A Comparison of Proctoring Regimens for Computer-Based Computer Science Exams

Chinedu Emeka
cemeka2@illinois.edu
University of Illinois
Urbana, Illinois, USA

Matthew West
mwest@illinois.edu
University of Illinois
Urbana, Illinois, USA

Craig Zilles
zilles@illinois.edu
University of Illinois
Urbana, Illinois, USA

Mariana Silva
mfsilva@illinois.edu
University of Illinois
Urbana, Illinois, USA

ABSTRACT

In this paper, we explore three different methods for administering computer-based tests at scale: (1) a dedicated Computer-Based Testing Center (CBTC), (2) Bring Your Own Device (BYOD) exams proctored in person in the classroom, and (3) BYOD exams proctored online via Zoom. We conducted two randomized crossover experiments to compare pairs of modalities against each other (CBTC vs BYOD-in-person and CBTC vs BYOD-online). We found that testing modality did not impact students' exam performance or students' preparation before exams. However, we observed that students preferred the modalities in which they had recently received the highest scores. Our results indicate that several different modalities can be effectively used to administer testing at scale for CS courses.

CCS CONCEPTS

• **Testing at scale**; • **Test delivery mode**; • **Synchronous versus asynchronous testing**; • **Computer-based testing**;

KEYWORDS

Testing at scale, learning, assessment, performance

ACM Reference Format:
Chinedu Emeka, Matthew West, Craig Zilles, and Mariana Silva. 2024. A Comparison of Proctoring Regimens for Computer-Based Computer Science Exams. In *Proceedings of the 2024 Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2024), July 8–10, 2024, Milan, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3649217.3653536

1 INTRODUCTION

During the COVID-19 pandemic, many faculty transitioned from paper-based exams to computer-based exams due to requirements of remote instruction [7, 11]. With the return to in-person instruction, many faculty have chosen to keep using computer-based exams. Computer-based exams present a number of advantages, including: they facilitate autograding, enable quicker feedback to students, enable more complex and diverse types of questions, and, for programming questions, they can provide a more authentic development environment. In addition, computer-based exams can facilitate scaling through reduced grading effort and paper management, which is particularly relevant to CS faculty due to increasing enrollments.

With the return to in-person instruction, faculty have additional choices about how to run their exams. During the pandemic, exams were typically run on the student's computer (called "Bring your own device" or BYOD exams) either not proctored or proctored using teleconferencing software, with or without a lock-down browser. In-person instruction allows BYOD exams to be conducted in classrooms (much as paper exams are conducted) with in-person proctoring.

Alternatively, students can be asked to take exams on institutional computers. In computer science, there is a long tradition of running "lab exams" where mid-term exams are offered during existing lab sections in computer labs, using the provided computers [2, 10, 14]. Alternatively, some institutions have dedicated computer-based testing centers (CBTCs), which are computer labs used exclusively for testing [19]. The advantage of running exams on institutional computers are that every student can be provided an identical computer setup known to support any required software and the computer can be secured to prevent undesired communication and file access.

Clearly, these different proctoring approaches represent different logistics, requirements on course personnel, access to technology, access to space, and administrative support. They also potentially have different impacts on students, provide different opportunities to cheat, and might lead to different student studying behavior.

In this paper, we compare and contrast three methods of delivering summative assessments for a large junior-level computer science course. Specifically, we consider (1) BYOD exams proctored via Zoom, (2) BYOD exams proctored in-person in a classroom, and (3) exams conducted in a Computer Based Testing Center. We investigate the impact of testing modality on student performance and on students' studying behavior. More specifically, we address the following research questions:

- **RQ1**: Does testing modality impact students' performance in Computer Science exams?
- **RQ2**: Does testing modality influence students' study behavior before their exams in Computer Science courses?

ITiCSE 2024, July 8–10, 2024, Milan, Italy

ACM ISBN 979-8-4007-0600-4/24/07
https://doi.org/10.1145/3649217.3653536

For RQ1, we will compare students' overall performance on exams for a set of testing modalities. Exams in our course are comprised of multiple-choice questions, fill-in-the-blank problems and programming questions, which are more complex. Therefore, we will also compare performance in the different modalities for various sub-classes of problems. We hypothesize that factors such as security of the exam location, comfort with the location, resources available, and the presence of course staff may affect performance. If one modality is more secure than another, or perceived to be more secure, this could mitigate the potential for cheating. Students' comfort with the exam environment may influence their performance. For BYOD exams, students use their own computers as opposed to the lab computers for the CBTC exams. Students may be more comfortable and efficient with their own computers, which may be reflected in performance data. The presence or absence of course staff may also affect students' comfort. The aforementioned factors may interact in complex ways to impact performance across testing modalities.

For RQ2, we operationalize studying as the number of questions completed on practice exams provided by the instructor, as described further in the Methods section. RQ2 is of interest because the varied methods of administering exams provide different affordances to students. As an example, CBTCs provide more flexibility to students in terms of when they schedule their exams and study for these exams. The properties of the various modalities may prompt students to prepare for exams in different ways. Therefore, we will investigate study behavior across the modalities.

We will only address these research questions for contexts similar to ours, since we are using only one course for our experiments.

2 RELATED WORK

Most of the prior work related to investigating students' performance across testing modalities has been focused on comparing paper-based tests to computer-based tests [15, 16, 21, 22, 26, 28]. This includes both in-person computer-based exams [22] and asynchronous online computer-based exams which may be completed at home [28]. Typically, the paper-based tests used in the studies are restricted to simpler question types such multiple-choice questions or short response questions, due to the high cost of manually grading these tests. Consequently, in the studies, the computer-based tests were similarly limited to multiple-choice questions or short response questions, in order to maintain parity of question types across modalities to facilitate comparisons. In general, these prior studies do not show a statistically significant difference in performance for paper-based versus computer-based tests [4, 15, 23]. This result holds across a variety of academic levels for high-stakes summative asssessments, such as a standardized test for K-12 students (specifically Grade 7 students) [21] and the Graduate Records Examination, for students who are seeking admission to graduate schools [22]. However, results have been somewhat mixed for some computer science exams. Prior research found a higher incidence of syntax errors on computer science exams for paper-based tests when compared to computer-based tests, although there were no statistically significant differences in performance across the modes [8, 17].

Students have reported a preference for online tests over paper-based tests when surveyed on the issue [5, 24, 30], even when there is no statistically significant performance difference between modalities [13]. Online exams allow for faster editing of exam answers [20] and provide more realistic assessment [18]. For instance, students preferred the ability to debug and test their code for programming exams, an opportunity that may be available for computer-based exams but not for paper-based exams [18]. Although online testing is preferred, students believe that cheating is easier for online examinations [1]. "Cheating in asynchronous, online examinations is at unconscionable levels" [5, 29]. There are many variables in online contexts which are hard to manage, such as internet access and communication among students. Likewise, Bring Your Own Device (BYOD) exams in class raise security concerns. Dawsom identified several vulnerabilities of BYOD exams, "hacks", which can undermine security [9]. For instance, students may be able to use keyboard shortcuts to maintain fast access to unauthorized resources on their computers. Recently, these unauthorized resources may include Large Language Models (LLMs), which have powerful capabilities [3] and whose outputs are sometimes difficult to distinguish from that of humans [12, 25].

The prior work comparing modalities leaves some important gaps in the literature. The types of questions that are used in the comparisons between modalities are often limited, which hampers the inferences that can be drawn from such comparisons. Additionally, for large-scale programming assessments, paper-based tests are often cumbersome to administer at scale. Computer-based tests can better facilitate testing at scale. Therefore, we focus on comparing various methods of administering computer-based tests. The prior research studies are also limited in that they look at a small number of high-stakes summative exams. In our work, we seek to compare different methods of administering computer-based tests across entire academic terms. As described in our Methods section, we are able to control for several factors such as difficulty of the exam. This allows us to provide more robust information on several feasible methods for frequent assessment at scale.

3 DATA AND COURSE CONTEXT

We conducted two studies at a large and competitive midwestern university in the United States. We used a computer science course on numerical analysis for the studies. The course is required for Computer Science students and is mostly taken by students in their second to fourth years of university studies.

With approval from our Institutional Review Board, we collected data from students for the two studies. Study 1 was conducted in the Spring 2023 academic term (January to May 2023) and Study 2 was conducted in the Fall 2023 academic term (August to December 2023). There was no extra credit or other reward offered to encourage participation in the studies, beyond possible improvements to CS courses. To be included in the research study, students had to (1) provide informed consent, and (2) complete all six exams in the course in the prescribed modalities, as described in our Methods section. Some students completed all six exams, but switched to a different exam section because of illness or conflicts. Those students were dropped from the study. In both studies, roughly 65% of all students met the requirements for participation.

The time limit for each exam was 50 minutes. All exams included 8-10 non-coding questions (multiple-choice questions and fill-in-the-blank numeric questions), and 2 coding questions. Exams were graded interactively, and students could re-attempt problems that they got wrong on a given exam within the 50-minute time period allocated for that exam. For the multiple-choice and fill-in-the-blank questions, students were allowed 2-3 retries after their initial attempt for partial credit. Students could attempt the coding questions an unlimited number of times and still receive full credit on those questions.

The exams were constructed as a series of "slots", where each slot is associated with either a single problem or a pool of problems of similar difficulty and concept coverage from which every student gets a random draw. In addition, each problem is implemented as item generators, creating a range of randomly parameterized question instances.

For BYOD-Zoom proctoring, students needed a second device, such as a phone or tablet, to capture their face, computer screen, and work area on Zoom. This device was solely for proctoring purposes, with students completing the exam on their primary device, in which they were not connected to Zoom. The Zoom sections maintained a ratio of about 40 students to one proctor. In-person BYOD exams occurred in the classroom with a 20-to-1 student-to-proctor ratio. CBTC exams had two proctors per room—rooms have between 40 and 80 computers—and overhead video cameras that can be reviewed later to analyze suspicious behavior observed by the proctors. Both BYOD modalities were offered synchronously during class time, while the exams at the CBTC were offered asynchronously during a period of three days.

Demographic data is shown in Table 1. The majority of the students were male. Most participants were in the third or fourth year of college and none of them were in their first year of college.

Study	N	Gender		Standing (year)		
		Male	Female	2nd	3rd	4th
Study 1	220	157	63	5	82	133
Study 2	218	166	52	5	93	120

Table 1: Demographics for the studies.

4 METHODS

We wanted to explore how exam modality influenced performance and study habits, if the modality did in fact have any impact. We therefore conducted two crossover studies, in two separate semesters. A balanced crossover study allows us to account for ordering effects (i.e. the order of the treatments potentially having an impact on outcomes). Additionally, crossover studies account for variations across students, since they can act as their own control, mitigating another possible source of variance.

In Study 1, students were randomly assigned into one of two groups, group A (*n=104 students*) and group B (*n=116*). The modalities for the exams were alternated for each group for each exam. Students in group A took their first exam in the CBTC, while students in group B took their first exam using the BYOD-Zoom format. For the second exam, the conditions were flipped, i.e., students in group A took their second exam using the BYOD-Zoom format, while students in group B took their second exam in the CBTC. The modality alternated in this manner for all six exams.

To check if groups A and group B were similar, we compared the average incoming Grade Point Average (GPA) of students in both groups. The incoming GPAs includes all grades from all prior college courses at the start of the academic term. There was no statistically significant difference in ability between students in the two groups, as measured by incoming GPA ($t = 0.90, p = 0.37$). The average GPA of students in the course was high; it was over 3.7 on a 4.0 scale.

In Study 2, students were randomly assigned to one of three groups, groups A (*n=64 students*), B (*n=71*), or C (*n = 71*). For any given exam, two groups completed the exam in the CBTC and one group completed the exam in a BYOD-in-person setup. Therefore, every student in the study took four exams in the CBTC and two BYOD-in-person exams. Study 2 originally consisted of three groups as opposed to two groups solely because the classroom was not large enough to securely accommodate half of the students for BYOD-in-person testing; we wanted to have an abundance of spacing between students in the classroom to mitigate cheating. Students in the three groups have similar ability, based on a comparison of the average incoming GPA for the three groups. The average GPA was high for all three groups; over 3.7 on a 4.0 scale. We compared the three groups using a one-way ANOVA: $F(2, 215) = 1.73, p = 0.18$. To simplify analysis, we will combine the two groups that used the CBTC for any given exam when we report our results.

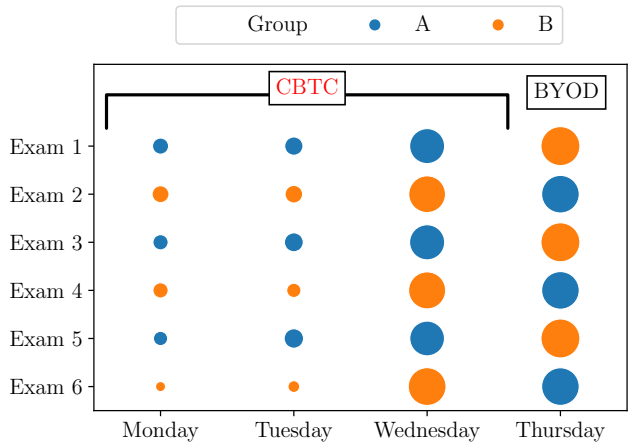
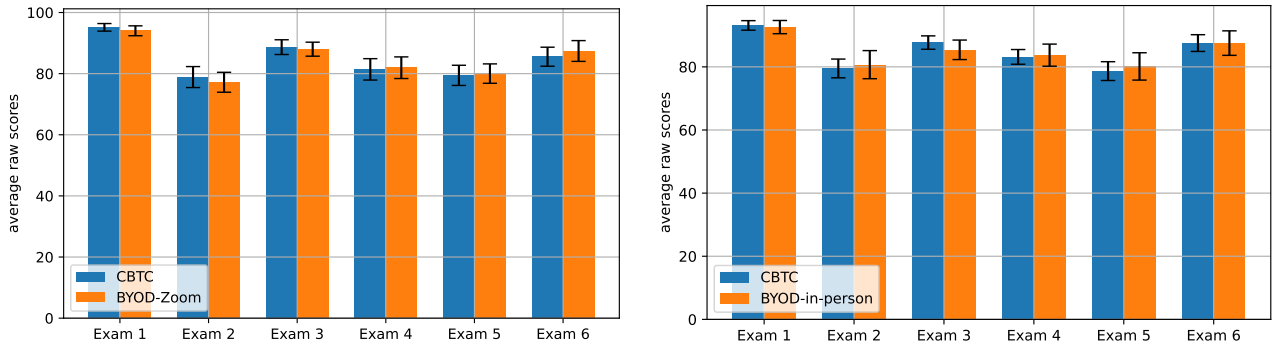


Figure 1: Exam Schedule for Study 1 (the schedule for Study 2 follows the same pattern). Students taking the BYOD exams took the exams on Thursdays during class time, while each student taking a CBTC exam selected an exam slot between Monday and Wednesday, inclusive.

In both crossover studies, each exam was held over a four-day period. Students who were assigned to the CBTC for a given exam could complete it on any of the first three days of that four-day exam period. Students assigned to a BYOD modality took the exam synchronously during class time, on the last day of the testing window. Figure 1 illustrates this schedule configuration for Study



(a) Study 1: CBTC vs BYOD-Zoom **(b) Study 2: CBTC vs BYOD-in-person**

Figure 2: Average performance on exams. The error bars are the 95% confidence intervals.

1, where the area of each circle indicates the number of students taking the exam on each exam day. Data from Study 2 results in a similar image, with small variations of the circle areas.

Given that CBTC exams are asynchronous, there was a potential for students to communicate among themselves over the exam period and share information with fellow students. By scheduling the BYOD exam for the last day of the testing window, we ensure that students in the synchronous format can also benefit from the dissemination of information that happens during asynchronous exams. If the synchronous exams were all held on the first day of the exam window, then that group would have no opportunity to benefit from communication. We note that students are actually not permitted to communicate about the exam content, but unfortunately, this may still occur. This has been described as "collaborative cheating" in the literature [6, 27]. If such communication did occur, we wanted to make sure both groups benefited equally; this allows us to isolate the impact of modality (i.e. BYOD versus CBTC) from potential benefits of collaborative cheating or communication.

4.1 Modality impact on performance

To measure the impact of exam modality, we fit an ordinary least squares (OLS) model that is suitable for quantifying the relationship between the exam environment and exam scores while controlling for confounding variables like GPA.

$$s_{ij} = \mu_j + \delta\, GPA_i + \beta \mathrm{A}_{ij} \tag{1}$$

where the left-hand-side value s_{ij} is the raw exam score that student i received in exam j and A_{ij} is an indicator variable that is 1 if student i took the exam j in a BYOD mode and 0 if at the CBTC. We used GPA as a control variable, where GPA_i is the incoming GPA of student i. Variables μ_j, δ, and β are the regression parameters that we want to estimate and can be interpreted as follows:

- μ_j: The mean score of exam j
- δ: The coefficient corresponding to the ability of student i
- β: The score advantage when taking a BYOD exam

To get the effect size and remove the impact of the difficulty of each exam, we also fit regression models using standardized z-scores:

$$z_{ij} = \delta\, zGPA_i + \beta \mathrm{A}_{ij} \tag{2}$$

where z_{ij} is the exam z-score that student i received in exam j and $zGPA_i$ is the incoming z-scored GPA of student i.

To better understand if a group of students seem to benefit more from one modality over the other, we obtained the score advantage per student from the following ordinary least squares (OLS) model:

$$z_{ij} = \alpha_i + \beta_i A_{ij} \tag{3}$$

where α_i represents the ability of student i and β_i can be interpreted as the score advantage of student i when taking a BYOD exam.

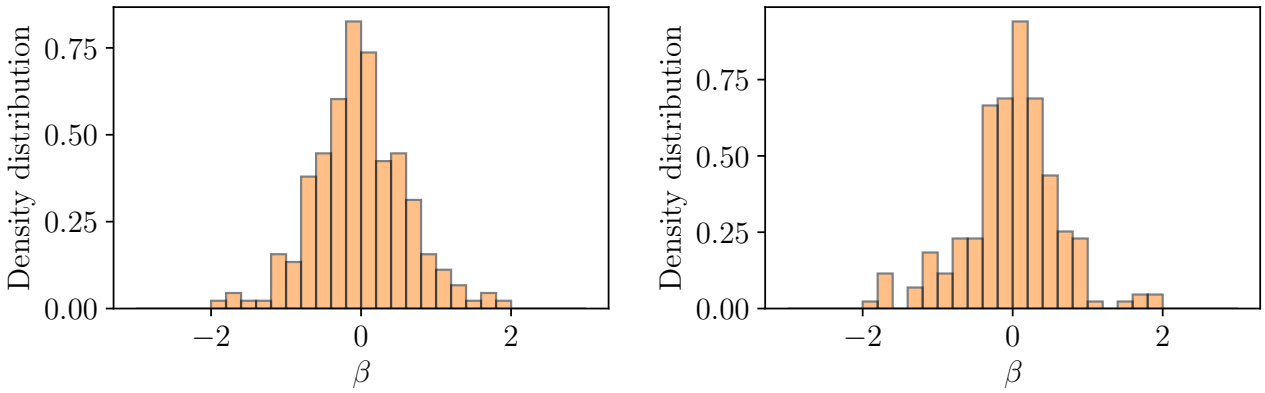
4.2 Modality impact on studying

To quantify studying behavior, we counted the number of problems students attempted on practice exams that were made available a week before each exam, for each modality. The practice exams were generated from the same randomly-parameterized question pools used to generate the actual exams, with the exception of the two coding questions. The practice exams were also delivered through the same learning management system as the actual exams. Students could generate as many practice exams as they wanted in order to prepare. Because of the practice exams' similarity to the actual exams, as emphasized by the course instructor, we believe that the number of questions a student earnestly attempts on practice exams is a good proxy for their overall studying for exams. We classify a problem as attempted earnestly if a student gets it right or spends more than 60 seconds on the problem.

5 RESULTS

5.1 No difference in aggregate performance across modalities

From the aggregated raw exam scores, there was no statistically significant difference in students' performance when modality was changed. Figure 2a shows the average exam scores for Study 1, indicating similar performance on exams in the BYOD-Zoom format and CBTC. Figure 2b shows the average exam scores for Study 2, which also reveals no statistically significant difference in performance between exams in the BYOD-in-person format and CBTC. In Section 5.2, we dig deeper into the data to determine if various sub-populations of students had different performance across the testing modalities.



(a) Study 1: CBTC vs BYOD-Zoom **(b) Study 2: CBTC vs BYOD-in-person**

Figure 3: Standardized histogram for the student score advantage β_i when taking exams in BYOD mode.

5.2 Computing the average score advantage based on exam modality

Using data from Study 1, the regression model in Eq.1 shows that students do not get any significant advantage when taking exams in BYOD mode with Zoom proctoring, compared to exams at CBTC ($\beta = -0.23$, 95% CI [−1.68,1.22], $p = 0.76$). Using standardized z-scores in a similar analysis, the advantage is found to be −0.02 standard deviations and not significant ($p = 0.65$). We find similar results when comparing BYOD with in-person proctoring and CBTC exams, with $\beta = 0.044$, 95% CI $[-1.69, 1.78]$, $p = 0.96$, and effect size equal to −0.0099 ($p = 0.87$).

5.2.1 Disaggregating by question type. Coding questions typically test application and problem-solving skills, while non-coding questions often assess theoretical understanding and knowledge recall. We repeated the regression analysis in Eq.1, but replacing s_{ij} with the score a student i obtained in question j. The results indicate that there is no significant impact on exam modality when looking separately at coding and no-coding questions.

5.2.2 Disaggregating by gender. When looking at only male and female groups, our regression models continued to indicate no significant difference between exam modalities.

5.2.3 Computing the score advantage for each student based on exam modality. The standardized β distribution (score advantage for students taking exams in BYOD mode) resulting from Eq.(3) is illustrated in Fig. 3a for Study 1 (with Zoom proctoring) and in Fig. 3b for Study 2 (with in-person proctoring). As expected, the average of the distributions is equal to −0.01 for both Study 1 and 2, consistent with the results obtained above for Eq.(2) (i.e. no effect).

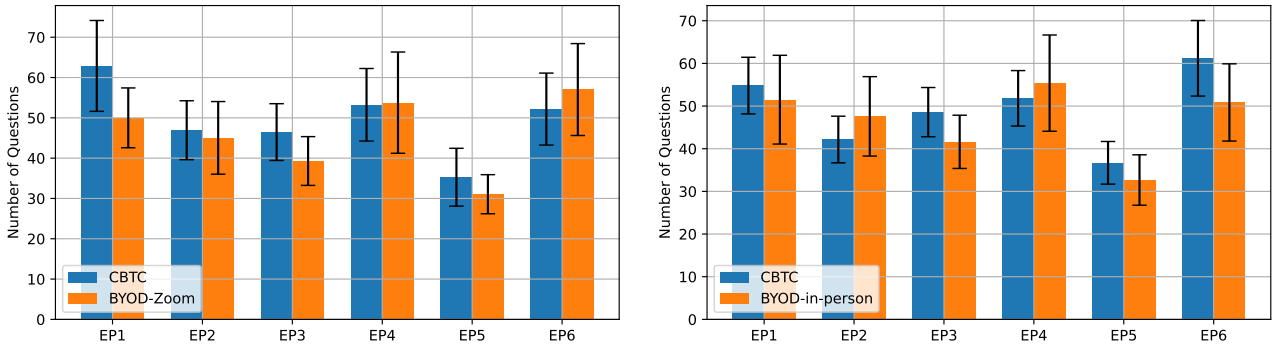
Figure 3b reveals that some students may be doing much better or much worse in BYOD. Our first hypothesis was that some students are just more "comfortable" in one of the exam modalities and hence do better. To investigate the effect of students' preference between modalities, we used results from a survey given to students included in Study 2. The survey asked students, "Do you prefer to work on the exams on your own laptop or on the CBTC computers?" Students had to select from the following options: -2 for *Strongly prefer CBTC computers*, -1 for *Prefer CBTC computers*, 0 for *No preference*, 1 for *Prefer my own laptop*, and 2 for *Strongly prefer my own laptop*. Overall, there were 51 (30.7%) students who either preferred or strongly preferred the CBTC, 70 (42.1%) students who preferred or strongly preferred the BYOD format and 45 (27.1%) students who had no preference. A number of students did not fill out the question which asked them to indicate their preference.

Our analysis indicated a weak correlation between students' scores and their preferences. Students who reported preferring to use their own devices had on average a standardized advantage of $\beta = 0.14$ (p=0.05), representing a significant advantage when taking exams in the BYOD format. Students who reported no preference or who preferred CBTC computers did not have any significant advantage in either modality. It was not clear at first the direction of causality, i.e. do students perform better in the modality of their preference, or do they report preference for the modality they perform better in? Since the survey was given to students in the middle of the academic term, we repeated the regressions analysis from Eq.3 using two different datasets: one including only the first 3 exams, which took place almost completely before the survey, and another one including the last 3 exams, which happened almost completely after the survey. The results are summarized in Table 2.

	Prefer CBTC	**No Preference**	**Prefer BYOD**
All exams	-0.07 (0.40)	-0.087 (0.16)	0.14 (0.053)
First 3 exams	**-0.29 (0.021)** *	-0.15 (0.10)	**0.24 (0.03)** *
Last 3 exams	0.14 (0.25)	-0.025 (0.79)	0.045 (0.68)

Table 2: Average standardized score advantage for students taking exams in BYOD mode resulting from regression models including "All exams", the "First 3 exams" and the "Last 3 exams", with associated p-values between parentheses. * indicates statistical significance at the 5% significance level.

We see that for the exams taken *before* the survey was administered (the first 3 exams), students who reported a preference for CBTC had on average a negative standardized score advantage ($\beta = -0.29$), meaning they scored higher in the CBTC relative



(a) Study 1: CBTC vs BYOD-Zoom **(b) Study 2: CBTC vs BYOD-in-person**

Figure 4: Number of practice questions studied for each exam modality. The error bars are the 95% confidence intervals. EPx stands for exam practice for exam number x.

to BYOD-in-person. Likewise, those who reported a preference for BYOD had on average a positive standardized score advantage ($\beta = 0.24$), meaning they did better in the BYOD-in-person modality. Both effects where statistically significant.

However, for the exams taken *after* the survey, these relationships completely disappeared. That is, students who reported a preference for the CBTC did not do better in it, and similarly for BYOD-preferring students. These results indicate that students prefer the exam modality in which they most recently scored higher, but *their preferences are not predictive for future exams.* This implies that it is random variation in their performance that is causing the preference, not the other way around.

Finally, we looked at the scores of the outliers (the students who appeared to do much better or much worse in one modality). We believe that the larger discrepancy between the average scores in both modalities for these students is due to them doing very poorly in one of the exams, likely due to the course policy that the lowest score is dropped from the final grade calculation.

5.3 No difference in exam practice across modalities

There was no statistically significant difference in the number of practice questions attempted earnestly when modality was changed. This result held for the CBTC versus BYOD-in-person modality (see Figure 4a) and the CBTC versus BYOD-Zoom modality (see Figure 4b). However, it is possible that the pacing of the studying differed. Preliminary unpublished data from interviews with students indicate that they prepared at different times for asynchronous versus synchronous exams. Some students reported that the flexibility provided by the asynchronous CBTC testing format allowed them to postpone exams until they felt comfortable with their preparation.

6 DISCUSSION

While we suspected that increased opportunities to cheat on BYOD exams might have led to increased performance, this was not evidenced by our results. There are several possible reasons for this finding in our context. First, in the CS course we examined, the instructor provided practice quizzes that were very similar to the actual exams. The benefits from the practice quizzes may have eliminated an advantage of one modality over the other. The practice quizzes may also have removed an incentive for cheating. Second, students are given access to the course notes through the assessment platform, during both the CBTC and BYOD exams. This may also have disincentivized cheating. Additionally, the large number of proctors for the BYOD conditions could have enhanced the security of the exams. Our findings may have been different if we had fewer proctors or a higher student to proctor ratio.

We are currently in the process of conducting interviews with students. These interviews will help us to discern whether specific features of the various modalities affect students' outcomes. Although we did not find that modality impacted performance in the aggregate, it is plausible that some properties of the modalities could independently have a beneficial or harmful impact on students' outcomes.

7 LIMITATIONS

An important limitation of our experiments is that we used only one course at a single competitive university. Results may vary for different types of CS courses or student populations. We hope to repeat the experiment with more courses to see if results hold.

A second limitation of our work is that the initial analysis of studying behavior did not capture the timing or pacing of students' studying. Rather, we only examined the amount of studying done. In the future, we intend to collect more data on when students prepare for exams for the various modalities. If one modality promotes distributed practice as opposed to massed practice (i.e. cramming), instructors may decide to transition to that modality to encourage consistent studying over time.

8 CONCLUSION

We conducted two studies where we varied the testing modality used for CS exams. We found no evidence that modality impacted performance or students' studying behavior. Our results indicate that there are several methods which can be equally effective for delivering computer science exams at scale.

REFERENCES

[1] Mohamed Abdelraouf Attia. 2014. Postgraduate students' perceptions toward online assessment: The case of the faculty of education, Umm Al-Qura university. In *Education for a knowledge society in Arabian Gulf countries*. Vol. 24. Emerald Group Publishing Limited, 151–173.

[2] Joao Paulo Barros, Luís Estevens, Rui Dias, Rui Pais, and Elisabete Soeiro. 2003. Using lab exams to ensure programming practice in an introductory programming course. In *Proceedings of the 8th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, (ITiCSE)*. 16–20.

[3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[4] Alan C Bugbee Jr. 1996. The equivalence of paper-and-pencil and computer-based testing. *Journal of research on computing in education* 28, 3 (1996), 282–299.

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

[6] Binglin Chen, Matthew West, and Craig Zilles. 2017. Do performance trends suggest wide-spread collaborative cheating on asynchronous exams?. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. 111–120.

[7] Ted M Clark, Christopher S Callam, Noel M Paul, Matthew W Stoltzfus, and Daniel Turner. 2020. Testing in the time of COVID-19: A sudden transition to unproctored online exams. *Journal of chemical education* 97, 9 (2020), 3413–3417.

[8] Jonathan Corley, Ana Stanescu, Lewis Baumstark, and Michael C Orsega. 2020. Paper or ide? the impact of exam format on student performance in a cs1 course. In *Proceedings of the 51st ACM technical symposium on computer science education*. 706–712.

[9] Phillip Dawson. 2016. Five ways to hack and cheat with bring-your-own-device electronic examinations. *British Journal of Educational Technology* 47, 4 (2016), 592–600.

[10] Chinedu Emeka and Craig Zilles. 2020. Student perceptions of fairness and security in a versioned programming exam. In *Proceedings of the 2020 ACM conference on international computing education research*. 25–35.

[11] Kelum AA Gamage, Erandika K de Silva, and Nanda Gunawardhana. 2020. Online delivery and assessment during COVID-19: Safeguarding academic integrity. *Education Sciences* 10, 11 (2020), 301.

[12] Vivian Emily Gunser, Steffen Gottschling, Birgit Brucker, Sandra Richter, and Peter Gerjets. 2021. Can users distinguish narrative texts written by an artificial intelligence writing tool from purely human text?. In *International Conference on Human-Computer Interaction*. Springer, 520–527.

[13] ÖZ Hüseyin and Tuba Özturan. 2018. Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests? *Journal of Language and Linguistic Studies* 14, 1 (2018), 67–85.

[14] Norman Jacobson. 2000. Using On-computer Exams to Ensure Beginning Students' Programming Competency. *SIGCSE Bull.* 32, 4 (Dec. 2000), 53–56. https://doi.org/10.1145/369295.369324

[15] Yassin Karay, Stefan K Schauber, Christoph Stosch, and Katrin Schüttpelz-Brauns. 2015. Computer versus paper—does it make any difference in test performance? *Teaching and learning in medicine* 27, 1 (2015), 57–62.

[16] Shohreh Kolagari, Mahnaz Modanloo, Reza Rahmati, Zahra Sabzi, and Ali Jannati Ataee. 2018. The effect of computer-based tests on nursing students' test anxiety: A quasi-experimental study. *Acta Informatica Medica* 26, 2 (2018), 115.

[17] Vesa Lappalainen, Antti-Jussi Lakanen, and Harri Högmander. 2017. Towards computer-based exams in CS1. In *International Conference on Computer Supported Education*. SCITEPRESS Science And Technology Publications.

[18] Ásrún Matthíasdóttir and Hallgrímur Arnalds. 2016. E-assessment: students' point of view. In *Proceedings of the 17th international conference on computer systems and technologies 2016*. 369–374.

[19] Patrick Moskal, Richard Caldwell, and Taylor Ellis. 2009. Evolution of a computer-based testing laboratory. *Innovate: Journal of Online Education* 5, 6 (2009).

[20] Jeremy Pagram, Martin Cooper, Huifen Jin, and Alistair Campbell. 2018. Tales from the exam room: Trialing an e-exam system for computer education and design and technology students. *Education Sciences* 8, 4 (2018), 188.

[21] John Poggio, Douglas R Glasnapp, Xiangdong Yang, and Andrew J Poggio. 2005. A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment* 3, 6 (2005), n6.

[22] Donald E Powers. 2001. Test anxiety and test performance: Comparing paper-based and computer-adaptive versions of the graduate record examinations (GRE©) general test. *Journal of Educational Computing Research* 24, 3 (2001), 249–273.

[23] Anna Agripina Prisacari and Jared Danielson. 2017. Rethinking testing mode: Should I offer my next chemistry test on paper or computer? *Computers & Education* 106 (2017), 1–12.

[24] Stacy MP Schmidt, David L Ralph, Bruce Buskirk, et al. 2009. Utilizing online exams: A case study. *Journal of College Teaching & Learning (TLC)* 6, 8 (2009).

[25] Elena Shalevska. 2023. AI LANGUAGE MODELS, STANDARDIZED TESTS, AND ACADEMIC INTEGRITY: A CHAT (GPT). *International journal of Education Teacher* 26 (2023), 17–25.

[26] Mark D Shermis and Danielle Lombard. 1998. Effects of computer-based test administrations on test anxiety and performance. *Computers in Human Behavior* 14, 1 (1998), 111–123.

[27] Mariana Silva, Matthew West, and Craig Zilles. 2020. Measuring the score advantage on asynchronous exams in an undergraduate CS course. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 873–879.

[28] Jeffrey R Stowell and Dan Bennett. 2010. Effects of online testing on student exam performance and test anxiety. *Journal of Educational Computing Research* 42, 2 (2010), 161–171.

[29] Daniel P Sullivan. 2016. An Integrated Approach to Preempt Cheating on Asynchronous, Objective, Online Assessments in Graduate Business Classes. *Online Learning* 20, 3 (2016), 195–209.

[30] Jeremy B Williams and Amy Wong. 2009. The efficacy of final examinations: A comparative study of closed-book, invigilated exams and open-book, open-web exams. *British Journal of Educational Technology* 40, 2 (2009), 227–236.