Some students may initially struggle to estimate the complexity of this material. Moreover, students and faculty are probably both more tired and stressed, conditions that likely affect their perceptions of challenge. Thus, students and instructors may fall out of sync in their perceptions as the semester progresses. An alternative possibility is that both students and faculty in our study experienced fatigue using the rating scales and may have completed them with less care later in the semester. If this were the case, however, it is unlikely that fatigue would only

affect correspondence for challenge ratings without affecting correspondence for ratings of overall teaching effectiveness.

In addition to the time of semester effect, students with lower GPAs were less likely to agree with their instructors about level of challenge than were students with higher GPAs. This finding should not be exaggerated, as both low- and high-GPA students exhibited significant agreement with their instructors' challenge ratings, and the difference between the two reached a marginal level of significance ($p = .055$). However, the fact that the relationship was less strong for lower GPA students suggests again that instructors would benefit from increased awareness of the experiences of students who may be less well integrated into the classroom environment. The finding may also suggest that lower GPA students are less skilled at judging how challenging classroom material is. It would be interesting to find out whether this difference actually *contributes* to these students' lower levels of academic success; perhaps they experience more difficulty in identifying the most challenging or complex material and thus are less sure of where to focus their studying.

Interestingly, the more "cognitive" student demographic variables of academic year and GPA predicted agreement more often than did the "personal variables" of student major and gender (although student race was indeed important). Newer students were less in sync with instructors than more experienced students for judgments of overall effectiveness, and students with lower GPAs were less in sync with instructors than students with higher GPAs for judgments of challenge. These findings suggest that some students have more "expertise" than others in judging classroom teaching. This study was conducted at a highly selective liberal arts college to which many students arrive with substantial preparation for postsecondary education. It would be interesting to examine the role of these cognitive variables at other kinds of colleges and universities where students bring a wider range of preparation. It seems likely that variables such as GPA and year in school would have an even larger impact on student–faculty agreement in more heterogeneous settings.

It is noteworthy that instructors made more stringent judgments of their own teaching than did their students. It is unclear whether this pattern reflects genuine underestimation of teaching effectiveness or whether it is due to modesty or social desirability on the part of instructors. The fact that instructors' ratings still varied systematically with student ratings, even if instructor ratings were lower overall, indicates that they are likely a reasonably adequate barometer of teaching effectiveness. It would be interesting for future research to examine the conditions under which instructors underestimate or overestimate the quality of their own teaching.

There were some limitations to the present study, including the small number of courses and instructors as well as the fact that the data came from one department at one college. Also, the low diversity of the instructors in the study made it impossible to assess the impact of student–instructor race match (see Basow & Montgomery, 2005). Finally, the findings of this study are correlational in nature and do not provide causal

information. Despite these limitations, the study was one of only a very few to examine student and instructor agreement about teaching quality, particularly with respect to specific days of class, and, as far as we know, the only study to examine student demographic variables and time of semester as moderators of the level of this agreement.

Conclusion

Our study asked whether students and faculty see eye to eye as they judge teaching on a day-to-day basis. Our results indicate that the answer is "often, but not always." Students and faculty tend to agree about teaching effectiveness, but they are less likely to agree when students are minorities or when they're less experienced. Students and faculty agree about challenge at the start of the semester, but they are less likely to agree at the end of the semester, and students with low GPAs are somewhat less likely to agree with their instructors about challenge as well. The findings of the present study add both support and cautions to discussions of the validity of teaching evaluations. In addition, the findings suggest that faculty would benefit from deeper knowledge about learning and views of effective teaching among minority students as well as among students in the early years of college and the stressful days at the end of the semester. Overall, the results indicate that although students and instructors tend to see eye to eye, there are interesting questions to explore about the circumstances when they do not.

Declaration of Conflicting Interests

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

Funding

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

References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.
- Al-Issa, A., & Sulieman, H. (2007). Student evaluations of teaching: Perceptions and biasing factors. *Quality Assurance in Education, 15*, 302–317.
- Badri, M. A., Abdulla, M., Kamali, M. A., & Dodeen, H. (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAE. *International Journal of Educational Management, 20*, 43–59.
- Barth, M. M. (2008). Deciphering student evaluations of teaching: A factor analysis approach. *Journal of Education for Business, 84*, 40–46.
- Basow, S. A., Codos, S., & Martin, J. L. (2013). The effects of professors' race and gender on student evaluations and performance. *College Student Journal, 47*, 352–363.
- Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education, 18*, 91–106.

- Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly, 30*, 25–35.
- Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment & Evaluation in Higher Education, 30*, 593–601. doi:10.1080/02602930500260688
- Bosshardt, W., & Watts, M. (2001). Comparing student and instructor evaluations of teaching. *Journal of Economic Education, 32*, 3–17.
- Brockx, B., Spooren, P., & Mortelmans, D. (2011). Taking the grading leniency story to the edge. The influence of student, teacher, and course characteristics on student evaluations of teaching in higher education. *Educational Assessment, Evaluation and Accountability, 23*, 289–306. doi:10.1007/s11092-011-9126-2
- Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco, CA: Jossey-Bass.
- Dee, T. S. (2004). The race connection: Are teachers more effective with students who share their ethnicity? *Education Next, 4*, 52–59.
- Drews, D. R., Burroughs, W. J., & Nokovich, D. (1987). Teacher self-ratings as a validity criterion for student evaluations. *Teaching of Psychology, 14*, 23–25.
- Ehrenberg, R. G., Goldhaber, D. D., & Brewer, D. J. (1995). Do teachers' race, gender, and ethnicity matter? Evidence from the National Educational Longitudinal Study of 1988. *Industrial and Labor Relations Review, 48*, 547–561. doi:10.1207/s15328023top1401_5
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*, 121–138.
- 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
- Griffin, T. J., Hilton, J. I., Plummer, K., & Barret, D. (2014). Correlation between grade point averages and student evaluation of teaching scores: Taking a closer look. *Assessment & Evaluation in Higher Education, 39*, 339–348. doi:10.1080/02602938.2013.831809
- Gurung, R. R., & Vespia, K. M. (2007). Looking good, teaching well? Linking liking, looks, and learning. *Teaching of Psychology, 34*, 5–10. doi:10.1207/s15328023top3401_2
- Hamermesh, D., & Parker, A. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*, 369–376.
- Kohn, J., & Hatfield, L. (2006). The role of gender in teaching effectiveness ratings of faculty. *Academy of Educational Leadership Journal, 10*, 121–137.
- Li, Y. (1993). *A comparative study of Asian and American students' perceptions of faculty teaching effectiveness at Ohio University*. Unpublished doctoral dissertation, Ohio University, Athens, OH.
- Marsh, H. W. (1980). The influence of student, course, and instructor characteristics in evaluations of university teaching. *American Educational Research Journal, 17*, 219–237.
- 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.
- Olivares, O. J. (2001). Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology, 26*, 382–399.
- Pan, D., Tan, G. H., Ragupathi, K., Booluck, K., Roop, R., & Ip, Y. K. (2009). Profiling teacher/teaching using descriptors derived from qualitative feedback: Formative and summative applications. *Research in Higher Education, 50*, 73–100. doi:10.1007/s11162-008-9109-4
- Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interaction in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics, 31*, 437–448.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal, 34*, 91–115. doi:10.1080/01411920701492043
- Renaud, R. D., & Murray, H. G. (2005). Factorial validity of student ratings of instruction. *Research in Higher Education, 46*, 929–953.
- Roche, L. A., & Marsh, H. W. (2000). Multiple dimensions of university teacher self-concept. *Instructional Science, 28*, 439–468.
- Santhanam, E., & Hicks, O. (2002). Disciplinary, gender and course year influences on student perceptions of teaching: Explorations and implications. *Teaching in Higher Education, 7*, 17–31.
- Smith, B. P. (2007). Student ratings of teacher effectiveness: An analysis of end-of-course faculty evaluations. *College Student Journal, 41*, 788–800. doi:10.1080/13562510120100364
- Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. *Women's Studies in Communication, 30*, 64–77.
- Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. London, England: Sage.
- Spooren, P. (2010). On the credibility of the judge: A cross-classified multilevel analysis on students' evaluation of teaching. *Studies in Educational Evaluation, 36*, 121–131. doi:10.1016/j.stueduc.2011.02.001
- Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*, 598–642.
- Ting, K. (2000). A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. *Research in Higher Education, 41*, 637–661.