The Quality of Online Instruction and Returns to Instructor Experience*

Xi Zhang †
Ann Atwater‡

September 18, 2024

Abstract

We examine student satisfaction and performance in online versus in-person sections at a large research university in the United States, exploring whether observed gaps are inherent to online instruction or can be mitigated with increased teaching experience. Using administrative data from over 40,000 course sections taught over eight years, we find that students evaluate online courses as worse than in-person courses, despite minimal differences in performance. This gap persists even when restricting the sample to courses taught using both modalities by the same professor in the same semester, and after matching on observable student characteristics. Lower evaluations are primarily driven by student perceptions of instructor availability, concern for students, and the ability to stimulate interest in the course. Although teaching experience improves evaluations in online sections, the gap between modes remains, suggesting fundamental challenges in online instruction beyond technological familiarity.

JEL Classification: I23, J24, J16
Keywords: Online Instruction, Teaching Evaluation, Return to Experience

*We would like to thank Perihan Saygin for her guidance throughout this project. We also wish to express our appreciation to Scott Kostyshak and Robert Ainsworth for their help in obtaining the student-level data. We also thank Richard Romano, Anne Boring, Min Fang, Gunnar Heins, Cecilia Peluffo, Mark Rush, Sarah Ayllon, WATE-Florida and SEA 2023 participants for their thoughtful comments. All errors are our own. This study was exempted by the University of Florida under IRB17843.

†Department of Economics, Finance, and Quantitative Analysis, Kennesaw State University.
‡Department of Economics, University of Florida.

1 Introduction

"Online instruction may be more economical to deliver than live instruction, but there is no free lunch." - David Figlio (Leopald, 2010)

Online instruction is presently a core component of higher education in the United States. The COVID-19 pandemic accelerated the adoption of online courses and prompted universities to consider the integration of online instruction for their residential students, who traditionally relied on in-person instruction. Therefore, evaluating the quality of online instruction compared to in-person instruction is crucial. While past research has been conducted on the quality of online education, it dominantly focused on student grades as the sole outcome of interest. However, educational quality transcends grades alone and appears in ephemeral course characteristics such as instructor engagement and ability to motivate students. The existing literature has additionally overlooked whether disparities observed between traditional and online modalities are enduring or temporary, influenced by factors like instructor unfamiliarity with the technology.

Our paper aims to address both concerns. We resolve the first by presenting a unified framework that assesses both student satisfaction and performance, with student satisfaction measured through end-of-semester course evaluations conducted by the university. Student satisfaction not only offers complementary insights beyond student performance but may also significantly influence the future demand for online course offerings. We resolve the second by examining how student satisfaction and performance evolve with instructors' experience with online instruction, as our data allow us to follow all instructors in the studied university along their teaching path spanning seven years.

We capitalize on a comprehensive administrative dataset combining course enrollment, teaching evaluations, and student tracking data from a large research public university, cov-

ering over 44,000 sections taught by 3,600 instructors between 2012 and 2019. This dataset enables us to compare student satisfaction and performance across online and in-person sections. Moreover, its richness permits us to narrow our comparison to specific subsets of large introductory courses taught by the same instructor within a semester, with some sections conducted in-person and others online. Furthermore, this dataset enables us to address concerns about student selection by matching observed student characteristics, especially on their previous academic performance and credits enrolled in the current semester. The student tracking data additionally empowers us to assess spillover effects of enrolling in online sections on students' performance in concurrent courses within the same semester, as well as on all courses in subsequent semesters. Lastly, we employ course enrollment data to assess instructors' experience in both online and in-person teaching before instructing a specific section. Following this, we examine the returns to online teaching experience by analyzing variations in student satisfaction and performance across semesters when the same instructor teaches the same course.

Our findings reveal that online sections consistently receive worse evaluations compared to in-person sections, even when taught by the same instructor under the same course code in the same semester. This evaluation gap persists even after employing matching techniques to account for bias arising from student self-selection into online or in-person courses. Additionally, we find nearly zero correlation between response rate and evaluation rating across online sections, and the significant gaps in student satisfaction are driven by sections with response rates over 80%. This finding lessens concerns that differential completion of surveys across modalities could drive our results. Further exploration of factors contributing to worse online evaluations reveals that the evaluation gap is primarily influenced by student views on instructor availability to assist students, concern for students, and ability to stimulate

interest in their course.

In terms of student performance, we found insignificant differences in grades between the current section and the subsequent semester across different instruction modalities, but a marginally significant decline in student performance in other concurrent-semester courses. If this marginal significance holds true, it might imply a negative spillover effect stemming from online instruction. One plausible interpretation of this finding is that students may allocate extra time to self-directed study to offset the perceived shortcomings of online instruction, enabling them to maintain comparable grades within online sections but not in concurrent courses.

Finally, our findings reveal positive and significant returns to instructors' online teaching experience in online sections, as evidenced by evaluations, albeit without corresponding improvements in student performance. Importantly, when incorporating data from in-person sections to compare returns to experience across modalities, the significant return in evaluation ratings diminishes. Consequently, the evaluation gap between online and in-person sections remains largely unmitigated through the acquisition of more online teaching experience alone. By synthesizing all evidence, we conclude that these observed gaps stem from intrinsic limitations of online instruction rather than from switching costs associated with transitioning from in-person to online instruction.

This paper draws from and contributes to several strands of literature. A few existing studies on student satisfaction with online courses have predominately relied on survey research, where a limited number of students (ranging from hundreds to a few thousand) were surveyed on their general impressions of online courses (Robertson et al., 2005; Burns, 2013; Cole et al., 2014; Platt et al., 2014). In contrast, our research takes a different approach by utilizing teaching evaluations conducted at the conclusion of each course, enabling us

to obtain timely and course-specific feedback based on actual student experiences. Some literature, such as Liu (2005) and Campbell and Sheridan (2011), has employed student evaluations to compare different modes of instruction. However, these studies were constrained by a limited sampling of courses, with these papers only observing two and five sections, respectively. Conversely, we have compiled a comprehensive dataset consisting of all (44,000+) evaluated course sections across various disciplines within the studied university. This breadth enables us to produce precise estimates and extrapolate to a broader population. Crucially, unlike those studies, we identify significant variances in student evaluations across different instructional modes.

Additionally, our study contributes further evidence to the existing literature on student performance in online instruction, given the mixed results from previous research. The meta-analysis commissioned by the United States Department of Education, Means et al. (2009), summarized 51 independent papers between 1996 and 2008 and suggested a positive correlation of online education on student outcomes.[1] Figlio et al. (2013) used a randomized controlled trial approach and found no significant impact of online instruction on student performance in an introductory economics course at a large public university, but Alpert et al. (2016) and Kofoed et al. (2021) also applied randomized controlled trials and found significantly worse student performance in online sections. Bettinger et al. (2017) uses an instrumental variable approach with observational data from a large, for-profit college chain and finds negative impacts from taking courses online which persist into later semesters. We interpret the disparity in results from different settings: evidence from selective universities, such as Figlio et al. (2013) yield negative but insignificant effect, while Bettinger and Long

[1]This positive correlation may be attributed to the fact that the "online" education in these studies encompasses not only online instruction but also cases where various online elements, such as e-learning, are integrated into traditional classroom settings. Additionally, many of the summarized pieces of evidence are suggestive rather than causal.

(2005), Xu and Jaggars (2011), Xu and Jaggars (2014), and Alpert et al. (2016) which study less selective universities, such as for-profit university chains and community college tend to yield significantly negative results. Our observational study from a selective university aligns with the experimental findings from Figlio et al. (2013)), which reinforces the internal validity of our matching methodology. Even though observational studies may not be comparable in terms of internal validity compared to experimental evidence, our study exhibited stronger external validity than would be possible in an experimental setting by examining across disciplines and across instructors. Moreover, the interpretation of the null effect of online sections on student performance in selective universities can be ambiguous: it could suggest either that students in selective universities are indifferent to instructional modalities, or it might suggest that student performance is not an adequate metric for capturing the difference, especially considering that students from selective universities often prioritize grades. This serves as another important reason behind our exploration of an additional dimension beyond student performance: student satisfaction. Finally, even though student performance in the following semester has been studied (Bettinger and Long, 2005; Krieg and Henson, 2016), we uncover new evidence measuring student performance in other concurrent courses to shed light on a previously underexplored negative spillover effect associated with online course enrollment.

Beyond the literature on online instruction, our paper contributes to the literature on returns to teacher experience by adding evidence at the postsecondary level and across instruction modes. Returns to instructor experience at the K-12 level has been covered extensively in the economics literature (Harris and Sass, 2011; Wiswall, 2013; Ost, 2014; Cook and Mansfield, 2016), but the understanding of returns to instructor experience at the postsecondary level remains limited. In contrast to the standardized test outcomes commonly

used in K-12 education research, our paper utilizes standardized teaching evaluations, which are more prevalent and relevant at the postsecondary level. While previous teacher value-added papers at the postsecondary level (Hoffmann and Oreopoulos, 2009; Carrell and West, 2010) used teaching evaluations as independent variables and explored their correlation with estimates of instructor value-added to discuss the validity of teaching evaluations. Our paper takes a different approach using teaching evaluations as the dependent variable and analyze how online teaching experience affects the evaluation gap between modes of instruction. Notably, our study is the first paper, as far as we know, examining the returns to experience under online instruction. While Vlieger et al. (2018) does discuss instructor value-added separately for in-person and online courses, our paper explores the dynamic return to online teaching experience. To precisely measure returns to experience, we adopt multiple methods, including the two-stage model proposed by Papay and Kraft (2015) to address for potential collinearity concerns between the time trend and teaching experience.

The rest of our paper is organized as follows. Section 2 details our setting and data. Section 3 outlines our methodology and presents the obtained results. Specifically, Section 3.1 examines student satisfaction in the context of online instruction and explores the underlying factors contributing to the satisfaction gap. Section 3.2 examines student performance under online instruction, and Section 3.3 investigates whether the observed gap can be alleviated through more online teaching experience. We conclude in Section 4.

2 Setting and Data

We conducted our analysis by merging course evaluation, administrative enrollment records, and longitudinal student tracking data from a large research university in the United States.

Within this section, we provide a detailed explanation of each component.

2.1 Course Evaluation and Enrollment Data

The university offered roughly 300 online courses[2] each Fall and Spring semester which resulted in over 20,000 student enrollments for those semesters, as depicted in Panel (a) of Figure 1. Despite comprising only a tenth of course offerings, these enrollments were one-sixth of total student enrollments for those semesters. In our study, online sections primarily refer to asynchronous online sections, such as recorded lectures posted on Canvas.[3]

The university gathers feedback from students regarding courses and instructors in nearly all course sections at the end of each semester.[4] Participation in these evaluations is voluntary, and responses are kept anonymous to safeguard student privacy. Evaluations are shared with instructors only after course grades have been finalized. However, students might take their anticipated final grades into consideration when responding to the evaluation survey. Therefore, we will control for their final course grades when comparing evaluation differences between online and in-person sections in our analyses.

The course evaluation data consists of three pieces of information: course identifiers, response data, and numerical evaluation ratings. The course identifier data includes the term the course was taught, the college and department that offered the course, course number, section number, course name, and instructor name. The response data includes the number of responses and the response rate of the evaluation. Finally, the evaluation

[2]Online sections include all sections designated by the university to consist of more than 50% online components. Therefore we consider online sections to include both hybrid (50%-79% online) and fully online (80%-100% online) sections.

[3]This classification is based on our examination of course syllabi and interviews conducted with both instructors and students.

[4]Courses that involve individual instruction (e.g., thesis or dissertation supervision), those conducted outside traditional classroom or laboratory settings, or courses with enrollments of 10 or fewer students are excluded from the student evaluation process.

comprises eight numeric questions, with seven focusing on specific aspects of the course and the eighth serving as the overall assessment score, which we utilize as the measure of students' overall satisfaction in the section for all subsequent analyses. Each of the eight questions is evaluated using a 5-point Likert scale, ranging from 1 for "Poor" to 5 for "Excellent". Table 1 summarizes the section-average responses to each question for both online and in-person sections. It indicates that online sections, on average, exhibit lower evaluation response rates and receive lower ratings on every evaluation question, except for the second question assessing the clarity of communication of ideas. In total, we obtained course evaluation data for 44,277 course sections of 3,214 undergraduate courses taught by 3,600 instructors from 108 departments between Fall 2012 and Spring 2019.[5] Of these sections, we observe 6,115 (13.8%) online sections and 38,162 in-person sections. A comprehensive comparison between online and in-person sections is provided in Table A1.

To examine the potential impact of instructors' teaching experience on narrowing the gap in student satisfaction and performance between online and in-person sections, we utilize administrative enrollment data. We quantify an instructor's online teaching experience by the number of semesters they have taught the same course online at the university.[6] Within our sample, the range of a professor's online teaching experience before instructing an online section varies from 0 to 19 semesters, with an average of 3.29 semesters. In terms of total teaching experience, encompassing both in-person and online teaching, the range spans from 0 to 31 semesters, with an average of 4.52 semesters.

[5]After Spring 2019, the university changed the format of teaching evaluations.

[6]Prior research on returns to instructor experience, such as Ost (2014) and Papay and Kraft (2015), shows that there is a positive return to experience but these returns generally diminish over time. In our study, if instructors had prior online teaching experience before joining the university, the significant gap we observed, along with the insignificant difference in instructors' returns to experience, may reflect a long-term equilibrium.

2.2 Student Tracking Data

We acquired longitudinal student tracking data between Fall 2012 and Spring 2016. This tracking data captures demographic information for each student (such as race, gender, age, and nationality) and their enrollment details on a semester basis (including list of courses taken by the student, grades earned in each course, term GPA, and graduation status).

This data first enables us to generate the distribution of student characteristics for each course section, allowing the use of the matching methodology later in this paper. As evaluation rating may be driven by students with extreme characteristics, we not only capture the average value within each section but also consider the distribution. As summarized in Table 2, we observe student information for 24,439 sections with 3,053 (12.5%[7]) sections taught online. Online sections, on average, have higher proportions of Black and White students, as well as more students who are US domestic students. Students in online sections tend to be older, are taking fewer concurrent courses, perform worse in the previous semester, and are more likely to have prior experience with online courses.

Furthermore, by utilizing student tracking data, we calculate average student grade points earned within each section as a metric for student performance. Specifically, we identify all students in the section and assign numerical values to their final letter grades based on the university's conversion scale: A (4), A- (3.67), B+ (3.33), B (3), B- (2.67), C+ (2.33), C (2), C- (1.67), D+ (1.33), D (1), and D- (0.67). Any other failing or non-punitive grades[8] are converted to 0. Subsequently, we averaged numerical grades across all students to obtain

[7]As student-level data is available only up to Spring 2016, and as online sections grew over subsequent years, the percentage of online sections with student-level data is lower than the percentage with evaluation data.

[8]If a course is evaluated using only a pass/fail system, both grades are converted into 0. As a result, the difference we estimated does not consider the distinctions between these online and in-person sections. Other non-punitive grades include withdrawal, deferral, and incomplete.

average student grade points for each course section. Similarly, we also calculate average student grade points earned in all other sections taken concurrently and all courses taken in the following semester as metrics to evaluate potential spillover effects of online learning. When comparing student performance, we observe that all three measurements—average student grade points in the current section, in other concurrent courses, and in the following semester—are lower for online sections as shown in Table 2.

To better contextualize our regression analysis later on, we plot the cumulative distribution functions of evaluation ratings and average student grade points in the section, separating between online and in-person course sections in Figure 2. Both panels in the Figure demonstrate a similar pattern: the probability of receiving a lower value is higher for online sections compared to in-person sections. When examining the average difference between online and in-person sections, the unconditional evaluation gap is 13.4% (4.19 in online sections vs. 4.29 in in-person sections), and the unconditional average grade point gap is 11.2% (3.08 in online sections vs. 3.16 in in-person sections).

3 Methodology and Results

In this section, we examine the differences between online and in-person instruction in student satisfaction and performance. We then investigate the potential role of instructor experience in mediating these differences.

3.1 Student Satisfaction under Online Instruction

In this subsection, we analyze student satisfaction regarding online instruction. Specifically, we will disentangle potential confounding factors that could impact course evaluation ratings

other than instructional mode, including variation in instructors and courses, variation in student characteristics, and the potential for selection bias from voluntary completion of evaluations.

To account for concerns related to variation in instructors and courses across different modalities, our data affords us the ability to incorporate instructor, course, and year-semester fixed effects. The variance essential for our identification arises from large introductory courses,[9] wherein identical courses were taught by the same instructors with some sections online and others in-person during the same semester. To mitigate the influence of instructor or course selection for online instruction on our results, we gradually introduce each fixed effect. Our model estimation is as follows:

$$Y_{scit} = \alpha + \beta_1 Online_{scit} + X_{scit} + \epsilon_{scit}, \tag{1}$$

where Y_{scit} is the average overall assessment rating of section s under course code c taught by instructor i in year-semester t, $Online_{scit}$ is an indicator variable for whether section s under course code c taught by instructor i in year-semester t was an online section, and X_{scit} a vector of controls that varies between models. We use standard errors clustered at the instructor level to allow for arbitrary dependence of ϵ_{scit} across sections within instructors. The coefficient of interest is β_1, which captures the average effect of online instruction on the average rating, while holding X constant.

We report $\hat{\beta}_1$ estimated from various specifications in Table 3. In column (1), we observe an unconditional average difference of -0.097 points between all online and in-person sections. In columns (2) to (4), We introduce year-semester fixed effects, department fixed effects, and

[9] A detailed list of courses contributing to our primary identification is provided in Table A6.

total enrollment to account for variation by semesters, departments, and class size.[10] Across all specifications, we find that the gap in evaluations remains statistically significant. To account for potential differences between instructors who teach online versus in-person and the courses that are taught online versus in-person, we further introduce instructor fixed effects in column (5) of Table 3, both instructor and course fixed effects in column (6), and instructor-by-course fixed effects in column (7). The observed significant difference in overall assessment scores between online and in-person sections is robust to the inclusion of these fixed effects. Specifically, the overall rating for online sections remains consistently at least -0.11 points lower compared to in-person sections, representing approximately 15.3% of the standard deviation of the overall score across all sections. Moreover, we present a comprehensive summary of the evaluation scores by distinct instructor categories (those exclusively teaching in-person, exclusively online, or both) and course types (courses exclusively delivered in-person, exclusively online, or in both formats) in Table A3. Notably, the significantly lower evaluation scores in online sections prove to be robust across different instructor and course types.

3.1.1 Student Selection into Online Sections

Given that students can choose between online and in-person sections, the observed difference in evaluations may not solely signify lower perceived quality in online courses but could be influenced by the characteristics of students opting for online enrollment. To address this potential selection bias, we identify all students enrolled in each of the 24,439 course sections, summarize the observed student characteristics,[11] generate propensity scores for each course

[10] We additionally provide a comprehensive breakdown of the overall evaluation gap by year in Figure A2, by semester in Figure A3, by colleges in Table A4, and by hybrid and fully online sections in Table A5.

[11] Refer to Table 2 for detailed student characteristics.

to be taught online, and then match each online section with five in-person sections possessing similar propensity scores (N5 propensity score matching).[12] This allows for the creation of a counterfactual for each online section such that treatment, being taught online, is equivalent to random conditional on observed student characteristics. Thus, any observed differences in evaluation ratings can be attributed to the mode of instruction.

The propensity score, $p(x)$, is the probability of a section to be an online section ($T = 1$) given student characteristics ($X = x$): $p(x) = Pr(T = 1|X = x)$. Table 4 presents the regression outcomes derived from the logit model employed in constructing the propensity scores. Our results indicate students enrolled in online sections are more likely to be Black and White,[13] dispersed in age, female, with fewer than 12 credits enrolled in the current semester, with lower previous academic GPAs, and with prior online learning experience.

The distributions of propensity scores for online and in-person sections are summarized in Figure 3, illustrating the overlapping region of propensity scores primarily among those with scores of 0.4 or less between online and in-person courses. To visualize the impact of matching, we plot the percentage difference in each covariate after matching in the top panel of Figure 4, where the matched online and in-person sections display a reduced degree of difference across all student characteristics compared to their unmatched counterparts. However, some significant disparities between the two groups still exist. Specifically, the matched online sections have slightly higher percentages of students younger than 18 years old and with fewer cumulative credits earned. Additionally, they have a lower share of students with a GPA below 3 in the previous semester. To mitigate significant differences in student characteristics, we further trim sections with propensity scores higher than 0.4,

[12]Results using alternative matching methods are reported later.

[13]The reference group for student race comprises individuals whose race is either not identified or identified as multiple races. Asians and Hispanics do not exhibit a significantly different propensity to enroll online compared to the reference group.

as these sections can barely be matched with in-person counterparts based on Figure 3. The bottom panel of Figure 4 illustrates the percentage difference in each covariate after trimming. Following this additional step, all student characteristics become fully balanced across the matched online and in-person sections.[14]

As we apply nearest five neighbor matching, our identification strategy is as follows: for every matched online section i, we assume its counterfactual overall assessment rating when taught in-person $\hat{Y_i(0)}$, is equal to the weighted sum of the overall assessment rating of the five nearest in-person sections with similar student characteristics. That is:

$$\hat{Y_i(0)} = \sum_{j \in C(i)} w_{ij} Y_j,$$

where $C(i)$ is the set comprised of the five in-person sections nearest to online section i, and w_{ij} is the weight of each such that $\sum_{j \in C(i)} w_{ij} = 1$. We focus on the average treatment effect on the treated (ATT), where the estimator can be written as

$$\hat{ATT} = \frac{1}{N^T} \sum_i [Y_i - \hat{Y_i(0)}],$$

where N^T is the number of matched online sections, 3,027 in the residential sample and 2,512 in the trimmed residential sample.

We summarize the estimates before and after matching on student characteristics in Table 5. Columns (1) to (3) show the results from the residential sample. Column (1) presents the results from the unmatched sample with all fixed effects considered in the previous section, serving as a benchmark for the other estimates. The estimated evaluation gap is larger

[14]For a detailed summary of student characteristics after matching in the full and trimmed residential samples, refer to Table A7.

here compared to the full sample presented in Table 3 because this analysis uses data with student tracking coverage and is from an earlier time.[15] Columns (2) and (3) use the sample of matched residential sections. We observe that the estimated gap in column (2) is larger than in column (1), which may suggest that student self-selection is favorable to online course evaluations. In columns (3), we further incorporate average student grade points to capture the impact of student performance on student evaluation and find a slight decrease in the estimated coefficient. Similar changes are observed in Columns (4) to (6), which show the results for the trimmed residential sample where we have fully balanced student characteristics between the matched online and in-person sections. Overall, the evaluation gap between instruction modes enlarges after matching on observed student characteristics, and persists after factoring in the influence of students anticipating their final grades.

In Appendix C, we present additional results using alternative matching methods, including kernel matching utilizing all in-person sections and assigning greater weights to in-person sections with more similar characteristics, propensity score matching without replacement in both the full and trimmed subsamples[16] and Mahalanobis matching which matches online sections with the nearest in-person sections without replacement based on their scale-free Euclidean distance. The balance plot of student characteristics across these matching methods is aggregated in Figure A4, and regression results on the overall assessment rating for online sections are reported in Table A8. These findings reinforce our earlier conclusion that online sections consistently receive lower evaluations compared to in-person sections, even when

[15] Column (6) of Table 3 and Column (1) of Table 5 employ the same method. However, the former utilizes data spanning Fall 2012 to Spring 2019, while the latter is based on data covering Fall 2012 to Spring 2016, the period for which student performance data is available. This results in a magnitude increase from 15% SD to 18.9% SD.

[16] Here, we also try trimming sections to improve the balance of covariates between matched online and in-person sections. As we adopt propensity score matching without replacement, in contrast to with replacement in the previous analysis, we trim sections with propensity score higher than 0.3.

they share the same course code, are taught by the same instructor in the same semester, are matched on student characteristics, and exhibit similar average student grades. Using these alternative methods, the estimated evaluation gaps vary between 14.7% and 22.9% of a standard deviation.

3.1.2 Student Selection into Response

Within our dataset, online sections exhibit, on average, lower response rates compared to in-person sections (37% vs. 46%). This raises concerns about the possibility that the observed gap in evaluation ratings may stem from selection into completing evaluations.[17] In this section, we investigate whether the lower evaluation ratings in online sections are predominantly influenced by selection bias from response rates or genuinely reflect lower satisfaction.

Figure 5 illustrates the average overall assessment ratings by response rate across all 44,277 sections (including 6,115 online sections). Two noteworthy observations emerge: Firstly, in in-person sections, a positive correlation is observed between evaluation response rates and ratings. This indicates that the average evaluation scores might increase if all in-person sections achieve a 100% response rate. In contrast, for online sections, evaluation ratings display minimal correlation with response rates, as evidenced by a low Pearson correlation coefficient of 0.05. This suggests that a higher response rate for all online sections would be unlikely to significantly impact the average evaluation ratings. Consequently, the actual evaluation gap could be more substantial than our calculated estimate if both in-person and online sections were to achieve a higher response rate. Secondly, sections with response rates below 80% do not exhibit significant rating differences between instruction modalities. However, the significantly lower evaluation of online sections, as documented in

[17] One plausible scenario for this could be if online sections are less cohesive, potentially reducing the likelihood of high-evaluating students completing course evaluations, as suggested in Galyon et al. (2016).

our main results, is predominantly driven by sections with response rates exceeding 80%, in which case we consider that the potential selection into response is limited. As such, we conclude that the lower evaluation rating in online sections is unlikely to be driven by selection into responding to evaluations.

3.1.3 Contributing Factors to the Lower Satisfaction

So far, we have identified a significant evaluation gap between online and in-person instruction. We will now delve into the factors contributing to this gap by analyzing the evaluation ratings for seven specific questions (outlined in Table 1). Our approach involves a minor adjustment to Model 1, incorporating each of these seven questions (Q1-Q7) as control variables against the overall assessment rating (Q8). If the coefficient estimate for "online" becomes insignificant after controlling a specific, it suggests that the evaluation gap between online and in-person instruction diminishes holding the specific aspect constant. Thus, the added specific serves as a contributing factor to the overall evaluation gap. As presented in Table 6, when we individually include the rating of each specific question in columns (1) to (7), we observe that controlling for the evaluation ratings in questions 4, 5, or 6 eliminates the significant gap between online and in-person sections. This suggests that the lower overall assessment of online sections can be primarily attributed to the lower ratings addressing instructor availability to assist students (Q4), concern for students (Q5), and stimulation of interest (Q6).

In summary, our analysis in this section reveals a persistent evaluation gap between online and in-person sections. This gap remains significant even after narrowing the analysis to the subsets of courses taught in both modalities, instructed by the same teacher in the same semester, and matched for student characteristics, especially their academic history, current

course load, and previous online experience.

3.2 Student Performance under Online Instruction

In this section, we complement our exploration of student satisfaction by examining differences in academic performance between online and in-person sections. We achieve this by estimating the impact of online instruction on three outcomes: the average student grade in a section, average grade points earned in the next semester, and average grade points earned in other concurrent-semester courses. Notably, the grade a student earned within a section is under the control of the instructor. However, any given instructor has limited ability to influence other course grades. As such, the second and third of these outcomes of interests are subject to less concern over grading leniency to mitigate negative impacts of online instruction. Student performance in other concurrent courses is calculated as the average grade points across all other courses students were simultaneously taking with the section, and student performance in the subsequent semester is calculated as the average grade points across all courses students took in during the semester following the current section.

We employ the same procedure outlined in the prior analysis of student satisfaction. This entails controlling for instructor, course, and semester fixed effects, coupled with matching on student characteristics but with student grades as the dependent variable of our model.

Overall, we do not observe a consistently negative impact of taking online courses on student performance, as summarized in Table 7. We present the estimated coefficient of taking online sections across various matching methods. Specifically, for student performance in the current section (Panel A), Kernel matching in column (1) and propensity score matching without replacement in columns (2)-(3) show significant negative effects of online

instruction on student grades. By contrast, propensity score matching without replacement using the trimmed subsample in column (5) or Mahalanobis matching in column (6) results in insignificant estimates for the effect of online courses on student performance. Across columns, we notice that as the number of observations in the matched sample decreases, both the magnitude and significance of the gap diminish. This indicates that the difference in student characteristics, rather than instruction mode, is the main factor contributing to the observed gap.

Turning our focus to students' performance in other courses taken during the same semester, we observe a significantly negative average grade effect for students enrolled in online sections after including instructor and course fixed effects and matching on student characteristics. To better understand potential variations between courses taken concurrently with online and in-person sections, we offer a summary of the types of courses students opt to take simultaneously with online versus in-person sections in Table A9. We observe that courses taken concurrently with an online section are more likely to be online and upper-level courses. Consequently, we incorporate these course type differences into the estimation to account for potential variations in course selection, presenting the results in Panel B of Table 7, which reveals a negative yet marginally significant performance gap in other concurrent courses.

Finally, we examine student performance in the following semester in Panel C of Table 7, and we do not identify any significant future grade impacts from online sections. To address potential unobserved students who may have graduated or dropped out, we assess the probability of graduation and dropout in the following semester between online and in-person sections. The analysis also reveals no statistically significant difference in both cases, as indicated in columns (7) and (8) in Table A10.

In summary, our findings yield minimal evidence supporting a difference in student performance between online and in-person sections. This outcome aligns with Figlio et al. (2013), which found no significant difference in student performance under online instruction in a selective university, but differs from Bettinger et al. (2017), who observed significant negative spillovers into later semesters for students taking online courses in a for-profit university chain. To reconcile this discrepancy in the literature, we suggest that variations in settings play a crucial role. Students at more selective universities may have a clear grade goal they aim to achieve regardless of instructional quality, indicating that the lack of observed differences in grades between online and in-person sections does not necessarily mean these students are indifferent to instructional modality. Instead, it suggests that grades may not be a precise metric for capturing the differences in instructional quality. These findings emphasize the importance of exploring alternative measurements, such as student satisfaction, to better understand the quality of online instruction, especially in selective universities. Moreover, contrasting the insignificant performance gap and the significant evaluation gap helps reinforce our argument that online instruction drives the lower evaluation, as even though there may be unobserved differences between online and in-person sections, these differences do not seem to drive any difference in student performance.

3.3 Online Teaching Experience

So far, we have found significant evaluation gaps between online and in-person instruction. In this section, we investigate whether the gap results from the costs of integrating new technology into instruction or reflects structural issues with online teaching technology. To answer this question, we investigate the return to instructor teaching experience. If the earlier results were driven by a lack of instructor familiarity with teaching technology, we

would expect increased experience to reduce negative impacts. A description and summary statistics of teaching experience are provided in Section 2.1. In Section 3.3.1, we will first assess how student satisfaction and performance in online sections change with instructors' prior online teaching experience. Then, in Section 3.3.2, we will compare the returns to experience in both online and in-person sections to determine whether increased experience can mitigate the gap.

3.3.1 Returns to Online Teaching Experience

To explore returns to online teaching experience in online sections, we adopted an intuitive specification as follows:

$$Y_{scit} = \alpha_1 OnlineExp_{scit} + Inst_i + Crs_c + X_{scit} + \epsilon_{scit}, \tag{2}$$

where Y_{scit} is the outcome of interest, evaluation ratings or student grade points, in section s under course code c taught by instructor i in year-semester t, $OnlineExp_{scit}$ is the number of semesters that instructor i has taught online before teaching section s under course code c in year-semester t. For controls, we include instructor fixed effects ($Inst_i$) and course fixed effects (Crs_c). The vector X_{scit} includes section-level controls, such as student characteristics, enrollment, and—specifically for the analysis of evaluation ratings—average student grade points and response rates. Year-semester fixed effects are also included to account for any semester-to-semester variation common across instructors. With all fixed effects in place, the model relies on within-instructor, within-course variation in Y based on experience across semesters. Standard errors are clustered at the instructor level to account for arbitrary dependence of ϵ_{scit} across sections taught by instructor i. The parameter of interest, β_1, reflects the average effect of a one-semester increase in online teaching experience on the

outcome of interest.

We summarize the detailed regression results in Table 8. The overall assessment rating increases significantly with the online teaching experience, whereas the effect on student performance is indistinguishable from zero. Each additional semester of online teaching experience leads to a 0.027 increase in the overall assessment rating (equivalent to a 3% standard deviation), holding average grade points constant. We further verified that this increase in evaluation ratings is not driven by a change in response rate, as no significant change in response rates was observed with increased teaching experience. To address concerns that instructors with low evaluation scores might be removed from teaching the course, we regressed instructors' prior evaluation ratings against whether they taught the same course again across semesters. The results, as shown in Table A12, indicate that there is no statistically significant correlation. Thus, the significant increase in the evaluation scores is likely not the result of the selection of instructors into teaching them.

In summary, we show that online sections tend to receive higher evaluation scores when the instructor possesses more online teaching experience, suggesting a positive return on online experience in terms of student satisfaction. However, we do not find evidence supporting experience improving student performance. This discrepancy could be attributed to certain improvements that are reflected in evaluations but not in grades, or instructors' ability to curve grades to maintain consistency across semesters.

3.3.2 Returns to Experience Across Modes

We have established that online teaching experience has a positive effect on the evaluation of online sections. Now, we aim to explore whether increased teaching experience can help reduce the evaluation gap between online and in-person sections. Specifically, we examine

the effects of the instructor's overall teaching experience and their online-specific teaching experience.

We introduce the independent variables "TotalExp", which measures the instructor's total teaching experience for the course across all modalities, and "Online", an indicator for online sections, along with their interaction term "Online×TotalExp" Additionally, we control for instructor, course, and semester fixed effects to ensure the variation used to identify the return to experience is within-instructor and within-course across semesters. The estimate on the interaction term "Online×TotalExp" is the key coefficient of interest, capturing whether online sections yield a different return to overall teaching experience compared to in-person sections. If, and only if, this coefficient is significantly positive, we can conclude that increased general teaching experience helps mitigate the evaluation gap between in-person and online sections. Similarly, to examine whether the evaluation gap between modalities decreases with more online-specific teaching experience, we include "OnlineExp", which measures the instructor's online-specific teaching experience, "Online", an indicator for online sections, and "Online×OnlineExp", their interaction term. With instructor, course, and semester fixed effects in place, the sign and significance of this interaction term will reveal whether an additional semester of online-specific teaching experience influences the evaluation gap between in-person and online sections.

Unlike the previous results used only data from online sections, the regression results presented in Table 9 incorporate data from both in-person and online sections. In column (1), we examine the effect of the instructor's total teaching experience and observe a negative coefficient for "Online," indicating a consistent disparity between in-person and online sections for instructors without prior teaching experience. This finding aligns with our earlier results. Additionally, the positive coefficient for total teaching experience suggests that

prior teaching experience improves evaluation ratings in in-person sections. However, the insignificant interaction term indicates that the marginal return of an additional year of total teaching experience does not significantly differ between in-person and online sections. As a result, the evaluation gap between the two modalities remains largely unchanged. In column (2), we focus on the effect of the instructor's online teaching experience. Again, we find an insignificant estimate for the interaction term, suggesting that additional years of online-specific teaching experience also do not significantly close the evaluation gap between in-person and online sections.

To assess the robustness of our results to different estimation methods, we also investigate the returns to online teaching experience on student satisfaction and performance using a two-stage regression model proposed by Papay and Kraft (2015), as outlined in Appendix Appendix E. This two-stage model first estimates time fixed effects by capturing common shocks across instructors. In the second stage, these estimated time effects are removed from the dependent variable, and instructor fixed effects are reintroduced. Table A13 establishes the baseline for the returns to online teaching experience in online sections, while Table A14 examines whether the returns to teaching experience differ between online and in-person sections, potentially mitigating the gap we identified. Our findings using this two-stage model are largely consistent with those from the main model.

In summary, our findings suggest a positive return to instructors' online teaching experience in terms of student satisfaction. However, the return of online teaching experience to the evaluations of online sections does not significantly differ from that of in-person sections. Thus, acquiring more online teaching experience does not reduce the evaluation gap between online and in-person sections. These results suggest that our previous findings capture inherent disparities within evaluation between online and in-person instruction rather than

short term costs of technology adoption.

4 Conclusion and Implication

Online instruction has become increasingly important over the past few decades. During the COVID-19 pandemic, universities were compelled to transition most coursework online. Many institutions are now considering maintaining online delivery of a significant fraction of their courses using the institutional capacity developed during the pandemic. Online instruction offers benefits such as greater flexibility and accessibility than traditional in-person instruction (Cowen and Tabarrok, 2014; Deming et al., 2015). However, questions persist about whether online instruction can attain the same quality as traditional in-person instruction.

To resolve those questions, we present evidence based on a comprehensive dataset from a public research university in the United States. The dataset covers 44,277 course sections of 3,214 undergraduate courses, taught by 3,600 instructors across 108 departments, spanning Fall 2012 to Spring 2019. To assess student satisfaction with courses, we utilize end-of-semester evaluation ratings, while student performance is measured using final course grades. Our methodology uses variation within instructors and courses to eliminate potential biases arising from course design and instructor differences. Additionally, we employ propensity score matching to address concerns related to student self-selection into online and in-person sections. In contrast to experimental studies, our paper offers greater external validity by providing evidence across a broad range of courses, departments, instructors, and semesters.

Consistent with prior research on the quality of online instruction at selective universities, our study finds minimal differences in student performance between online and in-person

sections. However, a marginally significant gap appears in the average grade points of courses that students take concurrently with online versus in-person sections. This gap may suggest a negative spillover effect, where students allocate more time to online courses to compensate for potentially less effective instruction.

Beyond student performance, our paper provides new evidence on student satisfaction with online instruction. Online instruction is rated less favorably than traditional in-person instruction, even after controlling for course code, instructor, average grade points in the section, and applying propensity score matching to account for observed student characteristics. We find that the lower evaluations are primarily driven by student perceptions of instructor availability, concern for students, and the ability to stimulate interest in the course. Overall, our findings suggest that while student performance in selective universities does not differ significantly between instructional modalities, their evaluation scores reflect a clear preference for in-person instruction.

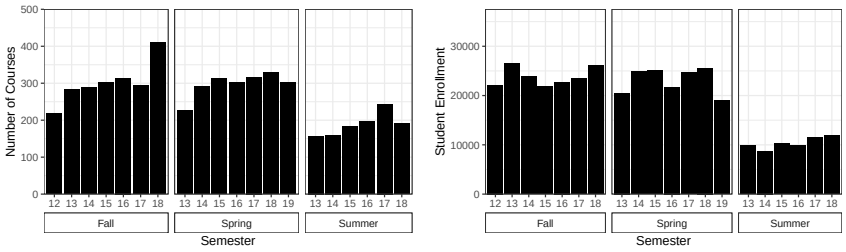
We also investigate whether instructor experience can help reduce the disparity between online and in-person sections. To assess this, we measure teaching experience by the number of semesters an instructor has taught a course. Our findings suggest that while online teaching experience does lead to improved evaluations, the evaluation gap between online and in-person sections is not reduced by overall teaching experience or online-specific experience. This indicates that the disparity may be intrinsic to the online format, extending beyond mere technological unfamiliarity.

As online instruction continues to develop as a core educational offering, our research has brought to light several concerning facts: online instruction may not be of similar quality as in-person instruction, and the quality gap may not diminish with more instructor experience. Previous literature has documented that the returns to experience are the largest

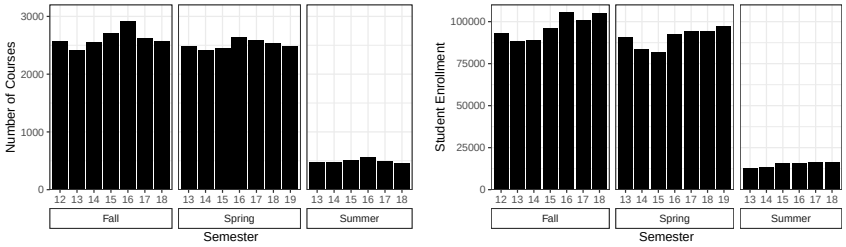
in the early stage of an instructor's career. As such, despite only capturing the beginning of a university's mass adoption of online teaching, our results finding no strong evidence to support mitigating returns to an online learning experience are significant. These findings highlight the complexity of evaluating online instruction and emphasize the need for further research into effective practices with the technology. Certain qualitative pedagogical discussions (Wiest, 2012; Rudd, 2014; Gold, 2019) may offer potential avenues of quantitative exploration. As institutions continue to adapt to the transition to online learning, it becomes increasingly essential to explore the associated challenges and potential solutions to ensure successful outcomes in this evolving educational landscape.

5 Figures and Tables

Figure 1: Course Offerings and Enrollments by Instruction Modes



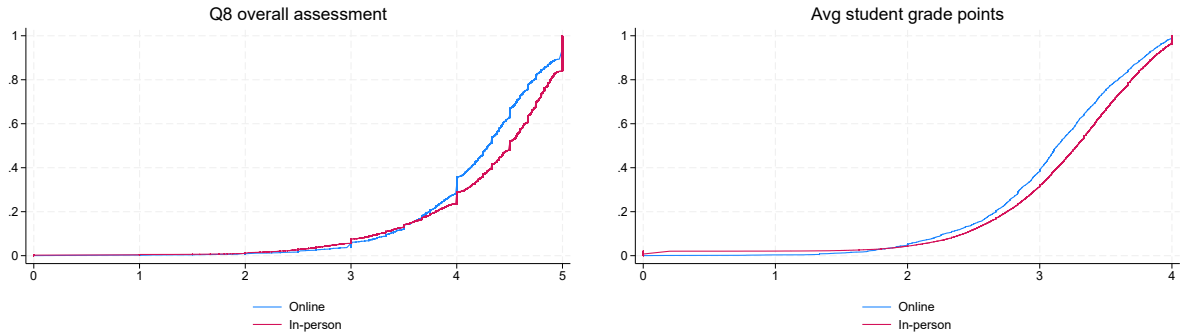
(a) Online course offerings and enrollments



(b) In-person course offerings and enrollments

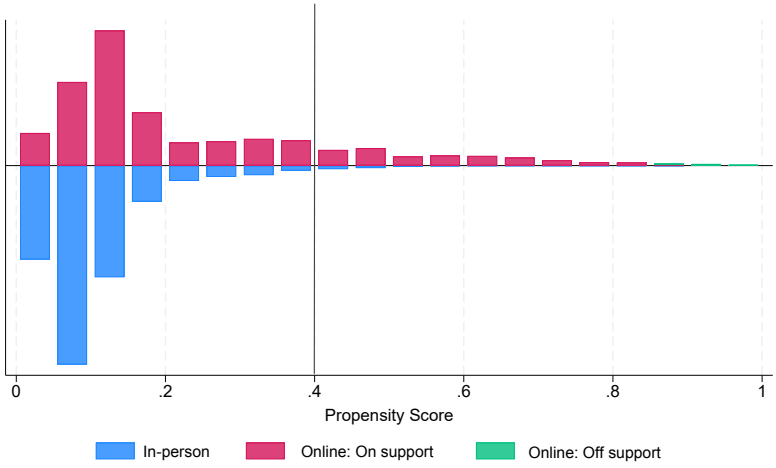
Notes: The figure plots course offerings and enrollments across years and semesters for online instruction in Panel (a) and in-person instruction in Panel (b). Course offerings refer to the number of courses offered each semester, and enrollments refer to the total number of enrolled students across all courses of each mode within a semester.

Figure 2: Empirical Cumulative Distribution Functions of Evaluation and Final Grade



Notes: The figure presents the empirical cumulative distribution functions of evaluation ratings and final grades between online and in-person sections. The evaluation ratings are measured using the overall assessment rating from Question 8 in the evaluation, while the final grades are measured using the average student grade points across all students who enrolled in the section.

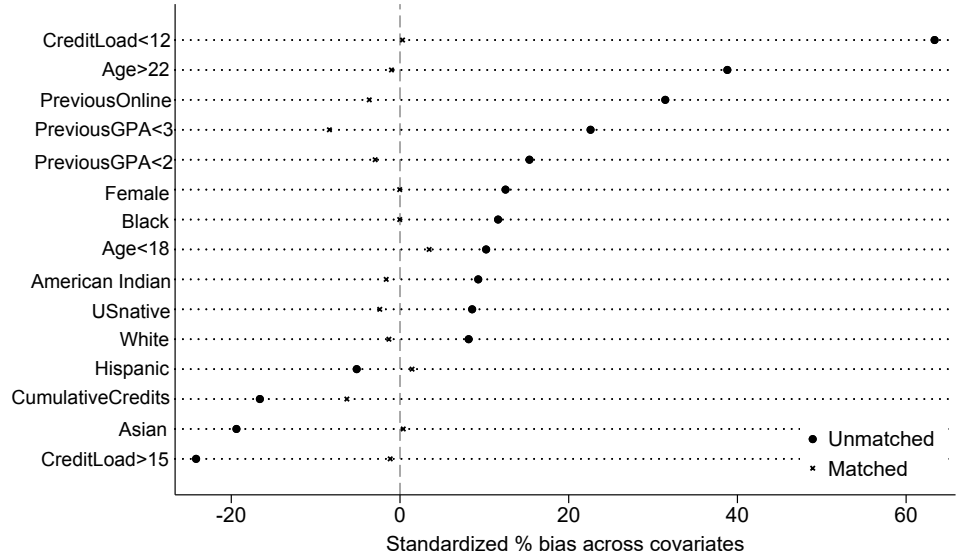
Figure 3: Propensity Score of a Section to be Taught Online



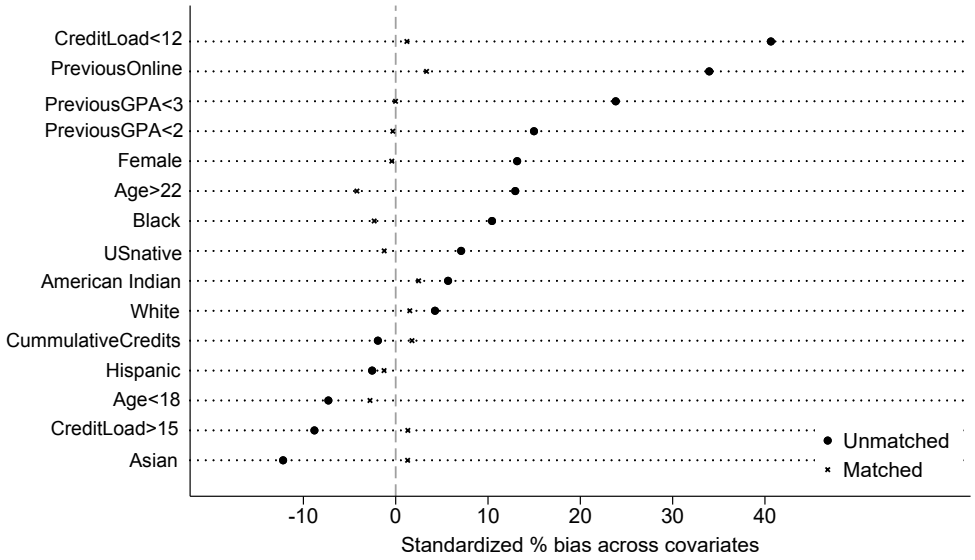
Notes: The histograms plot the estimated probability of online and in-person sections to be taught online based on student characteristics. Probabilities are estimated using the logit specification shown in Table 4.

Figure 4: The Balancing of Student Characteristics After Matching

(a) Residential Sample

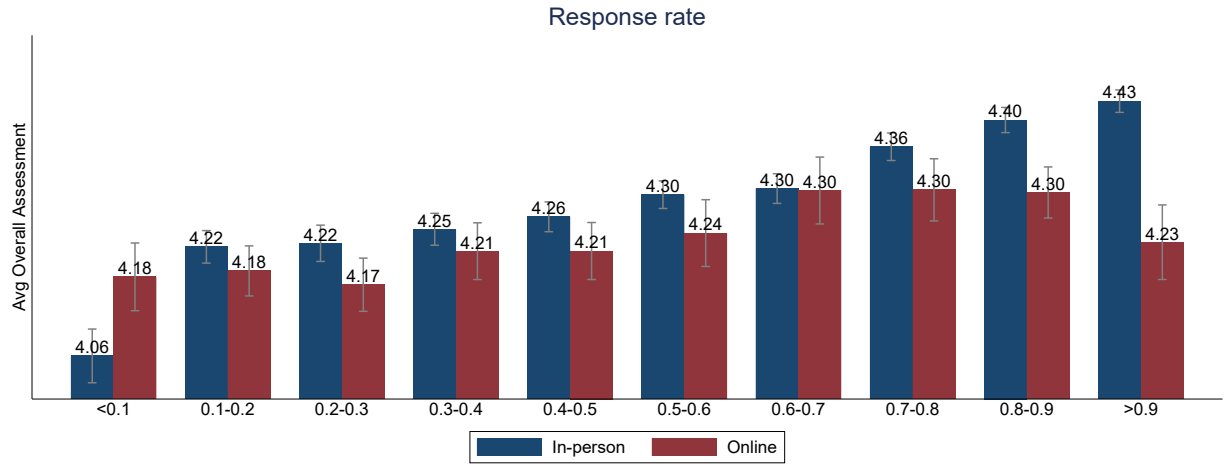


(b) Trimmed Residential Sample



Notes: The figure presents the balance of each covariate between matched and unmatched sections in the residential sample in Panel (a) and in the trimmed residential sample comprised of sections with propensity scores lower than 0.4 in Panel (b). Table A7 presents a detailed summary of student characteristics after matching each sample.

Figure 5: Overall Assessment Rating and Response Rate



Notes: The figure illustrates the relationship between the average overall student assessment rating and the response rate in the full sample, where we partition all sections into 10 categories based on their evaluation response rates. Each category represents a range of response rates as plotted. The bar in the figure represents the average overall assessment rating, while the line represents the 95% confidence interval.

Table 1: Summary Statistics of Course Evaluation (Section level)

	Online		In person		Difference
	Mean	sd	Mean	sd	b
Response rate	0.37	(0.21)	0.46	(0.24)	-0.09***
Q1 Clarity of objectives	4.26	(0.74)	4.31	(0.66)	-0.05***
Q2 Communication of ideas	4.20	(0.64)	4.21	(0.74)	-0.01
Q3 Expression of expectation	4.21	(0.76)	4.30	(0.67)	-0.09***
Q4 Availability to assist students	4.05	(0.83)	4.28	(0.72)	-0.23***
Q5 Concern for students	4.28	(0.63)	4.38	(0.70)	-0.10***
Q6 Stimulation of interest	4.13	(0.67)	4.27	(0.74)	-0.14***
Q7 Facilitation of learning	4.12	(0.66)	4.20	(0.76)	-0.08***
Q8 Overall assessment	4.19	(0.64)	4.29	(0.74)	-0.10***
Observations	6115		38162		44277

Notes: The table summarizes and compares the response rate and evaluation ratings for all eight numerical questions in online and in-person sections. * p<0.10, ** p<0.05, *** p<0.010

Table 2: Summary Statistics of Student Characteristics (Section level)

	Online		In person		Difference
	Mean	sd	Mean	sd	b
Race & Ethnicity					
% of Black or African American	0.08	(0.08)	0.07	(0.09)	0.01***
% of White	0.59	(0.15)	0.58	(0.15)	0.01***
% of Asian	0.09	(0.08)	0.10	(0.09)	-0.02***
% of Hispanic/Latino	0.19	(0.10)	0.19	(0.11)	-0.01**
% of American Indian or Alaska Native	0.01	(0.03)	0.01	(0.02)	0.00***
Other	0.04	(0.06)	0.04	(0.06)	-0.00*
Age					
% with age below 18	0.01	(0.05)	0.00	(0.01)	0.00***
% with age between 18 and 22	0.74	(0.29)	0.84	(0.20)	-0.10***
% with age above 22	0.26	(0.29)	0.16	(0.20)	0.10***
Average age	22.51	(3.79)	21.10	(1.87)	1.40***
Nationality					
% of US	0.94	(0.07)	0.93	(0.07)	0.01***
% of non-US	0.06	(0.07)	0.07	(0.07)	-0.01***
Current semester workload					
% with less than 12 credits	0.29	(0.35)	0.10	(0.22)	0.19***
% with 12 to 15 credits	0.58	(0.30)	0.74	(0.22)	-0.15***
% with more than 15 credits	0.13	(0.12)	0.16	(0.15)	-0.03***
Average credits	11.76	(3.05)	13.34	(1.75)	-1.59***
Previous academic performance					
Average gpa in previous semester	3.23	(0.31)	3.32	(0.31)	-0.09***
Average cumulative credits earned	43.66	(23.77)	47.91	(27.19)	-4.24***
% with previous gpa lower than 2	0.06	(0.07)	0.05	(0.07)	0.01***
% with previous gpa lower than 3	0.25	(0.15)	0.22	(0.15)	0.03***
Previous online experience					
% with previous online experience	0.69	(0.32)	0.58	(0.34)	0.10***
Student grades					
Avg student GPA	3.08	(0.57)	3.16	(0.69)	-0.08***
% of students with A or A-	0.50	(0.26)	0.53	(0.27)	-0.03***
% of students with B+, B or B-	0.28	(0.16)	0.29	(0.19)	-0.00
% of students with C+, C or C-	0.11	(0.11)	0.09	(0.10)	0.02***
% of students with D+, D or D-	0.02	(0.04)	0.02	(0.04)	0.01***
Avg grade points in other concurrent courses	3.02	(0.44)	3.16	(0.31)	-0.15***
Avg grade points in the following semester	3.19	(0.37)	3.28	(0.35)	-0.09***
Observations	3053		21386		24439

Notes: This table provides a summary of student characteristics categorized by instruction mode. The student tracking data covers only sections taught between Fall 2012 and Spring 2016. The racial category "Other" includes cases where the student's race is either unspecified or reported as multiple. "Avg student grade points", "Avg student grade points in other concurrent courses", and "Avg student grade points in the following semester" refer to the average grade points of all students in the current section, their average performance in other courses taken during the same semester, and their average performance across all courses taken in the following semester, respectively. Numerical grade points are converted using the university's conversion scale: A (4), A- (3.67), B+ (3.33), B (3), B- (2.67), C+ (2.33), C (2), C- (1.67), D+ (1.33), D (1), and D- (0.67). Other failing grades and non-punitive grades were converted to 0. * p<0.10, ** p<0.05, *** p<0.010

Table 3: Overall Assessment Rating - All

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Online sections	-0.097***	-0.12***	-0.20***	-0.19***	-0.16***	-0.11***	-0.11***
	(0.031)	(0.031)	(0.024)	(0.024)	(0.020)	(0.023)	(0.024)
Year × Semester FE		Yes	Yes	Yes	Yes	Yes	Yes
Department FE			Yes	Yes	Yes	Yes	Yes
Nb of Enrollment				Yes	Yes	Yes	Yes
Instructor FE					Yes	Yes	
Course FE						Yes	
Instructor × Course FE							Yes
Observations	44277	44277	44277	44277	44277	44277	44277
Adjusted R2	0.0021	0.014	0.11	0.11	0.41	0.46	0.48

Notes: Standard errors are reported in parentheses and clustered at the instructor level with 3,600 clusters. The dependent variable for all estimations is the average overall assessment score in each section, with a mean of 4.27 and a standard deviation of 0.72 across all sections. The independent variable "Online sections" is a binary variable that takes a value of 1 for online sections, including hybrid and fully online sections, and 0 otherwise. Various specifications are reported in separate columns, including the year-semester when the section was taught, the department providing the course, enrollment, instructor and course fixed effects, and instructor by course fixed effects. * p<0.10, ** p<0.05, *** p<0.010.

Table 4: Online Section and Student Characteristics (Logit)

	Coefficient	Marginal Effect
% Black or African American	1.33**	0.13**
% White	1.08**	0.10**
% Asian	-0.023	-0.0022
% Hispanic/Latino	0.32	0.031
% American Indian or Alaska Native	2.83**	0.27**
% US	0.14	0.013
% with age below 18	4.94***	0.48***
% with age above 22	1.40***	0.14***
% female	0.83***	0.079***
% with less than 12 credits	1.94***	0.19***
% with more than 15 credits	0.29	0.028
Average cumulative credits earned	-0.021***	-0.0020***
% with previous gpa lower than 2	1.49***	0.14***
% with previous gpa lower than 3	1.00***	0.097***
% with previous online experience	1.58***	0.15***
Observations	24439	
Pseudo R^2	0.124	

Notes: Reported are the estimated coefficients and marginal effects from a logit regression on the probability of a section to be taught online. Marginal effects measure the change in the probability from a one-unit change in the variable.

Table 5: N5 Propensity Score Matching: Online Sections and Overall Assessment Scores

	Residential Sample			Trimmed Residential Sample		
		Matched			Matched	
	(1)	(2)	(3)	(4)	(5)	(6)
Online sections	-0.14***	-0.17***	-0.16***	-0.14***	-0.16***	-0.14***
	(0.031)	(0.040)	(0.038)	(0.034)	(0.045)	(0.043)
Year × Semester FE	Yes	Yes	Yes	Yes	Yes	Yes
Nb. of Enrollment	Yes	Yes	Yes	Yes	Yes	Yes
Course FE	Yes	Yes	Yes	Yes	Yes	Yes
Instructor FE	Yes	Yes	Yes	Yes	Yes	Yes
Ex-post average student grade			Yes			Yes
Clusters	2491	2061	2061	2446	2006	2006
Observations	24439	11152	11152	23496	10259	10259
Adjusted R2	0.46	0.41	0.42	0.47	0.43	0.44
Mean overall assessment rating	4.23	4.22	4.22	4.22	4.19	4.19
SD of overall assessment rating	0.74	0.71	0.71	0.74	0.69	0.69

Notes: The table presents the estimated impact of online instruction on overall assessment rating before and after matching across different samples. The dependent variable is the overall assessment rating (maximum:5). Standard errors in all columns are clustered at the instructor level. Columns (1)-(3) present the results before and after matching the residential sample, while columns (4)-(6) apply the same matching method to the fully matched sample - the residential sample when trimming out course sections with a propensity score equal to and greater than 0.4. The removal of unmatched sections results in a decrease in the number of observations from column (1) to column (2), as well as from column (4) to column (5). * p<0.10, ** p<0.05, *** p<0.010.

Table 6: Overall Assessment Rating Controlling for Specific Questions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Online sections	-0.044***	-0.032***	-0.019*	0.014	0.0013	-0.00087	-0.018**
	(0.011)	(0.0089)	(0.011)	(0.012)	(0.012)	(0.010)	(0.0072)
Q1 Clarity of objectives	0.84***						
	(0.0090)						
Q2 Communication of ideas		0.85***					
		(0.0079)					
Q3 Expression of expectation			0.83***				
			(0.0099)				
Q4 Availability to assist students				0.74***			
				(0.014)			
Q5 Concern for students					0.84***		
					(0.0087)		
Q6 Stimulation of interest						0.83***	
						(0.0092)	
Q7 Facilitation of learning							0.85***
							(0.0065)
Year × Semester FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nb. of Enrollment	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Instructor FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	44277	44277	44277	44277	44277	44277	44277
Adjusted R2	0.80	0.86	0.82	0.78	0.84	0.85	0.88

Notes: Standard errors are reported in parentheses and clustered at the instructor level with 3,600 clusters. The dependent variable for all estimations is the average overall assessment score of the evaluation in each section, with a mean of 4.27 and a standard deviation of 0.72 across all sections. The independent variable "Online" is a binary variable that takes a value of 1 for online sections and 0 otherwise. Additionally, we add the average rating of each specific question from 1 to 7 into the regression. All columns control for the year-semester when the section was taught, the number enrolled, and instructor and course fixed effects. * p<0.10, ** p<0.05, *** p<0.010.

Table 7: Matching: Online Instruction on Student Performance

	Kernel Matching	Propensity Score Matching				Mahalanobis Matching
		with Replacement		without Replacement		
		Residential	Trimmed Residential	Residential	Trimmed Residential	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: performance in the current section						
Online sections	-0.081***	-0.073***	-0.087***	-0.10***	-0.053	-0.043
	(0.026)	(0.028)	(0.030)	(0.038)	(0.043)	(0.028)
Observations	24428	11152	10259	6054	4373	5137
Panel B: performance in other concurrent sections						
Online sections	-0.025	-0.035	-0.034*	-0.046*	-0.039*	-0.046*
	(0.023)	(0.023)	(0.018)	(0.025)	(0.024)	(0.026)
Observations	24405	11130	10256	6032	4371	5115
Panel C: performance in sections in the following semester						
Online sections	-0.025	-0.010	-0.014	-0.020	-0.0096	-0.021
	(0.022)	(0.024)	(0.024)	(0.021)	(0.029)	(0.032)
Observations	21373	9586	8826	5205	3695	4423
Course FE	Yes	Yes	Yes	Yes	Yes	Yes
Instructor FE	Yes	Yes	Yes	Yes	Yes	Yes
Year × Semester FE	Yes	Yes	Yes	Yes	Yes	Yes
Nb. of Enrollment	Yes	Yes	Yes	Yes	Yes	Yes
All student characteristics	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table displays the estimated coefficients of the online section indicator on average grade points across all students in the current section (Panel A), their average performance across other courses taken during the same semester (Panel B), and their average performance across all courses taken in the following semester (Panel C) after various matching on student characteristics. Column (1) uses kernel matching, Columns (2)-(3) use the nearest five neighbors propensity score matching in the residential and trimmed residential samples, Columns (4)-(5) are matched without replacement in the residential and trimmed residential samples. Column (6) uses Mahalanobis matching. Refer to the detailed matching description in Appendix C. All columns account for the year-semester in which the section was taught, the number of enrolled students, and include instructor and course fixed effects. Additionally, when analyzing student performance in other concurrent courses in Panel B, we controlled for different course types, specified in Table A9, to account for any variations in course selection alongside online sections. * $p<0.10$, ** $p<0.05$, *** $p<0.010$.

Table 8: Returns to Online Teaching Experience in Online Sections

	Evaluation		Avg Grade Points		
	Q8	% response	current section	other concurrent sections	following semester
OnlineExp	0.027**	-0.011	0.006	-0.009	0.007
	(0.013)	(0.007)	(0.011)	(0.013)	(0.011)
Course FE	Yes	Yes	Yes	Yes	Yes
Instructor FE	Yes	Yes	Yes	Yes	Yes
Year×Semester FE	Yes	Yes	Yes	Yes	Yes
Nb of Enrollment	Yes	Yes	Yes	Yes	Yes
All Student Chars.	Yes	Yes	Yes	Yes	Yes
Avg Grade Points	Yes	Yes			
Eval Response Rate	Yes				
Observations	3053	3053	3053	3032	2616
Adjusted R2	0.36	0.48	0.66	0.26	0.34

Notes: Standard errors are enclosed in parentheses and clustered at the instructor level. The regression results presented here only utilize data from online sections. The dependent variable in each column is specified and the independent variable equals the number of semesters the instructor has taught online before a given online section. All columns control for instructor, course and semester fixed effects, enrollment, all student characteristics summarized in Table 2, and the response rate of the evaluation and average student grade points in the section only when the dependent variable is the overall assessment rating in Q8. * $p<0.10$, ** $p<0.05$, *** $p<0.010$.

Table 9: Evaluation Gap by Teaching Experience

	Overall Assessment Rating Q8	
	(1)	(2)
Online	-0.10**	-0.15***
	(0.047)	(0.034)
TotalExp	0.009***	
	(0.003)	
Online × TotalExp	-0.005	
	(0.005)	
OnlineExp		-0.002
		(0.011)
Online × OnlineExp		0.012
		(0.009)
Course FE	Yes	Yes
Instructor FE	Yes	Yes
Year × Semester FE	Yes	Yes
Nb of Enrollment	Yes	Yes
All student characteristics	Yes	Yes
Avg Grade Points	Yes	Yes
Eval response rate	Yes	Yes
Observations	24439	24439
Adjusted R2	0.47	0.47

Notes: Standard errors are reported in parentheses and clustered at the instructor level. The dependent variable is the overall assessment rating in the evaluation. The independent variable "Online" is a binary variable that takes a value of 1 for online sections and 0 otherwise, "TotalExp" measures instructor's overall teaching experience across all modalities, "OnlineExp" equals the number of semesters the instructor has taught online before a given section. "Online × TotalExp" and "Online × OnlineExp" are their interaction terms. All columns control for instructor and course fixed effects, enrollment, response rate of the evaluation, the average grade points, and all student characteristics summarized in Table 2. * p<0.10, ** p<0.05, *** p<0.010.

References

W. T. Alpert, K. A. Couch, and O. R. Harmon. A Randomized Assessment of Online Learning. *American Economic Review*, 106(5):378–382, May 2016. ISSN 0002-8282. doi: 10.1257/aer.p20161057.

E. P. Bettinger and B. T. Long. Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students. *American Economic Review*, 95(2):152–157, May 2005. ISSN 0002-8282. doi: 10.1257/000282805774670149.

E. P. Bettinger, L. Fox, S. Loeb, and E. S. Taylor. Virtual Classrooms: How Online College Courses Affect Student Success. *American Economic Review*, 107(9):2855–2875, Sept. 2017. ISSN 0002-8282. doi: 10.1257/aer.20151193.

B. A. Burns. Students' Perceptions of Online Courses in a Graduate Adolescence Education Program. *MERLOT Journal of Online Learning and Teaching*, 9(1), Mar. 2013.

M. Campbell and J. Sheridan. Assessment Of Student Performance And Attitudes For Courses Taught Online Versus Onsite. *Journal of Applied Business Research*, 18, Jan. 2011. doi: 10.19030/jabr.v18i2.2114.

S. Carrell and J. West. Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *Journal of Political Economy*, 118(3):409–432, 2010. ISSN 0022-3808. doi: 10.1086/653808.

M. T. Cole, D. J. Shelley, and L. B. Swartz. Online instruction, e-learning, and student satisfaction: A three year study. *The International Review of Research in Open and Distributed Learning*, 15(6), Oct. 2014. ISSN 1492-3831. doi: 10.19173/irrodl.v15i6.1748.

J. B. Cook and R. K. Mansfield. Task-specific experience and task-specific talent: Decomposing the productivity of high school teachers. *Journal of Public Economics*, 140:51–72, Aug. 2016. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2016.04.001.

T. Cowen and A. Tabarrok. The Industrial Organization of Online Education. *American Economic Review*, 104(5):519–22, May 2014. doi: 10.1257/aer.104.5.519.

D. J. Deming, C. Goldin, L. F. Katz, and N. Yuchtman. Can Online Learning Bend the Higher Education Cost Curve? *American Economic Review*, 105(5):496–501, May 2015. doi: 10.1257/aer.p20151024.

D. Figlio, M. Rush, and L. Yin. Is It Live or Is It Internet? Experimental Estimates of the Effects of Online Instruction on Student Learning. *Journal of Labor Economics*, 31(4): 763–784, 2013. ISSN 0734306X, 15375307.

C. E. Galyon, E. C. T. Heaton, T. L. Best, and R. L. Williams. Comparison of group cohesion, class participation, and exam performance in live and online classes. *Social Psychology of Education*, 19(1):61–76, Mar. 2016. ISSN 1573-1928. doi: 10.1007/s11218-015-9321-y.

S. Gold. A constructivist approach to online training for online teachers. *Online Learning*, 5(1), Mar. 2019. ISSN 2472-5730, 2472-5749. doi: 10.24059/olj.v5i1.1886.

D. N. Harris and T. R. Sass. Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7):798–812, Aug. 2011. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2010.11.009.

F. Hoffmann and P. Oreopoulos. Professor Qualities and Student Achievement. *The Review of Economics and Statistics*, 91(1):83–92, 2009. ISSN 0034-6535.

M. S. Kofoed, L. Gebhart, D. Gilmore, and R. Moschitto. Zooming to Class?: Experimental Evidence on College Students' Online Learning During COVID-19. *IZA Institute of Labor Economics Discussion Paper Series*, 2021.

J. M. Krieg and S. E. Henson. The Educational Impact of Online Learning: How Do University Students Perform in Subsequent Courses? *Education Finance and Policy*, 11 (4):426–448, 2016. ISSN 1557-3060.

W. Leopald. Rushing Too Fast to Online Learning?: Northwestern University News, June 2010.

Y. Liu. A Comparison Study of Online versus Traditional Student Evaluation of Instruction. pages 3586–3591. Association for the Advancement of Computing in Education (AACE), June 2005. ISBN 978-1-880094-56-3.

B. Means, Y. Toyama, R. Murphy, M. Bakia, and K. Jones. Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies. Technical report, US Department of Education, May 2009.

B. Ost. How Do Teachers Improve? The Relative Importance of Specific and General Human Capital. *American Economic Journal: Applied Economics*, 6(2):127–151, Apr. 2014. ISSN 1945-7782. doi: 10.1257/app.6.2.127.

J. P. Papay and M. A. Kraft. Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, 130:105–119, Oct. 2015. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2015.02.008.

C. A. Platt, A. N. W. Raile, and N. Yu. Virtually the Same?: Student Perceptions of the Equivalence of Online Classes to Face-to-Face Classes. *Journal of Online Learning and Teaching*, 10(3), 2014.

J. S. Robertson, M. M. Grant, and L. Jackson. Is online instruction perceived as effective as campus instruction by graduate students in education? *The Internet and Higher Education*, 8(1):73–86, Jan. 2005. ISSN 1096-7516. doi: 10.1016/j.iheduc.2004.12.004.

D. P. Rudd. The value of video in online instruction. *Journal of Instructional Pedagogies*, 3, Feb. 2014.

P. D. Vlieger, B. Jacob, and K. Stange. Measuring Instructor Effectiveness in Higher Education. In *Productivity in Higher Education*, pages 209–258. University of Chicago Press, May 2018.

L. Wiest. Effective Online Instruction in Higher Education. *Quarterly Review of Distance Education*, 13:11–14, Jan. 2012.

M. Wiswall. The dynamics of teacher quality. *Journal of Public Economics*, 100:61–78, Apr. 2013. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2013.01.006.

D. Xu and S. S. Jaggars. The Effectiveness of Distance Education Across Virginia's Community Colleges: Evidence From Introductory College-Level Math and English Courses. *Educational Evaluation and Policy Analysis*, 33(3):360–377, 2011. ISSN 0162-3737.

D. Xu and S. S. Jaggars. Performance Gaps Between Online and Face-to-Face Courses: Differences Across Types of Students and Academic Subject Areas. *The Journal of Higher Education*, 85(5):633–659, 2014. ISSN 0022-1546.

Appendix A Summary Statistics

Table A1: Summary Statistics of Course Enrollment and Evaluation (Section level)

	Online		In person		Difference
	Mean	sd	Mean	sd	b
Enrollment					
Nb. of Enrollment before withdrawal	71.61	(121.29)	37.71	(62.57)	33.89***
Nb. of female students enrolled	42.15	(69.28)	20.67	(36.41)	21.48***
Course level					
Level 1 course	0.11	(0.31)	0.10	(0.30)	0.01**
Level 2 course	0.28	(0.45)	0.29	(0.45)	-0.01
Level 3 course	0.33	(0.47)	0.30	(0.46)	0.03***
Level 4 course	0.28	(0.45)	0.31	(0.47)	-0.03***
Evaluation response and ratings					
Response rate	0.37	(0.21)	0.46	(0.24)	-0.09***
Q8 Overall assessment	4.19	(0.64)	4.29	(0.74)	-0.10***
Within-section SD of Q8	0.80	(0.46)	0.68	(0.47)	0.12***
% of 1s received in Q8	0.03	(0.09)	0.03	(0.09)	0.00*
% of 2s received in Q8	0.05	(0.09)	0.04	(0.10)	0.00**
% of 3s received in Q8	0.15	(0.17)	0.12	(0.17)	0.03***
% of 4s received in Q8	0.23	(0.19)	0.20	(0.19)	0.02***
% of 5s received in Q8	0.53	(0.28)	0.59	(0.30)	-0.06***
College distribution					
Liberal Arts and Sciences	0.38	(0.49)	0.51	(0.50)	-0.13***
Agriculture	0.19	(0.39)	0.10	(0.30)	0.08***
Medicine	0.15	(0.36)	0.08	(0.27)	0.07***
Business	0.08	(0.28)	0.04	(0.18)	0.05***
Journalism	0.06	(0.24)	0.05	(0.22)	0.01**
Engineering	0.04	(0.20)	0.13	(0.34)	-0.09***
Arts	0.04	(0.20)	0.04	(0.20)	0.00
Education	0.03	(0.17)	0.01	(0.09)	0.02***
Architecture	0.02	(0.13)	0.03	(0.18)	-0.02***
Cross-College	0.00	(0.03)	0.00	(0.05)	-0.00*
Military	0.00	(0.00)	0.00	(0.02)	-0.00***
Observations	6115		38162		44277

Notes: This table provides a summary of course characteristics and evaluation ratings based on the instruction mode. The course levels are determined by the 1st digit of the course code, where a higher level indicates a more advanced course. The "Medicine" college encompasses Health and Human Performance, Medicine, Nursing, Pharmacy, Public Health, and Health Professions, as well as Veterinary Medicine. * p<0.10, ** p<0.05, *** p<0.010.

Appendix B Supplements for Satisfaction Analysis

Distribution of Overall Assessment Rating

To complement our primary finding that the average overall assessment rating for online sections is significantly lower than for in-person sections, we examine the distribution of the overall assessment rating within online sections. As detailed in Table A2, the lower overall assessment rating in online sections stems from a higher frequency of receiving ratings from 1 to 3, coupled with a lower frequency of 5s. As indicated by both the magnitude and significance, the main contributors to the overall lower satisfaction are the increased occurrence of 3s and the decreased occurrence of 5s.

Table A2: Within-Section Distribution of Overall Assessment Rating

	P(Q8=1)	P(Q8=2)	P(Q8=3)	P(Q8=4)	P(Q8=5)
Online sections	0.006*	0.009***	0.03***	0.007	-0.05***
	(0.0028)	(0.0029)	(0.0044)	(0.0049)	(0.0090)
Year × Semester FE	Yes	Yes	Yes	Yes	Yes
Nb of Enrollment	Yes	Yes	Yes	Yes	Yes
Instructor FE	Yes	Yes	Yes	Yes	Yes
Course FE	Yes	Yes	Yes	Yes	Yes
Observations	44277	44277	44277	44277	44277
Adjusted R2	0.22	0.20	0.24	0.14	0.48

Notes: Standard errors are reported in parentheses and clustered at the instructor level with 3,600 clusters. The dependent variables are the probabilities of receiving 1 through 5 for the overall rating, respectively. The independent variable is an indicator variable denoting whether the section is online. In all columns, we include controls for the year-semester when the course was taught, the number of enrolled students, and instructor and course fixed effects. * p<0.10, ** p<0.05, *** p<0.010.

Breakdown by Instructor and Course types

In Table A3, we provide a detailed breakdown of observations based on different instructors and course types. In our data, 2,586 instructors taught only in-person sections, 251 instructors taught only online sections, and 763 instructors taught both online and in-person sections. To address any potential differences in course content or design between online and in-person sections, we further categorize the sections taught by the 763 instructors with experience in both formats into three groups by course type: 1,017 courses taught only in-person, 215 courses taught only online, and 569 courses taught in both formats. We then present the within-section means and standard deviations of the overall assessment rating across instructor and course types. We consistently observe that online sections receive worse evaluations regardless of instructor or course type.

Table A3: Summary Statistics: Evaluation Ratings by Instructor and Course Type

	Instructors teach only in-person	Instructors teach only online	Instructors teach both			
			Course taught only in-person	Course taught only online	Course taught both	
					Sections in-person	Sections online
	(1)	(2)	(3)	(4)	(5)	(6)
Q8 Overall assessment	4.28	4.21	4.32	4.20	4.28	4.18
	(0.76)	(0.64)	(0.72)	(0.62)	(0.65)	(0.65)
Difference	-0.07**		-0.13***		-0.09***	
Instructors	2586	251	763			
Courses	2459	133	1017	215	569	
Observations	25586	576	7528	1148	5048	4391

Notes: The table summarizes the overall evaluation ratings (1-5 scale) across sections by instructor and course type, along with the differences between modes. Standard errors across sections are reported in parentheses.* p<0.10, ** p<0.05, *** p<0.010.

Breakdown by Colleges

Figure A1: College Level Use of Online Courses for Residential Students

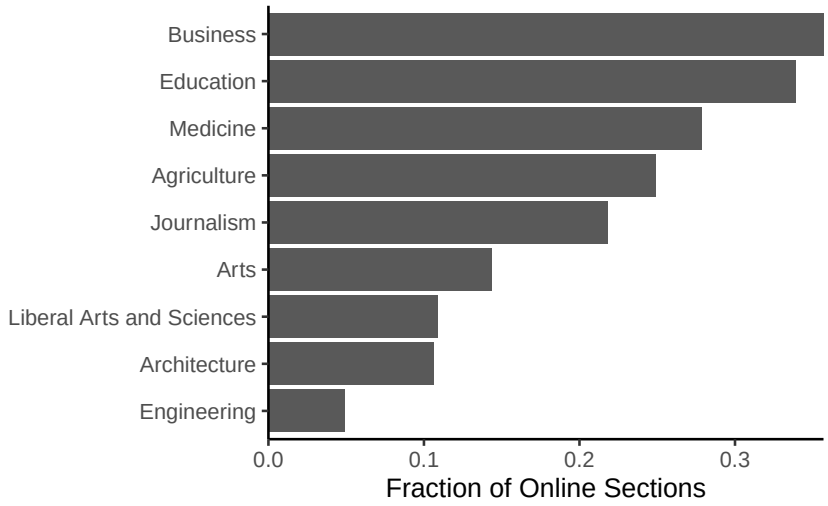


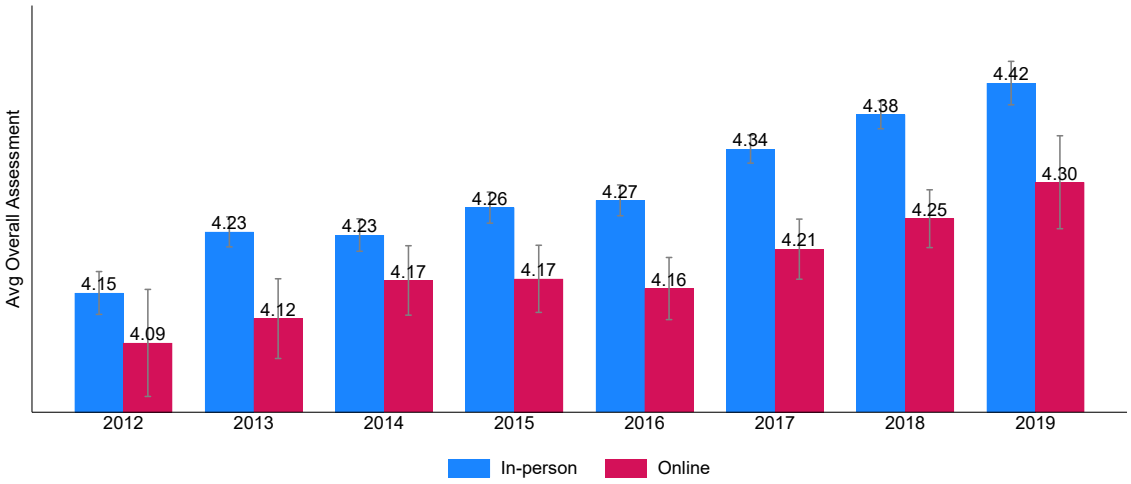
Table A4: Overall Assessment Rating by Colleges

	Journalism	Agriculture	Liberal Arts & Sciences	Medicine	Engineering	Architecture	Education	Business	Arts
Online sections	-0.19***	-0.16***	-0.14***	-0.097**	-0.094*	-0.19	-0.18	0.044	0.040
	(0.072)	(0.048)	(0.044)	(0.043)	(0.049)	(0.30)	(0.15)	(0.050)	(0.091)
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Instructor FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year × Semester FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nb. of Enrollment	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2392	5009	21957	4020	5213	1397	524	1861	1789
Adjusted R2	0.47	0.35	0.46	0.48	0.50	0.37	0.47	0.42	0.31

Notes: Standard errors are reported in parentheses and clustered at the instructor level with 3,600 clusters. The dependent variable for all estimations is the overall assessment rating in each section, with a mean of 4.27 and a standard deviation of 0.72 across all sections. The independent variable "Online" is a binary variable that takes a value of 1 for online sections, including all sections with an online component of more than 50%, and 0 otherwise. All columns control for the year-semester when the section was taught, the number enrolled, and instructor and course fixed effects. We exclude Military and Cross-college courses from the analysis because they barely offer online instruction as summarized in Table A1. * $p<0.10$, ** $p<0.05$, *** $p<0.010$.

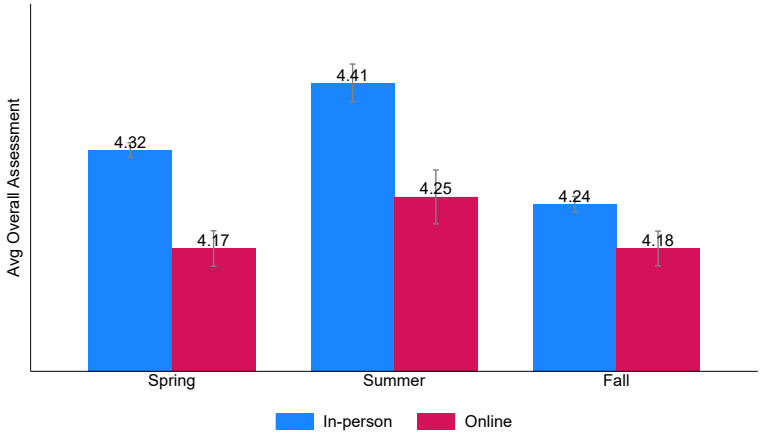
Breakdown by Years and Semesters

Figure A2: Overall Assessment Rating Across Years



Notes: The figure illustrates the relationship between the average overall student assessment rating and year. The bar in the figure represents the average overall assessment rating, while the line represents the 95% confidence interval.

Figure A3: Overall Assessment Rating Across Semesters



Notes: The figure illustrates the relationship between the average overall student assessment rating and the teaching semester. The bar in the figure represents the average overall assessment rating, while the line represents the 95% confidence interval.

Breakdown by Hybrid and Fully Online Courses

We have classified our online sections into two distinct categories: hybrid sections, which encompass 50% to 80% online components, and fully online sections, which comprise more than 80% online components. Our main findings in Table 3 are driven by fully online sections but not hybrid sections, as presented in Table A5. However, the interpretation of these results can be approached in different ways. It is possible that the perceived quality of hybrid sections is comparable to that of in-person sections, given the potential for interaction during class time. An alternative explanation could be that the students who respond to evaluations are a mixture of those who completed the course in-person and those who did so online.

Table A5: Evaluations for Hybrid and Fully Online Sections

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Hybrid sections	-0.023	-0.032	-0.041	-0.024	-0.064	-0.039	-0.011	-0.015
	(0.043)	(0.057)	(0.047)	(0.066)	(0.050)	(0.057)	(0.070)	(0.071)
Fully online sections	-0.083***	-0.098***	-0.12***	-0.19***	-0.14***	-0.14***	-0.12***	-0.12***
	(0.021)	(0.024)	(0.022)	(0.025)	(0.024)	(0.025)	(0.026)	(0.025)
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Instructor FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year × Semester FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nb. of Enrollment	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	44277	44277	44277	44277	44277	44277	44277	44277
Adjusted R2	0.38	0.45	0.37	0.39	0.41	0.44	0.45	0.46

Notes: Standard errors are reported in parentheses and clustered at the instructor level with 3,600 clusters. The dependent variables across estimations are the scores in Question 1 to 8 in the evaluation. The independent variable "hybrid sections" is a binary variable that takes a value of 1 for sections with online components between 50% and 80%, while "Fully online sections" is a binary variable that takes a value of 1 for section with more than 80% online components. All columns control for the year-semester when the section was taught, the number enrolled, and instructor and course fixed effects. * $p<0.10$, ** $p<0.05$, *** $p<0.010$.

Course List for Identification

Table A6: Top 15 Courses Taught Using Both Instruction Modes

Course Title	College	Sections/Observations
Introduction to Statistics 1	Liberal Arts and Sciences	471
Introduction to Biochemistry and Molecular Biology	Medicine	187
Theatre Appreciation	Arts	162
Precalculus: Algebra and Trigonometry	Liberal Arts and Sciences	156
What Is the Good Life	Liberal Arts and Sciences	105
Analytic Geometry and Calculus II	Liberal Arts and Sciences	101
Analytic Geometry and Calculus I	Liberal Arts and Sciences	95
Applied Human Anatomy with Lab	Medicine	88
Basic College Algebra	Liberal Arts and Sciences	72
Wildlife Issues	Liberal Arts and Sciences	65
Principles of Entrepreneurship	Business	51
Analytic Geometry and Calculus III	Liberal Arts and Sciences	50
Business Finance	Business	41
Survey of Calculus 1	Liberal Arts and Sciences	40
Introduction to Soils in the environment	Agricultural and Life Sciences	38

Notes: This table provides a concise summary of the top 15 courses contributing to our main identification, offered in both instructional modalities. Column 1 indicates the course title, Column 2 specifies the college offering the course, and Column 3 represents the total number of sections in our dataset, ranked in descending order.

Appendix C Supplements for Matching

Table A7: Balance of Student Characteristics in (Trimmed) Residential Sample

	Residential Sample			Trimmed Residential Sample		
Variable	Online	In-person	%bias	Online	In-person	%bias
Race&Ethnicity						
% Black or African American	0.085	0.085	0	0.083	0.085	-2.3
% White	0.589	0.591	-1.3	0.582	0.58	1.5
% Asian	0.086	0.086	0.4	0.093	0.092	1.3
% Hispanic/Latino	0.19	0.188	1.4	0.193	0.194	-1.3
% American Indian or Alaska Native	0.009	0.009	-1.6	0.008	0.008	2.5
Nationality						
% US	0.94	0.942	-2.4	0.939	0.94	-1.2
Age						
% below 18	0.003	0.002	3.5**	0.001	0.001	-2.8
% above 22	0.254	0.256	-1	0.179	0.187	-4.2
Gender						
% female	0.575	0.575	0	0.575	0.576	-0.4
Current credits load						
% below 12 credits	0.284	0.283	0.3	0.196	0.193	1.2
% above 15 credits	0.13	0.131	-1.2	0.151	0.149	1.3
Previous academic performance						
% with previous GPA below 2	0.064	0.066	-2.9	0.063	0.063	-0.3
% with previous GPA below 3	0.252	0.264	-8.4***	0.25	0.25	0
Avg. cumulative credits gained	43.95	45.559	-6.3**	47.575	47.134	1.8
Previous online experience						
% with previous online learning experience	0.686	0.698	-3.6	0.69	0.679	3.3

Notes: The table reports the balance of student characteristics between online and in-person sections after propensity score matching with five nearest neighbors in the full and trimmed samples (the fully balanced subsample that includes sections with propensity scores lower than or equal to 0.4).* p<0.10, ** p<0.05, *** p<0.010.

Alternative Matching Methodologies

In this section, we explore alternative matching methodologies beyond the nearest five neighbors propensity matching used in the main analysis. We examine (1) kernel matching, (2) nearest propensity score matching with no replacement in both full and trimmed samples, aimed at enhancing balance across all student characteristics between online and in-person sections, and (3) Mahalanobis matching, which involves matching with no replacement based on scale-free Euclidean distance. We begin by presenting the balance of all covariates for each alternative matching method in Figure A4, followed by the regression results on the overall assessment rating after employing each alternative matching method in Table A8.

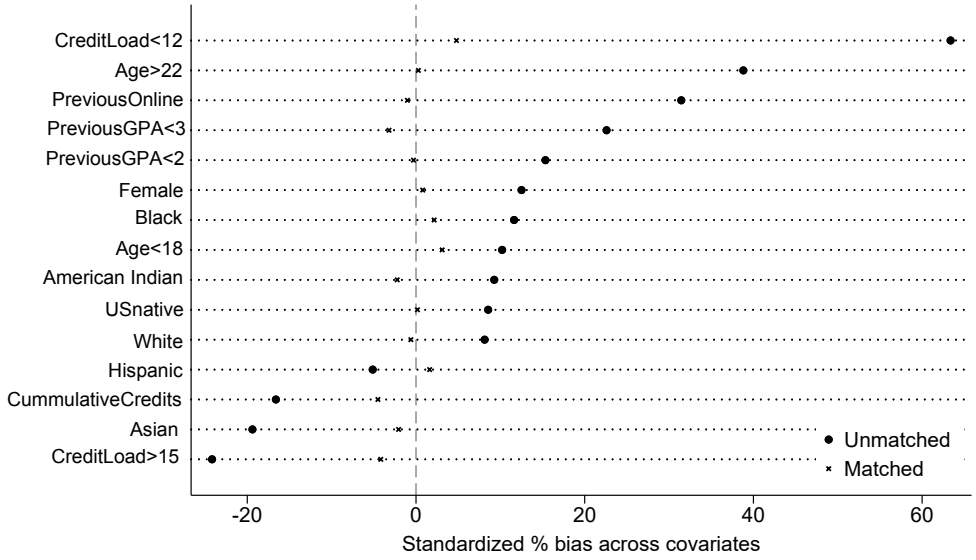
Table A8: Alternative Matching: Online Instruction and Overall Assessment Rating

	Kernel Matching	Propensity Score Matching w/o replacement		Mahalanobis Matching
		Residential	Trimmed Residential	
	(1)	(2)	(3)	(4)
Online sections	-0.16***	-0.16***	-0.12**	-0.10**
	(0.036)	(0.047)	(0.056)	(0.042)
Year × Semester FE	Yes	Yes	Yes	Yes
Nb. of Enrollment	Yes	Yes	Yes	Yes
Course FE	Yes	Yes	Yes	Yes
Instructor FE	Yes	Yes	Yes	Yes
Ex-post student grade	Yes	Yes	Yes	Yes
Observations	23431	4995	3328	4207
Adjusted R2	0.45	0.41	0.42	0.42
Mean overall assessment rating	4.22	4.23	4.18	4.19
SD of overall assessment rating	0.71	0.7	0.68	0.68

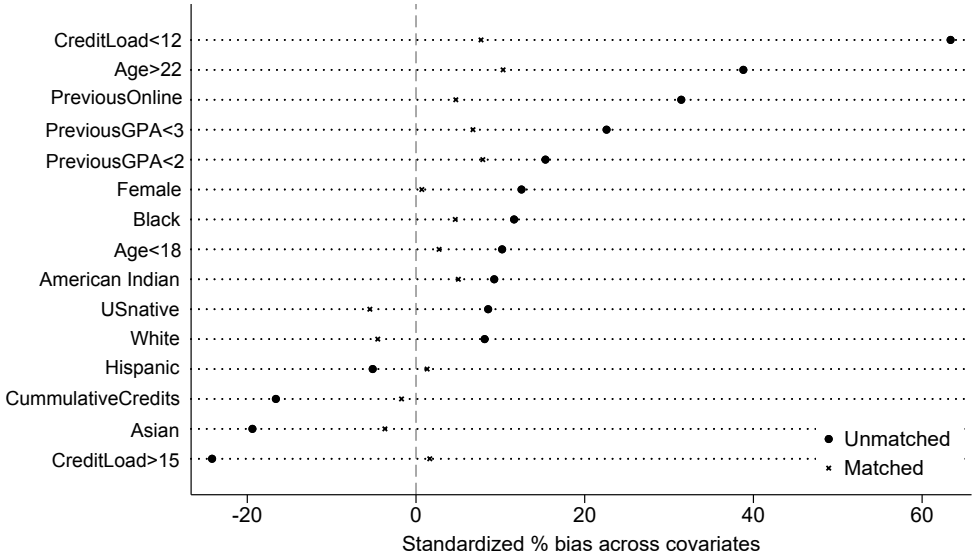
Notes: The table reports the estimated effects of online sections on overall assessment scores using alternative matching methods. Column (1) employs kernel matching, assigning more weight to in-person sections with similar characteristics. Columns (2)-(3) utilize propensity score matching without replacement; Column (2) presents results from the residential sample, while Column (3) represents the trimmed residential sample (excluding sections with propensity scores over 0.3). Column (4) employs Mahalanobis matching, matching with no replacement based on scale-free Euclidean distance. Across all columns, we control for the year-semester of section instruction, enrollment, instructor and course fixed effects, and ex-post student grade. * p<0.10, ** p<0.05, *** p<0.010.

Figure A4: Alternative Matching: Balance of Student Characteristics

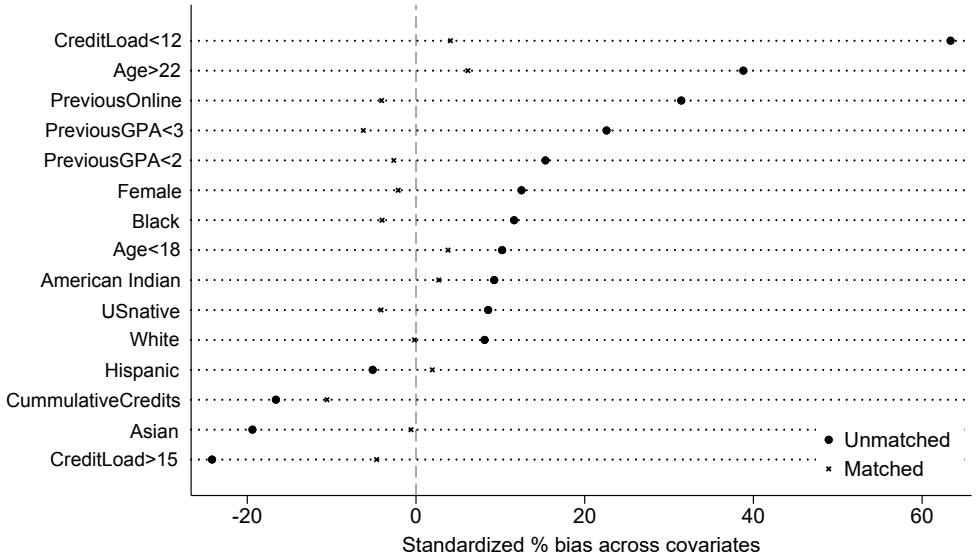
(a) Kernel Matching



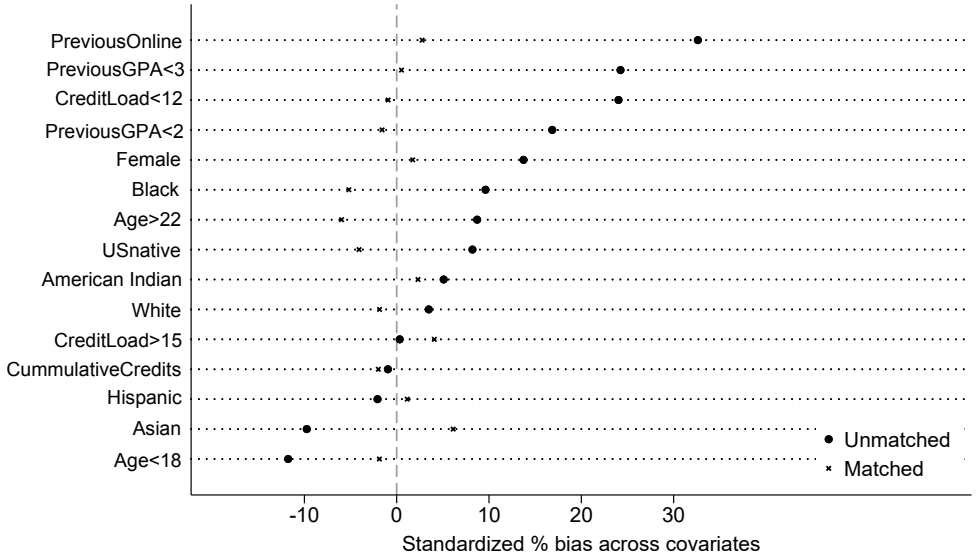
(b) Mahalanobis Matching



(c) Propensity Score Matching w/o Replacement



(d) Propensity Score Matching w/o Replacement in Trimmed Sample



Notes: This figure compares the standardized bias in percentage in the matched and unmatched sections across student characteristics between online and in-person sections when applying various matching methods: (a) kernel matching, (b) and (c) five-nearest neighbor propensity score matching without replacement in both full and trimmed samples (with the trimmed sample excluding sections with propensity scores greater than or equal to 0.4), and (d) Mahalanobis matching, which uses scale-free Euclidean distance for matching without replacement.

Appendix D Supplements for Performance Analysis

Table A9: Summary of Concurrent Courses Taken with Online vs In-person Sections

	Online		In person		Difference
	Mean	sd	Mean	sd	b
Avg grade points	3.02	(0.44)	3.16	(0.31)	-0.15***
% of level 1 course	0.07	(0.09)	0.09	(0.11)	-0.02***
% of level 2 course	0.28	(0.21)	0.32	(0.23)	-0.05***
% of level 3 course	0.37	(0.19)	0.31	(0.18)	0.06***
% of level 4 course	0.28	(0.22)	0.27	(0.24)	0.01
% of online course	0.39	(0.26)	0.19	(0.12)	0.20***
Observations	3032	3032	21382	21382	24414

Notes: This table presents a summary of course characteristics for other courses taken concurrently with either online or in-person sections in the same semester. * p<0.10, ** p<0.05, *** p<0.010.

Table A10: Online Sections on Avg Grade Points in the Following Semester

	Students in the Sample						Students left Sample	
	Avg Grade Points						Graduated	Dropout
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Online sections	-0.087***	-0.083***	-0.085***	-0.050***	-0.028	0.0081	-0.0012	0.0095
	(0.015)	(0.015)	(0.016)	(0.017)	(0.025)	(0.024)	(0.0041)	(0.0058)
Year × Semester FE		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nb. of Enrollment			Yes	Yes	Yes	Yes	Yes	Yes
Instructor FE				Yes	Yes	Yes	Yes	Yes
Course FE					Yes	Yes	Yes	Yes
All student characteristics						Yes	Yes	Yes
Observations	21738	21738	21738	21738	21738	21738	24439	24439
Adjusted R2	0.011	0.024	0.024	0.24	0.36	0.41	0.84	0.93

Notes: This table displays the estimated coefficients of the online section indicator on average student grade points across all courses taken in the semester following the online section in columns (1) to (6). The impact on the percentage of students who left our sample due to graduation or dropout is detailed in columns (7) and (8). All student characteristics are summarized in Table 2. Standard errors in all columns are clustered at the instructor level. * p<0.10, ** p<0.05, *** p<0.010.

Student-level Analysis

As our data supports student-level analysis in examining student performance, we will summarize our findings here. It is important to note that student-level analysis here is limited to comparing grades when the same student takes different online and in-person courses in the same semester within the same department. We cannot further constrain the course and instructor for the online and in-person sections, as typical undergraduate students, who did not fail a course, take a course only once.

We apply the following specification:

$$Y_{icdt} = \alpha + \beta Online_{icdt} + X + \epsilon_{icdt},$$

where Y_{icdt} is the grade for student i in course c provided by department d at semester t, and $online_{icdt}$ is an indicator that equals 1 if student i has taken an online course c by department d at semester t. X refers to a vector of controls, which we will gradually introduce student, semester and department fixed effects. As the level of the course students choose to taken online or in-person may differ, we also consider course level in X. ϵ_{icdt} is the robust error term. β is the estimate of interest, which captures the impact of online sections on student final grade.

Table A11: Student-level Analysis: Online Section on Student Performance

	(1)	(2)	(3)	(4)	(5)
Online Sections	0.038***	0.030***	0.13***	0.14***	0.14***
	(0.0024)	(0.0024)	(0.0023)	(0.0023)	(0.0027)
Course Level		Yes	Yes	Yes	Yes
Student FE			Yes	Yes	Yes
Year×Semester FE				Yes	Yes
Department FE					Yes
Observations	1283677	1283677	1283677	1283677	1283677
Adjusted R2	0.00018	0.0014	0.24	0.24	0.34

Notes: This table presents the estimated coefficients of the online section indicator on students' final grades using a student-level analysis. We provide the unconditional estimate in Column (1), introduce the level of the course taken online or in-person in Column (2), add student fixed effects in Column (3), include semester fixed effects in Column (4), and add department fixed effects in Column (5). The number of observations comprises 67,484 students. * p<0.10, ** p<0.05, *** p<0.010.

We summarize the estimates in Table A11, and we observe that after incorporating student, semester, and department fixed effects, online sections yield significantly higher grades compared to in-person sections taken by the same student in the same semester from the same department. Notably, there is a big difference in comparison between this analysis and the previous section: here, we are comparing different courses taken by the same student

in the same semester, whereas in the previous section-level analysis, we compared the same course taught by the same instructor but not taken by the same student. The positive and significant results using student-level analysis reinforce our previous findings from section-level analysis. We argue that when a student evaluates all courses they took in the same semester, it is unlikely for them to retaliate in the evaluation if they received higher grades in online sections. This discovery helps alleviate concerns about section-level analysis being susceptible to the influence of selection bias.

Appendix E Supplements for Returns to Experience

Table A12: The Effect of Previous Evaluation Rating on Continuity of Online Teaching

	Continuity of Online Teaching	
	(1)	(2)
Q8 rating	0.001	0.004
	(0.007)	(0.008)
Previous Q8 rating		-0.002
		(0.009)
Observations	2,494	2,395

Notes: This table presents the coefficients of evaluation rating when regressed on the continuity of online teaching. The dependent variable is a binary variable that takes a value of 1 when the same online course remains with the same instructor in a later semester, and 0 otherwise. The two independent variables are the overall assessment rating that the instructor received in the current section and the rating from the last time the same course was taught. * p<0.10, ** p<0.05, *** p<0.010.

Alternative Estimation

To examine whether our results are sensitive to estimation methods, we also follow the K-12 literature and investigate returns to online teaching experience using a two-stage regression model proposed by Papay and Kraft (2015). The two-stage model aims to first estimate time fixed effects by capturing common shocks across instructors. Subsequently, in the second stage, we remove the estimated time effects while reintroducing instructor fixed effects. This approach allows us to isolate the within-instructor within-course variation in online teaching experience, eliminating the influence of time trends. Specifically, we estimate the model

$$Y_{scit} = \alpha_1 OnlineExp_{scit} + \alpha_{2t} YS_t + X_{scit} + \epsilon_{scit},$$
$$\hat{Y_{scit}} = Y_{scit} - \hat{\alpha_{2t}} YS_t = \beta_1 OnlineExp_{scit} + Inst_i + X_{scit} + \eta_{scit},$$

where Y_{scit} is the outcome of interest, evaluation ratings or student grade points, in section s under course code c taught by instructor i in year-semester t, $OnlineExp_{scit}$ is the number of semesters that instructor i has taught online before teaching section s under course code c in year-semester t, and X_{scit} is a vector of controls across sections. Additionally, YS_t is the year-semester fixed effects, and α_{2t} captures any semester-to-semester variation across instructors in the dependent variable other than from changes in online teaching experience. We extract all estimated α_{2t} and subtract them from the dependent variable in the second stage. The parameter of interest is β_1, which captures the average effect of a one-semester

increase in online teaching experience on the outcome of interest. We use standard errors clustered at the instructor level to allow for arbitrary dependence of ϵ_{scit} and η_{scit} across sections given by instructor i.

The results here are consistent with those observed in the main paper using Model 2. We see a positive return to online teaching experience in online sections (Table A13), but this return does not differ between online and in-person sections, leaving the evaluation gap unmitigated (Table A14).

Table A13: Returns to Online Teaching Experience in Online Sections (Second Stage)

	Evaluation	$\widehat{AvgGradePoints}$		
	$\hat{Q}8$	current section	other concurrent sections	following semester
OnlineExp	0.019**	0.005	-0.006	0.0002
	(0.0076)	(0.0060)	(0.0057)	(0.0044)
Course FE	Yes	Yes	Yes	Yes
Instructor FE	Yes	Yes	Yes	Yes
Nb. of Enrollment	Yes	Yes	Yes	Yes
All student char.	Yes	Yes	Yes	Yes
Avg Grade Points	Yes			
Eval response rate	Yes			
Observations	3053	3053	3032	2616
Adjusted R2	0.37	0.66	0.24	0.33

Notes: Standard errors are enclosed in parentheses and clustered at the instructor level. The regression results presented here only utilize data from online sections. The dependent variable in each column is specified and adjusted by the year-semester fixed effects estimated in the first stage. The independent variable equals the number of semesters the instructor has taught online before a given online section. All columns control for instructor and course fixed effects, enrollment, all student characteristics summarized in Table 2, and the response rate of the evaluation and average student grade points in the section only when the dependent variable is the overall assessment rating in Q8. * p<0.10, ** p<0.05, *** p<0.010.

Table A14: Evaluation Gap by Teaching Experiences (Second Stage)

	Overall Assessment Rating $\hat{Q}8$	
	(1)	(2)
Online	-0.15***	-0.10**
	(0.034)	(0.046)
OnlineExp	-0.0068	
	(0.0096)	
Online×OnlineExp	0.012	
	(0.0091)	
TotalExp		0.0022
		(0.0042)
Online×TotalExp		-0.0048
		(0.0052)
Course FE	Yes	Yes
Instructor FE	Yes	Yes
Nb. of Enrollment	Yes	Yes
All student char.	Yes	Yes
Avg Grade Points	Yes	Yes
Eval response rate	Yes	Yes
Observations	24439	24439
Adjusted R2	0.47	0.47

Notes: Standard errors are reported in parentheses and clustered at the instructor level. The dependent variable is the overall assessment rating in the evaluation and adjusted by the year-semester fixed effects estimated in the first stage. The independent variable "Online" is a binary variable that takes a value of 1 for online sections and 0 otherwise, "OnlineExp" equals the number of semesters the instructor has taught online before a given section and "Online × OnlineExp" is the interaction term between the two. "TotalExp" equals the total number of semesters the instructor has taught the same course before a given section, both online and in-person, and "Online×TotalExp." is the interaction term. All columns control for instructor and course fixed effects, enrollment, response rate of the evaluation, the average grade points, and all student characteristics summarized in Table 2. * p<0.10, ** p<0.05, *** p<0.010.