# Learning to Cheat: Quantifying Changes in Score Advantage of Unproctored Assessments Over Time

**Binglin Chen, Sushmita Azad, Max Fowler, Matthew West, Craig Zilles**
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{chen386, sazad2, mfowler5, mwest, zilles}@illinois.edu

## ABSTRACT

Proctoring educational assessments (e.g., quizzes and exams) has a cost, be it in faculty (and/or course staff) time or in money to pay for proctoring services. Previous estimates of the utility of proctoring (generally by estimating the score advantage of taking an exam without proctoring) vary widely and have mostly been implemented using an across-subjects experimental designs and sometimes with low statistical power.

We investigated the score advantage of unproctored exams versus proctored exams using a within-subjects design for $N = 510$ students in an on-campus introductory programming course with 5 proctored exams and 4 unproctored exams. We found that students scored 3.32 percentage points higher on questions on unproctored exams than on proctored exams ($p < 0.001$).

More interestingly, however, we discovered that this score advantage on unproctored exams grew steadily as the semester progressed, from around 0 percentage points at the start of semester to around 7 percentage points by the end. As the most obvious explanation for this advantage is cheating, we refer to this behavior as the student population "learning to cheat". The data suggests that both more individuals are cheating and the average benefit of cheating is increasing over the course of the semester. Furthermore, we observed that studying for unproctored exams decreased over the course of the semester while studying for proctored exams stayed constant. Lastly, we estimated the score advantage by question type and found that our long-form programming questions had the highest score advantage on unproctored exams, but there are multiple possible explanations for this finding.

### Author Keywords

assessment; cheating; CS1; proctoring; STEM; exams

## INTRODUCTION

Assessment is a remarkably important part of education. Ideally it motivates engagement, evaluates progress, provides diagnostic feedback, and can be a learning experience in its own right [Waugh and Gronlund 2012, Laverty et al. 2016, Hartwig and Dunlosky 2012]. Educational assessments can also have a significant and lasting impact on students' lives through grades and standardized test scores. The high stakes of some educational assessments can serve as a motivation to cheat for some students.

Previous work has led to multiple theories about under what circumstances students might cheat. For example, Fraud Triangle Theory [Cressey 1953] suggests that three elements must be present for fraud (like cheating) to occur: (1) *perceived pressure/incentive/motive*, which in the context of cheating is typically based on a student's inability to earn their desired grade without cheating, (2) *perceived opportunity*, which occurs when insufficient steps are taken to make assessments secure, and (3) *rationalization*, a moral or ethical argument used to justify the behavior while retaining one's self image.

Mazar, et al. characterize the decision to cheat as the interaction of two processes [Mazar et al. 2008]. The first process performs a rational cost-benefit analysis that weighs the potential benefits against the potential risks and their likelihoods; higher expected value increases the likelihood of cheating. The second process evaluates how cheating will harm one's conception of their self; the more consistent the cheating is (or can be rationalized to be) with one's self-concept the more likely it will occur. The theory suggests that cheating occurs when an individual perceives that the expected benefit from cheating outweighs the negative impacts to one's self-concept. In both of these theories, the perception of the risk of getting caught and a student's ability to rationalize the behavior play an important role.

A student's ability to cheat and the risk of being caught is greatly influenced by the precautions an instructor has undertaken to prevent cheating. The most prevalent such precaution is proctoring, where test takers are observed while taking a test to prevent disallowed communication and the use of disallowed materials. Proctoring, however, can be costly in terms of faculty/staff time or in terms of money to pay others to perform proctoring (as is not uncommon for online courses). Furthermore, proctoring creates a logistical burden for both students and faculty. In an ideal world, proctoring would be unnecessary, but there is significant evidence to suggest that we do not live in such an ideal world, as discussed in the related work described in Section 2.



*L@S 2020* Atlanta, Georgia, USA



DOI: TBD

In this paper, we perform an analysis of student data from an on-campus course that included multiple proctored exams and multiple unproctored exams. In this course, exams were constructed individually for each student by randomly drawing questions from large question pools. Because many of the same questions appeared on multiple exams, we were able to estimate the apparent difficulty of a given question when it appeared on an unproctored exam and, separately, when it was on a proctored exam.

Not only did we observe that questions appeared easier on the unproctored exams, but the degree to which they appeared easier increased as the semester progressed. We suspect that this is the result of the class "learning to cheat" as they (1) become more comfortable that they won't get caught cheating on the unproctored exams and (2) can rationalize cheating through the (likely correct) belief that a significant fraction of the rest of the class is also cheating.

Specifically, this paper makes the following contributions:

1. We present the results of a within-subjects experiment to estimate the score advantage on unproctored exams using a class with 510 students, 5 proctored exams, 4 unproctored exams, and a total 128,136 observations.

2. We present, to our knowledge, the first quantitative observations of a student population learning to cheat in aggregate within a single semester course. In particular, we observe the distribution of student score advantages on unproctored exams to shift in the positive direction, indicating both an increased number of cheaters and an increased effectiveness of cheating over the course of the semester.

3. We support the assertion that the score advantage is derived from cheating by demonstrating that the aggregate amount of studying for unproctored exams decreases as the semester progresses, while it remains constant for proctored exams.

4. We show that specific types of questions are more prone to cheating, and we consider question features that might drive these differences. Although our data does not permit us to determine the relative importance of question features for cheating, we develop hypotheses that can potentially drive future work.

This paper is organized as follows. Section 2 describes related work. Section 3 describes the course in which the data was collected and the collected data. Section 4 describes a series of logistic regressions used to estimate the score advantage the students had on unproctored exams. In Section 5, we interpret the result and discuss limitations of our results, and, in Section 6, we conclude.

## RELATED WORK

Cheating and cheating in higher-education has been well studied. In this section, we review self-reported data from students about their cheating behavior and previous studies comparing scores on proctored vs. unproctored exams. Feinman's recent dissertation [Feinman 2018] provides a more exhaustive literature review relating to cheating and proctoring. We only found one previous study that documents a student population's learning to cheat over time, which we discuss at the end of this section.

Studies find that a significant number of students admit to having had cheated, but that the overall rate of cheating appears to have been relatively stable over the 1969–1996 time period [Whitley 1998]. Exam cheating, however, which historically was one of the most infrequent kinds of cheating, saw significant growth in this period [McCabe et al. 2001]. Many studies find GPA to be negatively correlated with cheating [McCabe et al. 2001, Haines et al. 1986, Lanier 2006].

Both faculty and students perceive that online courses are easier to cheat in than face-to-face courses, but that perception diminishes with increased experience with online courses [Kennedy et al. 2000]. The majority of studies of self-reported behavior have found higher rates of cheating in online exams [Lanier 2006, Vician et al. 2006, Stephens et al. 2007, Watson and Sottile 2010], including higher rates of using unallowed materials [Stephens et al. 2007], giving help to others [Lanier 2006], and getting help from other students [Watson and Sottile 2010]. Other surveys have found similar rates of cheating [Grijalva et al. 2006] or lower rates in online courses [Stuber-McEwen et al. 2009]. In one study, 46% of students reported having knowledge of other students getting help on online quizzes or exams [Watters et al. 2011]. In another, students reported that they believe academic integrity means different things in online and classroom environments [Cole and Swartz 2013].

Previous studies comparing unproctored to proctored exams have mixed results, with many showing a significant score advantage on unproctored exams while others finding statistically equivalent scores. We provide a sample of each. In all but two of these studies, exam scores on conserved exams were compared *across* cohorts of students (either across sections during the same semester or across semesters) under different testing conditions. Some of these studies use regressions to try to control for other characteristics of the students (e.g., GPA, year of student, gender) while others make no attempts to compare the populations.

Daffin and Jones performed a within-subject design study that compared exam scores across a range of 14 online psychology classes ($N = 1694$) [Daffin Jr and Jones 2018]. In each semester, a single exam was selected to be remotely proctored (the rest were unproctored), but which exam was selected was rotated across semesters. In the semesters where a given exam was unproctored, scores were 16.5 percentage points higher (more than a standard deviation higher) on average. In addition, while the exams were limited to 60 minutes in both conditions, students in the unproctored condition took an average of 20 minutes longer to complete the exam (48.2 vs. 27.7 minutes).

An on-campus undergraduate engineering dynamics course also used a within-subject design. This course included four quizzes (in addition to four proctored exams) and used a controlled crossover experimental design to randomly assign students to two groups that alternated between two treatments:

(1) taking the quizzes online, asynchronously and unproctored, and (2) taking the quizzes asynchronously proctored at their Evaluation and Proficiency Center [DeMara et al. 2016]. Students ($N = 276$) averaged 37.5 percentage points higher (84% higher) in the unproctored format [Nader et al. 2019].

Davis et al. compared final exam scores in two upper division accounting courses under three conditions: an online course with no proctoring, an online course with remote proctoring, and an on-campus course classroom with face-to-face proctoring. Unproctored students had scores that were 15% higher than remotely proctored and 9% higher than face-to-face proctored students ($N = 261$) [Davis et al. 2016].

Richardson and North compared the exams scores in four online business administration courses between semesters where exams were unproctored and a semester where proctored testing services were arranged [Richardson and North 2013]. In 19 of the 22 unproctored exams, the scores were statistically significantly higher than the same exam in the proctored semester with an average score advantage of 7.9 percentage points (0.63 standard deviation, $N = 333$).

Hollister and Berenson compared the exam scores of two sections of the same "Introduction to Computing for Business" course, where one section took proctored exams at a set time and the other section took them unproctored asynchronously over a 4-day window, but with the same 60 minute time limit. The unproctored groups scored 1 percentage point higher, but it wasn't statistically significant ($N = 173$) [Hollister and Berenson 2009].

Ladyshewsky compared the performance on a group of 50 multiple choice questions used in a graduate-level management and leadership business course between 136 students proctored at a set time (first 4 offerings) vs. 114 students that took exams asynchronously (last 5 offerings). There were no obvious trends to the scores, but scores in unproctored offerings were lower on average [Ladyshewsky 2015].

Stack compared exam scores in a criminological theory course between semesters where students were proctored at a fixed time to semesters where students were unproctored, but took the exam online at a fixed time using a lock down browser. Across the ten sections compared ($N = 287$), the unproctored students did slightly worse, but the result wasn't statistically significant [Stack 2015].

Beck compared mid-term and final exam scores in an introductory course with a live proctored section, a remotely proctored, asynchronously tested section, and an unproctored, asynchronous tested section. On both exams, scores on the unproctored exams ($N = 19$) were less than 1 point higher than the proctored sections ($N = 80$), but the results were not statistically significant.

The only previously published evidence we found of a population learning to cheat in exams was Martinelli et al.'s study of mathematics standardized testing in 88 high schools in Mexico under a variety of incentive conditions [Martinelli et al. 2018]. In this work, the authors statistically estimated the number of cheaters by counting the number of pairs of students taking the exam in the same room with a high degree of overlap in their answers relative to similar amounts of overlap in the whole test taking population. They found that conditions that provided students with monetary incentives based on their score had higher levels of cheating than those that did not. Furthermore, they observed that the number of cheating students in incentivised cohorts of students grew each year that they took the test, from 9% in 10th grade, to 25% in 11th grade, to 31% in 12th grade (on average) compared to 7% to 7% to 8.5% in student cohorts with no incentives. While this previous work estimates the growth in the number of cheaters over a period of years on an annual standardized test, our work observes the growth in score advantage by a course's population within a single semester.

## METHODS

This data collection was performed in an introductory programming course for non-technical majors during the Fall 2019 semester. The course had 602 students (247 female and 355 male) complete the semester. Because the course is specifically required by the College of Business, the majority of students (337) were business majors with a mix of other majors and undeclared making up the remainder. Students take this course early in their college career; the class was 205 freshmen, 265 sophomores, 83 juniors, 48 seniors, and one graduate student.

### Course organization

The course is organized as a flipped course, with one 90-minute lecture and one lab section per week. Before lectures, a reading and an assignment consisting of true/false and multiple choice questions based on the reading is to be completed. After lecture, a collection of 20–25 short answer (e.g., "what is the value of x after this code executes", "write a line of code that inserts the value x in a list at position y") and multi-line programming questions are due weekly.

All of these assignments are run through the PrairieLearn learning management system [West et al. 2015]. PrairieLearn provides students immediate feedback on the correctness of their answers and allows the writing of question generators, which use randomness to produce a large collection of potential question instances. On homework assignments, when students answer incorrectly they are given feedback on their answer or shown a correct solution, and given an opportunity to attempt another version of the question. Students can repeat questions until full credit on the homework is earned. PrairieLearn is also used for the class's computerized exams; on exams multiple attempts at each question are often permitted, but generally for decreased credit after each incorrect attempt.

These computerized exams, when proctored, are run in our campus's Computer-Based Testing Facility (CBTF) [Zilles et al. 2019]. The CBTF is a proctored computer lab that allows students to schedule their exams at times convenient to them during an alloted range of days. Students have access to a Python interpreter and Python language documentation on CBTF computers, but the networking is controlled to prevent unwanted communication or web browsing.

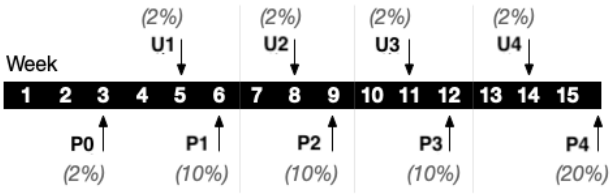


**Figure 1. During the fifteen week semester, 5 proctored exams were run in the Computer-Based Testing Facility (P0 to P4) and 4 unproctored exams (U1 to U4) were run in each of the weeks preceding proctored exams. The contribution of each exam to the final grade is shown.**

Throughout the fifteen week semester and finals period nine summative computer-based exams were performed, as shown in Figure 1. Five of the exams, referred to as the *proctored exams*, were proctored in the CBTF. These proctored exams were each worth 10% of the final grade, except the first (E0, worth 2%) which was intended to be a gentle introduction for students to the CBTF and the comprehensive final exam (E4, worth 20%). The other four exams, referred to as *unproctored exams*, were unproctored and could be taken wherever the student chose. These unproctored exams were each worth 2% of the final grade. For communication with students, the unproctored exams were called "quizzes" and the proctored exams were called "exams". All of the exams were run asynchronously for a three-day period, Thursday through Saturday. All of the exams had an enforced 50-minute nominal duration, except the final which allowed 3 hours.

All of the exams were heavily randomized. Each exam consisted of a series of 20–30 slots (the final had 41 slots). For each slot, there was a pool of questions of similar difficulty related to a given learning objective (e.g., writing conditional statements) and each student was given a question randomly drawn from the pool. Assessments consisted of a mix of roughly (by points) 25% true/false and multiple choice questions, 45% short answer questions, and 30% multi-line programming questions. Each true/false and multiple choice slot drew from a separate pool of 20 to 100 distinct questions; pools for short answer and programming question usually consisted of 5 to 12 different questions, with no overlap between pools. In addition, the short answer questions were randomly parameterized (e.g., changing the code to be read or the single line of code to be written). The goal for the exams was for them to sample from a pool much larger than any student could memorize without learning the material.

The majority of the questions on the exams were drawn from questions that were on pre-lecture and homework assignments. As such, the summative exams were intended to merely verify that students had learned to do the things that were on the homework. Approximately 70% of questions were shared between proctored and unproctored exams, with 46% appearing first on an unproctored exam and then later on a proctored exam, and the remainder being the reverse. For every exam, students were provided practice exam generators that allowed them to take as many practice exam as they wanted. The practice exam were generated from the same pools as the actual exam generator, except that 10–15% of the slots for the actual exams were replaced with pools of "hidden" questions that were used only on actual exams.

I understand that this quiz is intended to be completed by myself without assistance from others and without the use of any materials besides (1) what is provided by the quiz and (2) a Python interpreter. I agree to complete the quiz without the help of others and without disallowed materials.

We won't accept scores of students that do not comply by this honor code.

- (a) I understand and I agree to comply by this honor code
- (b) I refuse to comply by this honor code

**Figure 2. The first question of every unproctored exam and its corresponding practice exam was this honor pledge.**

In addition to the course credit and proctoring differences, unproctored exams (and their corresponding practice exams) included an honor pledge as the first question, as shown in Figure 2. This policy was also explained in lecture. After completing their unproctored exams, students could continue to view them until the end of the unproctored exam period.

### Data collected

We collected the scores on each question for each student for each of the exams. After removing individuals that didn't complete all of the exams, there were 510 students in the data set. There were 347 question generators used on exams. Some of these question generators represented a bundle of 10–30 topically related true/false or multiple choice questions. Because the exams were randomized, each student generally saw a different subset of the questions on each exam. In total, our data set had 128,136 student-question score observations. Each of these scores range from 0 (incorrect) to 1 (completely correct), with some questions granting partial credit.

## RESULTS

### Score advantage without proctoring

We first consider the overall score advantage in unproctored exams. Because question scores are bounded between 0 and 1 and typically have Bernoulli or other highly non-normal distributions, we use logistic regression models of question scores in terms of student, question, and environment coefficients.

The logistic model we fitted to study the overall score advantage in unproctored exams is of the form

$$z = \sigma\left(\sum_{i=1}^{m} \alpha_i s_i + \sum_{j=1}^{n} \beta_j q_j + \gamma u\right) \tag{1}$$

over all the observations, where $\sigma$ is the logistic function, $m$ is the total number of students, $n$ is the total number of question generators, and $z, s_i, q_j, u$ are observed values from each observation, defined as follows:

- $z$: the score the student got in the observation, a real number between 0 and 1,

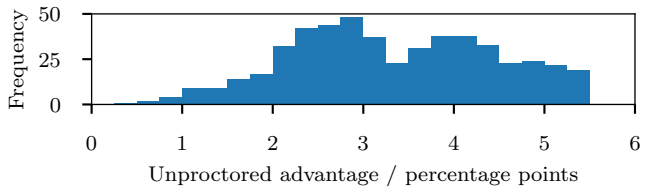


**Figure 3. The distribution of $c_i$, which can be interpreted as the distribution of per-student unproctored advantage.**

- $s_i$: 1 if the observation is associated with student $i$, 0 otherwise,
- $q_j$: 1 if the observation is associated with question generator $j$, 0 otherwise,
- $u$: 1 if the observation is from an unproctored exam, 0 otherwise,

and $\alpha_i, \beta_j, \gamma$ are the coefficients that we want to estimate, which can be interpreted as:

- $\alpha_i$: the ability of student $i$,
- $\beta_j$: the difficulty of question $j$,
- $\gamma$: the advantage of answering questions in unproctored exams.

Since the coefficients estimated in a logistic model do not directly measure percentage score, the second step is to convert the logistic model results to percentage scores that are readily interpretable. To achieve this, we utilize the scores predicted by the fitted logistic model over all students and over all questions under both proctored and unproctored conditions. Specifically, letting $\hat{z}_{ij0}$ and $\hat{z}_{ij1}$ be the predicted scores of student $i$ answering question $j$ under proctored and unproctored conditions, respectively, we denote $c_i$ as the advantage student $i$ has under the unproctored condition, which is computed as follows:

$$c_i = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{z}_{ij1} - \hat{z}_{ij0}\right). \tag{2}$$

We plotted the distribution of $c_i$ in Figure 3. As the figure shows, the per-student unproctored advantages range from 0 to 6 percentage points. The mean unproctored advantage is 3.32 (95% CI [3.22, 3.42]) percentage points, which is significantly more than zero ($p < 0.001$) and corresponds to an effect size of $d = 0.21$ standard deviations of exam scores.

#### Time dependence (learning to cheat)

To understand how the unproctored advantage changes over the course of the semester for each student, we fitted a logistic model of the form

$$z = \sigma\left(\sum_{i=1}^{m}\alpha_i s_i + \sum_{j=1}^{n}\beta_j q_j + \sum_{i=1}^{m}\sum_{\ell=1}^{r}\delta_{i\ell} s_i u_\ell\right), \tag{3}$$

over all of the observations, where $r$ is the total number of unproctored exams—which equals 4 in our case—and $u_\ell$ is the observed value from each observation, which is 1 if the observation is associated with the $\ell$th unproctored exam. The coefficient $\delta_{i\ell}$ can be interpreted as the advantage of student $i$ answering questions in the $\ell$th unproctored exam.

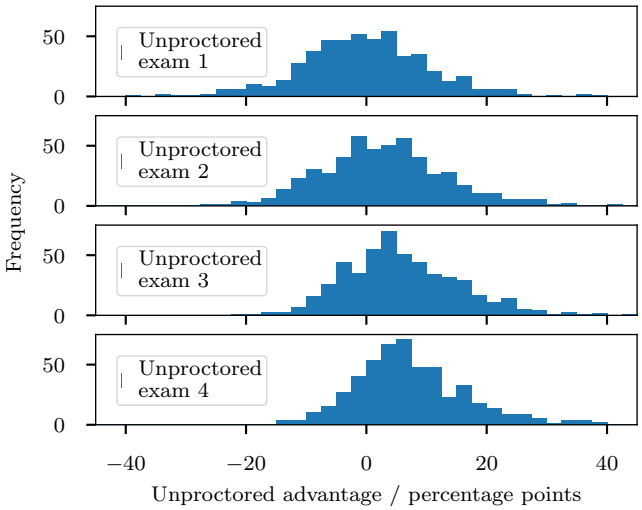


**Figure 4. The distribution of $d_{i\ell}$ for each $\ell$, which can be interpreted as the distribution of per-student unproctored advantage in unproctored exam $\ell$.**

To obtain the unproctored advantage of each student for each unproctored exam, we again utilize the scores predicted by the fitted logistic model over all students and over all questions under both proctored and unproctored conditions. Specifically, let $\hat{z}_{ij0}$ and $\hat{z}_{ij\ell}$ be the predicted scores of student $i$ answering question $j$ on proctored exams and on the $\ell$th unproctored exam, respectively. We then denote by $d_{i\ell}$ the unproctored advantage student $i$ has in the $\ell$th unproctored exam, which is computed as follows:

$$d_{i\ell} = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{z}_{ij\ell} - \hat{z}_{ij0}\right). \tag{4}$$

We plotted the distribution of $d_{i\ell}$ for each $\ell$ separately in Figure 4. As the figure shows, there is a clear trend where the mass of the distribution shifts to the right as $\ell$ increases, which suggests that the unproctored advantage is increasing over the course of the semester. We computed the mean of $d_{i\ell}$ for each $\ell$ and plotted them in Figure 5. The plot shows that the mean unproctored advantage in unproctored exam 1 is around 0, and it rapidly increases to around 3 percentage points in unproctored exam 2, is around 6 points in unproctored exam 3, and finally reaches 7 points in unproctored exam 4. This seems to suggest that cheating in unproctored exams was not prevalent at the beginning of the semester, but students soon found cheating in unproctored exams profitable and capitalized on it. The final unproctored advantage of 7 percentage points corresponds to an effect size of $d = 0.42$ standard deviations of exam scores and about 2/3 of a letter grade, and is higher than that for Unproctored exam 1 ($p < 0.001$).

#### Relationship to aggregate studying behavior

To further explore our learning-to-cheat hypothesis, we analyzed how much studying students undertook before proctored

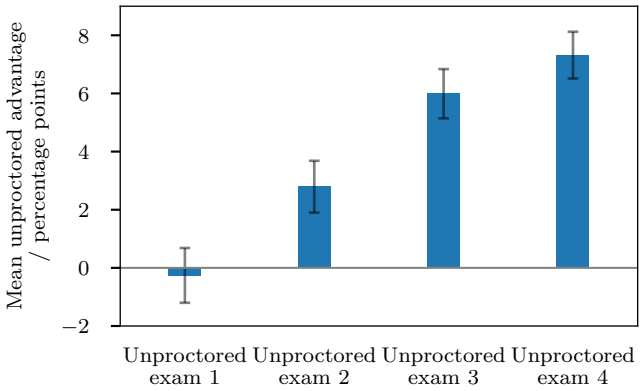


**Figure 5. The mean of $d_{i\ell}$ for each $\ell$, which can be interpreted as the mean of unproctored advantage in unproctored exam $\ell$. Error bars are 95% confidence intervals.**

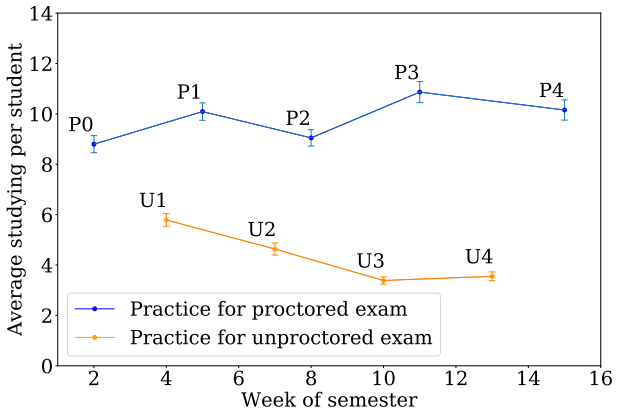


**Figure 6. Studying measured by mean submissions per practice exam question per student. Using median gives a qualitatively similar plot. Error bars are 95% confidence intervals.**

exams versus before unproctored exams as the semester progressed. Because all of the relevant homework and practice exams were offered through the PrairieLearn platform, we can use the amount of interaction with the platform as a proxy for the amount of student studying. Specifically, we computed the mean number of submissions made per student to the practice exam (offered before the exam), normalized by the number of questions per exam. This normalization is done to control for variation in the length of practice exams.

The studying data is plotted in Figure 6. We see that the amount of studying for proctored exams was relatively constant, with all values being within 12% of the mean studying across all proctored exams. Furthermore, there is no clear pattern to the amount of proctored exam studying; if anything, it increased as the semester progressed. In contrast, the amount of studying for unproctored exams clearly decreased as the semester progressed. There was a 39% drop from the amount studied for U1 to the amount studied for U4 (from an average of 5.8 practice exams to 3.6 practice exams). Furthermore, the evolution of unproctored exam studying correlates to the evolution of unproctored advantage shown in Figure 5; there is a large change from U1 to U3 and less change from U3 to U4.

| Question type | Avg. score | Weight | Randomization |
|---|---|---|---|
| Short answer | 90% | 2 | high |
| True/False | 81% | 1 | medium |
| Multiple choice | 86% | 2 | medium |
| Programming | 78% | 3 | low |

**Table 1. Question type characteristics: average score, relative weighting on exams, and relative degree of randomization.**

### Score advantage by question type

The questions on our exams can be largely categorized into four categories: true-false, multiple choice, short answer, and programming questions. Short answer questions generally involve writing a single line of code relating to a single concept. In contrast, we use the label *programming* for questions that generally require multiple lines of code and typically involve integrating multiple concepts together.

Table 1 summarizes characteristics of each type of question, including average score. When on exams, the question types are not equally weighted. Programming questions are generally worth about three times as much as true/false questions and both multiple choice and short answer questions are worth twice as much as true/false questions. Also, the different question types provide different degrees of question randomization on exams. Programming questions generally have the least diversity, since there are usually only 5–12 distinct instances in these question pools. Multiple choice and true/false have more diversity, with usually 20–100 unique instances occupying a given exam slot. Short answer questions typically have the most diversity, because these slots randomly pick from 3–8 question generators, each of which are randomly parameterized to produce one of typically hundreds if not thousands of versions.

To see if the unproctored score advantage favored particular kinds of questions, we fitted a logistic model of the form

$$z = \sigma\left(\sum_{i=1}^{m} \alpha_i s_i + \sum_{j=1}^{n} \beta_j q_j + \sum_{k=1}^{p}\sum_{\ell=1}^{r} \varepsilon_{k\ell} t_k u_\ell\right), \qquad (5)$$

where $p$ is the total number of question types and $t_k$ is the observed value from each observation, which is 1 if the observation is associated with the $k$th question type. The coefficient $\varepsilon_{k\ell}$ can be interpreted as the advantage students have on questions with the $k$th type in the $\ell$th unproctored exam.

To obtain the unproctored advantage of each question type for each unproctored exam, we again utilize the scores predicted by the fitted logistic model over all students and over all questions under both proctored and unproctored conditions. Specifically, let $\hat{z}_{ij0}$ and $\hat{z}_{ij\ell}$ be the predicted scores of student $i$ answering question $j$ in proctored exams and the $\ell$th unproctored exam respectively. Then we denote by $e_{k\ell}$ the unproctored advantage students have on questions with the $k$th type in the $\ell$th unproctored exam, which is computed as follows:

$$e_{k\ell} = \frac{1}{\sum_{j=1}^{n} v_{jk}} \sum_{j=1}^{n} v_{jk}\left(\hat{z}_{ij\ell} - \hat{z}_{ij0}\right), \qquad (6)$$

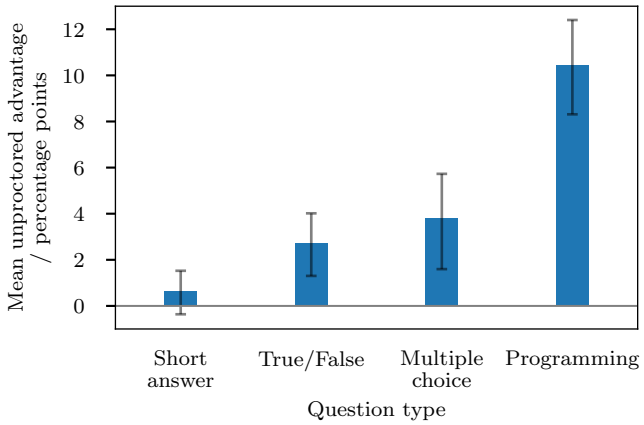


**Figure 7. The unproctored advantage per question type. Error bars are 95% confidence intervals.**

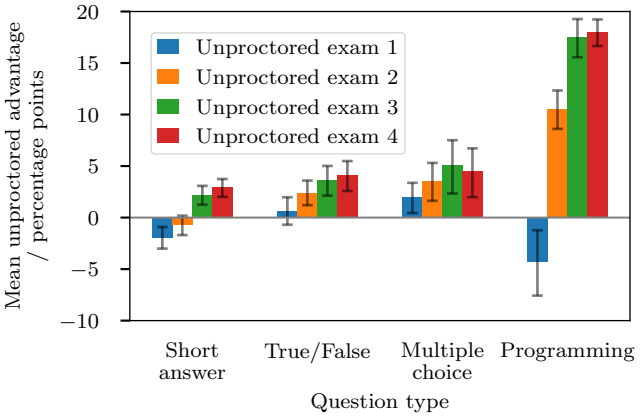


**Figure 8. The unproctored advantage per question type, per unproctored exam. Error bars are 95% confidence intervals.**

where $v_{jk}$ is 1 if the $j$th question is of the $k$th type and 0 otherwise.

Figure 7 shows the overall unproctored advantage per question type. For true/false, multiple choice, and programming questions the unproctored advantage is statistically significantly positive, with programming questions having the largest effect size by far. Furthermore, all question types exhibit the learning-to-cheat trend as shown in Figure 8. In fact, while the positive unproctored advantage observed overall for short answer questions isn't statistically significant, this appears to be because it has the smallest effect size and begins negative in U1. By U3 and U4, even the short answer questions have a statistically significantly positive unproctored advantage.

## DISCUSSION AND LIMITATIONS

There are multiple possible explanations why students could be performing better on unproctored exams than on proctored exams in our Computer-Based Testing Facility (CBTF). Besides cheating, the most compelling one is that students are more comfortable taking unproctored exams in the environment of their choice and on their own computer. Advocates of take-home exams cite reduction of student anxiety as a primary advantage of the format [Bengtsson 2019]. It is unclear, however, whether this reduction in anxiety would translate to increased advantage as the semester progresses.

Instead, we believe that cheating is the dominant effect leading to the measured unproctored advantage. Three aspects of existing fraud and cheating theories can explain the observed learning-to-cheat effect. First, as students have more experience with a given testing context, they can better assess what security precautions are being taken, both from first-hand experience as well as from communicating with fellow students. This knowledge translates into greater perceived opportunity to cheat and a reduced perceived risk of cheating. In particular, if cheaters are not caught, they and anyone they share this information with are likely to be emboldened.

Second, repeated exposure to a testing context allows students to refine their cheating strategies. What might begin with students helping each other take the unproctored exam may develop into students preparing resources (e.g., searchable repositories of previously used questions) for use during the unproctored exam. Such refined strategies would increase the expected value of cheating, leading to a larger effect size.

Third, students will be more able to rationalize cheating if they know or believe that other students are successfully cheating without being punished.

In addition to matching the theory, two other pieces of data support the claim that cheating is the dominant cause of the unproctored advantage. First, the reduction of unproctored exam studying relative to the amount of proctored exam studying, even while the unproctored advantage is rising, suggests that students are trading off cheating for study effort on the unproctored exams. Second, the particular pattern of how different question types have different unproctored advantages provides further evidence. Questions that are harder, have higher weight, and are less randomized tend to have higher unproctored advantages. This is what we would expect if cheating were the cause.

Beyond the data analyzed in this paper, we have observed instances of learning-to-cheat behavior in other contexts. For example, the operators of the CBTF on our campus report having to introduce new security measures over time. One such measure is that students' attempts to bring unauthorized paper into the CBTF led to the use of colored scratch paper, but later the CBTF switched to varying the color of scratch paper used each hour in unpredictable ways when students were found trying to sneak cheat sheets in on colored paper.

We also saw similar trends with students "gaming the system" on a practice activity within the class. In this activity, there was no penalty for incorrect answers and, when an answer was submitted, the activity would show the correct answer (so that the student could learn) and randomly cycle to another version of the question. In the first week of this activity, we found that a small fraction of students were providing nonsensical answers in order compel the activity into sharing its solutions and then waiting for it to cycle back to one of the previously seen versions. Each subsequent week, we saw a larger and larger fraction of students engaging in this behavior, until we introduced a mechanism to discourage this behavior.

### Why so much cheating on programming questions?

When we examined the unproctored score advantage by question type, programming questions had by far the largest effect sizes. There are four aspects of programming questions that may contribute to this large effect size relative to other question types. At present, we don't know which effects are most important.

First, programming questions are worth more points than other question types. Programming questions were worth $1.5\times$ to $3\times$ as much as other question types, and these higher stakes could be a driving factor behind student behavior.

Second, programming questions were, on average, the most difficult question type, with an average score of 78%. As such, these questions offer the most potential benefit from cheating. In general, the unproctored advantage effect size is loosely inversely correlated with average score (e.g., short answer questions had the smallest effect size and the highest average score).

Third, programming questions had the least amount of randomization on the exams. It would be most feasible and the least effort for students to build a database of questions and answers for these questions, which they could use during unproctored exams. The much larger pool sizes (true/false, multiple choice) and random generation (short answer) involved in other question types likely increases the effort of such a strategy.

Fourth and finally, students may be most prone to cheat on programming questions because they may be able to most correctly assess when they don't know the answer for that question type. Notably, the two students that were caught bringing cheat sheets into the CBTF had them covered with only answers to programming questions. For the other question types, the student might (incorrectly) have confidence that they can correctly answer the question and not go to the effort of cheating. True/false and multiple choice questions may be perceived as guessable. Our short answer questions rely more heavily on recognition and recall than problem solving, so a student can often produce a plausible attempt even if they can't produce the correct answer. Programming questions, however, would be very challenging to correctly guess.

### Limitations

The primary limitation of this work is that, while we can estimate the score advantage students have on the unproctored exams, we have no actual evidence of the cause of the score advantage. The increase in the score advantage's correlation to the decrease in unproctored exam studying is just that, a correlation. Furthermore, even if it is cheating, we have no means of determining how this learning to cheat effect manifests.

In addition, this study has all of the limitations found in studies in the context of a single course at a single institution. We should question whether these results will generalize to other (non-computing) subjects, other student populations (e.g., STEM students), and other institutions (e.g., those with honor codes, or those with higher/lower selectivity). Furthermore, the course studied is perhaps uncommon in the construction of its exams (heavily randomizing, but having large overlap with the homework questions) and in how it provides practice exam generators to students.

Additionally, there are reasons to believe that these results might not directly generalize to online programs. We suspect that much of the cheating contributing to these results involves collaboration between students and self-justification of cheating by knowing about the cheating behavior of other students. Unlike in a large on-campus course, students in online programs might not know other students in the program or not in the manner/depth in which discussions relating to cheating would occur. As such, cheating in the courses in such online programs might manifest differently, for example by the use of multiple accounts to harvest solutions [Valiente et al. 2016, Northcutt et al. 2015, Ruipérez-Valiente et al. 2019].

## CONCLUSIONS

We presented an analysis of the score advantage of unproctored versus proctored exams with $N = 510$ students in an on-campus CS1 course. We used a within-subjects experimental design that leveraged questions shared between proctored and unproctored exams and found that students scored 3.32 percentage points higher on unproctored exams than proctored exams ($p < 0.001$).

Most notably, these unproctored score advantages increased substantially as the course progressed. Because these score advantage increases occurred despite corresponding decreases in unproctored exam studying (while proctored exam studying stayed relatively constant), we believe that the primary source of this effect is from the course's student population "learning to cheat" in aggregate. We find this effect to be readily explainable in the context of existing cheating theory, as students likely gained confidence that they can successfully cheat with repeated exposure and may be able to justify their behavior if interactions with fellow students lead them to believe that others are cheating as well.

While the unproctored advantage was indistinguishable from zero on the first unproctored exam, by the end of the semester it grew to be an average of 7 percentage points. This effect size cannot be neglected, as it corresponds to 2/3 of a letter grade and $d = 0.42$ standard deviations of exam scores. In addition, a non-trivial number of students had unproctored advantages in excess of twenty percentage points. The per-student distribution of unproctored advantages showed that the entire distribution moved up over the semester, rather than there being a clearly identifiable subpopulation of students who were contributing most of the effect.

We also investigated how the type of question affected the unproctored advantage, and found that harder questions with less randomization and higher weight had higher unproctored advantages, which further supports the conclusion that cheating was the primary cause of the advantage. Furthermore, all question types showed consistent increases in unproctored advantage as the semester progressed.

The substantial time-dependence of cheating behavior that we found is an important consideration for the design of cheating interventions and experiments. Importantly, data from single assessments or short-time interventions has a risk of being

misleading, as it takes time for students to learn to cheat within a novel environment. Our results for exam cheating are consistent with prior work in contexts such as homework answer copying [Palazzo et al. 2010].

In the course studied, these unproctored exams were scheduled the week before exams with the intention of encouraging additional exam preparation and serving as formative assessments students could use to gauge their preparation for the exams. As such, little course credit was allocated to the unproctored exams, with the bulk of the course assessment credit allocated to the proctored exams. While proctoring these exams places a burden on both the institution and the student, we are convinced now more than ever that this burden is necessary.

## ACKNOWLEDGMENTS

This work was partially supported by NSF DUE-1347722, NSF CMMI-1150490, NSF DUE-1915257, and the College of Engineering at the University of Illinois at Urbana-Champaign under the Strategic Instructional Initiatives Program (SIIP).

## REFERENCES

Lars Bengtsson. 2019. Take-Home Exams in Higher Education: A Systematic Review. *Education Sciences* 9, 4 (2019), 267.

Michele T Cole and Louis B Swartz. 2013. Understanding academic integrity in the online learning environment: A survey of graduate and undergraduate business students. *ASBBS Proceedings* 20, 1 (2013), 738.

Donald R Cressey. 1953. *Other people's money; a study of the social psychology of embezzlement.* Free Press.

Lee William Daffin Jr and Ashley A Jones. 2018. Comparing Student Performance on Proctored and Non-Proctored Exams in Online Psychology Courses. *Online Learning* 22, 1 (2018), 131–145.

Ann Boyd Davis, Richard Rand, and Robert Seay. 2016. Remote proctoring: The effect of proctoring on grades. In *Advances in Accounting Education: Teaching and Curriculum Innovations*. Emerald Group Publishing Limited, 23–50.

Ronald F. DeMara, Navid Khoshavi, Steven D. Pyle, John Edison, Richard Hartshorne, Baiyun Chen, and Michael Georgiopoulos. 2016. Redesigning Computer Engineering Gateway Courses Using a Novel Remediation Hierarchy. In *2016 ASEE Annual Conference & Exposition*. ASEE Conferences, New Orleans, Louisiana. https://peer.asee.org/26063.

Lena Feinman. 2018. *Alternative to Proctoring in Introductory Statistics Community College Courses*. Ph.D. Dissertation. Walden University.

Therese C Grijalva, Clifford Nowell, and Joe Kerkvliet. 2006. Academic honesty and online courses. *College Student Journal* 40, 1 (2006).

Valerie J Haines, George M Diekhoff, Emily E LaBeff, and Robert E Clark. 1986. College cheating: Immaturity, lack of commitment, and the neutralizing attitude. *Research in Higher education* 25, 4 (1986), 342–354.

M. K. Hartwig and J. Dunlosky. 2012. Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin and Review* 19 (2012), 126–134.

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

Kristen Kennedy, Sheri Nowak, Renuka Raghuraman, Jennifer Thomas, and Stephen F Davis. 2000. Academic dishonesty and distance learning: Student and faculty views. *College Student Journal* 34, 2 (2000).

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

Mark M Lanier. 2006. Academic integrity and distance learning. *Journal of criminal justice education* 17, 2 (2006), 244–261.

J.T. Laverty, S.M. Underwood, R.L. Matz, L.A. Posey, J.H. Carmel, M.D. Caballero, C. L. Fata-Hartley, D. Ebert-May, S. E. Jardeleza, and M. M. Cooper. 2016. Characterizing college science assessments: The three-dimensional learning assessment protocol. *PLoS ONE* 11, 9 (2016), e0162333. https://doi.org/10.1371/journal.pone.0162333

César Martinelli, Susan W Parker, Ana Cristina Pérez-Gea, and Rodimiro Rodrigo. 2018. Cheating and incentives: Learning from a policy experiment. *American Economic Journal: Economic Policy* 10, 1 (2018), 298–325.

Nina Mazar, On Amir, and Dan Ariely. 2008. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research* 45, 6 (2008), 633–644.

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

Marino Nader, Ronald F DeMara, Adrian Tatulian, and Baiyun Chen. 2019. Quantitative Impact on Learning Achievement by Engaging High Integrity Testing using Lockdown Assessment for Online Delivery. In *ASEE Southeast Section Conferences*. ASEE Conferences, Raleigh, NC.

Curtis G. Northcutt, Andrew D. Ho, and Isaac L. Chuang. 2015. Detecting and preventing "multiple-account" cheating in massive open online courses. *Computers & Education* 100 (2015), 71–80.

David J. Palazzo, Young-Jin Lee, Rasil Warnakulasooriya, and David E. Pritchard. 2010. Patterns, correlates, and reduction of homework copying. *Phys.*

*Rev. ST Phys. Educ. Res.* 6 (Mar 2010), 010104. Issue 1. https://doi.org/10.1103/PhysRevSTPER.6.010104

Ronny Richardson and Max North. 2013. Strengthening the trust in online courses: A common sense approach. *Journal of Computing Sciences in Colleges* 28, 5 (2013), 266–272.

José A. Ruipérez-Valiente, Pedro J. Muñoz-Merino, Giora Alexandron, and David E. Pritchard. 2019. Using Machine Learning to Detect "Multiple-Account" Cheating and Analyze the Influence of Student and Problem Features. *IEEE Transactions on Learning Technologies* 12 (2019), 112–122.

Steven Stack. 2015. The impact of exam environments on student test scores in online courses. *Journal of Criminal Justice Education* 26, 3 (2015), 273–282.

Jason M Stephens, Michael F Young, and Thomas Calabrese. 2007. Does moral judgment go offline when students are online? A comparative analysis of undergraduates' beliefs and behaviors related to conventional and digital cheating. *Ethics & Behavior* 17, 3 (2007), 233–254.

Donna Stuber-McEwen, Phillip Wiseley, and Susan Hoggatt. 2009. Point, click, and cheat: Frequency and type of academic dishonesty in the virtual classroom. *Online Journal of Distance Learning Administration* 12, 3 (2009), 1–10.

José A. Ruipérez Valiente, Giora Alexandron, Zhongzhou Chen, and David E. Pritchard. 2016. Using Multiple Accounts for Harvesting Solutions in MOOCs. In *ACM Conference on Learning at Scale 2016*.

Chelley Vician, Debra D Charlesworth, and Paul Charlesworth. 2006. Students' perspectives of the influence of web-enhanced coursework on incidences of cheating. *Journal of Chemical Education* 83, 9 (2006), 1368. https://doi.org/10.1021/ed083p1368

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

Michael P Watters, Paul J Robertson, and Renae K Clark. 2011. Student Perceptions of Cheating in Online Business Courses. *Journal of Instructional Pedagogies* 6 (2011).

C. Keith Waugh and Norman E. Gronlund. 2012. *Assessment of Student Achievement (10th Edition)*. Pearson.

Matthew West, Geoffrey L. Herman, and Craig Zilles. 2015. PrairieLearn: Mastery-based Online Problem Solving with Adaptive Scoring and Recommendations Driven by Machine Learning. In *2015 ASEE Annual Conference & Exposition*. ASEE Conferences, Seattle, Washington.

Bernard E Whitley. 1998. Factors associated with cheating among college students: A review. *Research in Higher Education* 39, 3 (1998), 235–274.

Craig Zilles, Matthew West, Geoffrey Herman, and Timothy Bretl. 2019. Every university should have a computer-based testing facility. In *Proceedings of the 11th International Conference on Computer Supported Education (CSEDU)*.