Journal of Computing in Higher Education
https://doi.org/10.1007/s12528-023-09378-x

Academic dishonesty and monitoring in online exams: a randomized field experiment

Maite Alguacil[1,2] · Noemí Herranz-Zarzoso[2] · José C. Pernías[3] · Gerardo Sabater-Grande[4]

Accepted: 20 June 2023

Abstract

Cheating in online exams without face-to-face proctoring has been a general concern for academic instructors during the crisis caused by COVID-19. The main goal of this work is to evaluate the cost of these dishonest practices by comparing the academic performance of webcam-proctored students and their unproctored peers in an online gradable test. With this aim in mind, we carried out a randomized field experiment using a simple video surveillance system through Google Meet during an online closed-book final exam of an Introduction to Microeconomics course. Given that all conditions except for webcam monitoring were identical, differences in between-subjects scores are attributed to academic dishonesty. After controlling for potential confounding factors, including gender, academic degree, instructor, previous score and whether students were repeaters or not, we found that those students who were proctored via webcam obtained statistically significant lower scores in the final exam than those who were not using this surveillance system with a low level of invasiveness. Inspection of the potential factors behind these differences in scores suggests that the poorer performance of proctored students is more related to academic dishonesty than to reasons involving anxiety or heterogeneity factors.

Keywords Randomized field experiment · Academic dishonesty · Monitoring · Academic performance

JEL Classification C93 · D03

Introduction

Academic dishonesty encompasses any behavior intended to deceive instructors including plagiarism, cheating, fabrication, and falsification. As stated by the OECD Directorate for Education and Skills "student academic dishonesty … is by far the most frequently discussed challenge in higher education today with regard

Extended author information available on the last page of the article

Published online: 21 July 2023

to the shift of examinations online" (OECD, 2020). In fact, Butler-Henderson and Crawford (2020) claimed that, as many scholars state, cheating is a prevalent and approved element of the student experience. There is no single factor explaining why students engage in academic dishonesty although they recognize it is morally wrong. Feelings of guilt are dismissed because academic dishonesty is justified by a set of factors[1] including lack of interest and/or poor time management, academic pressure, and failure to understand academic conventions. Besides, students generally accept cheating, perceive these dishonest practices as a necessary way to succeed, and trust that they will not be caught. Additionally, the fast development of digital technology and online courses has produced a "distancing" effect that makes students mitigate their feelings of guilt by using innovative tools for cheating (Blau et al., 2021).

Concern for dishonest practices in online exams was especially noticeable during the COVID-19 health crisis, as academic instructors in many countries were forced to move from face-to-face to online exams. Janke et al. (2021) found high rates of self-admitted academic dishonesty reported by German students of higher education during the COVID-19 crisis. In this vein, Comas-Forgas et al. (2021) revealed a significant increase in Internet searches for information on cheating in online exams during the COVID-19 pandemic in Spain. Although at the time of writing this paper, teaching and exams in higher education are again being held face-to-face in most developed countries, there is, on the one hand, the threat of new waves of the pandemic and, on the other, a greater predisposition for online learning tools (Vazquez et al., 2021). In this scenario, understanding to what extent a remote proctoring system might discourage students from misconduct during exams would mitigate the widespread fear on the part of instructors that those online exams are an easy way for students to improve their score.

In this paper, we investigate for the first time, through a randomized field experiment, how remote proctoring by webcam may deter academic dishonesty practices in a closed-book online exam. Our empirical strategy allows us to control for individual heterogeneity and background factors, such as student's gender, academic degree, instructor and type of student (repeater or non-repeater). As a novelty, we also explore the factors potentially responsible for the results obtained.

Our findings confirm that students remotely monitored in an online exam obtained significantly lower scores than those without a webcam surveillance system, under otherwise identical conditions. This result holds even after considering potential individual heterogeneity and different academic backgrounds. We also present evidence of a relationship between the outperformance of unproctored students and dishonest academic conduct.

The rest of the paper is organized as follows. The next section introduces the relevant literature. The following section explains the experimental setting, with a description of the assessment system and the sample composition. The data analysis section presents the main descriptive statistics and the estimation results. We then show the limitations and propose future research. The final section concludes.

[1] See Brimble (2016) in order to analyze the motivations behind students' academic dishonesty in higher education.

Literature background

Dendir and Maxwell (2020) argued that online courses could be more susceptible to academic dishonesty because assessments often take place in unproctored settings where students can use unauthorized resources during their evaluation, communicate with other people or even ask someone else to take the test for them. Two types of papers have reported that academic dishonesty is indeed a significant problem in unproctored environments. The first studies rely on student self-reports[2] and the second comprise those presenting empirical evidence based on academic performance. As Howard (2019) argued, the shortcoming of survey-based studies is that they might not be reliable because academic dishonesty is a delicate issue and students may not be sincere in the answers they give in surveys. In the same line, Vazquez et al. (2021) claimed that studies that rely on survey responses to expose dishonesty depend on the questionable assumption that cheating students will be honest in reporting their conduct.

Focusing exclusively on the second type of studies, we find that empirical work in this field has traditionally addressed this issue through two different approaches: (i) some studies compare student performance in supervised face-to-face environments versus unsupervised online environments; (ii) others, in contrast, analyze the scores of the proctored and unproctored groups in the same online environment. In this latter case, however, students were not randomly assigned.

Examples of the first group of studies that offer positive evidence of cheating behavior are Carstairs and Myors (2009), Fask et al. (2014) and Brallier and Palm (2015). In contrast, Yates and Beaudrie (2009), Gold and Mozes-Carmel (2009), Beck (2014) and Ladyshewsky (2015) found no significant differences between student grades on face-to-face monitored versus online unmonitored exams. In both cases, this literature suffers from two major drawbacks. On the one hand, comparing academic performances between the two groups may not be accurate because differences in the potential outcomes could be partly due to a change in the testing environment, irrespective of the effect of proctoring (Dendir & Maxwell, 2020). On the other hand, self-selection bias can arise, given that no random assignment to the treatment group is applied.[3]

The second group of studies mentioned above (those that compare unproctored and proctored online exams) are not exempt from this last objection either. Harmon and Lambrinos (2008), Prince et al. (2009), Alessio et al. (2017), Daffin and Jones (2018) and Dendir and Maxwell (2020) presented experiments in which the treatment group was non-randomly assigned, obtaining significant rates of cheating when the online exams were not proctored.

[2] Berkey and Halfond (2015) found that 84% of 141 students surveyed claimed that academic dishonesty in online test-taking was a relevant question. Similar results are observed in previous studies by Etter et al. (2007), King et al. (2009), Stuber-McEwen et al. (2009) and Watson and Sottile (2010).

[3] This problem is not exclusive to papers comparing performance in face-to-face and online environments but is intrinsic to them because subjects are not randomly assigned to a test environment.

To our knowledge the only studies that have used randomized field experiments within this literature are Vazquez et al. (2021) and Hylton et al. (2016). Vazquez et al. (2021) found that, whereas live proctors[4] in a face-to-face exam had a significant effect on students' scores, web-based proctors[5] in an online class did not reduce students' grades to any significant extent. But, in this work, the analysis of cheating is limited to students' collaborative behavior. Specifically, Vazquez et al. (2021) analyzed academic dishonesty in open-book exams, where students have full access to the book, notes and Internet, but are restricted to a non-collaborative performance.[6] In this paper, we instead focus on cheating on closed-book exams, in which students are not allowed to use any additional aids or share any information.

More similar to our experimental design is the work by Hylton et al. (2016). Through a randomized field experiment, these authors studied the deterrent effect of webcam-based proctoring on cheating during online exams in an undergraduate course at a private university in Jamaica. They found no significant differences between the grades obtained by proctored and unproctored students. However, in this study, the authors failed to control for potential divergences between the two groups of students in terms of their background.

Experimental design

In early January 2021, the alarming evolution of the COVID-19 pandemic in Spain caused the final exams of some first-term courses at the Universitat Jaume I to change from face-to-face to online mode. This unexpected scenario provides a unique opportunity to study potential student misconduct behavior in online exams and assess to what extent simple remote proctoring methods could prevent it.

The subject Introduction to Microeconomics is taught at the Universitat Jaume I in the first semester of the first year of three different undergraduate degrees: Economics, Finance and Accounting, and Business Administration. Most of the students who take this subject have just started college. The course grade requires the completion of exercise assignments during the course (30% of the total grade) and a final exam (70%). Two weeks before the official date of the final exam (January 28), we informed all students enrolled in this course that, contrary to what was initially programmed, the final exam would be held online.

[4] In this experiment all students took the same 50-min online exam, but subjects assigned to the proctored exam had to take their online exam in the presence of a proctor in a classroom at a specific time on a particular day. Students assigned to the non-proctored group took the exam online at a time of their choosing within a 30-h window. Hence, *ex ante* time conditions were not identical for the two groups.

[5] In this experiment, all students took the same 180-min online exam, and they could choose the time of their exam within a 30-h window. However, whereas subjects assigned to the proctored exam were monitored in real time via their webcams, students assigned to the non-proctored group were not subject to video surveillance. Thus, *ex post* time conditions were not identical for the two groups.

[6] Cheating could occur if students who completed the exam earlier chose to share exam information with those who took it later. However, self-selection problems arose when Vazquez et al. (2021) compared the grades obtained by different samples depending on the time when students took the exam.

After obtaining the corresponding permission from the Ethics Committee of the Universitat Jaume I to carry out a field experiment during the online final exam of Introduction to Microeconomics, we randomly assigned each of the students enrolled in this course to one of two groups.[7] While the students in the treatment group had to keep their webcam on during the exam, the students in the control group did not. The random assignment process followed was not revealed to the students, and they were only informed that they might be asked to turn on their webcams.

The exam consisted of 14 multiple-choice questions that had to be answered in a maximum of 60 min.[8] Each question had 4 possible answers, only one of which was correct. Students received full credit for each correct answer, no credit for questions they did not attempt, and a penalty for incorrect answers equal to one-third of the points for a correct answer. In addition, we implemented some features to make dishonest behavior more difficult: (i) we presented the questions to each student in random order; (ii) each question had 4 similar variants and one was randomly chosen for each student; and (iii) we did not allow the students to go back to previous questions, even those that they had not answered.

Before starting the exam, we required all students to join a Google Meet session that, if necessary, would be used as a channel to communicate with the instructors. We also informed them that, in order to verify compliance with the basic rules of authenticity and authorship, they might be asked to keep their webcam on throughout the exam and to show a document proving their identity. It is worth mentioning that we did not design our proctoring setting to verify that students were unsupported by internet resources during the exam. Proctoring was restricted to monitoring the online exam in real time through a webcam, limiting the levels of opportunity to engage in misconduct but not totally preventing it.

Data analysis

Descriptive statistics and univariate tests

The data used in this study contains information from 443 students who took the exam of the subject Introduction to Microeconomics (90.2% of the students enrolled in this subject actually took the exam).[9] From this sample, we excluded 7.1% of the students (4.7% of the sample did not give their consent and 2.4% started the exam late), so the final sample comprised 412 students, of whom 317 were non-repeaters

[7] The appendix describes the random assignment process in detail.

[8] Although students had already taken practice tests under these same rules during the course, they were informed again a week before the final exam.

[9] We excluded some exceptional cases from the sample, such as exchange students or students with special needs. We consider that these students take the exam differently from the rest. In the first case, they have a different background and, in the second, they take the exam in conditions adapted to their needs.

Table 1 Descriptive statistics and group homogeneity tests

Variable	All students ($N = 412$)	Webcam Off ($N = 220$)	Webcam On ($N = 192$)	p-value
Midterm	0.00 (1.00)	0.06 (1.02)	−0.07 (0.98)	0.031
Female	189 (46%)	99 (45%)	90 (47%)	0.703
Repeater	95 (23%)	51 (23%)	44 (23%)	0.949
Finance	126 (31%)	79 (36%)	47 (24%)	0.012
Economics	83 (20%)	46 (21%)	37 (19%)	0.679
Business	203 (49%)	95 (43%)	108 (56%)	0.008
Lab groups				0.786

Mean (SD) for Midterm; N (%) for the remaining variables. The p-values for the variable 'Midterm' come from a Wilcoxon rank test for the homogeneity of the means of the webcam-off group and the webcam-on group. The p-values for the remaining variables correspond to Pearson Chi-square tests

and 95 repeaters.[10] Exam conditions between groups were identical during the experiment for the participants, with the exception of remote monitoring: whereas some participants were asked to allow video surveillance (220 students), others received no such request during the exam (192 students).

Next, in Table 1, we present some descriptive statistics calculated for the complete sample of all students, and for the subsamples of the treatment group (Webcam Off) and the control group (Webcam On). The variables included may influence the students' performance in the final exam and refer to their academic background and demographic characteristics. Table 1 also presents the p-values of the statistical tests that check the homogeneity between the treatment and control groups obtained from our random assignment process.

The *Midterm* variable records the standardized grades of the assignments done throughout the term for the subject Introduction to Microeconomics. Lee et al. (2020) found that the expected benefit of cheating is lower for students who have previously obtained better grades. Thus, we can predict that students with a poorer academic performance in the past have higher incentives to cheat. Accordingly, we can expect higher scores in the midterm exams to be negatively associated with academic dishonesty in the final exam. The statistics in Table 1 confirm the existence of a small but significant difference between the average midterm scores for the treatment group and the control group. Thus, to exclude any chance of divergence observed in academic performance being a consequence of this between-group heterogeneity, we control for this fact in the causal analysis.

In addition, different authors (see, for example, Pekkarinen, 2015; Iriberri & Rey-Biel, 2019; Espinosa & Gardeazabal, 2020) have shown that women's performance

[10] Repeater students are those that attended the course previously (students who had failed at least one previous exam opportunity).

in multiple-choice exams is poorer than that of their male counterparts. Furthermore, as Montolio and Taberner (2021) pointed out, these differences in performance may be more pronounced in high-stakes tests, such as final exams.[11] According to the statistics of the dummy variable *Female* shown in Table 1, somewhat less than 50% of the students who took the final exam were women and there were no significant differences between the groups in terms of the percentage of women. Therefore, in our sample, the possibility of there being differences in the results of the treatment and control groups due to an imbalance in the proportion of women is limited.

We further evaluate the potential group heterogeneity in terms of the proportion of students repeating the subject. The academic results of these students may reflect their greater difficulties in passing the subject, but also their greater experience at university or, simply, the fact that they are older. To do so, we define a dummy variable, *Repeater*, which takes the value 1 for repeaters. As Table 1 shows, almost a quarter of the students had already taken the subject in previous years. We found no significant differences in the proportions of repeaters in the treatment group and the control group.

Next, we focus on potential divergences in students' motivation to deal with the subject under study (how "close" it is to the academic program in which the student is enrolled). In this sense, we can reasonably think that students enrolled in the Degree in Economics find a greater motivation to study Microeconomics than students enrolled in the Degree in Finance and Accounting or the Degree in Business Administration. The dummy variables *Finance*, *Economics* and *Business* indicate the degree in which each student in the sample is enrolled. The statistical tests in Table 1 show significant differences in the proportions of Finance and Business students across the treatment and control groups.

Finally, we deal with a potential heterogeneity due to differences in students' instructors. Given that some of the classes in the subject were taught in small laboratory groups, we constructed a set of 21 dummy variables indicating which lab group each student attended (the students of the same laboratory group shared the same instructor). The p-value in the last row of Table 1 indicates that there were no significant differences across the laboratory groups in the proportions of students assigned to the treatment and control groups.

As a first approximation to our causal analysis, Fig. 1 illustrates the differences between the standardized scores (represented by the variable *Grade*) obtained by students who were monitored by a webcam and those who were not remotely monitored. The location of the boxplots clearly shows that the distribution of scores for the treatment group is shifted to the left relative to the distribution for the control group. That is, students who had to keep their webcam on during the final exam tended to get lower grades than students who did not.

This deterring effect of remote proctoring on *Grade* is also confirmed in Table 2 through the Wilcoxon Rank Sum Test. The *p*-value obtained indicates a significant difference in the mean of the variable *Grade* between the treatment and the control

[11] According to these authors, male students are found to surpass female students in multiple-choice exams as female students feel more stressed, leading them to skip more questions than their male peers.

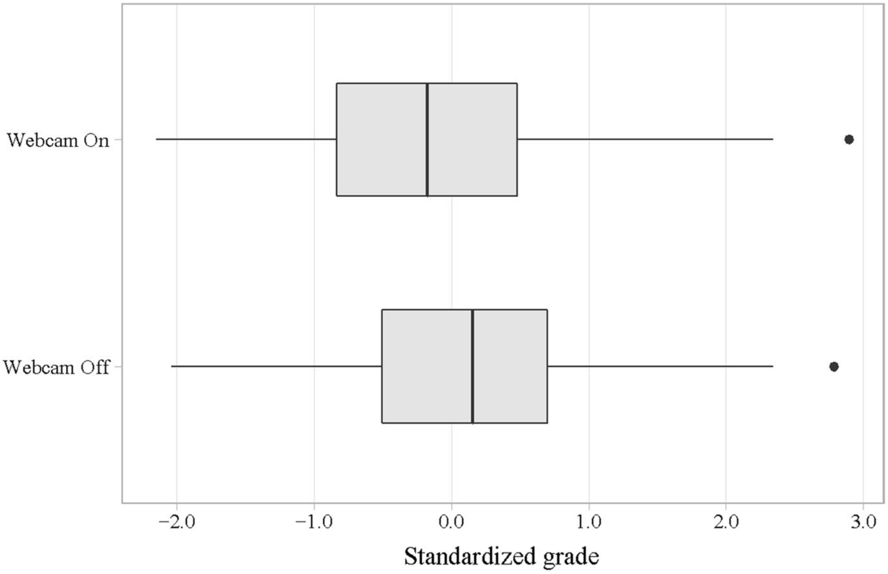

Fig. 1 Distribution of standardized grades by groups. Box plot comparing the distribution of standardized grades between the treatment group (Webcam On) and the control group (Webcam Off)

Table 2 Exam outcomes by group

Variable	All students ($N=412$)	Webcam Off ($N=220$)	Webcam On ($N=192$)	p-value
Grade	0.00 (1.00)	0.14 (1.02)	−0.16 (0.96)	0.003
Duration	54.72 (7.16)	54.98 (6.24)	54.42 (8.10)	0.863
Completed	346 (84%)	184 (84%)	162 (84%)	0.838
Skipped	2.73 (1.87)	2.73 (1.83)	2.72 (1.92)	0.848
Right	5.88 (2.47)	6.20 (2.53)	5.52 (2.34)	0.004
Wrong	5.05 (2.36)	4.72 (2.25)	5.42 (2.44)	0.003

Mean (SD); n (%). The p-values come from Wilcoxon rank tests for the homogeneity of the means of the webcam-off group and webcam-on group, except for the variable ‘Completed’, where a Pearson Chi-square test is used to test differences between the groups

groups. In particular, the difference in mean scores was approximately one-third of a standard deviation of *Grade*.

In this table, we also analyze the relevance of a set of factors potentially linked to an anxiety effect provoked by the fact of being under surveillance, as pointed out by Hylton et al. (2016). In particular, these authors revealed that monitored participants may be inclined to rush through a test due to the added anxiety resulting from the monitored environment and, consequently, obtain worse grades than their non-monitored peers. To investigate this hypothesis, Table 2 shows some statistics of the variables *Duration* and *Completed*. The first of these variables measures the minutes that each student took to complete the exam (with a maximum value of 60, even for those participants who could not answer all the questions). The second one, *Completed*, is a dummy variable that takes the value 1 for those students who were able to complete the exam during the stipulated time, and zero otherwise. As can be seen in this table, on average, the students took less than 55 min to complete the exam with no significant differences between the treatment group and the control group. A total of 84% of the students managed to complete the exam on time, regardless of whether they were monitored or not. This evidence rules out the hypothesis that differences in *Grades* stem mainly from the fact that monitored students felt more pressure and less free to use the time allowed.

Other authors have also highlighted the possibility that students under greater pressure underperform because they tend to omit more questions.[12] In Table 2, we show the mean value and the standard deviation for the variable *Skipped* (defined as the number of questions skipped by each student), and the variables *Right* and *Wrong* (defined as the number of correct and incorrect answers, respectively). These statistics reveal that the treatment and control groups are similar in terms of the number of skipped questions, while there are significant differences in the number of correct and incorrect answers.

The above results clearly reveal that the monitored students did not suffer higher levels of anxiety. Indeed, our findings confirm that the poorer performance of those students under webcam-based proctoring is indeed the result of more wrong answers and fewer right answers than the control group.

Regression results

In this section, we try to establish to what extent students' heterogeneity can explain the differences observed in their scores in the final exam. For this purpose, we estimate the following regression models:

$$\text{Grade}_i = \alpha_0 + \alpha_1 \text{Webcam}_i + x_i'\beta + u_i$$

where the dependent variable $Grade_i$ is the standardized score in the final exam of student i. The main parameter of interest is α_1, the slope of the dummy variable $Webcam_i$. This variable takes the value 1 for students in the treatment group and 0 for the participants in the control group. Finally, u_i is a regression disturbance term.

Table 3 presents the estimates of the above regression model using the whole sample. Firstly, in Model 1, we evaluate the total impact of using the video surveillance system through the webcam on our dependent variable, $Grade_i$, through the estimation of a simple regression where $Webcam_i$ appears as the only explanatory variable. Next, to control for other relevant factors, we present the estimates of the extended models which add a set of control covariates collected in the vector x_i.

Estimates from the simple regression model show that the effect of web monitoring on grades is statistically significant and negative, representing almost one-third of a standard deviation of the dependent variable. The effect of the variable *Webcam*

[12] See, for instance, Pekkarinen (2015), Iriberri and Rey-Biel (2019), and Montolio and Taberner (2021).

Table 3 Regression models: all students ($N = 412$)

	Model 1	Model 2	Model 3
Webcam	-0.300^{***}	-0.262^{***}	-0.248^{**}
	(0.097)	(0.097)	(0.100)
Midterm		0.126^{***}	0.133^{**}
		(0.046)	(0.047)
Female		-0.301^{***}	-0.285^{***}
		(0.098)	(0.098)
Repeater		0.246^{**}	0.293^{**}
		(0.110)	(0.118)
Economics		-0.224	-0.207
		(0.141)	(0.141)
Business		-0.150	-0.156
		(0.117)	(0.115)
Constant	0.140^{**}	0.322^{***}	0.758^{***}
	(0.068)	(0.112)	(0.228)
Lab groups	No	No	Yes
			$F_{20, 385} = 1.9$
			p-value = 0.007
R^2	0.022	0.069	0.140
Adjusted R^2	0.020	0.056	0.082

OLS estimates and heteroskedasticity-consistent standard errors (HC2) in parentheses. The dependent variable of all models is Grade, the standardized grade in the final exam. Model 3 includes 20 dummies for the different laboratory groups. The F-statistic shown in the table tests the joint significance of the laboratory dummies by means of a heteroskedasticity-robust Wald test. Significance of estimates is marked with: $^{***}p < 0.01$; $^{**}p < 0.05$; * $p < 0.1$

is also negative and significant when we include other potential confounding variables. In Model 2, we add the academic background of the students represented by the following variables: (i) *Midterm* (the grades obtained in the course assignments throughout the semester); (ii) *Repeater* (which takes the value 1 for repeater students and 0 otherwise); and (iii) a set of dummy variables controlling for the specific academic degree (*Economics* and *Business*, with *Finance* as the reference category). Additionally, we consider the student gender (by means of *Female*, which takes the value 1 for women). Model 3 also includes 20 dummies for the laboratory groups (omitted from the table) that depict the different instructors.[13] As shown in Table 3, the joint significant F-test strongly confirms the existence of differences in student performance between laboratory groups.

Although smaller than in Model 1, the estimates of the webcam coefficient in Models 2 and 3 remain sizable, being close to a quarter of a standard deviation. With respect to the other covariates, our findings align well with intuition and previous literature, with the exception of the specific academic degree. As expected,

[13] Although the coefficients of these variables have been omitted due to lack of space, we include a joint significance test on these dummies at the bottom of the table.

Table 4 Regression models: non-repeater students (N = 301)

	Model 1b	Model 2b	Model 3b
Webcam	−0.390***	−0.338***	−0.330***
	(0.116)	(0.112)	(0.118)***
Access		0.267***	0.258
		(0.060)	(0.076)**
Midterm		0.181***	0.159
		(0.060)	(0.065)
Female		−0.388***	−0.376***
		(0.111)	(0.116)
Economics		−0.446***	−0.426**
		(0.164)	(0.169)
Business		−0.309**	−0.302**
		(0.135)	(0.139)
Constant	0.122	0.496***	0.703***
	(0.084)	(0.129)	(0.252)
Lab groups	No	No	Yes
			$F_{20,\,274} = 1.0$
			p-value = 0.403
R^2	0.036	0.152	0.209
Adjusted R^2	0.033	0.135	0.134

OLS estimates and heteroskedasticity-consistent standard errors (HC2) in parentheses. The dependent variable of all models is Grade, the standardized grade in the final exam. Model 3b includes 20 dummies for the different laboratory groups. The F statistic shown in the table tests the joint significance of the laboratory dummies by means of a heteroskedasticity-robust Wald test. Significance of estimates is marked with: ***$p < 0.01$; **$p < 0.05$; * $p < 0.1$

students with a higher score in the semester assignments obtained a better outcome in the final exam. We also found that repeaters' performance is better than that of those who took the exam for the first time (non-repeaters). In line with Montolio and Taberner (2021), our estimates confirm that male students outperform their female peers in multiple choice tests. However, contrary to our initial assumption that students from the Economics degree, based on their own motivation, should perform better on an Introduction to Microeconomics exam, the estimates do not reveal significant differences for students enrolled in the different degrees considered.

In short, the regression models in Table 3 confirm the negative effect of web monitoring on final exam grades, this being robust to the inclusion of other explanatory variables that take into account a set of student characteristics. However, previous literature usually employed a more generic proxy of this academic background, such as the score of the university entrance exam. For the sake of greater robustness and for comparison purposes, we further estimate previous models using this student's academic background indicator as a control variable. In particular, we employ the variable *Access* (the standardized university entrance score) as a proxy of this academic background. In this case, we also restrict the sample to non-repeater students.[14] This allows us to avoid the heterogeneity bias that can arise from combining both students who have just entered university (non-repeaters) and students in their second or later years (repeaters).[15] The results are shown in Table 4 (models 2b and 3b).

The new estimates confirm our previous findings, corroborating the existence of significant differences between the scores of the treatment group and the control group. The Webcam coefficients are higher in absolute values than those obtained with the complete sample. This result might be taken as evidence that the effect of web monitoring is more pronounced in first-year students. The results for the explanatory variables related to previous academic performance and the gender of the students are similar to those obtained in Table 3. However, in contrast to the estimates of the complete sample, the coefficients of Models 2b and 3b indicate that the performance of Economics and Business students was significantly poorer than that of Finance students. Finally, we did not find any significant differences in the mean scores of the laboratory groups, which can be explained by the fact that repeaters are usually more concentrated in some labs groups, such as those with classes taking place in the evening.

Discussion

Our results show that final exam scores were lower when students were monitored remotely. These findings are based on data obtained from a clean design in which all experimental conditions between the groups being compared were identical except for the treatment condition (webcam monitoring). Our procedure for randomly assigning students to treatment and control groups avoided potential self-selection effects in our sample that might have obscured causal interpretations. The negative effect of monitoring on grades is a robust finding that holds even when we control for potential confounding factors, such as gender, prior academic performance, degree, and course groups, or when we use a subsample that only includes first-year students. The consideration of other covariates, such as the time it took students to complete the exam, the number of questions skipped, or the proportion of students who completed the exam on time, allows us to exclude the hypothesis that participants subjected to remote monitoring suffered higher levels of anxiety or stress. Therefore, a more plausible explanation of the differences in scores would be related to the use of material that is not allowed or other dishonest practices by unproctored students. Our analysis supports the hypothesis that it is possible to deter dishonest behavior in online tests by using a very lax monitoring system with a low degree of intrusiveness.

[14] Our sample includes 301 subjects accessing the university through an entrance exam.

[15] In previous regressions, we have considered this fact by including a dummy variable distinguishing both types of students. However, it can be argued that it might not be sufficient to fully control for this kind of heterogeneity.

Limitations and future research

However, this study also presents some limitations. Our primary concern is related to mitigating the potential impact of heterogeneity on our results. Despite using a randomized design and including in the statistical analysis most of the personal traits available in our database, we are aware that some unobserved variables related to the socioeconomic level of the participants (such as age, income, nationality, employment status, religion, etc.) may be partially responsible for some of the effects reported here.

The main practical implication of our results is that it is possible to deter dishonest behaviors using minimally invasive methods of webcam surveillance. However, the effectiveness of this option depends on credibility; that is, whether students believe that monitoring can detect unauthorized materials or communications with their peers or third parties. Our data contain some evidence that only first-year students were affected by surveillance. Further work would be needed to determine more precisely whether simple monitoring protocols would be effective beyond first-year courses.

Another aspect that deserves more careful study is the possible gender bias of monitoring via webcam. Our regressions show that females scored worse in the final exam than their male peers. The reason why this occurs is not apparent, but there is a possibility that web monitoring imposes an additional burden on women. Future research should focus on determining to what extent remote monitoring is gender biased and how it could be mitigated.

Conclusions

This paper presents the results of a randomized field experiment evaluating the effect of monitoring on a closed-book exam in an online environment. Despite using a soft supervision system based on webcam surveillance, we found significant differences between students' scores depending on whether or not they were subject to video monitoring. Specifically, the mean score for the treatment group, composed of the students under surveillance, was one-third of a standard deviation lower than the mean score for the control group. The experimental conditions between the treatment and control groups were identical except for webcam monitoring. We also control for different potential factors driving these differences in scores. The first relevant mechanism to consider is the possibility that video surveillance may increase students' anxiety during the exam. Contrary to the results of other authors, we did not find any significant differences between supervised and unsupervised students in terms of the time it took them to complete the exam. Furthermore, the proportion of students who completed the exam on time and the number of questions skipped on the test are similar in both groups. These findings rule out the idea that differences in exam scores stem from the fact that non-monitored students felt more relaxed and freer to use the time allotted.

Students' heterogeneity could also be a potential explanation of the differences in the scores observed between the treatment and control groups. In order to isolate the causal impact of webcam-based surveillance, we control for students' characteristics such as academic background, gender, academic degree, course instructors, and repeater or non-repeater status. After considering all these sources of observed individual heterogeneity, we obtained significant and robust differences in academic performance between proctored and unproctored students. These findings can be taken as evidence in favor of differences in scores being due to dishonest practices in taking the test without supervision rather than a result of anxiety or heterogeneity factors.

To conclude, we can state that it is possible and inexpensive to prevent, or at least to minimize, cheating in online exams. Our results reveal that when students had to keep their webcam on, they were less likely to cheat. Therefore, instructors may want to consider using web cameras during off-site exams to deter cheating. This low-touch anti-cheat intervention can be an effective alternative to more intrusive and expensive alternative mechanisms, such as artificial intelligence-based software.

Appendix

Five days before the final exam of Introduction to Microeconomics in the 2020–2021 academic year, one of the authors electronically signed the following procedure and posted it on the Moodle site of the Universitat Jaume I. This document was hidden from students but is available on request.

1. We assigned each student enrolled in the course a unique numerical identifier. The student identifiers were built by taking the last six digits of their ID or equivalent document, as recorded in the databases of the Universitat Jaume I.
2. All students whose identifier was above the median were assigned to the "Even" group. The remaining students formed the "Odd" group.
3. We determined which students had to turn on their cameras using the winning number of the first prize of the "Sueldazo del fin de semana de la ONCE" lottery on Sunday, January 24, 2021 (https://www.juegosonce.es/historico-resultados-coupones-once):

   - If the last digit of the winning number were even, the students in the "Even" group would be asked to turn on their cameras.
   - If the last digit of the winning number were odd, the students in the "Odd" group would be asked to turn on their cameras.

Acknowledgements This work has been supported by Project PID2021-123053OB-I00 funded by MCIN/AEI/10.13039/501100011033/ and by ERDF A way of making Europe; and by the Universitat Jaume I [UJI-B2021-23; UJI-B2020-57] and the Generalitat Valenciana (AICO/2021/005).

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Alessio, H. M., Malay, N., Maurer, K., Bailer, A. J., & Rubin, B. (2017). Examining the effect of proctoring on online test scores. *Online Learning, 21*(1), 146–161.

Beck, V. (2014). Testing a model to predict online cheating: Much ado about nothing. *Active Learning in Higher Education, 15*(1), 65–75.

Berkey, D., & Halfond, J. (2015). Cheating, Student Authentication and Proctoring in Online Programs. *New England Journal of Higher Education*.

Blau, I., Goldberg, S., Friedman, A., & Eshet-Alkalai, Y. (2021). Violation of digital and analog academic integrity through the eyes of faculty members and students: Do institutional role and technology change ethical perspectives? *Journal of Computing in Higher Education, 33*(1), 157–187.

Brallier, S. A., & Palm, L. J. (2015). Proctored and unproctored test performance in traditional and distance courses. *International Journal of Teaching and Learning in Higher Education, 27*(2), 221–226.

Brimble, M. (2016). Why students cheat: An exploration of the motivators of student academic dishonesty in higher education. In Bretag T. (ed.). *Handbook of Academic Integrity*. Springer.

Butler-Henderson, K., & Crawford, J. (2020). A systematic review of online examinations: A pedagogical innovation for scalable authentication and integrity. *Computers & Education, 159*, 104024.

Carstairs, J., & Myors, B. (2009). Internet testing: A natural experiment reveals test score inflation on a high-stakes, unproctored cognitive test. *Computers in Human Behavior, 25*(3), 738–742.

Comas-Forgas, R., Lancaster, T., Calvo-Sastre, A., & Sureda-Negre, J. (2021). Exam cheating and academic integrity breaches during the COVID-19 pandemic: An analysis of internet search activity in Spain. *Heliyon*, 7(10).

Daffin, L. W., Jr., & Jones, A. A. (2018). Comparing student performance on proctored and non-proctored exams in online psychology courses. *Online Learning, 22*(1), 131–145.

Dendir, S., & Maxwell, R. S. (2020). Cheating in online courses: Evidence from online proctoring. *Computers in Human Behavior Reports, 2*, 100033.

Espinosa, M. P., & Gardeazabal, J. (2020). The gender-bias effect of test scoring and framing: A concern for personnel selection and college admission. *The B.E. Journal of Economic Analysis & Policy, 20*, 20190316.

Etter, S., Cramer, J. J., & Finn, S. (2007). Origins of academic dishonesty: Ethical orientations and personality factors associated with attitudes about cheating with information technology. *Journal of Research on Technology in Education, 39*(2), 133–155.

Fask, A., Englander, F., & Wang, Z. (2014). Do online exams facilitate cheating? An experiment designed to separate possible cheating from the effect of the online test taking environment. *Journal of Academic Ethics, 12*(2), 101–112.

Gold, S. S., & Mozes-Carmel, A. (2009). A comparison of online vs. proctored final exams in online classes. *Journal of Educational Technology, 6*(1), 76–81.

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

Howard, D. (2019). *Online testing integrity in a general education math course: A correlational study*. Doctoral Dissertation. American College of Education.

Hylton, K., Levy, Y., & Dringus, L. P. (2016). Utilizing webcam-based proctoring to deter misconduct in online exams. *Computers & Education, 92–93*, 53–63.

Iriberri, N., & Rey-Biel, P. (2019). Competitive pressure widens the gender gap in performance: Evidence from a two-stage competition in mathematics. *The Economic Journal, 129*, 1863–1893.

Janke, S., Rudert, S. C., Petersen, Ä., Fritz, T. M., & Daumiller, M. (2021). Cheating in the wake of COVID-19: How dangerous is ad-hoc online testing for academic integrity? *Computers and Education Open, 2*, 100055.

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).

Ladyshewsky, R. K. (2015). Post-graduate student performance in 'supervised in-class' versus 'unsupervised online' multiple choice tests: Implications for cheating and test security. *Assessment and Evaluation in Higher Education, 40*(7), 883–897.

Lee, S. D., Kuncel, N. R., & Gau, J. (2020). Personality, attitude, and demographic correlates of academic dishonesty: A meta-analysis. *Psychological Bulletin, 146*(11), 1042.

Montolio, D., & Taberner, P. A. (2021). Gender differences under test pressure and their impact on academic performance: A quasi-experimental design. *Journal of Economic Behavior & Organization, 191*, 1065–1090.

OECD. (2020). *Remote Online Exams in Higher Education During the COVID-19 Crisis*. OECD Publishing.

Pekkarinen, T. (2015). Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance. *Journal of Economic Behavior & Organization, 115*, 94–110.

Prince, D. J., Fulton, R. A., & Garsombke, T. W. (2009). Comparisons of proctored versus non-proctored testing strategies in graduate distance education curriculum. *Journal of College Teaching & Learning, 6*(7), 51–63.

Stuber-McEwen, D., Wisely, P., & Hoggatt, S. (2009). Point, click, and cheat: Frequency and type of academic dishonesty in the virtual classroom. *Online Journal of Distance Learning Administration, 12*(2), 1–10.

Truszkowski, D 2019, *Proctored Versus Non-Proctored Testing: A Study for Online Classes*. Doctoral Dissertation. American College of Education.

Vazquez, J. J., Chiang, E. P., & Sarmiento-Barbieri, I. (2021). Can we stay one step ahead of cheaters? A field experiment in proctoring online open book exams. *Journal of Behavioral and Experimental Economics, 90*, 101653.

Watson, G. R., & Sottile, J. (2010). Cheating in the digital age: Do students cheat more in online courses? *Online Journal of Distance Learning Administration*, 13(1).

Yates, R. W., & Beaudrie, B. (2009). The impact of online assessment on grades in community college distance education mathematics courses. *The American Journal of Distance Education, 23*(2), 62–70.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Maite Alguacil is Full Professor in the Department of Economics and board member and co-founder of the Instituto of International Economics at the Universitat Jaume I, where she obtained her PhD in Economics. She holds a MSc in Economics and International Economics at the University of Nottingham. She coordinates the research group on International Economics: Global Value Chains, Migration, Innovation and Adaptation to Climate Change (CAMINA). Her research fields of interest are migration, trade and technology transfer, GVC and location of multinationals. Her work has been published in several academic journals, including International Review of Economics & Finance, Economic Modelling, Journal of International Development, etc.

Noemí Herranz-Zarzoso is Assistant Professor at the Analysis Department of Economics at University of Valencia. She completed her PhD at University Jaime I. Her research interest lie in the experimental and behavioral economics field. She has published in journals like Scientific Reports, Frontiers in Psychology, Journal of Behavioral and Experimental Economics, etc.

José Pernías works at the Department of Economics of the University Jaume I. His main research interest is applied econometrics. His work has been published in several journals, including American Economic Review, Journal of Industrial Economics, Transportation Research (Part A), Tourism Management, etc.

Gerardo Sabater-Grande is Associate Professor (tenured) of the Department of Economics and member and co-founder of the Laboratorio de Economía Experimental (LEE) of the University Jaume I of Castellón. He completed his Ph.D. at the University Jaume I and his undergraduate studies at the University of Valencia. His research interests lie in the area of experimental and behavioral economics. He has published in journals like Scientific Reports, Journal of Risk and Uncertainty, Frontiers in Psychology, Journal of Economic and Behavior Organization, Journal of Behavioral and Experimental Economics, Ecological Economics, etc.

Authors and Affiliations

Maite Alguacil[1,2] · Noemí Herranz-Zarzoso[2] · José C. Pernías[3] · Gerardo Sabater-Grande[4]

✉ Maite Alguacil
alguacil@uji.es

Noemí Herranz-Zarzoso
noemi.herranz@uv.es

José C. Pernías
pernias@uji.es

Gerardo Sabater-Grande
sabater@uji.es

[1] Institute of International Economics & Economics Department, Universitat Jaume I., Av. Vicent Sos Baynat, s/n, 12071 Castellón de la Plana, Castellón, Spain

[2] Department of Economic Analysis, University of Valencia, Av. Taronjers, s/n, 46022 Valencia, Spain

[3] Economics Department, Universitat Jaume I., Av. Vicent Sos Baynat, s/n, 12071 Castellón de la Plana, Castellón, Spain

[4] LEE & Economics Department, Universitat Jaume I., Av. Vicent Sos Baynat, s/n, 12071 Castellón de la Plana, Castellón, Spain