ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-513

Students' Grade Satisfaction Influences Evaluations of Teaching: Evidence from Individual-level Data and an Experimental Intervention

Vladimir Kogan
Ohio State University

Brandon Genetin
Ohio State University

Joyce Chen
Ohio State University

Alan Kalish
Ohio State University

Student surveys are widely used to evaluate university teaching and increasingly adopted at the K-12 level, although there remains considerable debate about what they measure. Much disagreement focuses on the well-documented correlation between student grades and their evaluations of instructors. Using individual-level data from 19,000 evaluations of 700 course sections at a flagship public university, we leverage both within-course and within-student variation to rule out popular explanations for this correlation. Specifically, we show that the relationship cannot be explained by instructional quality, workload, grading stringency, or student sorting into courses. Instead, student grade satisfaction -- regardless of the underlying cause of the grades -- appears to be an important driver of course evaluations. We also present results from a randomized intervention with potential to reduce the magnitude of the association by reminding students to focus on relevant teaching and learning considerations and by increasing the salience of the stakes attached to evaluations for instructor careers. However, these prove ineffective in muting the relationship between grades and student scores.

VERSION: January 2022

Suggested citation: Kogan, Vladimir, Brandon Genetin, Joyce Chen, and Alan Kalish. (2022). Students' Grade Satisfaction Influences Evaluations of Teaching: Evidence from Individual-level Data and an Experimental Intervention. (EdWorkingPaper: 22-513). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/spsf-tc23

Students' Grade Satisfaction Influences Evaluations of Teaching: Evidence from Individual-level Data and an Experimental Intervention

Vladimir Kogan
Department of Political Science
Ohio State University
kogan.18@osu.edu

Brandon Genetin
Department of AED Economics
Ohio State University
genetin.4@osu.edu

Joyce Chen
Department of AED Economics
Ohio State University
chen.1276@osu.edu

Alan Kalish
Office of Academic Affairs
Ohio State University
kalish.3@osu.edu

January 5, 2022

Abstract

Student surveys are widely used to evaluate university teaching and increasingly adopted at the K-12 level, although there remains considerable debate about what they measure. Much disagreement focuses on the well-documented correlation between student grades and their evaluations of instructors. Using individual-level data from 19,000 evaluations of 700 course sections at a flagship public university, we leverage both within-course and within-student variation to rule out popular explanations for this correlation. Specifically, we show that the relationship cannot be explained by instructional quality, workload, grading stringency, or student sorting into courses. Instead, student grade satisfaction—regardless of the underlying cause of the grades—appears to be an important driver of course evaluations. We also present results from a randomized intervention with potential to reduce the magnitude of the association by reminding students to focus on relevant teaching and learning considerations and by increasing the salience of the stakes attached to evaluations for instructor careers. However, these prove ineffective in muting the relationship between grades and student scores.

Keywords: students evaluations of teaching, student voice, student learning

Word Count: 4,668

Starting in the 1970s, student surveys have become widely used as one of the primary methods of evaluating teaching in higher education. Increasingly, such surveys are also being incorporated into state accountability systems and teacher evaluation frameworks in primary and secondary education, as complements or alternatives to test-score based metrics (Bacher-Hicks et al. 2019, Schneider et al. 2021).[1] However, student evaluations remain controversial, with disagreement among both researchers and practitioners about the validity and interpretation of the resulting measures.

In recent years, much of the concern has focused on the widely documented disparities in evaluations given by students to minority and female instructors.[2] The evidence that these disparities are driven at least partially by biases appears to be overwhelming (see Chávez and Mitchell 2020, Kreitzer and Sweet-Cushman 2021). But the high-profile conversations about bias in student evaluations overshadow a longer-run—and arguably more fundamental—debate about which dimensions of the student experience survey-based evaluations actually capture. At the heart of this debate is the well-documented correlation between the grades that students receive and the ratings they give to their instructors.

One school of thought points to this correlation as evidence that student evaluations accurately measure teaching quality. From this perspective, higher grades reflect greater learning, and students who learn more rate their instructors more positively as a result. As Linse (2017) summarizes, "Faculty who teach well, have grading practices that are accurate reflections of students' learning, and have grade distributions with a peak near the high end of the grading scale, may receive higher ratings—and deservedly so."

The more cynical view, on the other hand, argues that students reward instructors

[1] In the K-12 context, the Tripod Student Perception Survey is one of the most widely used instruments (Wallace, Kelcey and Ruzek 2016).

[2] Similar concerns have been raised about bias against other disadvantaged groups, including non-native English speakers.

who minimize the workload and effort required to obtain good grades.[3] Often described as the "leniency hypothesis," this perspective concludes that students appreciate easy classes, rewarding them with higher evaluations, and such classes also tend to produce the highest overall grades (e.g., Greenwald and Gillmore 1997). As Braga, Paccagnella and Pellizzari (2014) note, the data is consistent with a world in which "students dislike exerting effort, especially the least able ones, and when asked to evaluate the teacher they do so on the basis of how much they enjoyed the course."[4] Proponents of the leniency hypothesis argue that the increased reliance on student evaluations as the primary measure of teaching quality explains the grade inflation observed at the college level in recent decades (Eiszler 2002, Johnson 2003, Denning et al. 2021).[5]

We contribute to this debate by using data on more than 19,000 evaluations of more than 700 course sections completed by students at the Ohio State University in spring 2021. In contrast to much of the existing research, which focuses on class-level averages of both grades and student evaluation scores, we can take advantage of individual-level data, including official end-of-course grades and individual students' course evaluations.[6] By comparing survey scores of students enrolled in the same class, we can hold constant all shared aspects of the learning experience—including workload, grading expectations, and most aspects of teaching quality. In addition, by examining how the same student evaluates different classes taken during the same semester, we can directly account for

[3] A yet third perspective argues that the correlation between grades and evaluations reflects student sorting into classes, with more motivated students both earning higher grades and expressing more appreciation of better teaching. As we discuss below, our empirical strategy accounts for such potential sorting.

[4] Among other evidence, Braga, Paccagnella and Pellizzari (2014) show that factors orthogonal to teaching quality but which affect the effort required to attend class—in particular, weather on the day students complete evaluations—influence student ratings of instruction.

[5] Many studies dispute the leniency hypothesis (e.g., Abrami et al. 1980, Marsh and Roche 2000, Centra 2003), with much of the debate focused on disagreement about appropriate methodologies and model specifications.

[6] Stumpf and Freedman (1979) examined both within-class and between-class associations between grades and evaluations, although nearly all subsequent studies have focused on course-level averages, largely due to lack of availability of individual-level data. For an important exception, however, see Johnson (2003).

potential non-random sorting of students into courses.

Our central finding is that that *neither* the learning story nor the leniency hypothesis appear to be the primary driver of the individual-level correlation between students' course grades and their teaching evaluations. Controlling for both course- and student-level fixed effects, which together account for nearly all of the hypothesized mechanisms linking grades and evaluations proposed in the literature, does not meaningfully reduce the size of the raw bivariate association. Instead, the results are most consistent with student course evaluations reflecting "grade satisfaction"—with those who earn higher grades, regardless of the reason, reporting more positive evaluations of their instructors.[7] The magnitudes of these relationships are substantively large and policy relevant. For example, depending on the specification, receiving an A- instead of an A grade reduces a student's overall class evaluation score by between 0.1 and 0.2 points on a 5-point scale, the equivalent to between 10 percent and 20 percent of a typical standard deviation in average course ratings at Ohio State.

In the second half of the analysis, we also present results from a randomized experiment examining interventions with the potential to reduce the influence of irrelevant considerations on student course evaluations. Specifically, students are randomly presented with different versions of introductory text incorporated into the standard evaluation survey used by the university, some reminding students about the influence of implicit biases and others emphasizing the high-stakes nature of the evaluations for instructor careers. Because the randomization is done within classes, we can combine it with our fixed-effects model to examine the extent to which these interventions can reduce the observed relationship between course grades and evaluations. Perhaps surprisingly, the experiment provides no

[7] Although Greenwald and Gillmore (1997) discuss grade satisfaction as being synonymous with grading leniency, the two mechanisms are theoretically distinct because instructors can increase grades—and thereby grade satisfaction—through many other means as well.

evidence that these reminders make any meaningful difference, with the effect of grades on reported course satisfaction virtually unchanged regardless of the survey wording used.

These findings suggest that student evaluations strongly incentivize instructors to ensure that students receive higher grades, although the normative implications of this incentive depend greatly on how instructors respond. If they improve the quality of their teaching, the result is a closer alignment between the interests of students and educators. On the other hand, if instructors increase grades by lowering standards, reducing workload, or simply teaching to the test, higher evaluations may come at the expense of student learning and longer-term success. Although rigorous evidence on how instructors respond to pressures created by the need to achieve high student evaluations is limited, several high-quality studies do provide reason to worry (Carrell and West 2010, Braga, Paccagnella and Pellizzari 2014). Tracking students over time, these studies find that students who take courses with more highly-reviewed instructors tend to perform worse in subsequent, follow-on classes. Overall, our findings—when combined with evidence from these studies—suggest that heavy reliance on student evaluations as a tool for evaluating teaching may work at cross-purposes with other ongoing efforts to improve student achievement, persistence and long-term academic success.

Background on Student Evaluations

The intellectual foundation for modern course evaluations is built on a series of meta-analyses conducted in the 1970s and 1980s (e.g., Cohen 1981, Feldman 1989). These analyses examined data from multiple sections of the same courses, thereby holding constant both the course syllabi and assessments used. The evidence revealed a positive relationship between the average grades (typically on final exams) earned by students in each section and the average ratings of the instructor as reported by students on end-of-course evalua-

tions. The correlation was interpreted by many as evidence that student surveys provided an accurate measure of teaching quality. Although this interpretation remains contested, both proponents and critics of student evaluations generally accept the existence of the underlying statistical relationship.[8]

Largely overlooked at the time, even some of these early studies provided suggestive evidence that the association between grades and evaluations actually captured student grade satisfaction rather than teaching quality or grading leniency. For example, when reviewing the existing multi-section research, Cohen (1981) noted that the correlation between student performance and their evaluations was "much higher" when students knew their final grades compared to when they didn't, rising from 0.35 to 0.85. In another study, researchers added several questions to the end-of-semester student survey at the University of Wisconsin asking about the legibility of the instructors' handwriting, voice audibility, and quality of classroom facilities. Although such factors may account for differences in student learning between courses, the researchers found a positive association between student responses to these questions and their expected grades even *within* classes, suggesting the presence of a "halo effect" tied to grades (Greenwald and Gillmore 1997).

OSU Survey Instrument

The current survey instrument used at the Ohio State University was developed and piloted in the 1980s as part of an effort to standardize the process for incorporating student feedback into the teaching evaluation process. Prior to the development of the university-wide survey, individual departments relied on their own questionnaires, of varying quality, or a single-question survey fielded by the university. The current instrument asks students to

[8]Some subsequent research has raised questions about the robustness of the earlier meta-analysis. More recent meta-analyses have found more mixed results (Clayson 2009), and Uttle, White and Gonzalez (2017) argue that the original studies were distorted due to small sample sizes and bias against the publication of null results.

rate their course and instructor on ten separate questions—three focused on the instructor's preparedness and clarity of presentation, three on perceived rapport and instructor commitment to student learning, three on the students' sense of their own learning, and a "global" question asking for the overall rating of the course.[9] Although the instrument was designed to separately measure distinct dimensions of the student experience thought at the time to most strongly predict student learning, subsequent evaluations showed that student responses on all of the questions are highly correlated and appear to load on a single underlying latent factor, which is in turn highly correlated with the "global" satisfaction question that we focus on in this study (Zhao and Gallant 2012).

Although initially done on paper in class, the survey is now completed online by students at the end of the semester, prior to the final exams, with completion rates of about 50 percent.

Since the initial development of the instrument, the university has carried out three separate evaluations that are summarized in a 1998 working group report. Most noteworthy, each of these evaluations found a significant association between the average course grade and the average student evaluation score. The most recent analysis reported that "average grade was the single best predictor of the mean item 10 [global question] ratings, accounting for 8.2% of the variance across sections" (Ohio State University 1998). Heavily quoting various writings of psychology researcher Wilbert McKeachie, the working group concluded that this correlation was difficult to interpret. Acknowledging that the relationship could be viewed as evidence that students reward grading leniency, the report also noted alternative and "benign" explanations—including the possibility that the correlation could reflect student selection into courses ("students with strong academic motivation

[9]Specifically, the question states: "Overall, I would rate this instructor as ... [Poor, Fair, Neutral, Good, Excellent]." Per the university's standard practices, we convert the categorical responses to numerical scores, ranging from 1 for "poor" to 5 for "excellent."

both do better in their course work and more fully appreciate the efforts of the instructor") or true differences in teaching quality("good teachers induce students to learn more (and therefore earn and receive good grades), and as a result their strong teaching performance is rewarded with higher student evaluation scores").

To our knowledge, only one other published study has used Ohio State data to explore the relationship between course evaluations and student achievement. Weinberg, Hashimoto and Fleisher (2009) examined students enrolled in over 400 sections of introductory and intermediate economics courses between 1995 and 2004. Specifically, the authors asked whether student evaluation scores of instructors teaching introductory courses predicted performance in subsequent intermediate economics courses. They found that the average evaluation scores of introductory course instructors did in fact predict student performance in follow-on courses, but also that this relationships was entirely mediated by student grades in the initial, introductory course. The authors interpreted these findings pessimistically, writing: "Student evaluations of teaching differ from the ideal construct because they are affected by grade leniency and do not reflect learning produced in a course." However, the same results could alternatively be read to suggest that introductory course grades fully capture the true differences in instructional quality, leaving little residual variation in learning to be explained by course evaluations once grades are accounted for. Thus, these results are arguably consistent with many of the alternative interpretations of teaching evaluations in the literature.

Data and Empirical Strategy

The data used in our analysis comes from Genetin et al. (2021), a randomized controlled experiment that sought to replicate and extend an earlier study performed at Iowa State University (Peterson et al. 2019). Peterson et al. (2019) showed that short introductory

text added to the university's course evaluation survey appeared to reduce the bias against female instructors (primarily from male students), although Genetin et al. (2021) failed to replicate this effect among the much larger Ohio State sample.

The Ohio State data was collected in spring 2021 for a subset of undergraduate courses taught that semester in either the Colleges of Arts and Sciences or the College of Food, Agricultural, and Environmental Sciences. All instructors in these colleges were invited to participate, and approximately 400 (16%) did so. Genetin et al. (2021) provide additional details about the recruitment method and show that the resulting sample was broadly representative of the classes and students enrolled in courses during this period.

After removing co-taught courses, for which students completed separate evaluations for each instructor, our final sample includes 19,158 evaluations completed by 14,051 unique students in 718 different class sections. For each observation, we observe the student's score on the "global" evaluation question, the official end-of-course grade, and the student's major. About 70 percent of the students were enrolled in only one participating course, although we observe multiple class grades and ratings for the remaining third of students, an additional source of variation we leverage in some of our analyses.

Typically, students select the grading method used in each course, choosing either letter grades that contributes to their grade point averages or a pass-fail grading scheme. While passing grades earn credits toward graduation, they do not affect the students' GPAs. In spring 2021, however, the university used a modified grading system designed to minimize the impact of the COVID-19 pandemic on students. Under this system, all students received letter grades. D grades were subsequently converted into a "pass" grade in the university database and failing grades were automatically converted into a "no pass" grade, neither of which were included in the GPA calculations.[10] Because the grade distribution is highly

[10]Although the pandemic clearly affected grading policy and practices at the institution, there is little reason to believe this should influence the interpretation of our results, which leverage variation in grades

skewed, with far fewer low grades than high grades, we combine grades below an A- into whole-grade categories (dropping pluses and minuses). In other words, each final grade is coded to be in one of the following groups: A, A-, B, C, Pass, and No Pass.[11]

We begin with a pooled bivariate model, regressing the students' course evaluation scores on an indicator variable for each of the possible grade categories, with As serving as the omitted baseline. This simple model provides the benchmark against which we evaluate the more sophisticated specifications.

Next, we add class section fixed effects, estimating the following specification:

$$\text{SEI}_{i,c} = \alpha_c + \sum_{\theta} \beta_\theta \mathbf{Grade}_{i,c} + \epsilon_{ct} \tag{1}$$

where $\text{SEI}_{i,c}$ denotes the evaluation score recorded by student i in class c, α_c captures the course-specific effect, and β represents the effect of each discrete grade category (θ). Because this specification examines only variation in student evaluation scores and grades *within* classes, it holds constant aspects of the class experience that are common to all students, including workload, grading leniency, and many aspects of teaching quality.[12] To the extent that grading leniency and effective instructional practices are responsible for the positive association between student grades and evaluations, controlling for class fixed effects should substantially attenuate the observed association.

We further supplement this model by adding student fixed effects (γ_i):

$$\text{SEI}_{i,c} = \alpha_c + \sum_{\theta} \beta_\theta \mathbf{Grade}_{i,c} + \gamma_i + \epsilon_{ct} \tag{2}$$

across students that remain net of the grading changes.

[11] Overall, As account for over 50 percent of the grades in our sample, A-s for 13 percent, Bs for 23 percent, Cs for 9 percent, and Pass and No Pass grades together representing less than 5 percent.

[12] It is possible that some aspects of teaching quality, such as instructor feedback and/or willingness to answer questions outside of class, may vary between students.

This two-way fixed effects specification leverages information only from students who we observe in multiple classes, allowing us to account for non-random sorting into courses due to fixed student attributes—such as intrinsic motivation, work ethic, or aptitude—that could affect both student performance and each student's subjective satisfaction. Again, to the extent that student sorting drives the observed association, controlling for student fixed effects should further attenuate the relationship between grades and evaluation scores. Across all specifications, we cluster the standard errors by course section.

While the student fixed effects account for many types of student sorting into classes, it is also possible that certain student characteristics deferentially affect performance and evaluations in some classes more than others. Of particular concern is student interest in the subject matter. For example, it is possible that students are inherently more interested in courses taken to satisfy major requirements, compared to general education or elective classes. Thus, student interest could affect both overall course performance and a student's subjective evaluation of the course. To account for this possibility, we supplement Equation 2 by adding additional variables capturing the students' self-reported reasons for taking each course, which are also collected as part of the evaluation survey.[13]

RCT Intervention

As noted above, the data incorporates a randomized experiment in which different subsets of students saw slightly different versions of the survey instrument. While students assigned to the control group completed the standard questionnaire, the remainder were first presented with some introductory test, modeled on the intervention used in Peterson et al. (2019). The first version reminded students about implicit biases and encouraged them to focus on the course content rather than irrelevant considerations:

[13] As an additional check, we also created indicators for whether each student had the same major as the modal enrollee in each course. Adding this control had no impact on the results.

> The Ohio State University recognizes that student evaluations of teaching are often influenced by students' **unconscious** and **unintentional** biases about the race and gender of the instructor. Women and instructors of color are systematically rated lower in their teaching evaluations than white men, even when there are no actual differences in the instruction or in what students have learned.
>
> As you fill out the course evaluation, please keep this in mind and make an effort to resist stereotypes about professors. Focus on your opinions about the content of the course (the assignments, the textbook, the in-class material) and not unrelated matters (the instructor's appearance).

The second version increased the salience of the stakes involved:

> Student evaluations of teaching play an important role in the review of faculty. Your participation in this process is essential; having feedback from as many students as possible provides a more comprehensive view of the strengths and weaknesses of each course offering, allowing instructors to improve their practices and increase learning. Moreover, your opinions influence the review of instructors that takes place every year and will be taken into consideration for decisions regarding promotion and tenure.

The final version combined both the implicit bias and the high-stakes treatment. Randomization was done at the student level and stratified by course. For courses with enrollment below 40, students were randomized between the control group and the implicit bias treatment only. In courses with 40 or more students, randomization took place across all possible conditions. To ensure that students enrolled in more than one course included in the study saw the same consistent version of the survey across across all participating classes, treatment assignment was fixed based on the random assignment in the student's

largest enrolling course.[14] Genetin et al. (2021) report statistical tests confirming that the randomization worked as intended, with both student and instructor characteristics balanced across different treatment conditions.

Although not the primary focus of the original study, we leverage these variations in survey wording to examine whether they moderate the relationship between student grades and their course evaluations.

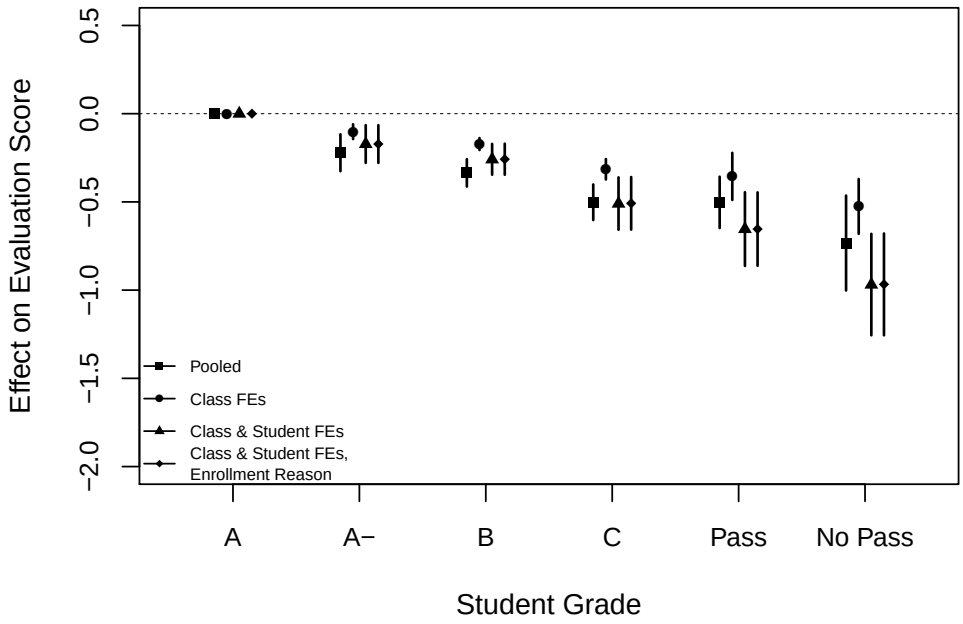
Results

Figure 1 visually presents the coefficients of interest and associated confidence intervals from each of the regression models described above. The first set of estimates, demarcated by the square symbols in the figure, comes from a pooled specification that simply regresses student course evaluation scores on their grades. Compared to students who receive an A in their class, students earning an A- rate their class 0.22 points lower on average on the five-point scale used on the university survey instrument. For students receiving a B, the average evaluation is 0.34 points lower. Scores fall further for both C and pass grades—corresponding to an average evaluation score that is approximately 0.5 points lower than reported by A students. Students who receive a failing grade rate their course more than 0.7 points lower.

The next three set of estimates incorporate class fixed effects (circle symbol), both class and student fixed effects (triangle), and finally both sets of fixed effects and additional controls for reason of enrollment in the course (diamond), respectively. While the precise point estimates vary somewhat across specifications, the general pattern is quite consistent, with students receiving lower grades giving their courses progressively lower evaluations.

[14] As a result, some students in classes with fewer than 40 students received either the high-stakes or the combo treatment based on randomization that took place in a larger course.

Figure 1: Relationship between end-of-semester grades and students' evaluations of instruction



Note: The figure plots the point estimates and associated 95 percent confidence intervals from four separate models regressing student class evaluations on grades. The first model pools all observations across classes and students. The second includes class fixed effects, effectively limiting the comparison to students enrolled within the same classes. The third includes both class and student fixed effects. The final model includes class and student fixed effects and the student-reported reason for taking each course.

Moreover, the two-way fixed effects results are nearly identical to the pooled estimates—regardless of whether controls for reason of enrollment are included. That accounting for both fixed course characteristics and student attributes does not meaningfully attenuate the observed relationship between grades and course evaluations means that the correlation cannot be explained by either grading leniency or quality of instruction—which are both absorbed in the course effects—nor by intrinsic differences among students. In other words, students who receive better grades tend to report a subjectively more positive experience in their courses, regardless of the underlying reasons that led to their high performance.

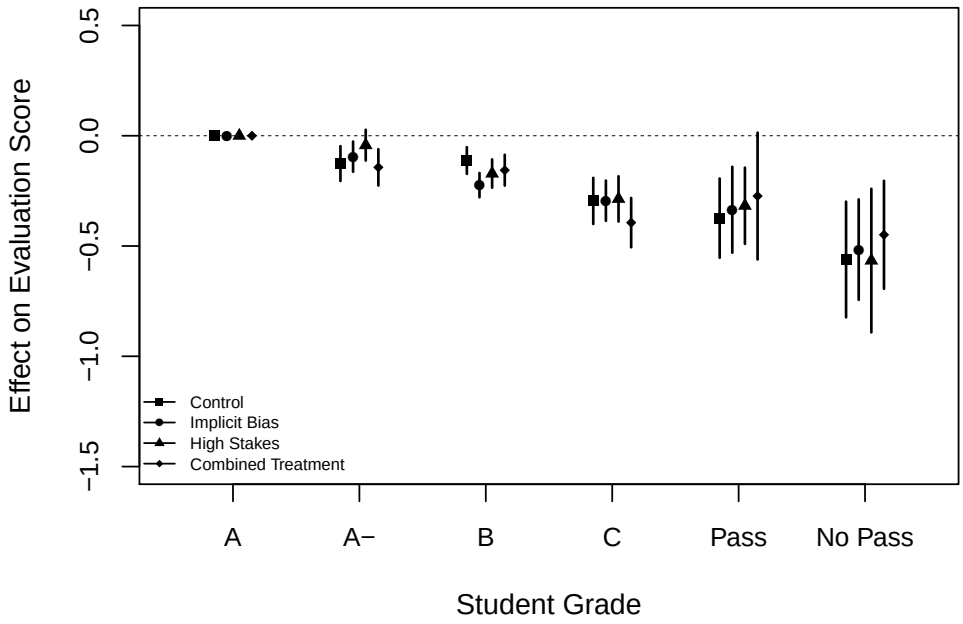
In addition to being highly significant, these differences are substantively large. Average scores on the overall satisfaction survey question typically have a class-level standard deviation between 0.9 and 1. The difference in average evaluation scores given by students who receive an A and an A-, representing a modest grade difference, ranges from 0.1 to 0.2, depending on the specification, corresponding to between 10 percent and 20 percent of a standard deviation. Thus, modest changes in average student grades could meaningfully impact an instructor's relative standing as measured on the student surveys.[15]

In Figure 2, we compare average ratings by the version of the introductory text students saw at the top of their survey. The point estimates and associated confidence intervals are estimated by interacting each student's grade with treatment assignment indicators. To maximize statistical power, the figure presents estimates from our class fixed effects specification. Similar results for the two-way fixed effects specification are provided in the supplemental appendix. Several of the individual coefficients are significantly different from one another. For example, students who receive an A- rate their course significantly more positively when provided the high-stakes introductory language compared to students completing the standard survey. By contrast, students who receive a B rate their course significantly more negative when provided the implicit bias text. However, none of the treatments have a consistent effect on evaluations—either in terms of statistical significance or even in terms of sign—across different grade categories. Indeed, the most striking aspect of the figure is how similar the mapping between grades and evaluations looks across different versions of the survey instrument.

One limitation with the experiment is that about one-third of students completed the evaluation survey using a mobile application. Due to technical issues, the mobile app did not have the capability to display the introductory texts. Thus, the results in Figure 2

[15] In the supplemental appendix, we disaggregate the results separately by student gender and race, finding no significant subgroup differences in the effects of student grades.

Figure 2: Impact of survey wording on relationship between end-of-semester grades and students' evaluations of instruction



Note: The figure plots the point estimates and associated 95 percent confidence intervals from a model that regresses student class evaluations on grades and includes class fixed effects. Each grade is interacted with the survey form seen by the student, producing the estimates presented in the figure.

represent intent-to-treat effects. In the supplemental appendix, we replicate the figure after dropping students who completed the mobile version of the survey, approximating treatment on the treated effects.[16] Although excluding mobile respondents reduces the precision of of the estimates, particularly for the lower grades, it does not substantively affect the main results.

In the supplemental appendix, we also examine the probability that students complete their course evaluations and find that higher achieving students are significantly more likely to do so. For example, compared to a student earning an A grade, a student earning an

[16]Treatment assignment did not significantly affect the mode of completion, so dropping mobile respondents does not introduce concerns about differential attrition or post-treatment bias.

A- is approximately 5 percent (4 percentage points) less likely to fill out the survey, and a student who fails the class is 70 percent (46 percentage points) less likely leave a course evaluation. However, this association appears to capture student-level characteristics—such as work ethic or time-management skills—that affect both academic achievement and the willingness to complete optional course evaluations. Adding student-level fixed effects, which nets out these student-level characteristics, large eliminates the relationship between grades and evaluation completion rates. One implication is that instructor grading practices is far more likely to affect *how* students evaluate their courses rather than *whether* students do so.

Conclusion

This study makes two contributions. First, we show that neither variation in teaching quality nor grading leniency appears to explain the observed correlation between student grades and teaching evaluations. Nor is this correlation due to student-level characteristics that may simultaneously affect achievement and students' subjective satisfaction in the classroom. Second, we find that changing survey wording to encourage students to prioritize relevant aspects of their course experience or emphasize the consequences of their evaluations does little to reduce the strength of this association. The central lesson of our findings is that using student evaluations for high-stakes personnel decisions encourages instructors to increase student grades.

The implications of these incentives depend greatly on how instructors respond. Ideally, we might hope that they would prioritize refining their craft to improve student academic outcomes—adopting instructional best-practices that increase learning. However, it's possible that instructors also engage in gaming behaviors designed that increase grades without enhancing learning. For example, some instructors already engage in innocuous tricks—

such as bringing sweets for the class on the day when students complete evaluations (Hessler et al. 2018)—to improve their ratings. Given the connection between student ratings and their grades, instructors might engage in more concerning strategems, such as lowering grading standards or teaching to the test, which can boost grades in the short term but potentially undermine longer-term student success.

To our knowledge, there is little rigorous causal research that documents how course instructors respond to the grading incentives created by student teaching evaluations. What little evidence exists, however, provides some grounds for concern. That higher teaching evaluations seem to lead to *worse* long-term outcomes for students, as documented in some contexts (Carrell and West 2010, Braga, Paccagnella and Pellizzari 2014), suggests that pressure to improve student ratings may lead to pedagogically unsound practices.

Many proponents of student surveys argue that including "student voice" in the teaching evaluation process may be desirable for other reasons, independent of its effects on learning. For example, doing so may increase students' sense of agency and perceived efficacy. This view has considerable merit, but our findings suggest that it is nevertheless important to examine the extent to which unintended consequences of the resulting incentives created for instructors offset some of these benefits. Medicine, another field where satisfaction surveys are widely used and are increasingly tied to reimbursement rates, provide a cautionary example of how well-meaning efforts to increase feedback can backfire. Some research has suggested that pressure to increase patient satisfaction as measured on surveys has contributed to over-prescribing of antibiotics (Martinez et al. 2018) and pain medication (Hessler et al. 2018), potentially contributing to antibiotic resistance and opioid addiction.

We end with two important notes on interpretation. First, our results should not be read to imply that student grades are the only—or even most important—predictor of student course evaluations. Overall, grades explain less than 5 percent of the variation

in student evaluation scores in our data and even less of the variation observed within course sections.[17] It is possible that students also incorporate other considerations more closely tied to instructional quality into their evaluations as well. However, evidence that instructor ratings at Ohio State do not predict student success in subsequent courses after controlling for grades (Weinberg, Hashimoto and Fleisher 2009) is inconsistent with this account. At minimum, our results suggest that supporters of student evaluations should face the burden of proof to show that evaluations measure and incentivize desirable teaching practices that promote students' long-term success. The correlation between student grades and evaluation scores—the primary evidence historically given in favor of this view—does not itself support this interpretation.

Second, it is unclear the extent to which our results apply beyond the post-secondary context and are relevant for the use of student surveys in K-12 education. Perhaps it is possible that younger students are better able to identify effective teaching practices and put greater weight on these considerations (relative to grades) in their responses when compared to college students. The only existing study that specifically examines this issue in the K-12 context is discouraging, however. Bacher-Hicks et al. (2019) find that student ratings of teachers fail to predict teacher contributions to student test score gains, as measured by value added models—and indeed show that the relationship is, if anything, negative. This suggests that our findings may have broader relevance in policy conversations about the role that student surveys and evaluations should play in educational accountability systems.

[17] Of course, individual-level survey data of the kind we study here are notoriously noisy. In many applications, even highly saturated models explain less than 20 percent of the variation recorded in survey responses. Aggregating up to the course level eliminates much of the noise, since the idiosyncratic factors largely cancel out, which probably helps explain why analyses using class-level measures typically find that course grades account for a larger share of the variation in class-average evaluations.

References

Abrami, Philip C., Wenda J. Dickens, Raymond P. Perry and Les Leventhal. 1980. "Do Teacher Standards for Assigning Grades Affect Student Evaluations of Instruction?" *Journal of Education Psychology* 72(1):107–118.

Bacher-Hicks, Andrew, Mark J. Chin, Thomas J. Kane and Douglas O. Staiger. 2019. "An Experimental Evaluation of Three Teacher Quality Measures: Value-Added, Classroom Observation, and Student Surveys." *Economics of Education Review* 32:101919.

Braga, Michela, Marco Paccagnella and Michelle Pellizzari. 2014. "Evaluating Students' Evaluations of Professors." *Economics of Education Review* 41:71–88.

Carrell, Scott E. and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy* 118(3):409–432.

Centra, John A. 2003. "Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work?" *Research in Higher Education* 44(5):495–518.

Chávez, Kerry and Kristina M.W. Mitchell. 2020. "Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity." *PS: Political Science Politics* 52(3):270–274.

Clayson, Dennis E. 2009. "Student Evaluations of Teaching: Are They Related to What Students Learn?: A Meta-Analysis and Review of the Literature." *Journal of Marketing Education* 31(1):16–30.

Cohen, Peter A. 1981. "Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies." *Review of Educational Research* 51(3):281–309.

Denning, Jeffrey T., Eric R. Eide, Kevin Mumford, Richard W. Patterson and Merrill Warnick. 2021. Why Have College Completion Rates Increased? An Analysis of Rising Grades. Working Paper 28710 National Bureau of Economic Research.
URL: *https://www.nber.org/papers/w28710*

Eiszler, Charles F. 2002. "College Students' Evaluations of Teaching and Grade Inflation." *Research in Higher Education* 43(4):483–501.

Feldman, Kenneth A. 1989. "The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies." *Research in Higher Education* 30:583–645.

Genetin, Brandon, Joyce Chen, Vladimir Kogan and Alan Kalish. 2021. "Mitigating implicit bias in student evaluations: A randomized intervention." *Applied Economic Perspectives and Policies* .

Greenwald, Anthony G. and Gerald M. Gillmore. 1997. "Grading Leniency Is a Removable Contaminant of Student Ratings." *American Psychologist* 52(11):1209–1217.

Hessler, Michael, Daniel M. Pöpping, Hanna Hollstein, Hendrik Ohlenburg, Philip H. Arnemann, Christina Massoth, Laura M. Seidel, Alexander Zarbock and Manuel Wenk. 2018. "Availability of cookies during an academic course session affects evaluation of teaching." *Medical Education* 52(10):1064–1072.

Johnson, Valen E. 2003. *Grade Inflation: A Crisis in College Education.* New York: Springer.

Kreitzer, Rebecca J. and Jennie Sweet-Cushman. 2021. "Evaluating Student Evaluations of Teaching: A Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform." *Journal of Academic Ethics* .
URL: *https://doi.org/10.1007/s10805-021-09400-w*

Linse, Angela R. 2017. "Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees." *Studies in Educational Evaluation* 54:94–106.

Marsh, Herbert W. and Lawrence A. Roche. 2000. "Effects of Grading Leniency and Low Workload on Students' Evaluations of Teaching: Popular Myth, Bias, Validity, or Innocent Bystanders?" *Journal of Education Psychology* 92(1):202–228.

Martinez, Kathryn A., Mark Rood, Nikhyl Jhangiani, Lei Kou, Adrienne Boissy and Michael B. Rothberg. 2018. "Association Between Antibiotic Prescribing for Respiratory Tract Infections and Patient Satisfaction in Direct-to-Consumer Telemedicine." *JAMA Internal Medicine* 178(11):1558–1560.

Ohio State University. 1998. "Analysis of the Student Evaluation of Instruction (SEI) Instrument at The Ohio State University, December 1998.".

Peterson, David A. M., Lori A. Biderman, David Andersen, Tessa M. Ditonto and Kevin Roe. 2019. "Mitigating gender bias in student evaluations of teaching." *PLOS One* 14(1):e0216241.

Schneider, Jack, James Noonan, Rachel S. White, Douglas Gagnon and Ashley Carey. 2021. "Adding 'Student Voice' to the Mix: Perception Surveys and State Accountability Systems." *AERA Open* 7(1):1–18.

Stumpf, Stephen A. and Richard D. Freedman. 1979. "Expected Grade Covariation With Student Ratings of Instruction: Individual Versus Class Effects." *Journal of Education Psychology* 71(3):293–302.

Uttle, Bob, Carmela A. White and Daniela Wong Gonzalez. 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.

Wallace, Tanner LeBaron, Benjamin Kelcey and Erik Ruzek. 2016. "What Can Student Perception Surveys Tell Us About Teaching? Empirically Testing the Underlying Structure of the Tripod Student Perception Survey." *American Educational Research Journal* 43(6):1834–1868.

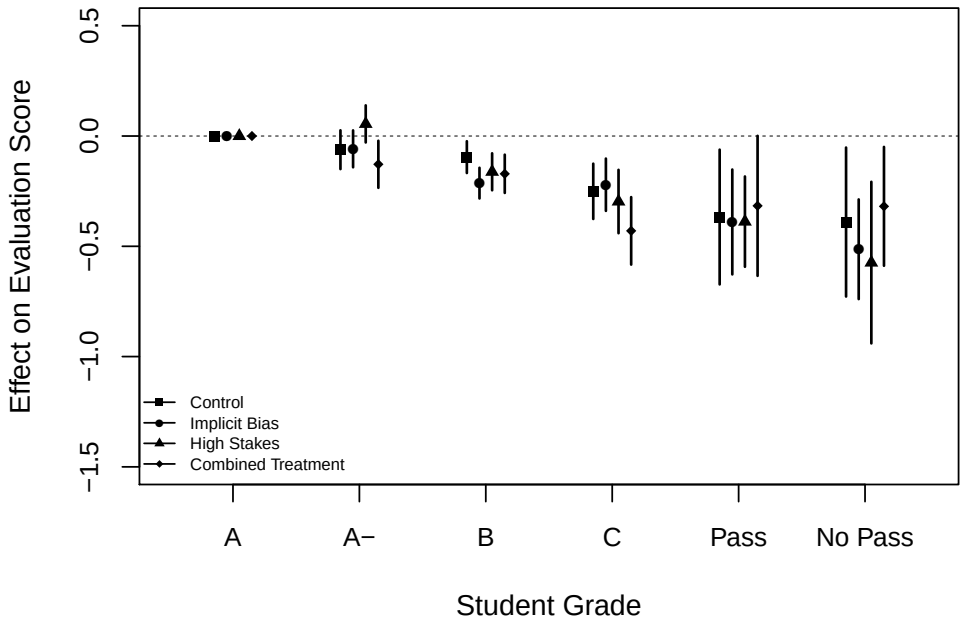
Weinberg, Bruce A., Masanori Hashimoto and Belton M. Fleisher. 2009. "Evaluating Teaching in Higher Education." *Journal of Economic Education* 40(3):227–261.

Zhao, Zing and Dorinda J. Gallant. 2012. "Student Evaluation of Instruction in Higher Education: Exploring Issues of Validity and Reliability." *Assessment and Evaluation in Higher Education* 37(2):227–235.

Online Appendix for Students' Grade Satisfaction Influences Evaluations of Teaching: Evidence from Individual-level Data and an Experimental Intervention

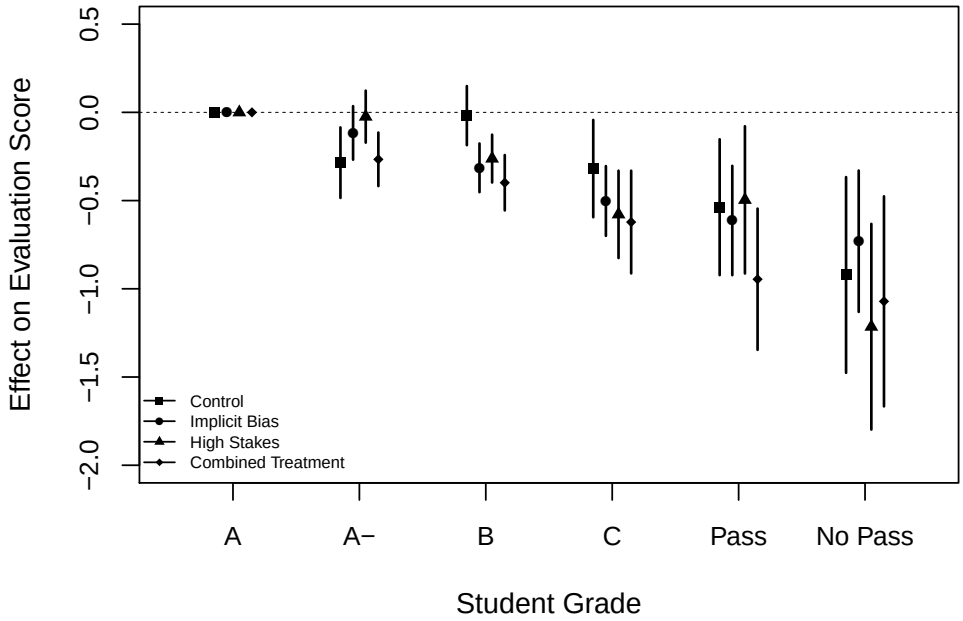
Supplemental Online Appendix

Figure A1: Impact of survey wording on relationship between end-of-semester grades and students' evaluations of instruction, using two-way fixed effect specification



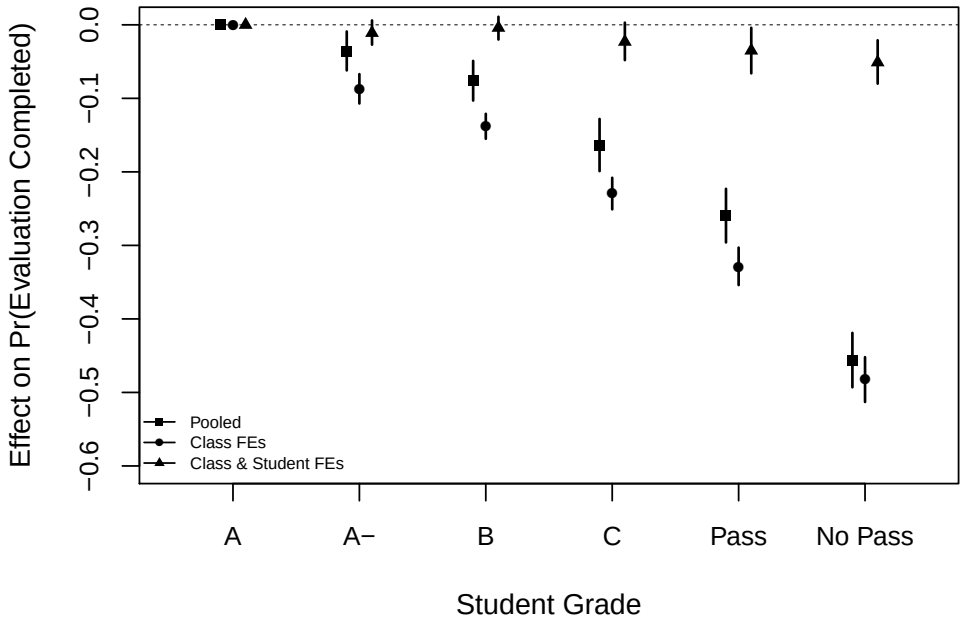
Note: The figure plots the point estimates and associated 95 percent confidence intervals from a model that regresses student class evaluations on grades and includes class and student fixed effects. Each grade is interacted with the survey form seen by the student, producing the estimates presented in the figure.

Figure A2: Impact of survey wording on relationship between end-of-semester grades and students' evaluations of instruction, excluding mobile app responses



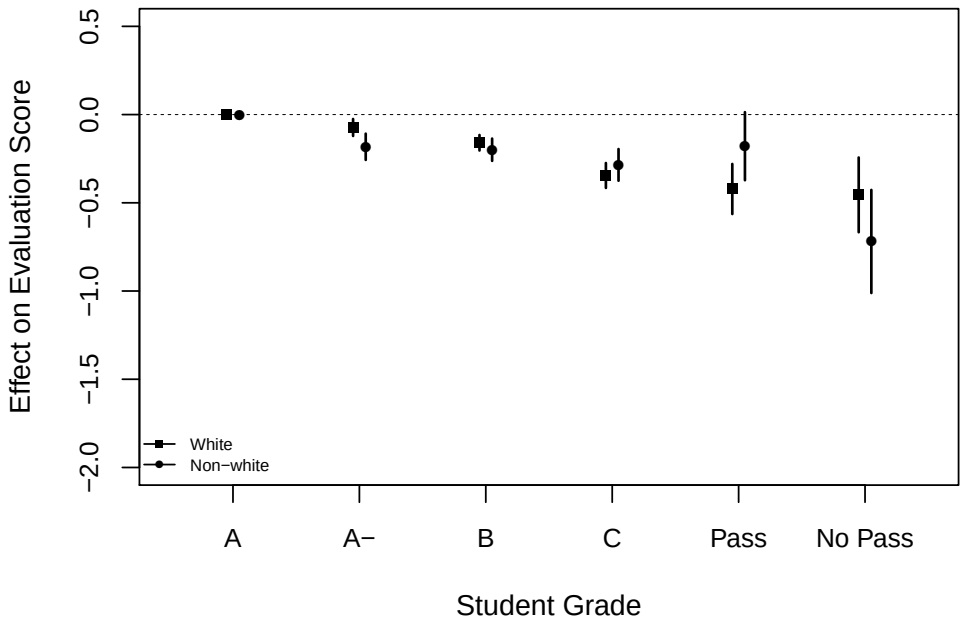
Note: The figure plots the point estimates and associated 95 percent confidence intervals from a model that regresses student class evaluations on grades and includes class fixed effects after excluding students who completed the course evaluation survey using the mobile application. Each grade is interacted with the survey form seen by the student, producing the estimates presented in the figure.

Figure A3: Relationship between end-of-semester grades and the probability that students completed course evaluation



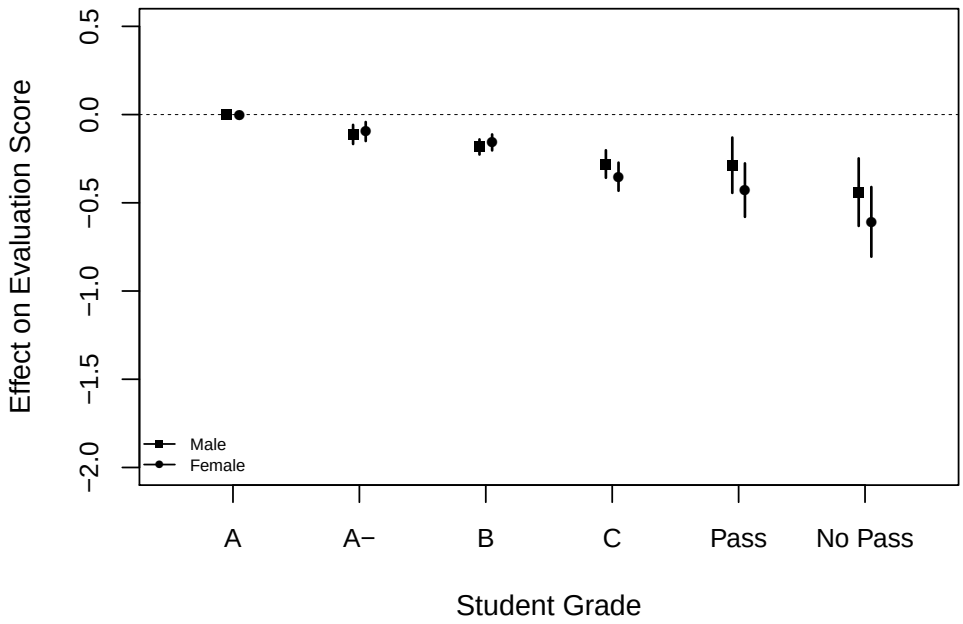
Note: The figure plots the point estimates and associated 95 percent confidence intervals from three separate linear probability models regressing an indicator variable for whether a student student completed an evaluation for a class on the grade in that class. The first model pools all observations across classes and students. The second includes class fixed effects, effectively limiting the comparison to students enrolled within the same classes. The third includes both class and student fixed effects.

Figure A4: Relationship between end-of-semester grades and students' evaluations of instruction, by student race



Note: The figure plots the point estimates and associated 95 percent confidence intervals from a model that regresses student class evaluations on grades and includes class fixed effects. Sample excludes students with no reported race.

Figure A5: Relationship between end-of-semester grades and students' evaluations of instruction, by student gender



Note: The figure plots the point estimates and associated 95 percent confidence intervals from a model that regresses student class evaluations on grades and includes class fixed effects. Sample excludes students who identify as neither male nor female.