The Reliability of Instructor Evaluations

Erin M. Buchanan[1], Jacob Miranda[2], and & Christian Stephens[2]

[1] Harrisburg University of Science and Technology

[2] University of Alabama

Author Note

The authors made the following contributions. Erin M. Buchanan: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing, Analysis; Jacob Miranda: Writing - Original Draft Preparation; Christian Stephens: Writing - Original Draft Preparation.

Correspondence concerning this article should be addressed to Erin M. Buchanan, 326 Market St., Harrisburg, PA 17010. E-mail: ebuchanan@harrisburgu.edu

Abstract

Student evaluations of teaching are regularly used within college classroom to gauge effectiveness of instruction, provide evidence for administrative decision making, and inform instructors of course feedback. The validity of teaching evaluations is often questioned, as they appear to be influenced by outside of teaching factors such as gender, race/ethnicity, grading, previous student achievement, and more. However, teaching evaluations do appear to be a reliable measure, often showing strong correlations for an instructor. In this study, we investigate over 30 years of teaching evaluations to determine the reliability of teaching evaluations across course, instructor, and time. Generally, instructors teaching the same course within the same semester showed the highest reliability estimates, with lower estimates for the same course in different semesters. The reliability of instructor's evaluations showed a small decrease over time. Finally, we investigated the impact of a validity measurement (perceived fairness) on reliability and found no evidence that this variable influence reliability estimates.

Keywords: reliability, teaching effectiveness, fairness, grading, evaluations

The Reliability of Instructor Evaluations

In the United States, college and university professors are evaluated to varying degrees on research productivity, service, and teaching effectiveness. These dimensions are often used for high-stakes administration decisions, including hiring, retention, promotion, pay, and tenure (Freishtat, 2014; Hornstein, 2017; Spooren et al., 2013; Stroebe, 2020). Depending on the institution, a major failure of one of these evaluative dimensions could jeopardize a professor's position within the department; thus, professors are urged to maintain high standards of research, service, and teaching. Indeed, the vast majority of the 9,000 professors polled by the American Association of University Professors believed the teaching evaluative dimension should be taken as seriously as research and service (Flaherty, 2015). The consequences of teacher effectiveness may motivate collegiate faculty into actively considering the quality of their classroom.

Teaching effectiveness can be defined as the degree to which student achievement is facilitated [i.e., how much have students learned in a particular course; P. A. Cohen (1981)]. Generally, assessments of teaching effectiveness come from student evaluations of teaching (SETs) or the course itself [e.g., "Student Opinion of Instruction", "Student Evaluations of Teaching", "Students Opinion of Teaching Effectiveness", "Students Evaluation of Faculty", "Overall Course Ratings", "Instruction Rating"; P. A. Cohen (1981); Flaherty (2020)]. Often these metrics are described as evaluating the "quality" of the individual or course (Gillmore et al., 1978; Marsh, 2007). Teaching effectiveness measures are intended to gauge multiple facets of teaching, such as an instructor's proficiency in communication, organization, presentation, and grading (Hattie & Marsh, 1996). Given the use of SETs in administrative decisions, both the reliability and validity of these measures should be demonstrated to ensure their utility. Thus, the question naturally arises: are SETs reliable and valid measures of teaching effectiveness?

Validity

Sheehan (1975)'s review of instructor evaluation literature found such measures contained multiple factors potentially conducive to bias. These include 1) student demographics: gender, class, age, previous achievement, 2) class type: subject matter, size, degree requirements, and 3) instructor qualities: gender, rank, gender-match to student, etc. Decades later, studies still underscore that sexism (MacNell et al., 2015; Mitchell & Martin, 2018), racism (Smith & Hawkins, 2011), and general biases pervade students' evaluations today in both traditional courses and possibly online ones as well(Heffernan, 2022; O'Sullivan et al., 2014; Rovai et al., 2006; Zheng et al., 2023) Individual factors may also yield some influence on SET ratings, including instructors' cultural background (Fan et al., 2019), attractiveness (Felton et al., 2008; Wright, 2000), position ranking (Johnson et al., 2013), and students' expected grade from the course (Chen et al., 2017; Crumbley et al., 2001; Marks, 2000). Biasing factors may even include the volume of the instructor's voice and how legible their instructor's writing is (W. E. Becker et al., 2012). Concerningly, Stroebe (2020) also highlights the danger of an incentive system tied to student ratings; in other words, instructors may be incentivized to be a less effective teacher (e.g., grade leniently, choose to teach courses based on student interest, etc.) rather than challenge students critically to boost their SET ratings. Concerns of bias have not dissipated over time (Boring et al., 2016; Dunn et al., 2014; Hornstein, 2017; Uttl et al., 2017). Recent meta-analyses suggest SETs may be entirely unrelated to material learned (Uttl et al., 2017) and their biasing aspects cannot be altered due to the complex interaction of factors included in their calculation Boring et al. (2016). While students' ratings may show some utility in indicating to other students which classes to pursue and with which professor (Stankiewicz, 2015), this usefulness may come at the cost of the professor's self-efficacy (Boswell, 2016). While SETs are conceptually valuable towards gaining insight on teacher effectiveness or course quality, the many outstanding issues suggest they may not be valid measures. Even so, some researchers argue that the complete removal of SETs from administrative consideration is the wrong course of action (Benton & Ryalls, 2016). A more appropriate solution may be to utilize multiple measures of teaching effectiveness simultaneously (e.g., subject-matter experts sit-in on lecture, peer reviews of course curriculum Benton & Young, 2018; Berk, 2018; Esarey & Valdes, 2020; Kornell & Hausman, 2016). However, the cost of implementing a more accurate, multi-pronged approach may be unrealistic given a university's budget and expectations of the instructor. Institutions may then opt to continue using SETs regardless of their validity.

Perceived fairness

Extant research broadly supports that SETs are influenced by students' grades. Some instructors may feel pressured into reducing the rigor of their course for the sake of attaining higher SET ratings (Greenwald & Gillmore, 1997; Marks, 2000). However, as pointed out by Wright (2000), students' expectations of their final grades may not affect their SET ratings nearly as much as their perceived fairness of their grades or the grading process that produced them. Professors who are consistent, representative, accurate, unbiased, and correctable in their grading may receive high SET ratings regardless of how much a student learns or what his/her final grade turns out to be (Horan et al., 2010; Leventhal, 1980). Students' grades may predict their SETs only so much as students perceive the grading processes as fair (Tata, 1999). Hence, students' perceptions of fairness may be more akin to comprehensive, and hopefully valid, assessments of the instructor rather than just face-value judgments of their grade. Perceived fairness may play a multifactorial role in its influence on SETs. For example, Tripp et al. (2019) found that students' perceived fairness of their instructors' grading processes affected their perceived fairness of their assigned grade, which then translated to their instructor evaluation ratings of teacher effectiveness. Further, perceived fairness of the course workload and difficulty may be inversely related to perceived fairness of the grading process as a challenging professor may be thought of as less fair (Marks, 2000). Access to grading criteria, frequency of feedback, and proactive instruction are other aspects of grading thought to explicitly affect perceived fairness (Pepper & Pathak, 2008). Therefore, the perceived fairness of those aspects must also be considered when determining the impact of perceived fairness on SET ratings, especially when different professors teach the same course or teach multiple courses in the same semester. The validity and reliability of SETs may then partially hinge on the consistency of students' perceptions of fairness.

Reliability

Past investigations of SETs concluded they are reliable measures (Arubayi, 1987; Marsh & Roche, 1997). Even some contemporary reviews have explored the reliability of SETs when controlling for various factors. For example, Benton and Cashin (2014) found SETs collected from the same class to be internally consistent when teaching effectiveness was assessed through several items. Even so, other data suggest that instructor, course, and student factors each contribute meaningfully to the variance of student evaluation ratings, which can influence their reliability (Feistauer & Richter, 2017). This result suggests SET ratings may be reliable over time if the aspects of a classroom remain constant. However, few data have explored the interactions of time with validity variables or how it affects reliability among SETs in relation to perceived fairness specifically. Thus, while previous research has explored teacher effectiveness over time (Marsh, 2007), our study extends this work by examining the reliability patterns of 30 years of SET data with respect to various moderating influences.

The current study

The current study is similar in scope to recent work (Boring et al., 2016; Fan et al., 2019) in its analysis of teacher evaluations collected over an extensive period. Boring et al. (2016)'s investigation on both French instructors and U.S. teaching assistants' gender ranged across five years. Similarly, Fan et al. (2019)'s investigated the topic across seven. Their utilization of multi-sections has been described as the gold standard for researching students' ratings. Thus, we aimed to follow their lead by analyzing the reliability of students' ratings provided the same or different instructor, course type, and/or semester of enrollment in addition to testing reliability over more than 30 years of data. We examined the impact of a potential validity variable on the reliability of ratings using perceived fairness of grading. Therefore, we sought to explore the following research questions:

Exploratory Research Questions:

1) What is the reliability of student evaluations?
2) Are student evaluations reliable across time?
3) Is the average level of perceived fairness of the grading in the course a moderator of reliability in student evaluations over time?
4) Does the average variability in instructor fairness rating moderate reliability of student evaluations over time?

The following was pre-registered as a secondary data analysis at: https://osf.io/czb4f. The manuscript, code, and data can be found on our Open Science Framework page at: https://osf.io/k7zh2/ or GitHub: https://github.com/doomlab/Grade-Lean. This manuscript was written with the *R* packages *papaja* (Aust et al., 2022), *rio* (J. Becker et al., 2021), *dplyr* (Wickham et al., 2020), *nlme* (Pinheiro et al., 2017), *ggplot2* (Wickham, 2016), *MuMIn* (Bartoń, 2020), *ppcor* (Kim, 2015), and *effectsize* (*Effectsize*, 2023).

Method

Data Source

The archival study was conducted using data from the psychology department at a large Midwestern public university. We used data from 2898 undergraduate, 274 mixed-level undergraduate, and 42 graduate psychology classes taught from 1987 to 2018 that were evaluated by students using the same 15-item instrument. 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. In the last several years of evaluations, online versions of these forms were used with faculty encouraged to give students time to complete them in class while they were outside the classroom.

The questionnaire given to students can be found at https://osf.io/4sphx. These items were presented with a five-point scale from 1 (*strongly disagree*) to 5 (*strongly agree*). For this study, the overall instructor evaluation question was "The overall quality of this course was among the top 20% of those I have taken." For fairness, we used the question of "The instructor used fair and appropriate methods in the determination of grades." The ratings were averaged for each course, and the sample size for each rating was included.

Planned Analyses

The evaluations were filtered for those with at least fifteen student ratings for the course (Rantanen, 2012). We performed a robustness check for the first research question on the data when the sample size is at least $n = 10$ up to $n = 14$ (i.e., on all evaluations with at least 10 ratings, then at least 11 ratings, etc.) to determine if the reliability estimates are stable at lower sample sizes. We first screened the dataset (two evaluation questions, sample size for course) for accuracy errors, linearity, normality, and homoscedasticity. The data is assumed to not have traditional "outliers", as these evaluations represent true averages from student evaluations. If the linearity assumption fails, we considered potential nonparametric models to address non-linearity. Deviations from normality were noted as the large sample size should provide robustness for any violations of normality. If data appears to be heteroscedastic, we used bootstrapping to provide estimates and confidence intervals.

This data was considered structured by instructor; therefore, all analyses below were coded in *R* using the *nlme* package (Pinheiro et al., 2017) to control for correlated error of instructor as a random intercept in a multilevel model. Multilevel models allow for analysis of repeated measures data without collapsing by participant [i.e., each instructor/semester/course combination can be kept separate without averaging over these measurements; Gelman (2006)]. Random intercept models are regression models on repeated data that structure the data by a specified variable, which was instructor in this analysis. Therefore, each instructor's average rating score was allowed to vary within the analysis, as ratings would be expected to be different from instructor to instructor. In each of the analyses described below, the number of students providing ratings for the course was included as a control variable to even out differences in course size as an influence in the results. However, this variable was excluded if the models did not converge. The dependent variable and predictors varied based on the research question, and these are described with each analysis below.

RQ 1

In this research question, we examined the reliability of student evaluations on the overall rating and separately on the fairness rating. We calculated eight types of reliability using course (same or different) by instructor (same or different) by semester (same or different). The dependent variable was the first question average with a predictor of the comparison question average, and both sample sizes (first sample size, comparison sample size). Instructor code was used as the random intercept for both ratings (i.e., two instructor random intercepts, first and comparison). The value of interest was the standardized regression coefficient for the fixed effect of question from this model. Given that the large sample size will likely produce "significant" p-values, we used the 95% CI to determine which reliability values were larger than zero and to compare reliability estimates to each other.

RQ 2

We used the reliability for the same instructor and course calculated as described in RQ1 at each time point difference between semesters. For example, the same semester would create a time difference of 0. The next semester (Spring to Summer, Summer to Fall,

Fall to Spring) would create a time difference of 1. We used the time difference as a fixed effect to predict reliability for the overall question only with a random intercept of instructor. We used the coefficient of time difference and its confidence interval to determine if there was a linear change over time. Finally, we plotted the changes over time to examine if this effect was non-linear in nature and discussed implications of the graph.

RQ 3

Using the reliability estimates from RQ 2, we then added the average rating for the fairness question as the moderator with time to predict reliability. Fairness was calculated as the average of the fairness question for all courses involved in the reliability calculation for that instructor and time difference. Therefore, this rating represented the average perceived fairness of grading at the time of ratings. If this interaction effect's coefficient does not include zero, we performed a simple slopes analysis to examine the effects of instructors who were rated at average fairness, one standard deviation below average, and one standard deviation above average (J. Cohen et al., 2003).

RQ 4

Finally, we examined the average standard deviation of fairness ratings as a moderator of time to predict reliability. This variable represented the variability in perceived fairness in grading from student evaluations, where small numbers indicated relative agreement on the rating of fairness and larger values indicated a wide range of fairness ratings. The variability in fairness ratings was calculated in the same way as the mean fairness, which was only for the instructor and semester time difference evaluations that were used to calculate the reliability estimate. This research question was assessed the same way as research question three.

Results

Data Screening

The overall dataset was screened for normality, linearity, homogeneity, and homoscedasticity using procedures from Tabachnick et al. (2019). Data generally met assumptions with a slight skew and some heterogeneity. The complete anonymized dataset and other information can be found online at https://osf.io/k7zh2. This page also includes the manuscript written inline with the statistical analysis with the *papaja* package (Aust et al., 2022) for interested researchers/reviewers who wish to recreate these analyses. The bootstrapped versions of analyses and robustness analysis can be found online on our OSF page with a summary of results. We originally planned to bootstrap all analyses; however, the compute time for research question 1 was extremely long due to the size and complexity of the multilevel models, and therefore, we did not bootstrap that research question.

Descriptive Statistics

3214 evaluations included at least 15 student evaluations for analysis. Table 1 portrays the descriptive statistics for each course level including the total number of evaluations, unique instructors, unique course numbers, and average scores for the two rating items. Students additionally projected their course grade for each class ($A = 5$, $B = 4$, $C = 3$, $D = 2$, $F = 1$), and the average for this item is included for reference. Overall, 231 unique instructors and 70 unique courses were included in the analyses below across 94 semesters.

RQ 1

Each individual evaluation was compared to every other evaluation resulting in 5163291 total comparisons. Eight combinations of ratings were examined using instructor (same, different), course (same, different), and semester (same, different) on both the overall and fairness evaluation ratings separately. One of the individual ratings was used to predict the comparison rating (i.e., question 1 was used to predict a comparison question 1 for the same instructor, different instructor, same semester, different semester, etc.), and the number of ratings (i.e., rating sample size) per question were used as fixed-effects covariates. The instructor(s) were used as a random intercept to control for correlated error and overall average rating per instructor. The effects were then standardized using the *parameters* package (Lüdecke et al., 2023). The data was sorted by year and semester such that "predictor" was always an earlier semester predicting a later semester's scores, except in cases of the the same semester comparisons. Therefore, positive standardized scores indicate that scores tend to go up over time, while negative scores indicate that scores tend to go down over time.

As shown in Figure 1, reliability was highest when calculated on the same instructor in the same semester and within the same course for both overall rating and fairness. This reliability was followed by the same instructor, same semester, and different courses. Next, the reliability for same instructor, same course, and different semesters was greater than zero and usually overlapped in confidence interval with same instructor, same semester, and different courses. Interestingly, the same instructor with different courses and semesters showed a non-zero negative relationship, indicating that ratings generally were lower for later semesters in different courses.

For different instructors, we found positive non-zero readabilities when they were at least calculated on the same semester or course. These values were very close to zero, generally in the .01 to .05 range. The reliabilities that were calculated on different courses, semesters, and instructors include zero in their confidence intervals. Exact values can be found in the online supplemental document with the robustness analysis in csv format. Robustness analyses revealed the same pattern and strength of results for evaluation reliabilities when sample size for evaluations was considered at $n = 10$, 11, 12, 13, and 14.

RQ 2

The paired evaluations were then filtered to only examine course and instructor matches to explore the relation of reliability across time. Reliability was calculated by calculating the partial correlation between the overall rating for the course first evaluation and the overall rating for the course second evaluation, controlling for the number of ratings within those average scores. This reliability was calculated separately for each instructor and semester difference (i.e., the time between evaluations, 0 means same semester, 1 means the next semester, 2 means two semesters later, etc.). The ratings were filtered so that at least 10 pairs of ratings were present for each instructor and semester difference combination (Weaver & Koopman, 2014). Of 36084 possible matched instructor and course pairings, 30728 included at least 10 pairings, which was 1009 total instructor and semester combinations.

The confidence interval for the effect of semester difference predicting reliability did not cross zero, b = -0.004, 95% CI [-0.005, -0.003], $R^2 = .04$. The coefficient, while small, represents a small effect of time on the reliability of instructor ratings. As shown in Figure 2, reliability appears to decrease across time.

RQ 3

The confidence interval for the interaction of semester time difference and average fairness did cross zero, b = -0.001, 95% CI [-0.007, 0.005], $R^2 = .04$. Therefore, there was no effect of the interaction of average fairness with semester differences in predicting reliability. Similarly, average fairness did not predict reliability overall, b = -0.041, 95% CI [-0.226, 0.143].

RQ 4

The confidence interval for the interaction of variability of fairness and semester time difference did cross zero, b = -0.010, 95% CI [-0.022, 0.002], $R^2 = .05$. The variability of fairness also did not predict reliability overall, b = 0.291, 95% CI [-0.091, 0.672].

Discussion

This investigation measured the reliability of SETs by calculating the reliability of evaluations across instructors, semesters, and courses. Our first research question asked what the reliability of SETs was given the instructor, course, or semester. Our data showed that SETs of the same instructor within the same course and same semester were the most reliable [rs ~ .50 — 75th percentile of known psychology correlations; Lovakov and Agadullina (2021)], followed by those collected from students enrolled in the same course, with the same instructor, but in different semesters (rs ~ .12 — 25th percentile of known psychology correlations). Our second question investigated if instructors' SETs became more reliable with increasing years of teaching experience; stated simply, we explored if experience across time matters. We extended previous meta-analyses on reliability to show that reliability appears to slightly, but significantly, decrease over time — a new finding in comparison to the work of Marsh (2007). Given the small size of this effect, reliability would decrease approximately .06 points in the time normally designated for tenure and/or promotion (i.e., -.004 x 3 semesters x 5 years). This small decrease may not impact the administrative process, but it is worth considering that decreases in reliability could be expected.

Last, we explored the relationship of a variable that we believed potentially impacts the validity of SETs: perceived fairness in grading. Perceived fairness did not appear to impact reliability scores, nor did it moderate with time to predict reliability scores. While variability in perceived fairness is found across and within instructor ratings, this variability also did not impact reliability information. In other words, our data does not support that instructors perceived as fair have higher or lower reliability of their SETs. Further, it did not seem to matter if all students agreed the instructor was fair (low variability in perceived fairness) or if they disagreed (high variability in perceived fairness) when predicting the reliability of SETs.

This study extends previous work with several new strengths (Benton & Cashin, 2014; Benton & Ryalls, 2016; Marsh, 2007; Zhao & Gallant, 2012). The data included in this manuscript represents over 30 years of SETs and was analyzed for reliability within and across courses, semesters, and instructors; thus, providing new insights into the expected level of reliability in different calculation scenarios. Sensitivity and bootstrapped analyses show that these results are robust even with a smaller number of evaluations used, supporting and extending work by Rantanen (2012). Last, we investigated the impact of validity variables on reliability, not just the overall validity of SETs based on various potential biases.

Given these results, what should instructors and administrators do with student evaluations of teaching? (Benton & Young, 2018) provide a comprehensive checklist of ways to assess teaching and interpret evaluations considering the long history of validity questions for SETs. Here, we add that it is important to understand that reliability will vary by course and semester as instructor variability is usually expected. It is tempting to think that the same instructor teaching the same course should reliably get the same SET ratings; however, we should consider that instructors will grow and change over time, which may contribute to lessened reliability across time (in addition to other known biases, such as age). Further, facets of the different courses taught likely contribute to the lessened reliability between courses taught by the same instructor (i.e., required statistics courses versus elective courses). As Benton and Young (2018) describe, the evaluation procedure should be useful, and it may not be fruitful to compare different years or even courses. SETs should therefore be contextualized to the course and semester in which they were received.

While this study provides valuable evidence about SET reliability, it only includes the SET ratings of one department, and our descriptive statistics suggest these ratings were often collected at ceiling on a 1 to 5 Likert type scale. Moreover, SETs are always biased by the students who are in class or fill out the online survey — information about missing student perceptions are never recorded. The concerns about the validity of SETs are still relevant, and it may be that reliability is interesting but not altogether useful if the scores are not valid representations of teaching effectiveness. However, open-ended feedback, paired with SET scores, are often a beneficial gauge for instructors to reflect on new practices or how a semester progressed. As universities struggle to balance demands of higher education cost and student enrollment, teaching effectiveness may be a critical target for administrators to ensure student engagement and retention. These results suggest that SETs can be reliable indicators of teaching effectiveness, but likely only within the same courses and semester. Thus, a multifaceted approach to assessing instructor effectiveness and improvement is a more appropriate measurement tool for long-term evaluations of instruction (Benton & Young, 2018).

References

Arubayi, E. A. (1987). Improvement of instruction and teacher effectiveness: are student ratings reliable and valid? *Higher Education*, *16*(3), 267–278. https://doi.org/10.1007/BF00148970

Aust, F., Barth, M., Diedenhofen, B., Stahl, C., Casillas, J. V., & Siegel, R. (2022). *Papaja: Prepare american psychological association journal articles with r markdown.* https://CRAN.R-project.org/package=papaja

Bartoń, K. (2020). *MuMIn: Multi-model inference.* https://CRAN.R-project.org/package=MuMIn

Becker, J., Chan, C., Chan, G. C., Leeper, T. J., Gandrud, C., MacDonald, A., Zahn, I., Stadlmann, S., Williamson, R., Kennedy, P., Price, R., Davis, T. L., Day, N., Denney, B., & Bokov, A. (2021). *Rio: A swiss-army knife for data i/o.* https://cran.r-project.org/web/packages/rio/

Becker, W. E., Bosshardt, W., & Watts, M. (2012). How Departments of Economics Evaluate Teaching. *The Journal of Economic Education*, *43*(3), 325–333. https://doi.org/10.1080/00220485.2012.686826

Benton, S. L., & Cashin, W. E. (2014). *Student Ratings of Instruction in College and University Courses* (M. B. Paulsen, Ed.; pp. 279–326). Springer Netherlands. https://doi.org/10.1007/978-94-017-8005-6_7

Benton, S. L., & Ryalls, K. R. (2016). *Challenging Misconceptions about Student Ratings of Instruction. IDEA Paper #58.* https://eric.ed.gov/?id=ED573670

Benton, S. L., & Young, S. (2018). *Best Practices in the Evaluation of Teaching. IDEA Paper #69.* https://eric.ed.gov/?id=ED588352

Berk, R. A. (2018). Start Spreading the News: Use Multiple Sources of Evidence to Evaluate Teaching. *The Journal of Faculty Development*, *31*(1), 73–81.

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research.*

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

Boswell, S. S. (2016). Ratemyprofessors is hogwash (but I care): Effects of Ratemyprofessors and university-administered teaching evaluations on professors. *Computers in Human Behavior*, *56*, 155–162. https://doi.org/10.1016/j.chb.2015.11.045

Chen, C. Y., Wang, S.-Y., & Yang, Y.-F. (2017). A Study of the Correlation of the Improvement of Teaching Evaluation Scores Based on Student Performance Grades. *International Journal of Higher Education*, *6*(2), 162–168. https://doi.org/10.5430/ijhe.v6n2p162

Cohen, J., Cohen, P., West, S. G., & Aiken, L. (2003). *Applied multiple regression / correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates.

Cohen, P. A. (1981). Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies. *Review of Educational Research*, *51*(3), 281–309. https://doi.org/10.3102/00346543051003281

Crumbley, L., Henry, B. K., & Kratchman, S. H. (2001). Students' perceptions of the evaluation of college teaching. *Quality Assurance in Education*, *9*(4), 197–207. https://doi.org/10.1108/EUM0000000006158

Dunn, K. A., Hooks, K. L., & Kohlbeck, M. J. (2014). Preparing Future Accounting Faculty Members to Teach. *Issues in Accounting Education*, *31*(2), 155–170. https://doi.org/10.2308/iace-50989

Effectsize: Indices of effect size. (2023). [Computer software]. https://cran.r-project.org/web/packages/effectsize/

Esarey, J., & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education*, *45*(8), 1106–1120. https://doi.org/10.1080/02602938.2020.1724875

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLOS ONE*, *14*(2), e0209749. https://doi.org/10.1371/journal.pone.0209749

Feistauer, D., & Richter, T. (2017). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education*, *42*(8), 1263–1279. https://doi.org/10.1080/02602938.2016.1261083

Felton, J., Koper, P. T., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues: student evaluations of professors on Ratemyprofessors.com. *Assessment & Evaluation in Higher Education*, *33*(1), 45–61. https://doi.org/10.1080/02602930601122803

Flaherty, C. (2015). Flawed Evaluations. In *Inside Higher Ed*. https://www.insidehighered.com/news/2015/06/10/aaup-committee-survey-data-raise-questions-effectiveness-student-teaching

Flaherty, C. (2020). Even "Valid" Student Evaluations Are 'Unfair'. In *Inside Higher Ed*. https://www.insidehighered.com/news/2020/02/27/study-student-evaluations-teaching-are-deeply-flawed

Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research*. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, *48*(3), 432–435. https://doi.org/10.1198/004017005000000661

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement*, *15*(1), 1–13. https://www.jstor.org/stable/1433721

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

Hattie, J., & Marsh, H. W. (1996). The Relationship Between Research and Teaching: A Meta-Analysis. *Review of Educational Research*, *66*(4), 507–542. https://doi.org/10.3102/00346543066004507

Heffernan, T. (2022). Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assessment & Evaluation in Higher Education*, *47*(1), 144–154. https://doi.org/10.1080/02602938.2021.1888075

Horan, S. M., Chory, R. M., & Goodboy, A. K. (2010). Understanding students' classroom justice experiences and responses. *Communication Education*, *59*(4), 453–474. https://doi.org/10.1080/03634523.2010.487282

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, *4*(1), 1304016. https://doi.org/10.1080/2331186X.2017.1304016

Johnson, M. D., Narayanan, A., & Sawaya, W. J. (2013). Effects of Course and Instructor Characteristics on Student Evaluation of Teaching across a College of Engineering: Student Evaluation of Teaching across a College of Engineering. *Journal of Engineering Education*, *102*(2), 289–318. https://doi.org/10.1002/jee.20013

Kim, S. (2015). *Ppcor: Partial and semi-partial (part) correlation.* https://cran.r-project.org/web/packages/ppcor/

Kornell, N., & Hausman, H. (2016). Do the best teachers get the best ratings? *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00570

Leventhal, G. S. (1980). *What Should Be Done with Equity Theory?* (K. J. Gergen, M. S. Greenberg, & R. H. Willis, Eds.; pp. 27–55). Springer US. https://doi.org/10.1007/978-1-4613-3087-5_2

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

Lüdecke, D., Makowski, D., Ben-Shachar, M. S., Patil, I., Højsgaard, S., Wiernik, B. M., Lau, Z. J., Arel-Bundock, V., Girard, J., Maimone, C., Ohlsen, N., Morrison, D. E., & Luchman, J. (2023). *Parameters: Processing of model parameters.* https://CRAN.R-project.org/package=parameters

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innovative Higher Education*, *40*(4), 291–303. https://doi.org/10.1007/s10755-014-9313-4

Marks, R. B. (2000). Determinants of Student Evaluations of Global Measures of Instructor and Course Value. *Journal of Marketing Education*, *22*(2), 108–119. https://doi.org/10.1177/0273475300222005

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology*, *99*(4), 775–790. https://doi.org/10.1037/0022-0663.99.4.775

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52*(11), 1187–1197. https://doi.org/10.1037/0003-066X.52.11.1187

Mitchell, K. M. W., & Martin, J. (2018). Gender Bias in Student Evaluations. *PS: Political Science & Politics*, *51*(3), 648–652. https://doi.org/10.1017/S104909651800001X

O'Sullivan, C., Bhaird, C. M. an, Fitzmaurice, O., & Fhlionn, E. N. (2014). *An irish mathematics learning support network (IMLSN) report on student evaluation of mathematics learning support: Insights from a large scale multi?institutional survey.* National Centre for Excellence in Mathematics; Science Teaching; Learning (NCEMSTL). https://mural.maynoothuniversity.ie/6890/

Pepper, M. B., & Pathak, S. (2008). Classroom contribution: What do students perceive as fair assessment? *Journal of Education for Business*, *83*(6), 360–368. https://doi.org/10.3200/JOEB.83.6.360-368

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., & Team, R. C. (2017). *Nlme: Linear and nonlinear mixed effects models*. https://cran.r-project.org/package=nlme

Rantanen, P. (2012). The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalisability of students' evaluation of teaching. *Assessment & Evaluation in Higher Education*, *38*(2), 224–239.

https://doi.org/10.1080/02602938.2011.625471

Rovai, A. P., Ponton, M. K., Derrick, M. G., & Davis, J. M. (2006). Student evaluation of teaching in the virtual and traditional classrooms: A comparative analysis. *The Internet and Higher Education*, *9*(1), 23–35. https://doi.org/10.1016/j.iheduc.2005.11.002

Smith, B. P., & Hawkins, B. (2011). Examining student evaluations of black college faculty: Does race matter? *The Journal of Negro Education*, *80*(2), 149–162. https://www.jstor.org/stable/41341117

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*(4), 598–642. https://doi.org/10.3102/0034654313496870

Stankiewicz, K. (2015). Ratings of Professors Help College Students Make Good Decisions. In *New York Times*. https://www.nytimes.com/roomfordebate/2015/12/16/is-it-fair-to-rate-professors-online/ratings-of-professors-help-college-students-make-good-decisions

Stroebe, W. (2020). Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis. *Basic and Applied Social Psychology*, *42*(4), 276–294. https://doi.org/10.1080/01973533.2020.1756817

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2019). *Using multivariate statistics* (Seventh edition). Pearson.

Tata, J. (1999). Grade distributions, grading procedures, and students' evaluations of instructors: A justice perspective. *The Journal of Psychology*, *133*(3), 263–271. https://doi.org/10.1080/00223989909599739

Tripp, T. M., Jiang, L., Olson, K., & Graso, M. (2019). The Fair Process Effect in the Classroom: Reducing the Influence of Grades on Student Evaluations of Teachers. *Journal of Marketing Education*, *41*(3), 173–184. https://doi.org/10.1177/0273475318772618

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Weaver, B., & Koopman, R. (2014). An SPSS macro to compute confidence intervals for pearson's correlation. *The Quantitative Methods for Psychology*, *10*(1), 29–39. https://doi.org/10.20982/tqmp.10.1.p029

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Kirill Müller. (2020). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Wright, R. E. (2000). Student Evaluations and Consumer Orientation of Universities. *Journal of Nonprofit & Public Sector Marketing*, *8*(1), 33–40. https://doi.org/10.1300/J054v08n01_04

Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education*, *37*(2), 227–235. https://doi.org/10.1080/02602938.2010.523819

Zheng, X., Vastrad, S., He, J., & Ni, C. (2023). Contextualizing gender disparities in online teaching evaluations for professors. *PLOS ONE*, *18*(3), e0282704. https://doi.org/10.1371/journal.pone.0282704

Table 1

Descriptive Statistics of Included Courses

Statistic	Undergraduate	Mixed	Master's
N Total	2898	274	42
N Instructors	223	40	10
N Courses	41	21	8
Average N Ratings	34.39	21.15	21.10
Average Overall	3.94	4.01	3.72
SD Overall	0.55	0.59	0.67
Average Fairness	4.46	4.50	4.19
SD Fairness	0.35	0.38	0.55
Average Grade	4.26	4.52	4.41
SD Grade	0.33	0.27	0.34

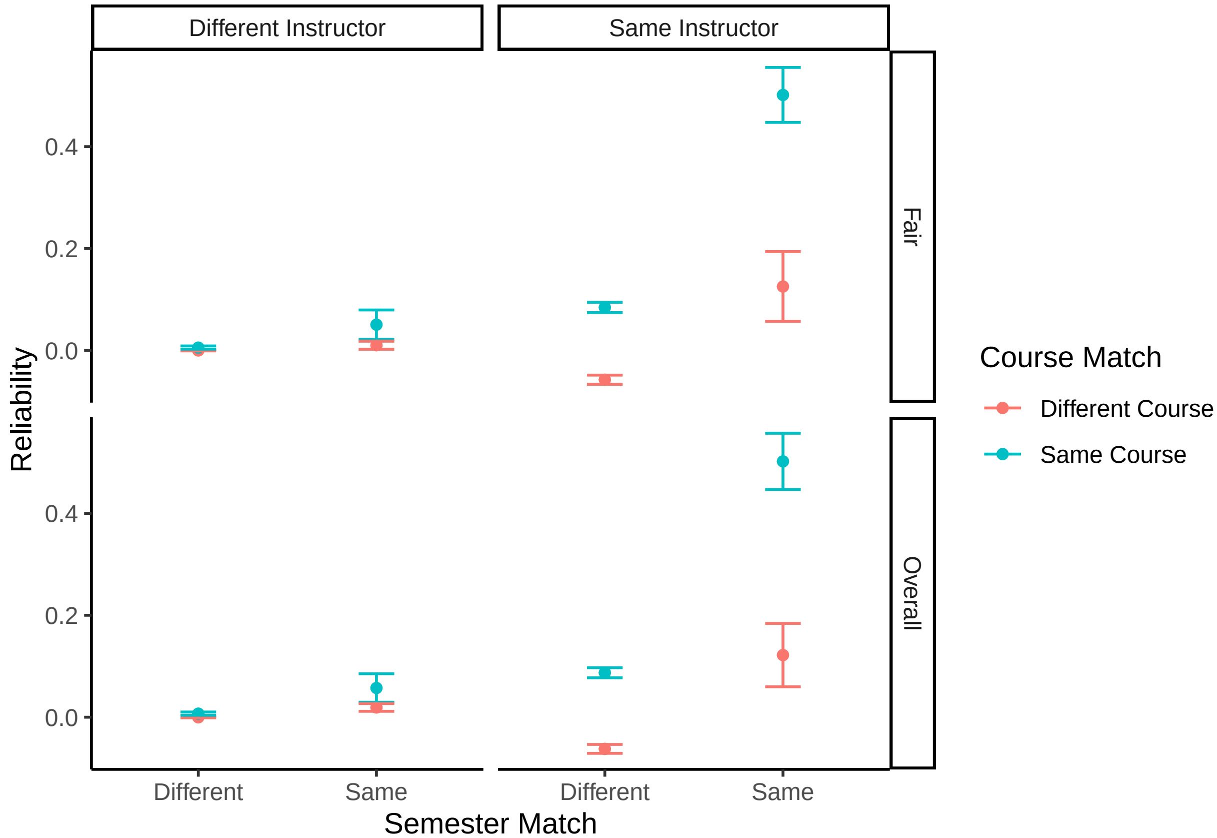

Figure 1

Reliability estimates for instructor, course, and semester combinations.

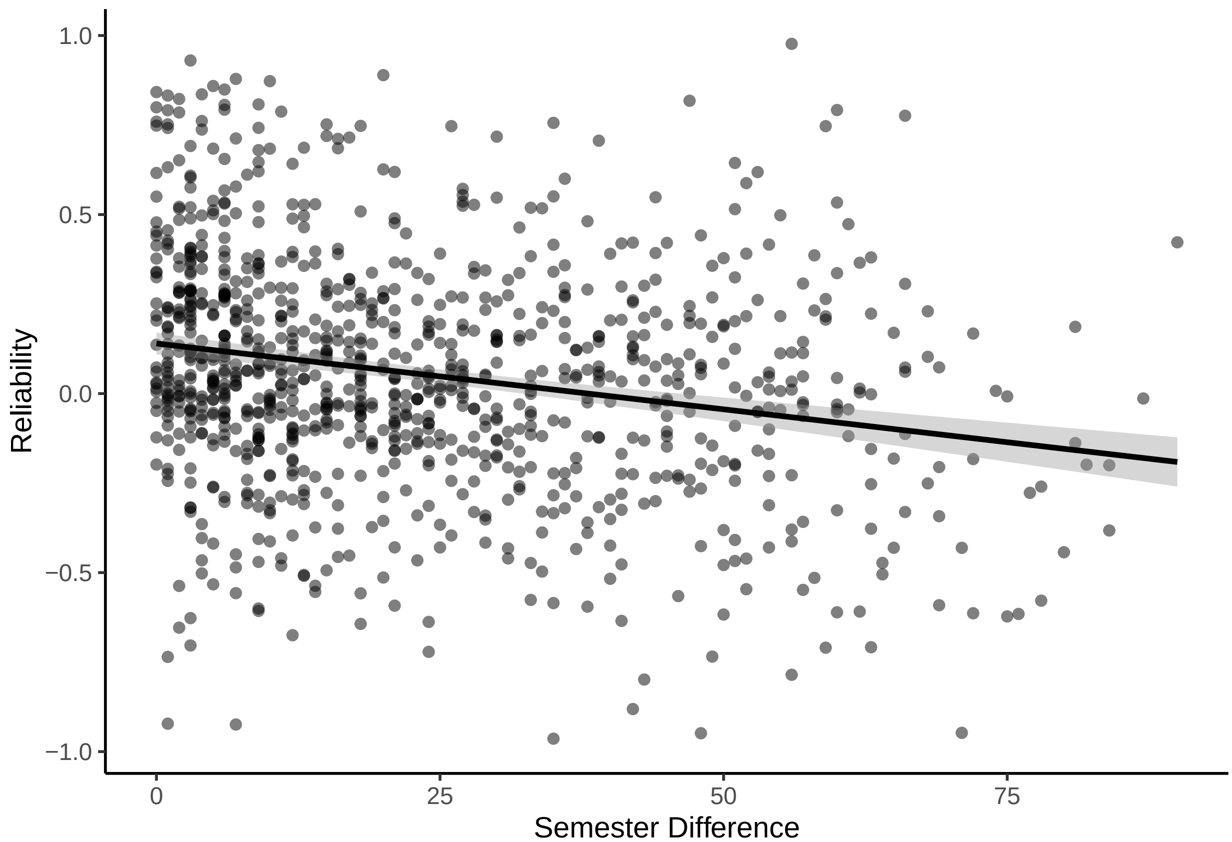

Figure 2

Reliability estimates for same instructor and course across time.