Academic Integrity and Exam Performance in the Online Environment

Phillip Hartley
Georgia Gwinnett College
E-mail: Hartley@ggc.edu

Luis Torres
Georgia Gwinnett College
E-mail: LTorres1@ggc.edu

P. Wesley Routon
Georgia Gwinnett College
E-mail: PRouton@ggc.edu

Abstract

The COVID-19 pandemic has led colleges and universities to augment their online offerings. For faculty and students, this brings concerns about academic integrity and performance. Findings of prior research conflict as to the prevalence of cheating and differences in student performance online versus in-person. Some scholars have studied methodologies for properly researching the phenomenon and ways to prevent cheating. One approach to preventing online cheating is electronic proctoring. In response to the discrepancies in the literature and the increased proliferation of online instruction, this work analyzes student exam performance in-person versus online versus online with video monitoring.

Keywords*:* Exam Format, Online Testing, Online Exams, Exam Proctoring, Distance Learning

Introduction

Online course offerings increased drastically across institutions of higher education in 2020 (Gallagher & Palmer, 2020). Digital learning became the norm when the COVID-19 Pandemic hit the world, and analysts expect most institutions will continue to use similar class modalities going forward (Burke, 2020). Besides the inherent challenges to providing remote instruction, evaluation and assessment are in new realms for both faculty and students. Concerns about academic integrity and the associated performance on exams are the focus of this study.

As schools transitioned to digital teaching, three primary evaluation modalities became widely available. In-person was still available in some critical cases (e.g, instruction needing specific space such as laboratories), online – students took tests in a digital remote setting, and monitored – several alternatives that used technology to proctor exam takers electronically. This last approach is designed to be a stricter alternative to prevent online cheating, but one prior study found no differences in performance with or without proctoring and no additional indication of cheating among the students who were not proctored (Hollister & Berenson, 2009). On the other hand, Vazquez et al. (2021), for instance, found differences when proctoring online.

Considering the need to maintain ethical behavior in an online learning environment and the range of available technologies to evaluate remotely, instructors are faced with a difficult decision. Furthermore, inconsistencies in the literature, which create additional uncertainty, served as the authors' main motivation to analyze student performance in the aforementioned settings (Daffin & Jones, 2018). Exam scores from multiple sections across four courses offered in the fall of 2019 and the spring and fall of 2020 serve as our data. Results show that students perform differently on tests with different delivery formats.

Theoretical Background

Assessment, which is crucial in the context of accreditation, continues to gain popularity as a topic of study in academics. Hence, the relationship between teaching and learning, assessed usually by assignments that include exams, has been the subject of several studies in higher education (e.g., Laborde et al., 2006; Sherin, 2002; Turney et al., 2009). Unethical behavior while completing these assessments, put simply, cheating, could be mitigated by incorporating technology in different evaluation methods.

Academic Integrity

Observation has improved with newly available technologies that facilitate monitoring exam takers in a remote setting (Ranger et al., 2020). Still, concerns about cheating are so prevalent that researchers have experimented with different methodologies for properly researching the phenomenon and ways to prevent it (e.g., D'Souza & Siegfeldt, 2017; Hollister & Berenson, 2009).

Academic honesty has been an ongoing topic of interest (e.g., Beck, 2014; Goff et al., 2020). Different demographic variables such as country of origin, age, academic level, and field of study have been used to investigate potential differences in academic misbehavior. Voelker et al. (2012) show that academic dishonesty is similar at both the undergraduate and graduate levels, though this changes when classes shift from the face-to-face environment to online. Specifically, students in an in-person environment seem to better understand plagiarism in the context of writing assignments and engage differently than students do online. The relationship between cheating and academic performance was also studied, and substantial differences between high and low performers were found. Academic misconduct was more present in low performers, specifically. One reason might be different average levels of understanding regarding plagiarism. Apparently, top performers exhibited better understanding, while results show that bottom performers did not understand that they were doing something wrong (Voelker et al., 2012).

In-Person vs. Online Performance

The idea of similar behavior in remote and in-person settings may be cumbersome to accept for instructors who are used to more traditional ways of evaluation, given online learners often have little supervision while completing assignments. Yet, at least one study found no evidence of differences in student performance online versus in-person and identified three factors potentially affected similarity online and in-person: (1) the computational context; (2) if it was lecture oriented; and (3) if the exam questions needed to apply methods learned in class (Spivey & McMillan, 2014). In contrast, Sevim-Cirak and Islim (2022) found that the students who took quizzes in the classroom had higher exam and retention test scores than those who took quizzes online. In addition, some studies have also found a positive relationship between online learning systems and effectiveness for improving student performance (Buttner & Black, 2014; Turney et al., 2009). Still other findings suggest the testing environment may hinder performance in an online setting (Fask et al., 2014). Thus, the body of research in this area demonstrates major inconsistencies.

Specifically regarding academic integrity, Kidwell and Kent (2008) found that online students are less likely to engage in academic misconduct than face-to-face students. However, other research suggests that cheating is widespread among students taking online exams (Golden & Kohlbeck, 2020). These discrepant findings suggest that further research is needed, but instructors of online classes must carry on despite the possibility of academic dishonesty.

Proctoring online examinations was developed as one potential solution to academic dishonesty, and some scholars insist it is needed (e.g., Reisenwitz, 2020). Proctoring has been found as an effective tool to mitigate unethical behavior with instructors, sometimes combining it with additional tools such as allowing open book exams (Dendir & Maxwell, 2020). Even though it has raised some ethical concerns about the use of artificial intelligence and student privacy, it has been adopted by a wide array of higher education institutions (Coghlan et al., 2021; Selwyn et al., 2021). Yet some instructors prefer not to proctor online exams (Dadashzadeh, 2021). Beyond the ethical concerns, proctoring can come with other drawbacks. In particular, proctoring comes with costs of time and money for both students and institutions (Cluskey et al., 2011). In addition, Akaaboune et al. (2021) suggest that proctoring affects lower-performing students' test performance more so than higher-performing individuals. Given these concerns, some scholars seek to identify and improve upon various methods for proctoring online exams, but the debates on the overall effectiveness of proctoring and the best method for doing so continue (Asep & Bandung, 2019; Atoum et al., 2017; Hylton et al., 2016; Selwyn et al., 2021).

Methods and Materials

Data for this work comprise test scores from multiple terms across which three different test modalities were utilized. In total, the data include 1,458 observations (exam grades) from 528 business students at a 4-year state college in the Southeastern United States. The terms were all full-length terms—either fall or spring instead of the shorter, more concentrated summer term. Test modalities included in-person exams given in the classroom and proctored by the professor, online exams without any form of monitoring, and exams administered online with video monitoring. The use of monitoring for online exams was at the discretion of course instructors. The sample includes four business courses ranging from core business curriculum to upper-level electives. Within each course, the instructors, textbooks, and test banks from which questions were drawn were consistent across the sample period. In the online conditions, instructors delivered exams via a learning management system that enabled the randomization of questions and answers, as well as restrictions on when and for how long students could access the tests. Exams comprised a combination of multiple choice and short essay questions, with the essay questions representing 30% of the exam grade. The exam scores included in the analysis reflect students' overall performance on these exams, which includes points earned from both multiple-choice and short essay questions. In our most complex regression models, we include section fixed effects. Since there is only ever one professor per section within the sample, this methodology addresses the unobserved heterogeneity across professors, who are likely to grade short essay answers differently.

In the case of monitored online exams, students were proctored via an automated platform that utilized video and audio recording as well as artificial intelligence that flags instances of possible integrity violations, including disappearing from view or appearing to focus attention offscreen. The final data set did not include any instances in which students were identified as cheating. Regardless of modality, students were not allowed to use textbooks, notes, or any other outside materials while taking exams. Course syllabi, which are considered contracts between the instructors and students, all included a statement of the academic integrity policies for the given course. Thus, any violation of these policies would constitute cheating.

Following a descriptive analysis and statistical hypothesis testing, we turn to multivariate techniques to best examine differences across exam delivery format. Specifically, we estimate regressions of the form:

$$\text{score}_{s,e} = \alpha + \beta\text{online}_e + \gamma\text{monitored}_e + \delta\mathbf{X}_{s,e} + \varepsilon_{s,e} \qquad (1)$$

where $\text{score}_{s,e}$ is student s's numerical score on exam e; α a constant term; online_e an indicator for exams that were administered online but without monitoring, with β its corresponding parameter; monitored_e an indicator for exams that were administered online and where digital monitoring was implemented, with γ its corresponding parameter; $\mathbf{X}_{s,e}$ a vector of control variables, with δ its corresponding vector of parameters; and $\varepsilon_{s,e}$ the usual (in this case, well-behaved) error term. Of key interest are estimates of β and γ, the estimated average impacts of an exam's delivery format on student performance. Since $\text{score}_{s,e}$ is continuous in nature (can take on any value 0 through 100 for all exams sampled), ordinary least squares is used initially. Then, to additionally adjust for unobservables, we estimate fixed effects regressions otherwise similar to equation (1). In some models, the fixed effects are at the course and term (semester) levels. In others, the fixed effects are at the course section level. To adjust for a trivial amount of heteroscedasticity, robust standard errors are always calculated.

The controls in $\mathbf{X}_{s,e}$ include the student's homework average in that specific course; the number of multiple-choice questions on the exam; the number of essay questions; the average score on that specific exam in that section; the standard deviation of those scores, as a measure of variability; and the chronology of the exam (e.g., a value of 3 for the third exam of the semester). An additional control was sometimes used: a student's score on their most recent exam in that specific course section. For example, for all Exam 3 observations, the value for this control is their Exam 2 score. Key results from one such model are presented in Appendix Table A1. We note that this additional control does not change our primary findings (estimates of β and γ) in any significant way, and that including it requires a significant reduction in sample size (all Exam 1 observations must be dropped as there is no prior exam, as do observations following an exam missed by a student). Thus, we consider the inclusion of this additional control a robustness check of our primary estimates.

When implemented, course and term fixed effects hold constant unobserved heterogeneity across the sampled courses and semesters (e.g., Angrist & Pischke, 2008; Kennedy, 2008; Wooldridge, 2010). Section fixed effects address even more unobserved heterogeneity, any across the sampled course sections. Among other things, this includes differences across instructors (since there is always one instructor per section within the sample), course design, course delivery, teaching approach, and how instructors facilitate exams. Since a section is always within both a particular course and term, section fixed effects additionally address heterogeneity across the sampled courses and semesters, as did the course and term fixed effects. To provide a robust set of estimates, we estimate and report results from models at varying levels of control. Models are then re-estimated within key subsamples of student scores.

We are additionally interested in whether the impacts of exam delivery differ across the distribution of student scores. That is for example, are those impacts noticeably different for high-scoring students than they are for low-scoring students? To that end, a quantile regression model of the form

$$score_{s,e} = \alpha_\tau + \beta_\tau \text{online}_e + \gamma_\tau \text{monitored}_e + \delta_\tau \mathbf{X}_{s,e} + \varepsilon_{\tau,s,e} \qquad (2)$$

with

$$Q_\tau(score_{s,e}|\mathbf{X}_{s,e}) = \alpha_\tau + \beta_\tau \text{online}_e + \gamma_\tau \text{monitored}_e + \delta_\tau \mathbf{X}_{s,e} \qquad (3)$$

is estimated. Here, τ denotes a conditional score quantile (Koenker & Bassett, 1978), and all other terms are the same as in equation (1). Quantile regression allows one to simultaneously estimate how a model's coefficients vary across the distribution of the dependent variable (Koenker & Hallock, 2001). Specifically in the context here, it allows us to estimate if and how much an exam's administration format impacts performance across the distribution of student scores. Compared to subsampling, quantile regression has several clear advantages, including that it makes use of all available information (the full sample) when calculating coefficient estimates. We choose $\tau = \{0.05, 0.25, 0.50, 0.75, 0.95\}$, which is a standard choice set in the literature that uses this methodology (e.g., Brown & Routon, 2016, 2018; Edie & Showalter, 1998; Koenker & Hallock, 2001). This allows us to estimate exam delivery effects at the 5^{th}, 25^{th}, median, 75^{th}, and 95^{th} conditional score quantiles, that is, from the worst performers (those near the 5^{th} percentile exam score) to the best performers (those near the 95^{th} percentile exam score).

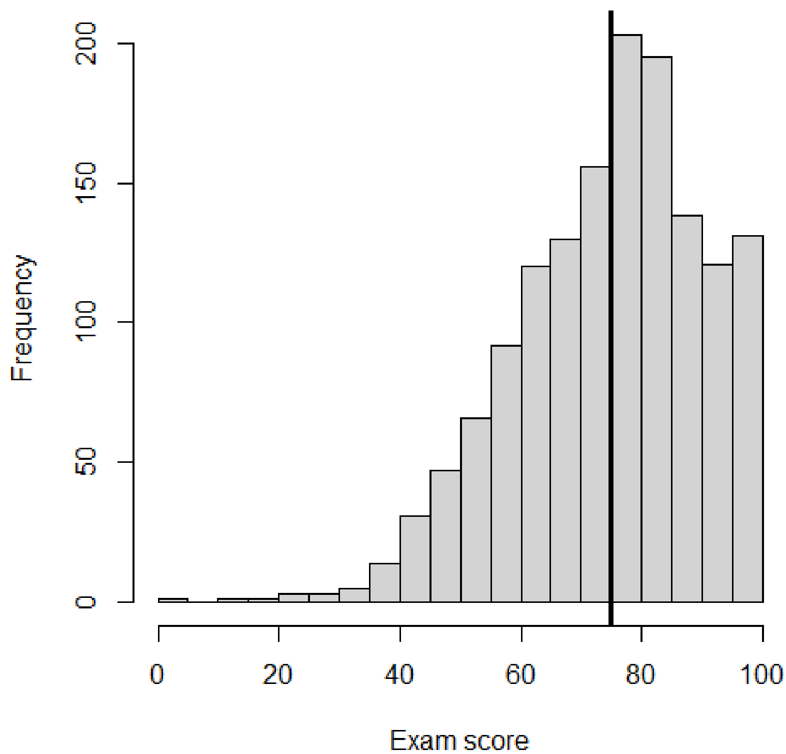
Results

An assessment of overall test performance reveals that the distribution of sampled scores is left-tailed with a mean of 74.99. Figure 1 presents this sampled distribution. Collectively, students are perhaps not performing exceedingly well on exams. However, group performance is in line with historical standards.

Analysis of students' performance by delivery format shows that scores on tests administered in-person (σ=15.320, where σ is standard deviation) or online with monitoring (σ=15.228) varied more than those administered online without monitoring (σ=13.472). Hence, students' scores were more consistent or similar in the condition of online testing without monitoring. Drawing any meaningful conclusion from these simple variability estimates alone is not feasible, but that there appears to be a difference motivates deeper examination.

Comparing mean scores of students across exam delivery format provides greater insight. Shown in Table 1, these comparisons reveal that students taking exams online without any form of proctoring scored, on average, almost 10.5 points higher than students taking exams in-person and more than 8 points higher than students who were monitored online. In contrast, the difference of less than 2 points in favor of students monitored while taking exams online versus those who took exams in-person was only significant at the 90% level (p = 0.061). Putting these mean differences in terms of a typical grading scale, this indicates students scored about a letter grade higher on online exams without proctoring relative to exams given in the classroom or online with proctoring.

Figure 1: Histogram of exam scores



Notes: Bold line is the mean (74.99). 25 bins used. n = 1,458.

Table 1: Hypothesis test results, *t*-tests of mean equality for exam scores

Sample 1	Sample 2	Mean 1	Mean 2	t	p
In-person	Online	72.912	84.495	10.469	$2.200e^{-16}$
In-person	Monitored	72.912	74.582	-1.874	0.061
Monitored	Online	74.582	84.495	8.285	$1.707e^{-15}$

Notes: Here, "online" refers to online exams that were not monitored/proctored, "monitored" refers to online exams that were monitored/proctored, and "in-person" refers to traditional in-class exams. n (in-person) = 800; n (online) = 195; and n (lockdown) = 463.

Table 2 presents the key results from five regression models estimated using the full sample. Moving from left to right in the table, the models have an increasing level of control, that is, more factors held constant through larger sets of variables within $\mathbf{X}_{s,e}$ or switching to fixed effects regression. At the highest available level of control, column (5) in Table 2, online exams without monitoring are estimated to increase student scores by 10.29 points above a similar in-person exam, on average. This estimate is statistically significant at the 99% confidence level. In contrast, there appears to be no significant difference between average student performance on in-person exams versus monitored online exams, with estimated differentials never reaching 1.7 exam points or the 95% confidence level.

Table 2. Key regression results, full sample models

	(1)	(2)	(3)	(4)	(5)
Full sample (n = 1,458)					
Online without monitoring	11.58***	11.68***	13.05***	11.83***	10.29***
	(1.20)	(1.13)	(1.13)	(1.52)	(1.56)
Monitored online	1.67*	1.34	1.30*	1.48*	-1.66
	(0.88)	(0.83)	(0.77)	(0.82)	(2.06)
Adjusted R^2	0.06	0.17	0.29	0.30	0.34
Controlling for…					
Student's non-exam class avg.		X	X	X	X
Exam characteristics			X	X	X
Course & term fixed effects				X	
Section fixed effects					X

Notes: Values are marginal effects with robust standard errors in parentheses. *$p < 0.10$; **$p < 0.05$; ***$p < 0.01$. Both marginal effects should be compared to traditional in-person (proctored) exams. Subsample sizes: n (in-person) = 800; n (online without monitoring) = 195; and n (Monitored online) = 463. Exam characteristics include (*i*) the number of multiple-choice questions; (*ii*) the number of essay questions; (*iii*) average score; (*iv*) the standard deviation of scores; and (*v*) chronology (e.g., 1 for Exam 1 of the semester). Columns (1)-(3) show results from OLS regressions, while columns (4)-(5) show results from fixed effect regressions.

Table 3 presents results from otherwise identical models estimated using three subsamples. The top panel shows how the results differ when the A grades (scores of 90 or above), the top exam performances, are excluded. The key estimates are consistently positive and significant within all models for the online without monitoring condition, and all but one model for the monitored online condition were non-significant. Looking at the most robust model, Model 5, online exams without monitoring are now estimated to increase scores by 7.92 points, while monitored online exams are again estimated not to yield a significantly different average performance as compared to in-person exams. The next subsample excludes the other extreme of student performance. The results for this analysis are shown in the second panel of Table 3. The panel shows that when F grades (scores below 60), the weakest exam performances are excluded, the estimated difference changes to 8.68 points for online exams without monitoring, while the difference is again statistically non-significant for monitored online exams.

The final panel of Table 3 shows results from models using a subsample in which both A and F grades are excluded. One can see that the Model 5 estimate for online exams without monitoring shrinks to 5.64 points. Meanwhile, the estimated difference in the monitored online condition remains statistically nonsignificant. Overall, the results in Table 3 imply that the benefit of online exams without monitoring, whatever it may be, remains when the tails of the distribution (the best and/or worst performing students) are no longer considered. That the marginal effects are smaller here than in the full sample models may simply stem from the fact that the range of the dependent variable has decreased; it could imply that a significant portion of the benefit is being had by the A and F students, or both. Particularly in the context of section fixed effects, the single grade subsample sizes (e.g., just "A" grades) were largely deemed too small for individual regression analysis. Thus, the results in Table 3 come from models where, at most, two letter grades are removed.

Table 3. Key regression results, subsample models

	(1)	(2)	(3)	(4)	(5)
"A" grades excluded (*n* = 1,182)					
Online without monitoring	7.86***	8.97***	10.91***	8.78***	7.92***
	(1.32)	(1.29)	(1.25)	(1.61)	(1.78)
Monitored online	1.08	0.84	1.05	1.64**	-0.44
	(0.84)	(0.82)	(0.78)	(0.83)	(2.17)
"F" grades excluded (*n* = 1,206)					
Online without monitoring	7.69***	7.99***	8.97***	9.05***	8.68***
	(0.88)	(0.84)	(0.90)	(1.17)	(1.23)
Monitored online	0.85	0.74	0.78	0.60	1.06
	(0.68)	(0.65)	(0.62)	(0.67)	(1.59)
"As" & "Fs" excluded (*n* = 930)					
Online without monitoring	4.64***	5.01***	6.40***	5.49***	5.64***
	(0.83)	(0.84)	(0.85)	(1.08)	(1.22)
Monitored online	0.33	0.29	0.39	0.61	1.22
	(0.57)	(0.57)	(0.55)	(0.60)	(1.45)
Controlling for…					
Student's homework avg.		X	X	X	X
Exam characteristics			X	X	X
Course & term fixed effects				X	
Section fixed effects					X

Notes: Values are marginal effects with robust standard errors in parentheses. $*p < 0.10$; $**p < 0.05$; $***p < 0.01$. Both marginal effects should be compared to traditional in-person (proctored) exams. Exam characteristics include (*i*) the number of multiple-choice questions; (*ii*) the number of essay questions; (*iii*) average score; (*iv*) the standard deviation of scores; and (*v*) chronology (e.g., 1 for Exam 1 of the semester). Columns (1)-(3) show results from OLS regressions, while columns (4)-(5) show results from fixed effect regressions.

While the subsample analysis provides an initial glimpse of differences in test delivery format across student performance, a more robust method for exploring differentials is (fixed effects) quantile regression. The key results from such a model are presented in Table 4. Differences across the distribution of scores are apparent and demonstrate how delivery format impacts can be expected to vary across students who perform at different levels. Usefully, as shown in the second row of the table, in the sample here, the literature standard set of taus (the 5th, 25th, 50th, 75th, and 95th percentiles) correspond to numerical scores within each of the 5 typical letter grades (F, D, C, B, and A, respectively), assuming a typical grading scale.

At the median of the distribution, "typical" students perform significantly better (+14.68 points) on online exams without monitoring and demonstrate relatively small (+1.35 points) non-significant differences in performance on monitored online exams, both compared to in-person exams. The gap between score estimates for online without monitoring and in-person exams narrows as you move toward the highest-performing group of students (+6.61 points) but is largest (+15.65 points) for the second lowest-performing group (what one may choose to call "D" students). The second-smallest difference was found for the lowest-performing group (+8.85 points). The only statistically significant estimated difference between performance on monitored online and in-person exams was found for the second highest performing group (+3.49 points), and it was considerably smaller than the estimate for the same subsample for the Online without monitoring condition (+11.66 points). Most notably, whatever benefit unmonitored online exams grant persists across the entire distribution of student performance, though at varying degrees.

Table 4. Key quantile regression results

Tau (τ):	0.05	0.25	0.50	0.75	0.95
Full sample score at that percentile:	48	65	77	86	97
Full sample (n = 1,458)					
Online without monitoring	8.85**	15.65***	14.68***	11.66***	6.61***
	(4.47)	(1.73)	(1.24)	(1.13)	(0.75)
Monitored online	-0.59	0.62	1.35	3.49***	0.44
	(2.11)	(1.06)	(1.04)	(0.92)	(0.75)

Notes: Values are marginal effects with standard errors in parentheses. *p < 0.10; **p < 0.05; ***p < 0.01. All marginal effects should be compared to traditional in-person (proctored) exams. Results come from a single fixed effects quantile regression that includes the full set of controls and section fixed effects, comparable to model (5) of Table 2.

Discussion

Overall exam performance within this sample exhibits a seemingly typical range and distribution of scores. Though this certainly does not prove students are *not* cheating, it also does not seem to indicate they are. In addition, this would seem to suggest that exams are, on average, generally fair assessments of students' knowledge. The overall distribution does not, however, consider differences across exam delivery format.

Ceteris paribus, if the three formats served as equally fair assessments, neither significant average differences in performance nor significant (estimated) marginal effects for any of the formats would exist. Those differences and effects *do* exist in this sample, so the next question is "Why?" The answer is unclear and untestable with available data. What is clear is that students do perform differently on tests across delivery format. A possible explanation is that some students consistently cheat on exams administered online without proctoring, but cheating is no more likely during online exams with proctoring than during exams given in the live setting. Given at least an approximate 10-point difference in average scores, a non-trivial differential, one must consider that students might cheat on online exams without proctoring.

Splitting the sample by exam grade indicates that students consistently experience an increase in test performance in the “online without monitoring” condition. The results of the quantile regression analysis offer a detailed look at performance differentials across the distribution of scores. From the results, it appears the benefit of the online format without monitoring is greatest for those nearer the middle of the pack, so to speak, and smallest for the highest performers.

When considering students at the extremes, both the lowest and highest performers, it is perhaps unsurprising that average performance within these groups is more similar across test delivery formats than it is within other students. Logically, the best performers can only do so much better, regardless of the format. However, the average discrepancy across formats for top performers is still quite substantial at over 6 points. With regard to the lowest performers, one would not expect the test format to affect scores as much as other things such as preparation or prior knowledge, so a relatively small estimated difference across formats may be expected. For these students, though, opportunity and temptation to take advantage of the unmonitored condition would both seem greatest of all. The relatively small estimated difference in scores across delivery formats for these students, coupled with the largest standard error of all the quantile estimates, might suggest less consistent response to the test format for lowest performers. As such, it may be that low performers either cheat without effectively improving the resulting grade beyond a failing score, or they may simply cheat less often.

Examining quantile regression results for the students more within the interquartile range (the 25th through the 75th percentiles), it appears that those students experience a boost to test scores ranging from about 12 points to over 15 points, about a letter grade and a half, in the online without monitoring condition. Maybe the median student is more able to avoid distractions in the online environment than in a classroom full of other students. Alternately, perhaps the digitization of exams primes higher levels of self-efficacy in today’s typical digital native students, leading to better performance. In either case, it could be that individual observation during an exam could cause these students performance anxiety leading to added stress and less positive outcomes. On the other hand, perhaps the median student is simply more able, and/or more inclined, to use disallowed outside resources such as notes, books, or electronic devices to aid their test performance. As noted above, these differences could stem from many different origins. Unfortunately, determining the actual causes is beyond both the scope of this research and what can be accomplished with available data.

One persistent concern in academics is that past students may share information about exams with current students, making cheating easier and more prevalent across time. If a significant amount of this activity was occurring, and assuming professors are not updating their exams, one may expect mean grades within a course to be generally increasing through time, all else equal, perhaps up to a certain plateau. Also, the distribution of grades could trend toward a bimodal distribution if this were true, such that students with access to shared information would perform much better on exams than those without access. Within the sample here, mean grades are statistically different across semesters but not consistently increasing within any of the four courses, and there is no apparent trend in their distributions. Additionally, partitioned R^2s show that the semester of the exam never explains more than 5.3% of the variation in grades, holding other factors constant. Ultimately, some students may indeed be sharing exam information across semesters, but we find no evidence this is happening at a significant level within this particular sample. Further, we control for the semester of the exam in all of our full models.

Broadly speaking, this research cannot confirm or dispute any suspicion that students may cheat on assessments given online. However, it can be said that the results presented here do not support the inverse relationship between academic performance and academic integrity espoused by a previous study described above (Voelker et al., 2012). Further, proctoring online exams appears to nullify any advantage, unfair or otherwise, students gain from taking exams virtually instead of in the classroom. In terms of exam performance, our analyses suggest that students perform just as well in the online environment as they do in-person. This finding may be of comfort to instructors who harbor misgivings about the increased frequency of online course offerings.

Conclusion

As classes and, therefore, exams shift increasingly to virtual formats, a concern educators may have is whether this arrangement is as conducive to class performance as the traditional live setting. The "silver lining" perspective on these findings is that, on average, students perform equally well, statistically speaking, on tests given online and monitored as they do on tests given in the classroom. This may be a source of reassurance for some, but the glaring question of what is driving the improvement in performance in the online without monitoring condition remains. Could it be that, in the unmonitored condition, high-performing students offer their skills as test-takers for hire? While this is possible, online testing platforms offer plagiarism detection through applications such as Turnitin, which allow instructors to identify text duplication. While this does not eliminate the possibility of students posing as another or others or working together on exams, the use of such applications in combination with longer form or essay questions would increase the burden on the test-taker(s) by requiring them to alter their answers each time they took the exam. Another possibility in the unmonitored condition is that students could use outside materials such as notes or textbooks while completing exams. In addition to the use of plagiarism detection as described above, the inclusion of large numbers of questions or questions that require more critical thinking and less memorization, along with time limits on exams, should limit the feasibility of using such materials during exams. Future research may seek to explore these possibilities, but our findings suggest the use of proctoring for online tests.

A noteworthy limitation of this analysis is that the sample is comprised solely of exams taken in business courses, lowering the external validity of the findings to the general pool of college students. This is particularly important given there is some evidence that business majors are differently prone to cheating behaviors, on average, with some researchers suggesting they are more likely to cheat (e.g., McCabe & Trevino, 1995) and other research suggesting they are less likely (e.g., Simkin & McLeod, 2010). This limitation could be addressed through a future analysis on a similar sample, but one that includes exams from a plethora of disciplines.

An additional avenue for future research is the investigation of exam completion timing. For example, if students are given a two-day window to complete an online exam, is there more evidence of cheating on Day 2 than Day 1, given that questions can be shared among students? More related to the analysis here would be the follow-up question of whether exam sharing is more prevalent in the unmonitored online modality than in the monitored online modality. Such a study would require data on when each student's exam was completed, not available in the present sample. Educators concerned with exam sharing can mitigate the practice by shortening availability windows and through the randomization of exam questions. Even in the absence of exam sharing, shorter availability windows and question randomization may reduce the score gaps between exam modalities, as they arguably make all forms of cheating more difficult.

Other specific areas for further investigation include comparisons of performance on various test modalities between classes with quantitative material versus those that are more conceptual in nature. In addition, while the data analyzed here did not facilitate this type of examination, it would be interesting to explore the potential for self-selection effects in terms of which students opt for face-to-face as opposed to online course modalities. Lastly, this research only included one means of monitoring students during online exams. Given that various methods of proctoring online exams are available, their relative effectiveness and respective impacts on student performance should be studied more.

REFERENCES

Akaaboune, O., Blix, L. H., Carrington, L. & Henderson, C. (2021). Accountability in distance learning: The effect of remote proctoring on performance in online accounting courses. *Journal of Emerging Technologies in Accounting*. https://doi.org/10.2308/JETA-2020-040

Angrist, J. D., & Pischke, J. S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Asep, H. S., & Bandung, Y. (2019). A design of continuous user verification for online exam proctoring on M-Learning. In *2019 International Conference on Electrical Engineering and Informatics (ICEEI)*, 284-289.

Atoum, Y., Chen, L., Liu, A. X., Hsu, S. D., & Liu, X. (2017). Automated online exam proctoring. *IEEE Transactions on Multimedia*, *19*(7), 1609-1624.

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

Brown, C. & Routon P. W. (2016). Military service and the civilian labor force: Time- and income-based evidence. *Armed Forces & Society*, *42*(3), 562-584.

Brown, C. & Routon P. W. (2018). On the distributional and evolutionary nature of the obesity wage penalty. *Economics & Human Biology, 28*(1), 160-172.

Burke, L. (2020). Moving into the long term. *Inside Higher Ed*, https://www.insidehighered.com/digital-learning/article/2020/10/27/long-term-online-learning-pandemic-may-impact-students-well, accessed online November 2, 2020.

Buttner, E. H. & Black, A. N. (2014). Assessment of the effectiveness of an online learning system in improving student test performance. *Journal of Education for Business*, *89*(5), 248–256.

Cluskey Jr, G. R., Ehlen, C. R., & Raiborn, M. H. (2011). Thwarting online exam cheating without proctor supervision. *Journal of Academic and Business Ethics*, *4*(1), 1-7.

Coghlan, S., Miller, T. & Paterson, J. (2021). Good proctor or "Big Brother"? Ethics of online exam supervision technologies. *Philosophy & Technology*, 1-26.

Dadashzadeh, M. (2021). The online examination dilemma: To proctor or not to proctor?. *Journal of Instructional Pedagogies*, 25.

Daffin Jr, L. W. & 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*, 100033.

D'Souza, K. A. & Siegfeldt, D. V. (2017). A conceptual framework for detecting cheating in online and take-home exams. *Decision Sciences Journal of Innovative Education, 15*(4), 370-391.

Edie, E. & Showalter, M. H. (1998). The effect of school quality on student performance: A quantile regression approach. *Economics Letters, 58*(3), 345–350.

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.

Gallagher, S. & Palmer, J. (2020). The pandemic pushed universities online. The change was long overdue. *Harvard Business Review*, https://hbr.org/2020/09/the-pandemic-pushed-universities-online-the-change-was-long-overdue, accessed online November 2, 2020.

Goff, D., Johnston, J. & Bouboulis, B. S. (2020). Maintaining academic standards and integrity in online business courses. *International Journal of Higher Education, 9*(2), 248-257.

Golden, J. & Kohlbeck, M. (2020). Addressing cheating when using test bank questions in online classes. *Journal of Accounting Education, 52*, 100671.

Hollister, K. K. & Berenson, M. L. (2009). Proctored versus unproctored online exams: Studying the impact of exam environment on student performance. *Decision Sciences Journal of Innovative Education, 7*(1), 271-294.

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

Kennedy, P. (2008). *A Guide to Econometrics* (6th ed.). Blackwell Publishers.

Kidwell, L. A. & Kent, J. (2008). Integrity at a distance: A study of academic misconduct among university students on and off campus. *Accounting Education: An International Journal, 17*(S1), S3–S16.

Koenker, R. & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

Koenker, R. & Hallock, K. (2001). Quantile regression. *Journal of Economic Perspectives, 15*(4), 143–156.

Laborde, C., Kynigos, C., Hollebrands, K., & Strässer, R. (2006). Teaching and learning geometry with technology. *Handbook of Research on the Psychology of Mathematics Education*, 275-304.

McCabe, D. L. & Trevino, L. K. (1995). Cheating among business students: A challenge for business leaders and educators. *Journal of Management Education, 19*(2), 205-218.

Ranger, J., Schmidt, N. & Wolgast, A. (2020). The detection of cheating on e-exams in higher education—the performance of several old and some new indicators. *Frontiers in Psychology*, *11*, 2390.

Reisenwitz, T. H. (2020). Examining the necessity of proctoring online exams. *Journal of Higher Education Theory and Practice, 20*(1), 118-124.

Selwyn, N., O'Neill, C., Smith, G., Andrejevic, M., & Gu, X. (2021). A necessary evil? The rise of online exam proctoring in Australian universities. *Media International Australia*, 1329878X211005862.

Sevim-Cirak, N. & Islim, O. F. (2022). Paper versus online quizzes: Which is more effective?. *Active Learning in Higher Education*, 1–18, 14697874221079737.

Sherin, M. G. (2002). When teaching becomes learning. *Cognition and Instruction, 20*(2), 119-150.

Simkin, M. G. & McLeod, A. (2010). Why do college students cheat? *Journal of Business Ethics*, *94*, 441-453.

Spivey, M. F. & McMillan, J. J. (2014). Classroom versus online assessment. *Journal of Education for Business, 89*(8), 450-456.

Turney, C. S. M., Robinson, D., Lee, M., & Soutar, A. (2009). Using technology to direct learning in higher education: The way forward?. *Active Learning in Higher Education, 10*(1), 71-83.

Vazquez, J. J., Chiang, P. E. & 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.

Voelker, T. A., Leonard, L. G. & Pentina, I. (2012). Plagiarism: What Don't They Know?. *Journal of Education for Business, 87*(1): 36–41.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). MIT Press.

Appendix

Table A1. Results from a model with an additional control (student's last exam score)

Reduced sample (n = 934)	
Online without monitoring	10.97***
	(2.63)
Monitored online	-0.70
	(3.02)
Adjusted R^2	0.53

Notes: Values are marginal effects with robust standard errors in parentheses. *p < 0.10; **p < 0.05; ***p < 0.01. Both marginal effects should be compared to traditional in-person (proctored) exams. The model includes the full set of controls and section fixed effects, as in model (5) of Table 2, plus a student's exam score on their most recent previous exam in that course section. The sample size is necessarily reduced since there are no previous exam scores for all Exam 1 observations and in the few cases where a student missed the previous exam.

Table A2. Descriptive summaries and ANOVA results

Exam modality	*n*	Mean	Median	Mode	Variance	Std. dev.
In-person	800	72.9	74.3	77	234.7	15.3
Monitored	463	74.6	76.0	78	231.9	15.2
Online	195	84.5	87.0	100	181.5	13.5

ANOVA

Source of variation	*df*	*SS*	*MS*	F	Critical F	*p*-value
Between groups	2	21,149	10,574	46.642	3.002	< 0.00001
Within groups	1,455	329,866	227			
Total:	1,457	351,015				

Notes: Monitored refers to online exams that made use of student monitoring software, while Online refers to online exams that did not make use of such software. *df* = degrees of freedom. *SS* = sum of squares. *MS* = mean sum of squares.