Journal of Behavioral and Experimental Economics 90 (2021) 101653

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

Can we stay one step ahead of cheaters? A field experiment in proctoring online open book exams

Jose J. Vazquez[a,*], Eric P. Chiang[b], Ignacio Sarmiento-Barbieri[c]

[a] Department of Economics, University of Illinois, Urbana-Champaign, Urbana, IL 61801, (217) 419-3518
[b] Department of Economics, Florida Atlantic University, Boca Raton, FL 33431, (561) 297-2947
[c] Department of Economics, Universidad de Los Andes, 111711 Bogota, Colombia, (571) 332-4493

ARTICLE INFO

JEL Codes:
A20
I23
A22

Keywords:
Academic integrity
Cheating
Online learning
Personality traits

ABSTRACT

As more institutions of higher learning expand their offerings of online courses, the use of online assessments has become an important topic of discussion. Although the use of online assessments can be very beneficial, instances of cheating in the absence of a proctor poses a cost in protecting academic integrity. This has led to the development of many proctoring solutions to address this challenge. This paper presents two field experiments used to analyze the effects of proctoring methods on exam scores: one involving a face-to-face class and the other involving an online class. Also, two proctoring methods were used: live proctors and web-based proctors. In each class, best practices were used to minimize cheating and students were informed in advance which exams were proctored. Our results show that students whose exams were not proctored scored over 11% higher on average than those whose exams were proctored. However, the results varied significantly: the use of live proctors in the face-to-face class had a much larger effect on test scores than web-based proctors in the online class. We compare variables affecting each testing environment to uncover possible determinants, including the ease of collaboration, test anxiety, and information sharing over the testing period.

1. Introduction

Online assessments offer instructors several key advantages over comparable traditional in-class exams, including lower administrative costs, greater variety of multimedia assessment tools, convenience for students, and faster analysis of results. But the main hesitation by instructors for implementing major exams online is the potential risk of academic dishonesty.

Many instructors believe that online assessments offer an invitation for students to cheat, and therefore avoid these exam formats due to a belief that excessive cheating may occur. Although anecdotal evidence suggests that cheating is pervasive in online exams, the literature has largely focused on traditional in-class exams as opposed to cheating in an online setting. This is especially relevant as institutions of higher learning dramatically expand their online course and degree offerings.

In this paper, we analyze cheating behavior in online exams by conducting two field experiments. Both are randomized controlled trials in large enrollment microeconomic principles classes in which students completed all exams online in a learning management system, one of which must be proctored. The first experiment took place in a face-to-face class where students took their proctored exam using their own laptop but in the presence of a live proctor in the classroom. The second experiment took place in a fully online class where students took their proctored exam in any location using a web-based proctor. The exam to be proctored was randomized and students were informed in advance.

Our results show that students who took exams without supervision scored over 11% higher on average than those taking the exam under supervision. However, we find that the difference in scores between proctored and non-proctored exams is much smaller when a web-based proctoring tool was used instead of an in-class live proctor. We explore several mechanisms that could potentially explain our results, including test anxiety, information sharing, and collaboration on exams. Although our data do not rule out the possibility of anxiety explaining these results, we present evidence that collaboration among peers is the primary explanation for the differences in scores between proctored and non-proctored exams. Therefore, in our setting in which students have full access to course materials during exams, collaboration is how students cheat despite rules prohibiting these actions during exams.

Moreover, we found larger proctoring effects on performance among lower achieving students and among underclass students, but

* Corresponding author.
E-mail addresses: vazquezj@illinois.edu (J.J. Vazquez), chiang@fau.edu (E.P. Chiang), i.sarmiento@uniandes.edu.co (I. Sarmiento-Barbieri).

https://doi.org/10.1016/j.socec.2020.101653
Received 10 May 2020; Received in revised form 17 August 2020; Accepted 27 November 2020
Available online 2 December 2020
2214-8043/

insignificant differences based on gender. The negative relationship between cheating and measures of student ability builds upon recent experimental research on student cheating by Yaniv et al (2017) by studying the costs and benefits students place on external rewards (e.g., grades) versus internal rewards (e.g., desire to achieve answers).

This paper explores various factors and student characteristics to gain insights into the nature of cheating behavior in online exams. The remainder of the paper is as follows: Section 2 reviews the literature on online testing and cheating behavior. Section 3 describes the field experiment design. Section 4 presents the empirical model. Section 5 presents the baseline results, while Section 6 examines the heterogeneous responses by student characteristics. Section 7 concludes and describes extensions for further research.

2. Background on online assessments

The implementation of online assessments became a necessity as institutions expanded distance learning and other online-enhanced courses over the past two decades. Even prior to the COVID-19 crisis in early 2020 that forced nearly all education to be moved online, most institutions of higher learning had already invested in tools and training to offer more courses and degrees online. Part of this phenomenon is explained by increased competition for students and space restrictions on campus. Moreover, attitudes toward online learning have become more favorable and learning outcomes between online and face-to-face formats have become increasingly comparable as online learning tools improve (see Bosshardt and Chiang, 2016; Figlio et al, 2013; and Means et al, 2010, for an overview of online versus face-to-face learning). Still, one issue that has not been studied in great detail is the role of online assessments in facilitating academic dishonesty.

Academic dishonesty has been widely studied in a variety of contexts, but primarily in traditional in-class settings. For example, Nowell and Laufer (1997), Kerkvliet and Sigmund (1999), and McCabe (2005) studied the factors affecting cheating among economics and business students, and Ward and Beck (1990) used a field experiment to study the relationship between gender and cheating. But these studies were conducted prior to the widespread use of online exams. The focus on cheating during online exams has received greater attention over the past decade (see King et al, 2009 and Harton et al, 2019).

Much of the existing studies on cheating in an online setting use either surveys in which students self-report their own actions and/or opinions on cheating, or collect experimental data based on a game or activity instead of an actual assessment. Survey-based studies allow researchers to gain insights into how students perceive cheating in online courses and on methods to minimize cheating, such as Watters et al. (2011) which focused on business courses at public universities in Texas and Arkansas. Similarly, Grijalva et al. (2006) used a randomized response survey to predict the probability that a student will cheat on an exam based on specific factors such as a student's expected grade, penalty for cheating, and whether a proctor was used.

Studies that rely on self-reported measures to detect cheating rely on the assumption that students who cheat on exams will then be honest in their survey responses, which is questionable. An alternative approach used in behavioral studies focuses on explaining trends in dishonest behavior based on individual actions instead of survey responses. Many of these studies use experiments to expose cheating behavior based on actual actions. For example, Schwieren and Weichselbaumer (2010) showed that competitive pressures lead to more cheating, Akin (2019) studied how individuals sort themselves into environments which then influences dishonest behavior, and Belot and van de Ven (2019) found that dishonest behavior depends more on present incentives than on past incentives, concluding that dishonesty is not persistent.

In a study by Yaniv et al (2017), students participated in a trivia quiz twice over two months, once with cheating overlooked and once monitored. They found that higher achievers actually cheat more, which contradicts many survey-based studies that argue that lower achievers cheat more because they have more to gain. An important consideration that is not captured in their experimental design is the role of opportunity cost, which becomes more evident in a real exam setting. Higher-achieving students are more efficient at studying and potentially have more to lose (e.g., merit scholarships, admission to graduate school, etc.) if caught cheating, which translates to a higher opportunity cost of cheating. In a study involving a real math competition among children in which the first stage was unmonitored while the second stage was monitored, Azar and Applebaum (2020) compared scores between the two stages and found that younger children, those who attended secular schools, and those with higher socioeconomic status were more likely to be dishonest. They also found that, contrary to popular belief, gender did not influence cheating behavior.

Harmon and Lambrinos (2008) used an experimental approach in analyzing the effect of proctoring using two online macroeconomics class. In one class, students took the final exam online without a proctor over a three-day period, while in the other students took the final exam with a proctor at a set time and location. Both exams were open book, open notes, and consisted of multiple-choice questions randomly selected from a test bank. They compared the fit of a linear regression of exam scores on GPA by studying the variations in the R-squared values between samples of proctored and non-proctored exams. This approach assumes that the variation in test scores that is not explained by human capital measures can be attributed to cheating. The limitation of this approach is the potential of selection bias since there was no randomization in the assignment of students to each class. We address this concern by presenting, to our knowledge, the first large-scale randomized trial of the effect of proctoring using online exams.

Levitt and Lin (2019) used a different and more direct approach in studying the effects of proctoring and the detection of cheating. Their strategy consisted of comparing correct and incorrect answers in a closed book final exam of student pairs based on their seating arrangement. They found that students who chose who they could sit next to were more likely to collaborate than students who were randomly placed next to one another. The authors found that the use of randomized seating along with proctoring significantly reduced cheating. A cost that arises from this strategy is the implementation of a seating chart, which can be time consuming in large classes.

Experimental studies to date have not adequately addressed student behavior when different proctoring methods and other prevention strategies are used to minimize cheating. The present study attempts to address these limitations by studying cheating using field experiments in both face-to-face and online classes. Over the past decade, proctoring technology has evolved significantly, offering more options at lower costs that utilize both human labor as well as algorithms using artificial intelligence to detect dishonest behavior during exams. Comparing the effect of proctors on cheating behavior and its impact on student grades are issues needing greater attention. We pursue these objectives in this paper using data collected from field experiments as opposed to survey instruments or laboratory experiments. We build on the limited studies that have used experimental data to evaluate the effects of proctoring methods in specific controlled settings.

3. Experiment design

This study was conducted in two large-enrollment sections of microeconomic principles at a large public university in Illinois. One was a face-to-face class that took place in spring 2017 and the other was an online class in the prior winter term. The two classes covered the same content but differed in the length of the term, number of students enrolled, and number of exams.

The face-to-face class lasted 16 weeks. Students were required to take three non-cumulative midterm exams (one of which is proctored) and one proctored cumulative final exam. To determine which midterm exam would be proctored, students were randomly placed into one of three groups: proctor exam 1 (PE1), proctor exam 2 (PE2), and proctor

Table 1
Descriptive statistics (face-to-face class).

	All (1)	PE1 (2)	PE2 (3)	PE3 (4)	Pr(>F) (5)
Age	18.880	18.809	18.881	18.942	0.550
	(0.050)	(0.089)	(0.089)	(0.083)	
GPA PreCourse	3.424	3.434	3.423	3.417	0.954
	(0.024)	(0.042)	(0.042)	(0.039)	
GPA PostCourse	3.426	3.463	3.402	3.414	0.520
	(0.022)	(0.040)	(0.040)	(0.037)	
ACT	29.077	28.752	29.165	29.287	0.417
	(0.172)	(0.305)	(0.305)	(0.287)	
Male	0.486	0.412	0.474	0.561	0.037
	(0.024)	(0.043)	(0.043)	(0.040)	
White	0.432	0.515	0.444	0.348	0.016**
	(0.024)	(0.042)	(0.042)	(0.040)	
Freshman	0.392	0.471	0.348	0.361	0.073*
	(0.024)	(0.042)	(0.042)	(0.039)	
Sophomore	0.413	0.368	0.437	0.432	0.427
	(0.024)	(0.042)	(0.042)	(0.040)	
Junior	0.131	0.125	0.141	0.129	0.924
	(0.016)	(0.029)	(0.029)	(0.027)	
Senior	0.061	0.029	0.074	0.077	0.175
	(0.012)	(0.021)	(0.021)	(0.019)	
Final Exam	89.884	90.025	89.400	90.181	0.661
	(0.368)	(0.653)	(0.655)	(0.611)	
Course Total	90.799	91.214	90.689	90.530	0.700
	(0.344)	(0.610)	(0.613)	(0.572)	
Observations	426	136	135	155	

* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

exam 3 (PE3), and were informed in advance which group they were in. Each midterm exam was 50 minutes in length and administered online using the campus LMS system (Moodle). Therefore, all students took the same online exam, with the only difference being that one group had to take their online exams in the presence of a proctor in a classroom. Students assigned to the proctored exam took the exam in class at a specific day and time using their own laptops, while those assigned to the non-proctored exam took the exam online at a time of their choosing within a 30-hour window. Since the time limit for each exam is much shorter than the 30-hour window of availability, cheating could occur if students who completed the exam earlier chose to share exam information with those who took it later. We investigate this possibility in our results.

The online class lasted four weeks. There were two non-cumulative midterm online exams, one of which was proctored using ProctorU, a third-party web-based proctoring service in which students are monitored in real time via their webcams. Students were randomly placed into two groups: proctor exam 1 (PE1) and proctor exam 2 (PE2). The time limit for each exam was 180 minutes, and students could choose the time of their exam within a 30-hour window. Similar to the face-to-face class, information sharing is a possible method of cheating.

Exams in both classes had the same structure. They consisted of multiple-choice questions presented in random order for each student. Exams in the face-to-face class had 25 questions while exams in the online class had 40 questions. Exams (including those that were proctored) were open-book, allowing students to use their notes, textbook, and the Internet. However, they were warned that the exam was to be completed independently and any collaboration was not allowed. Before the start of each exam, students must read a set of instructions which included the following stern message: “You may NOT collaborate with any other person on this exam using any means (including the use of phones, email, or social media applications). Also, you may NOT discuss the questions in this exam with any other student until after the exam window closes.”

One of the important conditions in our study is that all students had access to the book and Internet while taking the exam. We did not study whether students cheat by searching for answers online because all students had the same access. However, we acknowledge that cheating

Table 2
Descriptive statistics (online class).

	All (1)	PE1 (2)	PE2 (3)	Pr(>F) (4)
Age	19.482	19.366	19.595	0.435
	(0.146)	(0.208)	(0.205)	
GPA PreCourse	3.385	3.380	3.389	0.938
	(0.054)	(0.077)	(0.077)	
GPA PostCourse	3.388	3.329	3.447	0.370
	(0.065)	(0.093)	(0.092)	
ACT	28.960	28.947	28.973	0.977
	(0.440)	(0.623)	(0.631)	
Male	0.518	0.512	0.524	0.917
	(0.055)	(0.079)	(0.078)	
White	0.398	0.439	0.357	0.452
	(0.054)	(0.077)	(0.076)	
Freshman	0.181	0.220	0.143	0.370
	(0.042)	(0.061)	(0.060)	
Sophomore	0.373	0.341	0.405	0.557
	(0.053)	(0.076)	(0.075)	
Junior	0.193	0.171	0.214	0.620
	(0.044)	(0.062)	(0.062)	
Senior	0.253	0.268	0.238	0.755
	(0.048)	(0.069)	(0.068)	
Course Total	92.190	90.831	93.518	0.112
	(0.844)	(1.189)	(1.175)	
Observations	83	41	42	

* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

through the use of Internet searches for exams that do not allow students to access these resources would be a valuable extension to our study.

Following IRB regulations, we asked students for consent to use their university academic and demographic records. Of the 863 registered students in the face-to-face class, 423 students (49%) gave consent. Of the 112 students in the online class, 72 (64%) gave consent. Tables 1 and 2 report descriptive statistics for the academic and demographic records obtained from the university registrar. We also checked for the randomization balance in these tables by presenting p-values for the test of the alternative hypothesis that at least one mean is different across treatment groups. Tables 1 and 2 confirm that the randomization was successful. With the exception of the variables *White* and *Freshman* in the face-to-face class, there were no statistically significant differences across observable student characteristics in either class.

4. Empirical model

To identify the average causal effects of taking the exam with a proctor, we estimate the following fixed effects model for the overall sample:

$$Score_{ij} = \beta_0 + \beta_1 Proctor_i + \theta X_i + \delta_j + u_{ij} \tag{1}$$

where $Score_{ij}$ is the exam score for student i in exam j. $Proctor_i$ is an indicator variable that takes the value of 1 if the student took the exam with a proctor. Although we have a randomized setup, we address the potential concern that student performance on an exam may be correlated with the presence of a proctor (i.e., the presence of a proctor may generate anxiety on the test taker). We use a vector X_i of student-level controls that include demographic characteristics (age, gender, ethnicity), class standing, and proxies for student ability (GPA before the course and ACT composite score). δ_j is an exam and class fixed effect that ensures that our comparisons are within exam and class between students who took exam j with and without a proctor and thus controlling for exams of different time lengths and number of questions. Finally, u_{ij} is the usual error term.

To provide further comparisons between the two classes, we estimate separate regressions for each midterm exam in each class using the following model:

$$Score_i = \beta_0 + \beta_1 Proctor_i + \theta X_i + u_i \tag{2}$$

Table 3
Mean score by exam and proctoring condition.

	All (1)	Proctored (2)	Non-proctored (3)	Diff (4)
Face-to-Face Class				
Exam 1	76.653	71.862	82.979	-11.117***
	(14.964)	(1.195)	(1.373)	(1.821)
Exam 2	82.783	76.119	87.350	-11.232***
	(15.312)	(1.231)	(1.019)	(1.598)
Exam 3	79.715	76.119	87.350	-15.307***
	(19.059)	(1.231)	(1.019)	(2.099)
Online Class				
Exam 1	75.783	73.232	78.274	-5.042*
	(11.288)	(1.728)	(1.708)	(2.430)
Exam 2	82.741	83.155	82.317	0.838
	(12.006)	(1.863)	(1.885)	(2.650)

Notes: Standard errors in parentheses.
* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

Table 4
Effect of proctor on exam score: pooled OLS estimates.

	Dependent Variable: Exam Score		
	(1)	(2)	(3)
Proctor	-11.064***		
	(2.044)		
Proctor x F2F		-12.679***	
		(1.782)	
Proctor x Online		-2.230*	
		(0.901)	
Proctor x F2F x Exam 1			-11.399***
			(0.376)
Proctor x F2F x Exam 2			-10.289***
			(0.349)
Proctor x F2F x Exam 3			-16.502***
			(0.298)
Proctor x Online x Exam 1			-4.146**
			(1.334)
Proctor x Online x Exam 2			-0.314
			(1.334)
Age	0.328	0.327	0.355
	(0.708)	(0.665)	(0.629)
Male	2.236	2.230	2.262
	(1.187)	(1.202)	(1.236)
White	-2.320**	-2.325**	-2.505**
	(0.730)	(0.703)	(0.612)
Freshman	1.416	1.393	1.306
	(2.493)	(2.484)	(2.467)
Sophomore	0.840	0.824	0.716
	(1.937)	(1.884)	(1.899)
Junior	-0.151	-0.160	-0.342
	(1.496)	(1.436)	(1.424)
GPA PreCourse	10.405***	10.408***	10.520***
	(0.940)	(0.969)	(1.014)
ACT	1.181***	1.182***	1.183***
	(0.086)	(0.090)	(0.092)
Observations	974	974	974
R^2	0.358	0.371	0.378

Notes: Robust standard errors in parentheses.
* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

where the variables are the same as in equation (1) except fixed effects are not included.

5. Results

We begin by exploring the effects of taking the midterm exam with a proctor on the average exam score itself. Table 3 presents the mean score by proctoring condition. Students who took the exam with a live proctor on Exam 1 scored on average 11.1% lower compared to students who took the exam without a proctor. This difference increased to 11.2% on Exam 2 and 15.3% on Exam 3. In the online class, which used a web-

Table 5
Effect of proctor on exam score: OLS estimates (face-to-face class).

	Dependent Variable: Exam Score		
	Exam 1 (1)	Exam 2 (2)	Exam 3 (3)
Proctor	-11.609***	-10.198***	-16.302***
	(1.616)	(1.490)	(1.735)
Age	1.530*	0.794	-1.415
	(0.843)	(0.943)	(1.035)
Male	1.619	-0.303	3.975**
	(1.606)	(1.457)	(1.825)
White	-2.237	-3.061**	-1.269
	(1.592)	(1.414)	(1.906)
Freshman	10.124**	2.580	0.408
	(5.088)	(4.068)	(4.098)
Sophomore	7.548	1.350	0.725
	(4.943)	(3.714)	(3.754)
Junior	3.312	0.457	0.190
	(5.007)	(3.628)	(3.729)
GPA PreCourse	9.070**	10.587***	12.425***
	(2.097)	(1.752)	(2.239)
ACT	1.318***	1.243***	1.211***
	(0.296)	(0.263)	(0.290)
Constant	-22.263	-1.078	35.573
	(22.435)	(24.902)	(24.830)
H0	Exam 1=Exam 2	Exam1=Exam3	Exam2=Exam3
F-stat	0.412	7.122	3.917
P-val	0.521	0.008	0.048
Observations	233	320	273
R^2	0.361	0.357	0.403
Adjusted R^2	0.336	0.339	0.383

Notes: Estimates come from estimating equation 1. Heteroskedasticity-robust standard errors in parentheses.
* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

Table 6
Effect of proctor on exam score: OLS estimates (online class).

	Dependent Variable: Exam Score	
	Exam 1 (1)	Exam 2 (2)
Proctor	-3.255	-1.138
	(2.086)	(2.346)
Age	-0.056	1.600
	(0.879)	(1.198)
Male	4.477**	4.657**
	(2.012)	(1.997)
White	-4.214**	-1.436
	(1.962)	(2.273)
Freshman	-9.360***	-2.394
	(3.105)	(4.222)
Sophomore	-3.459	-1.165
	(2.716)	(3.809)
Junior	-2.114	1.714
	(3.206)	(3.269)
GPA PreCourse	7.248***	9.474***
	(2.264)	(2.935)
ACT	0.860***	1.091***
	(0.291)	(0.324)
Constant	31.977	-12.499
	(24.690)	(29.762)
H0	Exam 1=Exam 2	
F-stat	0.455	
P-val	0.501	
Observations	74	74
R^2	0.426	0.383
Adjusted R^2	0.345	0.296

Notes: Estimates come from estimating equation 1. Heteroskedasticity-robust standard errors in parentheses.
* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

Table 7
Effect of proctor on exam score: by time of exam (face-to-face class).

	Proctored (1)	Earlier (2)	Later (3)	(1)-(2) (4)	(1)-(3) (5)
Exam 1					
Score	71.862	86.291	78.452	-14.429***	-6.590***
	(1.178)	(1.852)	(2.025)	(2.099)	(2.418)
GPA	3.434	3.487	3.362	-0.052	0.072
	(0.042)	(0.066)	(0.073)	(0.078)	(0.085)
ACT	28.752	28.611	28.413	0.141	0.339
	(0.295)	(0.463)	(0.502)	(0.542)	(0.598)
Exam 2					
Score	76.119	90.169	84.937	-14.050***	-8.818***
	(1.223)	(1.559)	(1.348)	(2.189)	(1.887)
GPA	3.423	3.513	3.413	-0.090	0.010
	(0.040)	(0.052)	(0.044)	(0.064)	(0.061)
ACT	29.165	29.253	29.303	-0.088	-0.137
	(0.306)	(0.397)	(0.338)	(0.500)	(0.453)
Exam 3					
Score	72.852	91.451	85.784	-18.599***	-12.932***
	(1.402)	(2.444)	(2.029)	(2.969)	(2.631)
GPA	3.417	3.418	3.331	-0.001	0.086
	(0.041)	(0.072)	(0.059)	(0.082)	(0.074)
ACT	29.287	29.120	29.082	0.167	0.204
	(0.294)	(0.510)	(0.422)	(0.583)	(0.519)

Notes: Standard errors in parentheses.
* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

based proctor, the difference between students who took the exam with a proctor and those who took it without a proctor was significantly smaller. Students who were proctored on Exam 1 scored 5% lower, while students who were proctored on Exam 2 scored 0.8% higher than those who took non-proctored exams.

Table 4 presents the baseline results according to the pooled OLS estimates of equation (1), with fixed effects. Column (1) shows the average effect of proctoring to significantly decrease exam scores across the two classes. Column (2) separates the results by class to show the proctoring effect being much greater in the face-to-face class, and Column (3) further separates the results by exam in each class. The effect of proctoring is shown to be insignificant only in the second exam in the online class.

There are different potential mechanisms that may be driving these results. Tables 5 and 6 present OLS estimates separately for each class based on equation (2). One important mechanism to consider is test anxiety. Previous research has shown that taking tests is a stressful task that may be exacerbated by the presence of a proctor, especially if the student is unfamiliar with the proctor (Derosa and Patalano, 1991). Taking an exam at home or in a more relaxed environment may lead to lower anxiety. In our experiment, the face-to-face class took their proctored exam with a familiar proctor (which should reduce anxiety) but in a less comfortable environment (which can increase anxiety), while the online class took their proctored exam using a less familiar web-based proctor (which can increase anxiety) but in a more comfortable environment of their choosing (which should reduce anxiety). If anxiety is the mechanism driving our results, we should see some differences in how each class performs on the proctored exam.

The results in Tables 5 and 6 support the possibility that anxiety could be affecting the scores between proctored and non-proctored exams. Specifically, the adjusted exam score differentials for the face-to-face class are significant at the 1% level, while they are not statistically significant at the conventional levels for the online class. Therefore, if anxiety played a role in reducing scores for the face-to-face class, it is likely the result of a more comfortable testing environment for online students. However, because we do not know where the online students took the exam (at home, at the library, at a coffee shop, etc.), the role of anxiety remains anecdotal at best. Azar and Applebaum (2020) argue that stress is not likely to cause lower scores in a proctored environment because students are accustomed to taking exams in classrooms, while test taking at home can be more distracting than in a classroom setting. Further research can better disentangle the role of anxiety in test performance.

Another mechanism that can affect our results is the 30-hour testing window which could allow students who take an exam earlier in the testing window to record and share information with students who take it later. To test this hypothesis, we divide our sample according to the time when students took the exam. Students in the face-to-face class who took the proctored exam all started at the same time, while those who took the exam without a proctor could take it anytime within the 30 hour window. Tables 7 and 8 present the average exam scores based on students who took the non-proctored exam earlier in the testing window (first 15 hours) compared to those who took it later in the testing window (second 15 hours).

Our data show that students who took an exam earlier mostly performed better than those who took it later. In the face-to-face class, students who took the non-proctored exam earlier scored higher in all three exams, and with greater significance in score differentials compared to those who took exams later. In the online class, students who completed the non-proctored exam earlier scored higher on the first exam, but about the same in the second exam. The score differential was insignificant for both exams among those who took the exam later. These results cast doubt on the mechanism that there was widespread sharing of exam information among students. The more likely explanation of the difference in scores based on time of the exam is selection.

Table 8
Effect of proctor on exam score: by time of exam (online class).

	Earlier		Later		(1)-(2)	(3)-(4)
	Proctored (1)	Non-proctored (2)	Proctored (3)	Non-proctored (4)	(5)	(6)
Exam 1						
Score	68.421	84.50	77.386	76.328	-16.079***	1.058
	(2.393)	(3.298)	(2.224)	(1.844)	(3.979)	(2.923)
GPA	3.313	3.391	3.435	3.388	-0.078	0.047
	(0.115)	(0.173)	(0.104)	(0.086)	(0.206)	(0.136)
ACT	27.941	29.875	29.762	28.724	-1.934	1.038
	(0.927)	(1.351)	(0.834)	(0.710)	(1.548)	(1.123)
Exam 2						
Score	80.600	81.618	86.912	82.813	-1.018	4.099
	(2.401)	(2.912)	(2.912)	(2.451)	(4.362)	(3.082)
GPA	3.347	3.339	3.446	3.410	0.007	0.035
	(0.102)	(0.118)	(0.118)	(0.102)	(0.158)	(0.154)
ACT	29.238	29.067	28.625	28.870	0.171	-0.245
	(0.848)	(1.004)	(0.972)	(0.811)	(1.402)	(1.183)

Notes: Standard errors in parentheses.
* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

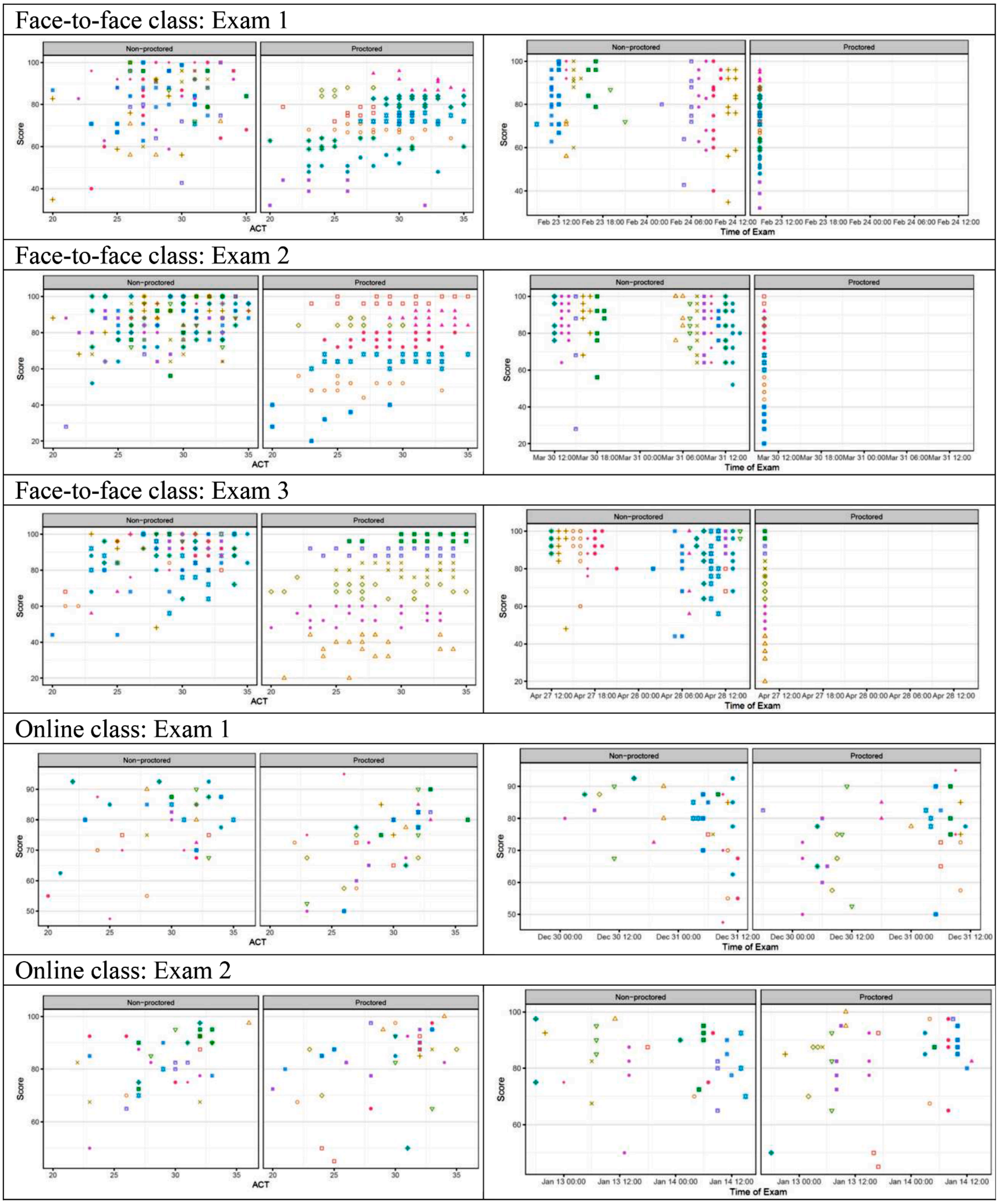

Figure 1. Cluster detection based on ACT and time of exam. Notes: Figures show clusters of students classified using K-means with 20 clusters. Clusters are created for each exam using the hour students started the exam, ACT score, age, GPA, gender, ethnicity, and class.

Diligent students tend to take exams earlier, which is also supported by the fact that students who took the exam earlier also have higher average GPA and ACT scores.

The third mechanism to be considered is the potential collaboration between students when taking an exam. Students taking the exam and helping each other appear to be an important factor driving our results. They account for the fact that we find significant differences in scores in the face-to-face class but not in the online class. Collaboration is more likely among students in a face-to-face setting where students can network more easily. Examples of networking include the fact that students tend to sit in the same area throughout the term, allowing them to make friends with those they see regularly. Also, students may already know others but were not aware they were in the same class until seeing each other in person. These are networking advantages that are not as common in online classes, where students work more independently in an asynchronous learning environment.

The tables presented thus far provide evidence that collaboration appears to be driving the difference in average scores between proctored and non-proctored exams. As noted earlier, all students had access to the same resources while taking exams (including the use of the Internet). Therefore, under these testing conditions, we can rule out any advantage from students having greater access to online resources than others.

To further investigate the role of collaboration in driving higher exam scores, we attempt to detect "collaborating groups" using a K-

Table 9
Effect of proctor on exam score: heterogeneity by ACT (face-to-face class).

	Dependent Variable: Exam Score		
	Exam 1	Exam 2	Exam 3
	(1)	(2)	(3)
Proctor x ACT Low	-18.434***	-18.261***	-24.892***
	(2.689)	(3.707)	(3.398)
Proctor x ACT Medium	-10.017***	-10.322***	-17.396***
	(1.996)	(2.138)	(2.639)
Proctor x ACT High	-7.370***	-4.567***	-11.949***
	(1.903)	(1.726)	(2.143)
Age	1.205	-0.009	-1.301
	(0.802)	(0.891)	(1.046)
Male	2.033	0.861	4.900***
	(1.645)	(1.430)	(1.830)
White	-2.770*	-3.392**	-1.791
	(1.578)	(1.392)	(1.882)
Freshman	8.499	-0.156	0.654
	(5.712)	(3.799)	(4.171)
Sophomore	7.531	-1.007	2.068
	(5.652)	(3.556)	(3.796)
Junior	5.234	-0.525	3.020
	(5.604)	(3.505)	(3.695)
GPA PreCourse	10.080***	11.576***	12.911***
	(2.053)	(1.911)	(2.383)
Constant	18.865	48.832**	66.291***
	(20.057)	(21.436)	(24.508)
Observations	228	317	269
R^2	0.335	0.337	0.396
Adjusted R^2	0.304	0.316	0.373

Notes: Heteroskedasticity-robust standard errors in parentheses.
* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

Table 10
Effect of proctor on exam score: heterogeneity by ACT (online class).

	Dependent Variable: Exam Score	
	Exam 1	Exam 2
	(1)	(2)
Proctor x ACT Low	-10.553**	-7.661*
	(4.501)	(4.068)
Proctor x ACT Medium	-4.919*	3.135
	(2.549)	(2.880)
Proctor x ACT High	0.597	0.681
	(2.372)	(3.346)
Age	-0.364	0.842
	(0.907)	(1.188)
Male	5.943***	6.082***
	(2.057)	(2.246)
White	-2.785	-2.984
	(2.140)	(2.532)
Freshman	-10.185***	-3.914
	(3.334)	(4.734)
Sophomore	-3.968	-1.750
	(2.586)	(3.941)
Junior	-1.746	1.487
	(3.021)	(3.313)
GPA PreCourse	7.719***	9.743***
	(2.227)	(2.969)
Constant	60.832**	33.284
	(24.386)	(26.752)
Observations	72	72
R^2	0.442	0.334
Adjusted R^2	0.351	0.225

Notes: Heteroskedasticity-robust standard errors in parentheses.
* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

Table 11
Effect of proctor on exam score: heterogeneity by gender (face-to-face class).

	Dependent Variable: Exam Score		
	Exam 1	Exam 2	Exam 3
	(1)	(2)	(3)
Proctor x Female	-11.441***	-10.890***	-17.297***
	(1.921)	(1.889)	(2.343)
Proctor x Male	-12.169***	-9.421***	-15.219***
	(1.901)	(2.039)	(2.076)
Age	1.652*	0.767	-1.543
	(0.853)	(0.937)	(1.014)
White	-2.007	-3.167**	-0.872
	(1.603)	(1.407)	(1.863)
Freshman	10.493**	2.524	-0.218
	(5.166)	(4.016)	(4.146)
Sophomore	7.801	1.380	0.288
	(4.999)	(3.668)	(3.826)
Junior	3.531	0.457	-0.506
	(5.068)	(3.559)	(3.811)
GPA PreCourse	8.805***	10.737***	12.094***
	(2.094)	(1.744)	(2.222)
ACT	1.428***	1.206***	1.338***
	(0.300)	(0.253)	(0.291)
Constant	-26.420	-0.109	37.803
	(22.630)	(24.792)	(24.540)
H0	Female = Male		
F-stat	0.125	0.332	0.603
Observations	233	320	273
R^2	0.359	0.358	0.395
Adjusted R^2	0.333	0.339	0.375

Notes: Heteroskedasticity-robust standard errors in parentheses.
* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

means clustering technique. A K-means clustering approach seeks to partition data in a way such that the within-group variation is as small as possible (James et al, 2013). Figure 1 presents the clusters for each exam and class. The left panels show the clusters based on ACT score (as a measure of ability) and the right panels show the clusters based on time of exam. The aim of this analysis is to determine whether students with similar ability taking the exam at the same time are achieving higher scores, which might suggest collaborative activity. This assumes that students who are similar in ability are more likely to collaborate with one another than students with differing abilities.

The patterns of clusters shown in Figure 1 support our hypothesis. Each shape identifies a cluster that a student belongs to. While the proctored exams show a more linear correlation between ACT and exam scores, the non-proctored exams show some clustering among those with similar ACT scores and those taking the exam at the same time. Although these results are not fully conclusive, they provide some additional evidence that collaboration played a role in the higher scores achieved in the non-proctored exams.

6. Heterogeneity

The average student effect of a proctor could potentially mask substantial heterogeneity. To explore heterogeneity we first focus on ACT scores which serve as a proxy for student ability. We divide students into three groups according to their ACT scores: below 26 (Low), from 26 to 30 (Medium), and greater than 30 (High). This roughly corresponds to the 80th percentile and lower, 80th to 95th percentile, and greater than 95th percentile, respectively, based on percentiles reported from ACT.

Tables 9 and 10 show a gradient in the effects of a proctor. Students with Low ACT scored on average 18.4% lower on the first exam, 18.3% lower on the second exam, and 24.9% lower on the third exam when the exam was proctored. Those in the Medium ACT group scored 10%, 10.3%, and 17.4% lower, respectively. Those in the High ACT group scored 7.4%, 4.6%, and 11.9% lower, respectively. Across the three categories of academic ability as measured by ACT scores, the absolute value of the negative score differential between proctored and non-proctored exams decreases as academic ability increases. This suggests that students with lower academic ability have more to gain from collaboration.

A similar pattern arises in the online course but the magnitude of the differences are smaller. Students in the Low ACT group scored on average 10.5% and 7.7% lower on the first and second proctored exams,

Table 12
Effect of proctor on exam score: heterogeneity by gender (online class).

	Dependent Variable: Exam Score	
	Exam 1	Exam 2
	(1)	(2)
Proctor x Female	-7.658***	-1.504
	(2.543)	(3.138)
Proctor x Male	0.921	-0.196
	(2.030)	(2.648)
Age	0.073	1.554
	(0.853)	(1.217)
White	-3.347*	-1.673
	(1.876)	(2.347)
Freshman	-9.261***	-3.247
	(2.882)	(4.365)
Sophomore	-3.388	-2.148
	(2.654)	(3.729)
Junior	-1.374	0.173
	(2.927)	(3.178)
GPA PreCourse	7.363***	8.736***
	(2.199)	(3.059)
ACT	0.936***	1.243***
	(0.280)	(0.356)
Constant	28.769	-10.368
	(24.336)	(30.267)
H0	Female = Male	
F-stat	13.710	0.159
Observations	74	74
R^2	0.457	0.353
Adjusted R^2	0.381	0.262

Notes: Heteroskedasticity-robust standard errors in parentheses.
* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

Table 13
Effect of proctor on exam score: heterogeneity by class standing (face-to-face class).

	Dependent Variable: Exam Score		
	Exam 1	Exam 2	Exam 3
	(1)	(2)	(3)
Proctor x Freshman	-11.171***	-13.475***	-20.089***
	(2.007)	(2.487)	(2.615)
Proctor x Sophomore	-11.446***	-8.906***	-13.732***
	(1.984)	(1.907)	(2.335)
Proctor x Junior	-13.159***	-8.020***	-14.441***
	(3.924)	(2.893)	(4.033)
Proctor x Senior	-7.408	-7.112	-14.605***
	(5.065)	(6.069)	(4.451)
Age	-0.077	-0.239	-1.878**
	(0.921)	(0.682)	(0.877)
Male	2.078	-0.038	3.735**
	(1.688)	(1.433)	(1.814)
White	-2.863*	-3.071**	-1.403
	(1.581)	(1.392)	(1.866)
GPA PreCourse	9.106***	10.437***	12.072***
	(2.105)	(1.781)	(2.273)
ACT	1.153***	1.090***	1.116***
	(0.303)	(0.242)	(0.274)
Constant	20.535	24.900	48.895**
	(21.584)	(16.961)	(19.693)
H0	Eq of Coef		
F-stat	0.345	1.049	1.431
Observations	233	320	273
R^2	0.345	0.364	0.416
Adjusted R^2	0.318	0.346	0.396

Notes: Heteroskedasticity-robust standard errors in parentheses.
* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

respectively. Students in the Middle ACT group scored 4.9% lower and 3.1% higher, respectively. And those in the High ACT group scored 0.6% and 0.7% higher, respectively, on proctored exams. A plausible reason for this result is that high ability students tend to be better prepared for proctored exams, allowing them to score well compared to non-proctored exams in which these students might choose to study less.

Table 14
Effect of proctor on exam score: heterogeneity by class standing (online class).

	Dependent Variable: Exam Score	
	Exam 1	Exam 2
	(1)	(2)
Proctor x Freshman	-10.756***	3.097
	(2.665)	(5.373)
Proctor x Sophomore	-1.875	-3.335
	(2.738)	(3.213)
Proctor x Junior	-0.073	0.592
	(4.938)	(3.065)
Proctor x Senior	-2.230	0.214
	(2.418)	(4.353)
Age	0.669	1.988**
	(0.843)	(0.995)
Male	5.258***	4.401**
	(1.932)	(1.885)
White	-4.789**	-2.053
	(2.029)	(2.208)
GPA PreCourse	8.244***	9.663***
	(2.187)	(2.961)
ACT	0.839***	1.037***
	(0.290)	(0.343)
Constant	11.805	-19.316
	(22.627)	(26.826)
H0	Eq of Coef	
F-stat	4.102	0.537
Observations	74	74
R^2	0.431	0.391
Adjusted R^2	0.351	0.305

Notes: Heteroskedasticity-robust standard errors in parentheses.
* Significant at 10% level; ** Significant at 5% level; *** Significant at 1% level.

Heterogeneity in test performance and cheating behavior can also be influenced by other individual characteristics. Specifically, gender has been well-studied as a potential driver of heterogeneity in exam performance. McCabe et al. (2001) showed that results are mixed when studying cheating behavior by gender, while Azar and Applebaum (2020) find no significant effect on cheating by gender among children. Tables 11 and 12 present heterogeneity by gender for the face-to-face and online courses, respectively. We find no statistical differences among genders with respect to collaboration on the exams when a live proctor was present. Male students scored between 10.8% and 17.3% lower, on average, on proctored exams, and female students scored between 9.4% and 15.2% lower, on average. For web-proctored exams, female students scored 7.6% lower while male students scored 0.9% higher on Exam 1, all else equal.

McCabe and Treviño (1997) found that freshman and sophomore students, especially in large enrollment classes, are more likely to rationalize cheating. We examine this hypothesis in Tables 13 and 14 for the face-to-face and online classes, respectively. We find evidence that collaboration is prevalent across all class standings, but the extent to which students are presumed to engage in this behavior appears to be greater among underclass students.

7. Conclusion

Despite significant anecdotal evidence of collaboration and other cheating behavior among college students during exams, little empirical evidence has been collected beyond self-reported survey data. This paper presents the results of two randomized field experiments assessing the effect of proctoring on exam scores in face-to-face and online settings. We find that students who were not subject to proctoring scored on average 11% higher compared to those who were required to use a proctor. Moreover, the difference in scores between proctored and non-proctored exams was much greater in the face-to-face class where greater networking opportunities can facilitate collaboration.

Our results suggest that collaboration is the most common way in which students cheat in online exams when students have access to their notes, book, and the Internet. Moreover, we find that students with lower ACT scores appear to engage in collaboration strategies to a greater extent, which is supported by the argument that lower ability students have more to gain while higher ability students have more to lose (a higher opportunity cost to cheating).

Although our evidence is indirect and requires further research to identify the whether collaboration is the true driver for the score differences between proctored and non-proctored exams, our results provide important insights on the effectiveness of different proctoring systems for online assessments. Further study on the manner by which students collaborate is needed to gain a better understanding of the root causes of cheating behavior. In addition, additional study needs to focus on closed-book exams, in which non-proctored exams offer students the opportunity to utilize the Internet to search for information and solutions. This comparison was not analyzed in our study where all students had full access to the book, notes, and Internet.

Lastly, if the primary cause of cheating behavior is collaborative behavior, it leaves open the question of whether it is truly a "bad" thing. Bloom (2009) suggests that collaborative testing not only produces higher test scores than individual testing but also improves retention of course content. In other words, if the objective of the course is to improve learning outcomes, one can argue that an instructor's time is better spent preparing pedagogical units (i.e., focus on teaching) than devising strategies to prevent students from "cheating" by collaborating with one another.

Declaration of Competing Interest

None

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.socec.2020.101653.

References

Akin, Z. (2019). Dishonesty, social information, and sorting. *Journal of Behavioral and Experimental Economics, 80*, 199–210.

Azar, O., & Applebaum, M. (2020). Do children cheat to be honored? A natural experiment on dishonesty in a math competition. *Journal of Economic Behavior and Organization, 169*, 143–157.

Belot, M., & van de Ven, J. (2019). Is dishonesty persistent? *Journal of Behavioral and Experimental Economics, 83*. Article 101464.

Bloom, D. (2009). Collaborative test taking: benefits for learning and retention. *College Teaching, 57*(4), 200–216.

Bosshardt, W., & Chiang, E. P. (2016). Lecture capture learning: Do students perform better compared to face-to-face classes? *Southern Economic Journal, 82*(3), 1021–1038.

Derosa, A. P., & Patalano, F. (1991). Effects of familiar proctor on fifth and sixth grade students' test anxiety. *Psychological Reports, 68*(1), 103–113.

Figlio, D., Rush, M., & Yin, L. (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.

Grijalva, T. C., Nowell, C., & Kerkvliet, J. (2006). Academic honesty and online courses. *College Student Journal, 40*(1).

Harmon, O. R., & Lambrinos, J. (2008). Are online exams an invitation to cheat? *Journal of Economic Education, 39*(2), 116–125.

Harton, H. C., Aladia, S., & Gordon, A. (2019). Faculty and student perceptions of cheating in online vs. traditional classes. *Online Journal of Distance Learning Administration, 22*(4).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning, with Applications in R*. New York: Springer.

Kerkvliet, J., & Sigmund, C. L. (1999). Can we control cheating in the classroom? *Journal of Economic Education, 30*, 331–343.

King, C. G., Guyette, R. W., & Piotrowski, C. (2009). Online exams and cheating: An empirical analysis of business students' views. *Journal of Educators Online, 6*(1).

Levitt, S. D. and Lin, M. J. 2019. "Catching cheating students." Economica, 10.1111/ecca.12331.

McCabe, D. L. (2005). Cheating among college and university students: A North American perspective. *International Journal for Academic Integrity, 1*(1), 1–11.

McCabe, D. L., Treviño, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: a decade of research. *Ethics and Behavior, 11*(3), 219–232.

McCabe, D. L., & Treviño, L. K. (1997). Individual and contextual influences on academic dishonesty: a multicampus investigation. *Research in Higher Education, 38*, 379–396.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2010). *Evaluation of evidence-based practices in online learning: A meta-analysis and review on online learning studies*. Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development.

Nowell, C., & Laufer, D. (1997). Undergraduate student cheating in the fields of business and economics. *Journal of Economic Education, 28*, 3–12.

Schwieren, C., & Weichselbaumer, D. (2010). Does competition enhance performance or cheating? A laboratory experiment. *Journal of Economic Psychology, 31*, 241–253.

Ward, D. A., & Beck, W. L. (1990). Gender and dishonesty. *Journal of Social Psychology, 130*(3).

Watters, M. P., Robertson, P. J., & Clark, R. K. (2011). Student perceptions of cheating in online business courses. *Journal of Instructional Pedagogies, 6*.

Yaniv, G., Siniver, E., & Tobol, Y. (2017). Do higher achievers cheat less? An experiment of self-revealing individual cheating. *Journal of Behavioral and Experimental Economics, 68*, 91–96.