Are in-person lectures beneficial for all students? A Study of a Large Statistics Class

Ellen S. Fireman, Zachary S. Donnini, Daniel J. Eck, and Michael B. Weissman

Abstract

Over 1000 students over four semesters were given the option of taking an introductory statistics class either by in-person attendance in lectures, augmented by online recorded lectures, or by taking the same class without the in-person lectures. The all-online students did slightly better on computer-graded exams. The causal effect of choosing only online lectures was estimated by adjusting for potential confounders using four methods. The four nearly identical point estimates remained positive but were small and not statistically significant. No statistically significant differences were found in preliminary comparisons of effects on females/males, U.S./non-U.S. citizens, freshmen/non-freshman, and lower-scoring/higher-scoring math ACT groups.

Keywords: online vs in-person lectures, statistics education, large online lectures, average treatment effect, causal inference
1 Introduction

Interest in using the increasing availability of good Internet access to expand the use of online education has accelerated greatly due to the Covid-19 epidemic. A common impression is that online education is somewhat inferior to in-person education [Loeb, 2020] largely based on secondary education results, e.g. Heppen et al. (2017). On the other hand, a meta-analysis based largely on college-level courses concluded that online components mixed with in-person components offered some advantages over in-person controls [Means et al., 2009]. The circumstances favoring one or the other method have not yet been well mapped-out.

It is not obvious a priori whether to expect online or in-person lectures to be more effective. Some reasons one might expect online lectures to be more effective, especially in any cumulative course, requiring understanding each step before moving on to the next, include that students can:

1. replay the parts with which they have difficulty,

2. fast-forward through parts that they find unnecessary,

3. listen when they’re in the mood,

4. take breaks if they have trouble concentrating for 50 or 80 minutes,

5. take just as much time as they need on in-lecture exercises,

6. make up lectures missed due to emergencies, and

7. use closed-captions if they prefer them to spoken English.

The disadvantages of online lectures include:

1. distractions in a non-classroom environment,

2. loss of the direct sense of personal involvement and interaction with other students, and
3. inability to ask questions during lecture.

The balance of factors is likely to depend on course content, structure, student characteristics, and lecture style.

Simple comparisons of outcomes for online and in-person students do not give the causal effect of the teaching mode, since without random assignment the students in each mode may systematically differ. Adjustments for differences between the students enrolled in different modes should be able to approximately correct for those differences if they are not too large, as is routinely done, e.g. in [Coates et al. (2004)]. There are a few well-controlled studies to check under what conditions each lecture mode is best, including a very few randomized controlled trials.

Throughout this paper we will follow the convention of labelling as “insignificant” any effects for which a 95% confidence interval includes zero, to avoid the distraction of considering effects of unknown sign. We do not, however, mean to imply a non-zero prior probability for any null hypothesis.

Most randomized studies have been conducted on economics classes. One (with a good review of previous work) found that in-person material had a significant positive effect on some but not all outcome measures, when compared to specially prepared online material that did not include videos of the in-person lectures (Arias et al., 2018). Another found that eliminating in-person material had negative effects on exam scores (Alpert et al., 2016). A blended online/in-person method had statistically insignificant negative effects compared to the in-person version (Alpert et al., 2016). Another study found small but statistically significant benefits of having more in-person classes (Joyce et al., 2014). Small, statistically insignificant, benefits of using in-person lectures, with indications that those benefits were concentrated in some subgroups, were found in another (Figlio et al., 2013).

One particularly relevant randomized study found no significant effect on test scores in a beginning statistics class with the treatment difference consisting of reducing in-person lecture hours by about a factor of three.
The authors note that although students were randomized, they were unable to randomize instructors. The authors emphasized that the success of their “hybrid” treatment, with reduced in-person components, may have depended on the use of interactive online materials.

Of the non-randomized studies with fairly stringent controls, one used instrumental variables (especially the intermittent availability of the in-person version of a class) to try to extract the causal effect of using a completely online version instead of a largely in-person version of a course taught at a large for-profit university. It found that the online effect on grades and subsequent performance in other courses was noticeably negative overall, especially for weaker students. Since in-person classes were only taught when teachers specifically volunteered, while online classes were offered even when no teacher volunteered to be involved, there may have been systematic differences between the treatment and control teachers.

The generalizability of the mixed conclusions of these varied studies to courses with different characteristics is unknown. Some of the key variables that they have suggested might moderate the online treatment effect are the nature of the treatment itself, which is far from uniform in the different studies, and the strength of the students.

Here we report results from four semesters of a large beginning statistics course at a large selective public university. The course had been taught in approximately its current form for several semesters using in-person lectures only and online homework and pre-lectures. It became convenient to offer a version that was almost identical to the standard in-person version but with students allowed to sign up to watch videos of the same lectures online rather than in-person. This paper describes an estimated causal effect on objective exam scores of using the new online version vs. the in-person version.

It was not possible to randomly assign students to the two versions, but
Fortunately most of the measured student characteristics in the two treatment groups were not dramatically different. We find that, although online students did slightly better than in-person students on objective exams, once adjusted for known prior differences between the student groups there was no statistically significant estimated causal difference between exam performance of students in online and in-person versions, with narrow enough 95% confidence intervals to exclude differences with much practical importance.

The possibility of important missing confounders is typically the biggest potential issue for a non-randomized study attempting to estimate causal effects. One obvious unmeasured confounder is prior commitment to this particular course, as opposed to more general conscientiousness. We suspected that students who chose to allot a scarce schedule slot to this course would be more likely to put effort into it than would students who avoided any constraints on their other courses by choosing the flexible online option. If so, our estimates of the average treatment effect (ATE) would tend to favor the in-person version, compared to the true ATE. Indications of such unmeasured confounding leading to overestimates of the advantages of attending in-person lectures have been seen in at least one other study, using a difference-of-difference method comparing randomized and non-randomized treatment and control groups, adjusting for covariates similar to the ones we used (Joyce et al., 2014). While we suspect that the effect would have the same sign in our observations, skipping lectures for which one has signed up is enough different from choosing not to sign up for them that we cannot use the prior result (Joyce et al., 2014) to make a quantitative estimate of the effect. Instead, we used a comparison of our main outcome, objective exam scores, with a more effort-weighted outcome, homework scores, to try to estimate whether any effect of different commitments in the two treatment groups was likely to be important.

With regard to prior preparation for statistics, routine anonymous online surveys with greater than 80% response rates in the Spring and Fall of 2019
showed very small differences between the fraction of online students (80%) and in-person students (77%) self-reporting having taken prior courses. Even neglecting the likely collinearity between this covariate and others used, no plausible effect magnitude of this covariate could have an important effect on our estimated ATE, so we did not attempt to analyze transcripts to sort out such effects.

Our intent at this stage was only to estimate the ATE over the whole collection of students. Our data were underpowered for a determination of different effects on most subgroups. Nevertheless, due to increased interest in and concern about using online methods together with the impracticality of getting more data on in-person lectures in the near future, we include a first look at interactions of the online treatment with several covariates for which we heard specific concerns: freshman status, ACT scores, gender, and U.S. citizenship. We found no statistically significant effects.

2 Treatment Methods

The treatment studied here did not consist of simply substituting online components for in-person components. Experience with online lectures in a larger statistics course with similar structure had indicated that students enrolled in an online version did at least as well as those enrolled for in-person lectures. (Unfortunately we do not have sufficiently uniform comparative data on outcomes in that course to use in a research publication.) Therefore once we started making the lectures available online we considered it unfair not to allow all of the students full access to them. Furthermore, in these large courses the number of students who miss lectures for a variety of reasons and then request access to the online versions is so large that it was easiest just to open up the online lectures to all. Therefore the treatment under study consists of removing enrollment in an in-person lecture while keeping the other course components, including online lectures, unchanged. This resem-
bles one previous study (Joyce et al., 2014), but contrasts with most previous work, e.g. Alpert et al. (2016) and Figlio et al. (2013), in which access to online resources was restricted for in-person students or in which substantially different presentations were used for the online group (Coates et al., 2004).

Almost all elements of the course were shared between the two treatment groups. These include an incomplete-notes workbook/text filled out during lectures, very frequent automated randomized homework exercises (24 in a 15 week course) linked to an online discussion board, practice exams, prelecture videos, access to a user-friendly statistical computing program (http://www.istics.net/DataProgram/), and little-used in-person office hours. The versions shared a website and received the same email notifications. The lectures were recorded at the in-person class and then posted online, usually within several hours, with occasional very minor edits.

The course material was reorganized between the first batch of two-semesters and the second to improve the logical flow and to let us drop an un-enforced recommendation that students had taken a previous statistics course. In between these two two-semester batches, there was a semester with different instructors in charge of the online and in-person versions, so we did not analyze data from that semester.

Other than the lecture delivery, the one difference between online and in-person versions was that a small amount of bonus credit was available to the in-person students who answered questions in lecture using an i-clicker, except for the first semester of the new version, Spring 2019, when new i-clicker exercises were not yet ready. Although both groups received bonus points for completing their lecture note workbooks, more were given to the online students to approximately balance the i-clicker points. Thus those students who registered for in-person lectures had some extra motivation to frequently attend. The algorithm for counting bonus points raised final grades by an amount approximately proportional to how far the non-bonus grade was from 100, so the bonus points primarily served to motivate students
who were not otherwise doing well in the course.

To offer students maximum flexibility, online students were given the option of arranging to go to lectures and obtain i-clicker points by individually registering their i-clickers. Several expressed interest, but none followed through. Few if any students without i-clickers were observed in lecture. Therefore dilution of the treatment effect by crossover from online to in-person treatment appeared to be negligible.

Data collection stopped in Spring 2020 since all students were transferred to the online version when Covid-19 hit, and we had to switch to an unproctored open-book exam format. This forced switch would have provided an opportunity to estimate unmeasured confounding effects, if it had happened at the start of the semester and with reliable exams, but neither of those conditions held.

### 3 Evaluation Methods

The only evaluation we looked at for this project was performance on computer-graded exams, to avoid any subconscious bias on the part of graders. These exams also are our best measure of learning outcomes, as opposed to effort. Almost all the students took the exams in-person in mixed groups of in-person and online registrants. A small number of the online students (roughly 3%) and occasional in-person students took the same exams remotely, proctored via a commercial service using webcam monitoring.

The evaluation method changed between the two-semester batches. The first batch used hand-graded midterm exams, which we did not look at for the purposes of this study, and cumulative computer-graded finals, the results used here. The second batch used three equal-weighted computer graded exams, with the semester average used here. We did separate analyses of these two batches. Despite the different forms of exam scores used, the standard deviations for all four semesters were very similar, ranging from
8.8 to 10.5. (All scores presented here are on a 100 point scale, with a difference of 10 representing one grade point.) There was more variation of exam means and standard deviations within two-semester batches than between them. To reduce statistical uncertainty, we combined all the data for an overall comparison of whether switching to all-online lectures affected objective exam scores in these two closely related batches of the same course.

We used four methods to adjust for differences between the students in the two versions in order to estimate the causal ATE of dropping in-person lectures. These were multiple linear regression (MLR), stabilized inverse propensity weight (IPW) [Lunceford and Davidian, 2004; Austin and Stuart, 2015], doubly-robust (DR) [Robins et al., 1994], and a nonparametric outcome highly-adaptive lasso (OHAL) method [Ju et al., 2020] as described below.

We included as covariates those relevant predictors to which we had access and whose values were set before the treatment started. The campus data support service supplied anonymized data files with the relevant covariates. These were student year in school (treated categorically), semester when the course was taken (also categorical), ACTmath score (including SAT equivalents), Gender, U.S./non-U.S. citizenship, overall ACT score, and the approximate median ACT score of the major in which the student was enrolled, obtained by averaging the scores of the 25th and 75th percentiles, available from a university web site for prospective students. (This ACT-major score was initially used as a proxy for ACT scores before those were available to us, but to our surprise remained a significant predictor even after individual ACT scores were included.)

High-school grade-point averages (HSGPA) were obtainable for 675 students of the 1105. The remainder of the students were from systematically different major groups, especially transfer students and most international students, who it was also important to include in the ATE estimate. Therefore we ran a full analysis omitting the HSGPA on the larger sample, which
also had the advantage of reducing statistical error. To estimate any correction for systematic differences in the traits measured by HSGPA, we tested how much inclusion of HSGPA changed the ATE on the subsample for which HSGPA was available.

The initial exploration was via MLR using least-squares fitting, using the same point-and-click program used in the class. The MLR method gives unbiased estimates if the linear effects model is correctly specified. Since the MLR residuals were non-Gaussian and not constant across the predictor space, unsurprising given the constraints on the ObjectiveExam conditional distribution (Faraway, 2016, page 281), we checked the confidence intervals using nonparametric bootstrap methods, and checked the ATE estimate using other methods. The first adjustment method was a standard stabilized IPW analysis (Lunceford and Davidian, 2004; Austin and Stuart, 2015). This method gives unbiased estimates of the ATE if the main-effect logistic regression model for the propensity of different types of students to take each version is correctly specified. Our propensity score model used logistic regression on the same covariates as the MLR model. We checked that the important predictive covariates were fairly well-balanced in the pseudo-sample generated by the IPW method (Austin and Stuart, 2015).

The third method, DR, corrects for any imbalances in the IPW pseudo-sample by using MLR estimates to give unbiased estimates if either the linear effect model or the logistic propensity model is correct (Robins et al., 1994). We checked the key bottom-line ATE estimates with the OHAL targeted minimum loss method that avoids potentially problematic parametric misspecification (Kang et al., 2007; Ju et al., 2020).

The ATE confidence intervals for the MLR, IPW, and DR estimates of the ATE were estimated using standard bootstrap methods (DiCiccio and Efron, 1996). Confidence intervals on the small adjustment for inclusion of HSGPA were determined by a paired bootstrap analysis, in which the change of the ATE estimate from including HSGPA in the model was determined for each
bootstrap sample. That allows the resulting small HSGPA adjustment to be made with little increase in the confidence interval width for the ATE. Confidence intervals for the OHAL procedure were obtained using cross-validated standard errors (Ju et al., 2020, page 115). Full descriptions including the R code and full results for these methods are presented in the Supplementary Appendix. Since all methods gave nearly the same point estimates and confidence intervals, we focus here on the MLR results, whose coefficients have simple intuitive interpretations.

We included all 1177 students for whom we had final exam grades in an initial calculation of the raw point difference between online and in-person. Of those, we dropped 66 students for whom no admission test scores were available as well as 6 students not enrolled for undergraduate degrees before doing a full analysis on the remaining 1105 students. This analysis sample included 91% of the in-person students for whom we had final grades and 93% of the online students with final grades. The overall raw score average difference between online and in-person groups was very similar in the full sample (1.19) and the analysis sample (1.38).

4 Results

Of the students on whom we have records (those registered at the official drop date) 4 of 506 (0.8%) in-person students either dropped out by petition or otherwise failed to take the final exam, while 11 of 708 (1.6%) online students did so. The difference is not large enough to reject the null hypothesis of random dropouts at a 95% confidence level. At any rate, the dropout rate in both groups was quite low.

The raw exam scores and their SDs for each semester are given in Table 1. The score scale is 0-100, with 10 points corresponding to one grade point.

Although the online students usually did better, one cannot draw any conclusions about the treatment effect without adjusting for differences be-
Table 1: Raw exam scores, their standard deviations, and additional summary statistics for each semester.

| Semester | n  | Mean score | SD  | n_{OL} | n_{IP} | Mean_{OL} | Mean_{IP} | Mean_{OL} - Mean_{IP} | SE(diff) |
|----------|----|-------------|-----|--------|--------|------------|------------|------------------------|---------|
| F17      | 271| 87.27       | 8.82| 135    | 136    | 86.52      | 88.01      | -1.49                  | 1.07    |
| Sp18     | 274| 85.36       | 10.31| 158    | 116    | 86.88      | 83.29      | 3.59                   | 1.27    |
| Sp19     | 267| 83.30       | 10.84| 201    | 66     | 84.64      | 79.19      | 5.45                   | 1.69    |
| F19      | 293| 86.65       | 10.16| 153    | 140    | 87.44      | 85.78      | 1.66                   | 1.18    |

Table 2: Summary information for the collected covariates taken across the online and in-person groups. The abbreviations OL and IP are, respectively, shorthand for online and in-person. The random null p-value assesses the significance of differences in covariate composition between the online and in-person groups using a permutation test.

| Trait     | n   | n_{OL} | n_{IP} | Mean | SD   | Mean_{OL} | Mean_{IP} | Mean_{OL} - Mean_{IP} | SE(diff) | Random null p-value |
|-----------|-----|--------|--------|------|------|-----------|------------|------------------------|---------|---------------------|
| ACT       | 1105| 647    | 458    | 30.5 | 3.5  | 0.38      | 0.21       | 0.065                  |         |                     |
| ACTmath   | 1105| 647    | 458    | 32.2 | 4.0  | 0.45      | 0.25       | 0.064                  |         |                     |
| ACTverbal | 1105| 647    | 458    | 58.8 | 8.0  | 0.60      | 0.49       | 0.216                  |         |                     |
| ACTmajor  | 1105| 647    | 458    | 30.2 | 2.6  | 0.48      | 0.16       | 0.003                  |         |                     |
| HSGPA     | 675 | 377    | 298    | 3.5  | 0.3  | 0.04      | 0.03       | 0.183                  |         |                     |
| % of OL   |     | 435    | 270    | 165  | NA   | 0.42      | 0.36       | 0.048                  |         |                     |
| % of IP   |     | 474    | 290    | 184  | NA   | 0.45      | 0.40       | 0.119                  |         |                     |
| International | 155 | 40    | 115    | NA   | NA   | 0.06      | 0.25       | \approx 0.000           |         |                     |
| Female    | 407 | 212    | 195    | NA   | NA   | 0.33      | 0.43       | \approx 0.001           |         |                     |
| Sophomore | 312 | 226    | 86     | NA   | NA   | 0.35      | 0.19       | \approx 0.000           |         |                     |
| Junior    | 231 | 169    | 62     | NA   | NA   | 0.26      | 0.14       | \approx 0.000           |         |                     |

Table 2 compares the values of the covariates in the two treatment groups. The US/international and male/female distributions were not too far from uniform. The ACT scores and HSGPA were rather well matched, with differences between the group means always less than 20% of the SD for the individuals. The ACTmajor scores, however, differed significantly more than would be expected by random assignment. The enrollment by college year was very substantially different. (Students said that advisors to incoming freshmen discouraged them from taking the online version.)

Table 3 gives the multiple regression, stabilized IPW, and doubly-robust estimates for the ATE for the two batches of two semesters and for the combination of the four semesters, along with the 95% confidence intervals.
Table 3: Estimates and 95% confidence intervals for the ATEs of online learning across semesters. Confidence intervals for ATE\textsubscript{MLR}, ATE\textsubscript{IPW}, and ATE\textsubscript{DR} are obtained from the percentiles of a nonparametric bootstrap.

| Semesters | ATE\textsubscript{MLR} | ATE\textsubscript{IPW} | ATE\textsubscript{DR} | ATE\textsubscript{OHAL} |
|-----------|-----------------|-----------------|-----------------|-----------------|
| F17+Sp18  | -0.01 (1.42, 1.43) | 0.16 (1.21, 1.58) | 0.05 (1.35, 1.49) |
| Sp19+F19  | 1.54 (0.09, 3.21)  | 1.66 (0.07, 3.45) | 1.36 (0.28, 3.01) |
| All       | 0.63 (0.44, 1.71)  | 0.75 (0.34, 1.86) | 0.58 (0.44, 1.63) | 0.63 (0.50, 1.75) |

All the covariates were used except ACT\textsubscript{verbal}, which was predictable with $R^2 = 0.89$ from the other covariates, and whose inclusion would have a tiny effect (+0.02) on the ATE\textsubscript{MLR}. The four overall ATE estimates are nearly identical. Although overall the point estimate for the online effect is positive, it is very small for practical purposes and not statistically significant at the 95% confidence level in this sample.

There are some small variations between the semesters, e.g. between Fall and Spring semesters, but we lack sufficient power to see if these are systematic. omission of the Sp19 data, for which there was no i-clicker bonus incentive to attend lectures, slightly lowers the ATE\textsubscript{MLR} point estimate, from 0.64 to 0.35, too small an effect to draw any conclusions. It is also possible that the differences between the treatment groups differed between Falls, where adviser guidance was important in course choice, and Springs, where peer advice probably played a bigger role. The ATE difference between inverse propensity weighting in fall and spring averages was just short of conventional statistical significance, leaving a weak hint that unmeasured confounders should be considered.

Table 4 shows the multiple regression coefficients for the covariates of the
Table 4: The least-squares point estimates for the predictive coefficients in the multiple linear regression model based on the 1105 student sample are shown. The 95% confidence intervals are calculated by a nonparametric bootstrap method, and are close to those obtained using a t distribution. The intercept represents a very hypothetical domestic male senior in the in-person F19 class with 0’s for all ACT scores.

| Variable  | slope   | 95% Confidence Interval         |
|-----------|---------|---------------------------------|
| Intercept | 29.48   | (22.52, 36.05)                  |
| Online    | 0.63    | (-0.36, 1.68)                   |
| Gender    | 0.41    | (-0.63, 1.43)                   |
| International | 0.86 | (-0.26, 2.01)                   |
| F17       | 0.05    | (-1.28, 1.33)                   |
| S18       | -1.09   | (-2.47, 0.22)                   |
| S19       | -2.60   | (-4.02, -1.17)                  |
| FR        | -1.90   | (-3.66, -0.10)                  |
| SO        | -0.27   | (-1.45, 1.21)                   |
| JR        | -0.50   | (-1.79, 0.81)                   |
| ACTMajor  | 0.52    | (0.31, 0.74)                    |
| ACT       | 0.22    | (-0.04, 0.49)                   |
| ACTMath   | 1.07    | (0.80, 1.32)                    |

MLR model, for which \( R^2 = 0.34 \). With few exceptions, the ATE was not very sensitive to omission of any of these covariates. Removal of all the ACT scores increases the ATE by about 0.5. Removal of the college class increases the ATE by about 0.3. Other covariates had even smaller effects on the ATE.

Inclusion of HSGPA in the model on the subsample of 675 students for which it was available increased \( R^2 \) from 0.32 to 0.38. It increased the ATE in this subsample by 0.11, 0.09 and 0.15 for MLR, IPW and DR methods, respectively, each with 95% confidence intervals of about ±0.4. Assuming that the traits measured by HSGPA have roughly similar effects and similar group differences in the 39% of the sample for which HSGPA was not available, the ATE estimates would then be 0.83, 0.73, and 0.73 for IPW,
DR, and MLR respectively, each with 95% CIs of ±1.2, assuming the errors add in quadrature. All these estimates of ATE in the overall sample remain insignificantly positive.

We explored interaction effects of the treatment with the variables suspected of being relevant to the effectiveness of the online treatment (gender, citizenship, freshman status and ACTmath). Adding these effects, either individually or all simultaneously, gave no significant interaction term. Stratifying by the same variables gave insignificantly positive online ATE_{MLR} for both US and non-US citizens, both freshmen and non-freshman, and both the upper and lower halves of the ACTmath distribution. The stratified estimated ATE_{MLR} was insignificantly negative for females and nominally significantly positive (p=0.035) for males. The difference between the stratified results for the genders was not significant. Due to the multiple comparisons the effect for males should not be considered to be shown to be positive with confidence.

Although the ACTmath SD was 4.1, enough range to easily see its predictive power, less than 1% of the students scored below 20, the approximate national median. Thus we have essentially no evidence on how well the online treatment would work in a course like this for students with ACTmath scores below 20.

Since all the methods of estimating ATE from our covariates gave nearly the same results, the one serious remaining issue is possible unmeasured confounders. Given the very close balance in self-reported prior statistics courses taken, that would be an implausible confounder. Students’ unmeasured commitment (UC) to the course seems the most obvious variable likely to affect learning outcomes, to differ between groups choosing different versions, and not to show up in more general covariates. That commitment should affect how much effort students put into doing the homework. In fact homework (HW) was less predictable than ObjectiveExam from the covariates to which we had access ($R^2 = 0.11$ for HW in the HSGPA group, in contrast to 0.37
for ObjectiveExam), suggesting that HW may reflect causes not picked up by our standard covariates. In contrast to ObjectiveExam, HW was much better predicted by HSGPA than by ACT scores, consistent with our intuition that HSGPA and HW might show relatively large effort-dependent contributions.

Including HW as a covariate would bias the ATE estimate for three reasons. Most importantly, HW could be on the causal path from OL to ObjectiveExam, so including it would remove part of the treatment effect, since students using exclusively online lectures may spend more time going back-and-forth between lecture segments and related homework problems, improving their understanding. It would also introduce some slight M-bias (Greenland, 2003), probably negative, since HW is a descendant of both UC and the other covariates. It could also bias any treatment effect toward zero simply by serving as a marker of overall treatment effects.

Nevertheless, HW should pick up any dramatic imbalance of UC between the treatment groups. Including HW reduced $\text{ATE}_{\text{MLR}}$ from 0.64 to 0.34 in the overall sample and from 0.82 to 0.41 in the HSGPA-available subset. These small differences in the direction of the expected bias show no sign of any motivational confounding issue.

5 Discussion

Before discussing the online effect, we note in passing some observations, not directly relevant to our research questions but perhaps useful in other contexts, about test score predictors. The ACT scores (particularly ACTmath) were the most important predictors of scores among the covariates to which we had access. It is interesting that ACTmajor, the approximate median of the ACT of the student’s major, remains a highly significant predictor even when ACTmath is included. The effect is strongest among non-freshmen, suggesting that it arises largely from the effects of motivation, ability, and interests on how students sort into majors after they start taking college
Since the treatment under study consisted simply of removing in-person enrollment from otherwise fixed course components, it might seem surprising that a negative effect was not found. The general presumption is that removing resources will usually have at least a small negative effect (Coates et al., 2004). Furthermore, most previous studies have found somewhat negative effects of switching to all-online lectures (Coates et al., 2004). There are several potential explanations.

First, although the point estimate of the effect was weakly positive the statistical CIs are large enough that we cannot rule out a small negative effect from dropping in-person enrollment. The bottom of the overall CI range, however, is not more negative than -0.05 grade points, and thus would be of little practical significance.

Second, although we have done the best we could to adjust for relevant confounders, the possibility of important missing confounders always remains. Although we suspected that prior commitment to this particular course might slant our estimates to favor the in-person version inclusion of a commitment-related proxy variable (HW score) showed no evidence of any major effects of this sort. Adjustments for the covariates to which we had access, which provided fairly good predictors, had a small effect on the ATE, so we doubt that others would qualitatively change the results.

Third, attending in-person lectures probably reduced the effort spent following the same lectures online. Thus the true treatment effect of dropping the in-person lectures could indeed have been close to zero or even weakly positive, as in our point estimate.

The absence of significant interaction effects with freshman status or with ACTmath scores may seem more surprising, since previous work has generally indicated that better-prepared or stronger students have better online effects (Bettinger et al., 2017). Most of the students taking this class were already comfortable with some mathematics. For the larger introductory statistics
course, for which informal results were qualitatively similar, students had a much broader range of math preparation. Nevertheless, almost all students even in that course had been selected into a highly-ranked public university, so we do not know how well these methods would work for students who could not get in to such a university. Given the restricted range of our sample, the lack of an interaction term with ACTmath is entirely compatible with previous results (Bettinger et al., 2017). We were also somewhat surprised there was no indication that non-US students did particularly well online.

Although our results were underpowered for looking at interaction effects of the treatment even with measured covariates, presumably there are some differences among students in the relative value of online and in-person lectures. If students tended to take the version most suited to produce good test results for themselves, the estimated ATE would not give an accurate prediction of the net effect of requiring all students to take one version or the other. Either required version would give lower results than estimated here. If the students’ choices tended to be mistaken, the effect of requiring a single version for all students would be higher than our estimate. After Covid struck, all students were in fact required to take the online version, but historical controls on expected test scores would have been too imprecise to estimate any such effects reliably even if drastic changes in the testing protocol had not also been necessary.

Since the “online” version here still offered in-person office hours, it is not exactly equivalent to a fully remote online version. These office hours were not used very much, however. Summer sessions of the course, taken remotely, function well without them. After Covid hit they were replaced with Zoom versions during the semester. We think that this change is not very important.

Concerning the generalizability of the results, some specific features of this statistics class may lead to better online results than would be found for many other classes:
1. This is a cumulative logical/mathematical course, which may make some online benefits especially relevant, especially the ability to go back over difficult steps.

2. The prior work to develop effective beginning statistics courses without in-person discussion sections had already created important online course elements: very frequent homework with randomized numbers, supplementary pre-lectures.

3. The incentive to fill out the incomplete notebook may have reduced the effect of environmental distractions by keeping students actively engaged in lecture.

4. It’s possible that nearly all of the students in this course were well enough prepared and web-savvy to avoid some detrimental effects of missing in-person lectures.

5. Unlike in most previous work on online delivery (Coates et al., 2004), we made no effort to develop anything special for the online class beyond the online materials already available for the in-person class. Perhaps the videos of the traditional in-person lectures were more useful than most newly-developed online material, for which optimization is based on less experience.

6. The results pertain to only a single lecturer. Other lecturers with different lecture styles may get a different balance of online and in-person outcomes.

7. Although we found no evidence that the students needed the in-person experience, we did not test whether the lecturer needed it. The course was developed over several semesters with live feedback, and the lectures were delivered to in-person classes. The lecturer here (ESF) subjectively feels that live feedback was important for the course devel-
opment. Anecdotally, many other lecturers tell us that they need the
in-person experience to lecture well.

Although we suspect that for many other STEM courses online lecture
delivery will be at least as useful as in-person lecture delivery, so long as the
lecturer can stay motivated and aware of what students are finding easy or
difficult, we reiterate that we are not claiming that in-person course com-
ponents add little value to online components as a general rule. For example,
we make no claims about the ability to replace course components such as
labs with online materials. We also are not sure how well online exam-taking
systems will hold up under massive use. We have made no effort to compare
with a conventional purely in-person version, with no online lecture backup,
but we suspect that it would be inferior to either of the versions we exa-
mined. One key aspect that should be investigated in future studies, when
they become possible, is the indirect effect of lecture mode on learning via
its effect on the instructors rather than via direct effects on the students.

References

Alpert, W. T., K. A. Couch, and O. R. Harmon (2016). A randomized
assessment of online learning. *American Economic Review* 106(5), 378–
82.

Arias, J., J. Swinton, and K. Anderson (2018). Online vs. face-to-face: A
comparison of student outcomes with random assignment. *e-Journal of
Business Education and Scholarship of Teaching* 12(2), 1–23.

Austin, P. C. and E. A. Stuart (2015). Moving towards best practice when
using inverse probability of treatment weighting (iptw) using the propen-
sity score to estimate causal treatment effects in observational studies.
*Statistics in Medicine* 34(28), 3661–3679.
Bettinger, E. P., L. Fox, S. Loeb, and E. S. Taylor (2017). Virtual classrooms: How online college courses affect student success. *American Economic Review* 107(9), 2855–75.

Bowen, W. G., M. M. Chingos, K. A. Lack, and T. I. Nygren (2014). Interactive learning online at public universities: Evidence from a six-campus randomized trial. *Journal of Policy Analysis and Management* 33(1), 94–111.

Coates, D., B. R. Humphreys, J. Kane, and M. A. Vachris (2004). “no significant distance” between face-to-face and online instruction: Evidence from principles of economics. *Economics of Education Review* 23(5), 533–546.

DiCiccio, T. J. and B. Efron (1996). Bootstrap confidence intervals. *Statistical science*, 189–212.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press.

Figlio, D., M. Rush, and L. Yin (2013). 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.

Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 14(3), 300–306.

Heppen, J. B., N. Sorensen, E. Allensworth, K. Walters, J. Rickles, S. S. Taylor, and V. Michelman (2017). The struggle to pass algebra: Online vs. face-to-face credit recovery for at-risk urban students. *Journal of Research on Educational Effectiveness* 10(2), 272–296.

Joyce, T. J., S. Crockett, D. A. Jaeger, O. Altindag, and S. D. O’Connell (2014). Does classroom time matter? A randomized field experiment of
hybrid and traditional lecture formats in economics. Technical report, National Bureau of Economic Research.

Ju, C., D. Benkeser, and M. J. van Der Laan (2020). Robust inference on the average treatment effect using the outcome highly adaptive lasso. *Biometrics* 76(1), 109–118.

Kang, J. D., J. L. Schafer, et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science* 22(4), 523–539.

Loeb, S. (2020). How effective is online learning? what the research does and doesn’t tell us. *Education Week*.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 23(19), 2937–2960.

Means, B., Y. Toyama, R. Murphy, M. Bakia, and K. Jones (2009). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89(427), 846–866.