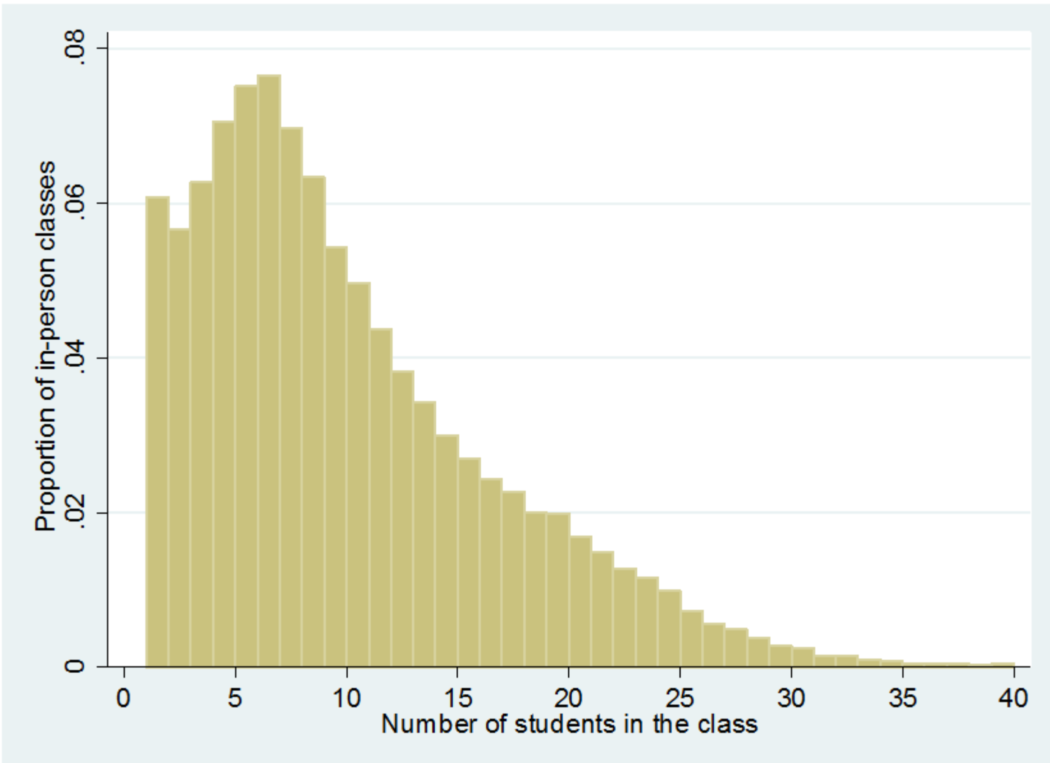
Virtual Classrooms: How Online College Courses Affect Student Success: Online Appendix

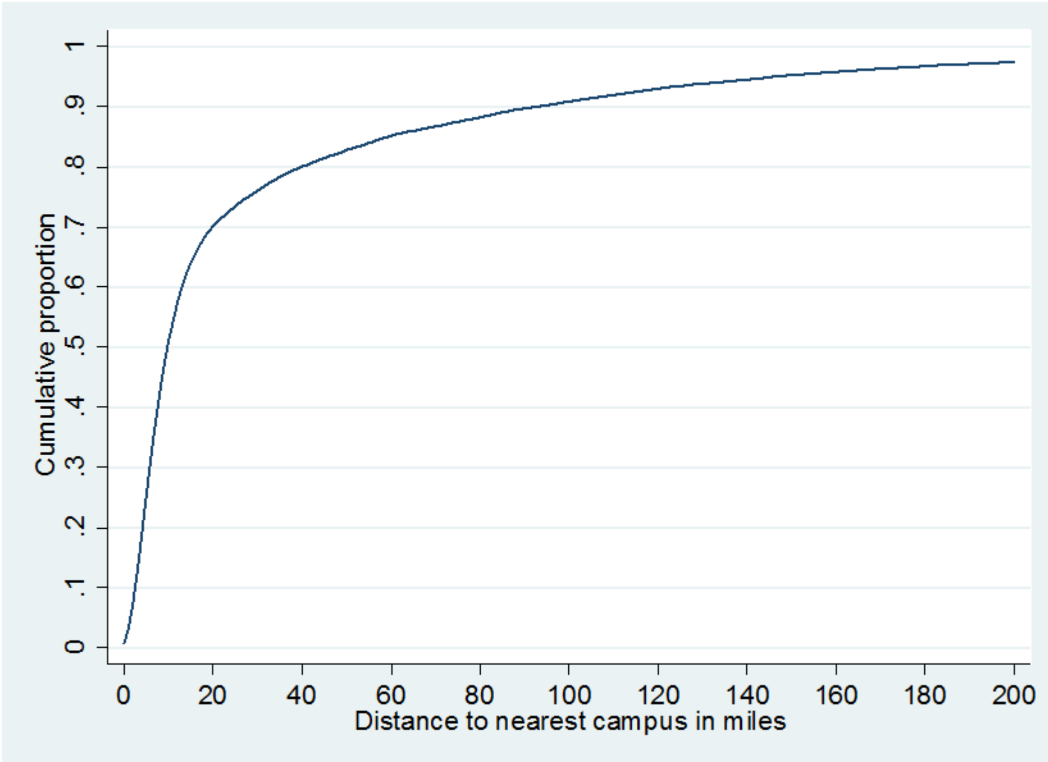
BY Eric P. Bettinger, Lindsay Fox, Susanna Loeb, and Eric S. Taylor

Appendix A. Additional Figures and Tables



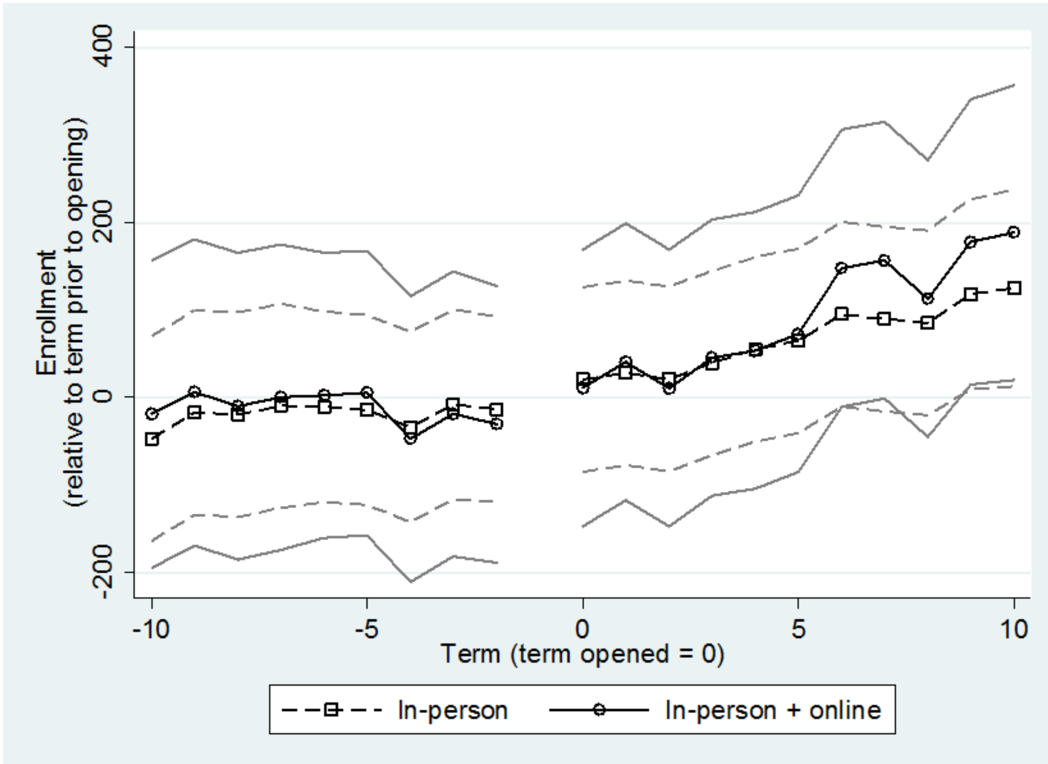
Appendix Figure 1—Histogram of class size for in-person classes

Note: Based on 112,628 in-person classes in the study data.



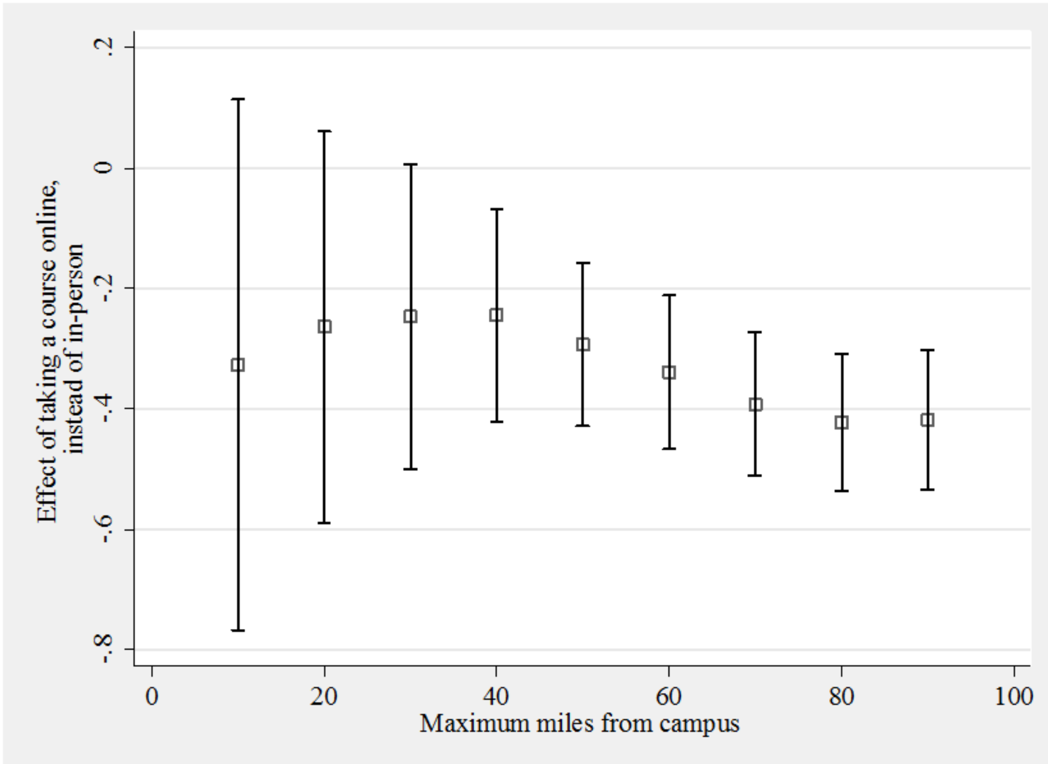
Appendix Figure 2—CDF of the distance in miles between student residence and nearest campus

Note: Based on 2,323,023 student-by-course observations. The x-axis is truncated at 200 miles to improve readability of the figure. 200 miles is the 97th percentile of the data.



Appendix Figure 3—Enrollment trends before and after a campus opening

Note: The square markers represent point estimates from a regression with campus-by-term observations. The dependent variable is the count of students enrolled in in-person courses. Students are assigned to the campus closest to the student's residence, even if that campus is not yet open. Thus the dependent variable counts enrollments for the assigned campus, not necessarily the campus where the student actually took the course. This enrollment count is regressed on a set of indicators for term relative to opening (leaving out the term just prior to opening, and existing campuses), campus FE, and calendar term FE (to control for university-wide trends). The circle markers repeat this process but for total enrollment: online plus in-person. They gray lines are 95% CIs.



Appendix Figure 4— Effect of taking a course online, instead of in-person, with the sample restricted to students within a certain distance of campus

Note: Each point is the local average treatment effect from a separate two-stage least squares regression estimated using students who live within the given (x-axis) distance of campus. The dependent variable is course grade. The specification includes one endogenous treatment variable, an indicator = 1 if the student took the course online. The excluded instrument is the interaction between (a) an indicator variable = 1 if the course was offered in-person at the student's home campus (defined as the nearest campus) and (b) the distance in miles from the student's home address to her home campus. All specifications include the main effects of (a) and (b). All specifications also include controls for gender, age, and separate fixed effects for course, term, home campus, and major. The bars show 95 percent confidence intervals which allow for clustering within campuses.

APPENDIX TABLE 1—STUDENT CHARACTERISTICS AND INSTITUTION TYPES (PERCENTAGES)

	Public 4-year	Private non-profit 4-year	Public 2-year	Private for-profit	The university we study
	(1)	(2)	(3)	(4)	(5)
Age					
23 or younger	69.6	71.2	49.1	31.6	28.1
24-39	24.8	19.0	36.4	50.7	52.5
40+	5.6	9.9	14.4	17.7	19.4
Female	53.9	56.6	55.7	64.1	44.5
Race/Ethnicity					
African-American	12.8	13.4	16.4	25.6	25.5
Asian	6.9	6.9	5.0	2.9	4.9
Hispanic or Latino	13.8	10.1	18.6	18.5	18.7
Other, More than one	4.4	4.5	4.3	4.6	4.8
White	62.2	65.1	55.8	48.5	46.1

Note: Columns 1-4 come from the 2011-12 National Postsecondary Student Aid Study (NPSAS:12) as reported by NCES QuickStats. Column 5 is from the university's annual report for 2012-13. Race/ethnicity is imputed for 16.9 percent of students assuming missing at random.

APPENDIX TABLE 2—COVARIATE BALANCE

	Dependent variable								
			Prior GPA		Number prior				
	Female	Age	online courses	in-person courses	online courses	in-person courses	Repeating course	Moved since last course	Terms since last course
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
A. With controls									
Took course online	0.018	-0.941	0.004	0.170	0.836	0.737	-0.010	0.001	-0.034
	(0.019)	(0.519)	(0.082)	(0.194)	(0.432)	(0.581)	(0.006)	(0.002)	(0.024)
B. Without controls									
Took course online	0.079	0.282	0.033	0.031	0.189	-1.291	0.022	0.004	0.011
	(0.038)	(0.885)	(0.105)	(0.236)	(0.822)	(1.026)	(0.007)	(0.002)	(0.017)
Observations	2,323,023	2,323,023	1,446,162	1,144,734	2,244,199	2,244,199	2,323,023	2,323,023	2,323,023

Note: Each column within panels reports the coefficient from a separate two-stage least squares regression. Dependent variables are described in the column headers. In Panel A, the estimation procedure is identical to that described in the note for Table 3 except that when "Female" is the outcome variable it is removed from the right hand side controls. The same is true for the other dependent variables. When a prior GPA variable is the outcome, all prior GPA variables are removed from the right hand side. Similarly, when number of prior courses is the outcome, all prior GPA controls are removed; some of the prior GPA controls are transformations of number of courses. In Panel B, estimates come from a simple 2SLS regression including no controls except the main effects of distance and the indicator for the course being offered. Standard errors allow for clustering within campuses.

APPENDIX TABLE 3—OLS ESTIMATES OF THE 'EFFECT' OF TAKING A COURSE ONLINE, INSTEAD OF IN-PERSON, ON STUDENT ACHIEVEMENT AND PERSISTENCE

	Dependent variable			
	Course grade (A=4...F=0)	GPA next semester	Enrolled next semester	Enrolled one year later
	(1)	(2)	(3)	(4)
Took course online	-0.381	-0.120	-0.053	-0.051
	(0.012)	(0.007)	(0.002)	(0.003)
Observations	2,323,023	2,106,090	2,360,645	2,360,645

Note: Each column reports the OLS coefficient from a separate regression. Dependent variables are described in the column headers. The specification includes one 'treatment' variable, an indicator = 1 if the student took the course online. All specifications also include controls for distance to home campus, availability of an in-person section, prior GPA, gender, age, and separate fixed effects for course, term, home campus, and major. Standard errors allow for clustering within campuses.

APPENDIX TABLE 4—EFFECT OF TAKING A COURSE ONLINE, INSTEAD OF IN-PERSON, ON STUDENT ACHIEVEMENT AND PERSISTENCE (LOCAL AVERAGE TREATMENT EFFECT) ADDITIONAL OUTCOMES

	Dependent variable			
	Course grade A or higher	Course grade B or higher	Course grade C or higher	Course grade D or higher (Passed course)
	(1)	(2)	(3)	(4)
Took course online	-0.122	-0.135	-0.101	-0.085
	(0.013)	(0.014)	(0.014)	(0.014)
Sample mean dep. var.	0.405	0.693	0.831	0.884
Observations	2,323,023	2,323,023	2,323,023	2,323,023

	Dependent variable		
	Withdrew from course	Credits next semester	Credits one year later
	(5)	(6)	(7)
Took course online	0.066	-0.623	-1.254
	(0.009)	(0.194)	(0.257)
Sample mean (st. dev.) for dep. var.	0.092	9.764	7.737
		(4.657)	(5.642)
Observations	2,601,742	1,980,377	1,520,954

Note. Each column, within panels, reports estimates from a separate two-stage least squares regression. The estimation procedure is described in the note for Table 3, only the dependent variables are different. Column 6 (7) is conditional on enrolling next semester (one year later). Standard errors allow for clustering within campuses.

APPENDIX TABLE 5—TREATMENT EFFECT HETEROGENEITY

	Dependent variable			
	Course grade (A=4...F=0)	GPA next semester	Enrolled next semester	Enrolled one year later
	(1)	(2)	(3)	(4)
A. By prior achievement				
Took course online	-0.363	-0.102	-0.061	-0.065
	(0.045)	(0.053)	(0.012)	(0.015)
Online * prior GPA	0.538	0.495	0.143	0.167
	(0.021)	(0.026)	(0.004)	(0.005)
Observations	2,323,023	2,106,090	2,360,645	2,360,645
B. Business-related majors				
Took course online	-0.446	-0.145	-0.068	-0.092
	(0.045)	(0.052)	(0.014)	(0.016)
Observations	1,007,534	900,335	1,006,664	1,006,664
C. Technology-related majors				
Took course online	-0.314	-0.083	-0.094	-0.094
	(0.068)	(0.087)	(0.018)	(0.022)
Observations	880,811	797,795	877,484	877,484
D. Health-related majors				
Took course online	-0.875	-0.267	-0.184	-0.095
	(0.170)	(0.173)	(0.047)	(0.068)
Observations	371,605	348,615	407,444	407,444
E. Courses required for the student's major				
Took course online	-0.523	-0.262	-0.053	-0.083
	(0.072)	(0.070)	(0.014)	(0.021)
Observations	1,072,736	987,179	1,070,763	1,070,763
F. Introductory and intermediate courses (below 300 level)				
Took course online	-0.391	-0.076	-0.094	-0.077
	(0.066)	(0.075)	(0.018)	(0.023)
Observations	1,460,463	1,325,177	1,511,154	1,511,154
G. Advanced courses (300 level or higher)				
Took course online	-0.470	-0.190	-0.029	-0.041
	(0.053)	(0.056)	(0.013)	(0.021)
Observations	862,560	780,913	849,491	849,491

Note. Each column, within panels, reports estimates from a separate two-stage least squares regression. The note for Table 3 describes the estimation procedure; however, each panel makes one change to the procedure. Panel A adds an endogenous variable: the interaction between taking the course online and prior GPA (measured in all courses). The main effect of taking a course online and the new interaction are instrumented for with two instruments: the main offered*distance instrument, and the interaction between the main instrument and prior GPA. In Panels B-D the estimation sample is restricted to students in each category of majors. In Panels E-G the estimation sample is restricted by type of course as described in the panel labels. Standard errors allow for clustering within campuses.

Appendix B. Characterizing and Estimating the Bias from Missing Data on "Never Takers"

A. Characterizing the Bias

Observations for "never takers" are missing in our data. The "never takers" in this setting are students who are unwilling to take course c online in term t. These "never takers" would prefer to take course c during term t, but will only do so if there is an in-person class at their home campus b. Thus the "never takers" will be observed in our data only when $Offered_{b(i)ct} = 1$, and missing from the data when $Offered_{b(i)ct} = 0$. By contrast, "always takers" and "compliers" are never missing from our data because courses are always offered online.

To simplify the notation, let (i) Y be the student outcome variable of interest, (ii) T be the treatment indicator called $Online_{ict}$ in Equation 1, (iii) Z be the instrument $\left(Offered_{b(i)ct} \times Distance_i\right)$, and (iv) W be a vector of all the remaining right hand side covariates included in our 2SLS first- and second-stages, including the main effects for $Offered_{b(i)ct}$ and $Distance_i$. Further let (v) m be an indicator = 1 if the observation is missing from our data as described above, and (vi) $\rho = \mathrm{E}[m = 1]$, the probability of being missing.

We can write the true effect of interest, δ in Equation 1, as a ratio of conditional covariances.

$$\delta = \frac{cov(Y, Z|W)}{cov(T, Z|W)} = \frac{\mathrm{E}[YZ|W] - \mathrm{E}[Y|W]\mathrm{E}[Z|W]}{\mathrm{E}[TZ|W] - \mathrm{E}[T|W]\mathrm{E}[Z|W]} \tag{B1}$$

From here on we drop the $|W$ notation to simplify, but $|W$ should be thought of as implicit in all (co)variances and expectations below.

We can also write δ as a function of weighted sums of expectations. In particular, the numerator in B1 can be written

$$\{\rho \mathrm{E}[YZ|m=1] + (1-\rho)\mathrm{E}[YZ|m=0]\} - \{\rho \mathrm{E}[Y|m=1] + (1-\rho)\mathrm{E}[Y|m=0]\} \times \{\rho \mathrm{E}[Z|m=1] + (1-\rho)\mathrm{E}[Z|m=0]\}. \tag{B2}$$

To simplify B2 first recall that the missing observations, $m = 1$, are missing because $Offered_{b(i)ct} = 0$. Thus when $m = 1$ it will always be the case that $Z = (0 \times Distance_i) = 0$. With this fact and a little algebra we can simplify B2 to

$$(1-\rho)(\mathrm{E}[YZ|m=0] - \{\rho \mathrm{E}[Y|m=1] + (1-\rho)\mathrm{E}[Y|m=0]\}\mathrm{E}[Z|m=0]). \tag{B3}$$

The denominator in B1 can be simplified the same way by replacing Y with T. Thus we can write the true effect of interest, δ, as

$$\delta = \frac{\mathrm{E}[YZ|m=0] - \{\rho \mathrm{E}[Y|m=1] + (1-\rho)\mathrm{E}[Y|m=0]\}\mathrm{E}[Z|m=0]}{\mathrm{E}[TZ|m=0] - \{\rho \mathrm{E}[T|m=1] + (1-\rho)\mathrm{E}[T|m=0]\}\mathrm{E}[Z|m=0]}. \tag{B4}$$

Contrast B4 with our empirical estimate $\hat{\delta}$ which is

$$\hat{\delta} = \frac{\mathrm{E}[YZ|m=0] - \mathrm{E}[Y|m=0]\mathrm{E}[Z|m=0]}{\mathrm{E}[TZ|m=0] - \mathrm{E}[T|m=0]\mathrm{E}[Z|m=0]}. \tag{B5}$$

Subtracting the true numerator in B4 from our estimate of the numerator in B5 leaves

$$\rho\{\mathrm{E}[Y|m=1] - \mathrm{E}[Y|m=0]\}. \tag{B6}$$

Notice, first, that the missing data bias in the numerator will be proportional to ρ, the share of "never takers". Second, that the numerator's bias will be positive if $\mathrm{E}[Y|m=1] > \mathrm{E}[Y|m=0]$ and negative if the inequality is reversed. Assuming taking a class online, instead of in-person, has a negative effect on student outcomes, then positive bias would mean our estimates, $\hat{\delta}$, understate the true negative effects of online classes.

Similarly, subtracting the true denominator in B4 from our estimate of the denominator in B5 leaves

$$-\rho \mathrm{E}[T|m = 0]. \tag{B7}$$

Expression B7 parallels B6, but further simplifies by noting that $\mathrm{E}[T|m = 1] = 0$. The missing observations, $m = 1$, are all "never takers" where $Online_{ict} = 0$ in all cases by definition. Again, first, notice that the denominator bias is proportional to ρ. Second, the denominator bias will always be negative; that is, the estimated denominator is too small relative to the truth. This negative bias would mean our estimates, $\hat{\delta}$, are too large in absolute value. Put differently, the denominator bias makes the first-stage too small, leading us to scale-up the reduced-form too much.

To summarize, first, the missing data bias is proportional to ρ. Second, our estimates will overstate the negative effects of taking a class online, instead of in-person, if $\mathrm{E}[Y|m = 1] < \mathrm{E}[Y|m = 0]$. If the inequality goes the other direction $\mathrm{E}[Y|m = 1] > \mathrm{E}[Y|m = 0]$ then the direction of the bias is ambiguous; bias in the numerator will understate the effects, but bias in the denominator will overstate the effects. The empirical tests discussed below and in section III.b suggest that potential bias created by "never takers" likely leads to a small underestimation of the effects on online courses.

B. Estimating the Bias

The expressions above provide one framework for estimating the potential bias from missing data. That estimate of bias itself requires a few input estimates. The first input is an estimate of ρ, the proportion of observations missing from our data because some students will never take a course online. A simple estimate of ρ could be obtained by fitting the specification

$$T_{ibct} = \rho * Offered_{b(i)ct} + \vartheta_{b(i)c} + \varepsilon_{ibct}, \tag{B8}$$

where T_{ict} is an indicator $= 1$ if student i whose home campus is b took course c, either online or in-person, during term t. The term $\vartheta_{b(i)c}$ is a campus-by-course fixed effect. Fitting B8 in our setting would be computationally intensive given the number of students, courses, and terms in our data. As an alternative, we (i) aggregate our data to course-by-campus-by-term observations, summing T_{ibct} over all i to get the total enrollment for course c in term t of students whose home campus is b; and then (ii) regress the log of total enrollment on the indicator variable $Offered_{bct} = 1$ if the course was offered in-person at campus b in term t. The aggregated regression also includes campus-by-course fixed effects. The resulting estimate is $\hat{\rho} = 0.328$. Adding controls for a non-parametric time trends in enrollment by course (course-by-term dummies) and by campus (campus-by-term dummies) does not substantially change the estimate ($\hat{\rho} = 0.314$).

The second input is an estimate of $\{\mathrm{E}[Y|m = 1] - \mathrm{E}[Y|m = 0]\}$, the difference in average course grade between missing and non-missing observations. By definition $\mathrm{E}[Y|m = 1]$ is unobserved; but we estimate $\mathrm{E}[Y|m = 1]$ using subsamples of observed data – subsamples which likely include many "never takers". First, we calculate the average grade for students in term t who have not taken an online course in a prior term $\tau < t$. Second, we calculate the average grade for students who *never* take an online course in any term in our data. The resulting estimates of $\{\mathrm{E}[Y|m = 1] - \mathrm{E}[Y|m = 0]\}$ are 0.013 and 0.315 respectively. These estimates are conditional on W. The unconditional differences are 0.110 and 0.403 respectively. Similarly, we also need an estimate of $\{\mathrm{E}[T|m = 1] - \mathrm{E}[T|m = 0]\}$. Using the same subsample strategy, the estimates are -0.338 and -0.418, respectively, conditional on W. Unconditionally $\mathrm{E}[T|m = 1] = 0$ and $\mathrm{E}[T|m = 0]$ is observed directly $= 0.591$.

Appendix Table B1 shows what the estimated "true" effect would be after adjusting for the bias described above. All of the calculations use $\hat{\rho} = 0.328$ and then different values of $\{\mathrm{E}[Y|m = 1] - \mathrm{E}[Y|m = 0]\}$ and $\{\mathrm{E}[T|m = 1] - \mathrm{E}[T|m = 0]\}$. The estimated "true" effects range between -0.374 and -0.512, suggesting our paper's main estimate of -0.440 may understate the negative effects by as much as 17 percent or overstate the effects by as much as 15 percent. The estimated "true" effects all fall within the 95 percent confidence interval of our paper's main estimate.

APPENDIX TABLE B1—EFFECT ESTIMATES FOR COURSE GRADE ADJUSTED FOR ESTIMATED BIAS FROM MISSING DATA; PERCENT DIFFERENT FROM MAIN ESTIMATE IN PARENTHESIS
ALL ESTIMATES USE $\hat{\rho} = 0.328$

Estimate of $\{\mathrm{E}[T\|m = 1] - \mathrm{E}[T\|m = 0]\}$	Estimate of $\{\mathrm{E}[Y\|m = 1] - \mathrm{E}[Y\|m = 0]\}$	
	0.013	0.403
-0.338	-0.401	-0.512
	(-8.83)	(16.58)
-0.591	-0.374	-0.478
	(-14.99)	(8.70)