UNPACKING THE INFLUENCE OF COMPUTER-BASED TESTING MODALITIES ON STUDENT STUDY BEHAVIOUR AND PERFORMANCE

R. Gulati, C. Zilles, M. West, M. Silva

University of Illinois, Urbana-Champaign (UNITED STATES)

Abstract

In this paper, we explore the relationship between computer-based exam modalities, study behaviors, and overall exam performance. Our research focuses on four distinct testing approaches: (1) asynchronous, proctored exams in computer labs with institution-provided, locked-down computers; (2) synchronous, in-class, proctored exams following a Bring-Your-Own-Device (BYOD) model, where students can use their own devices to complete the exams; (3) synchronous, remote BYOD exams with Zoom proctoring; and (4) synchronous, remote and unproctored BYOD exams. Conducted over four semesters, the study involved a sophomore/junior-level numerical methods course with high enrollment from computer science, math, and engineering majors at a large R1 university in the United States. The course's summative assessments comprised six 50-minute exams that were auto-graded, providing immediate feedback to students. To prepare for these exams, students were granted access to practice exams one week prior to each actual exam, which remained available until the exam date. The platform's log data, capturing the sequence and duration of questions attempted by students, revealed variations in study behavior across the different exam modalities. Specifically, we observed an increase in study time correlating with the level of exam security, with high-security exams leading to the most study effort, and low-security exams the least. In addition, we identify two study strategies (Mock-masters and All-rounders) that are correlated with better exam performance than the class as a whole, and students are more likely to use the best of these strategies (All-rounder) on the most secure exams. In addition, our results are consistent with prior research relating to distributed retrieval practice.

Keywords: Computer-based testing, exam security, auto-graders, mastery learning, study behaviors.

1 INTRODUCTION

The evolution of assessment methodologies in higher education, particularly within STEM disciplines, has been significantly influenced by advancements in technology. The rise of computer-based testing (CBT) in educational assessments has dramatically changed how instructors can deliver robust and secure exams and, consequently, how students prepare for these exams. Various modalities have emerged, ranging from fully proctored to unproctored, asynchronous to synchronous, and exams delivered using secure institutional computers to students' own devices.

The shift to computer-based testing is the result of two main forces: the Covid-19 pandemic and the near ubiquity of students having their own computing devices (e.g., laptops and tablets). The remote instruction forced by the pandemic led many faculty to undertake the substantial effort to computerize their materials and many have embraced the lower administrative overhead of running computer-based exams. The ubiquity of student devices has two implications: (1) that (most) students can be expected to provide their own computer to take exams in Bring-Your-Own-Device (BYOD) [1] context, and (2) since students prefer to work on their own machines, traditional computer labs are less utilized, making it easier for institution to convert these spaces into Computer-based Testing Facilities (CBTF) [2].

The BYOD model has gained popularity due to its flexibility and the fact that it can be implemented independently by a single faculty member in a wide variety of contexts. Previous research by Gulati et al. [3] explored the security of BYOD exams under various proctoring regimes by comparing the empirical exam scores in a collection of crossover studies. The CBTF model requires more institutional support (to set up and run the testing center) but provides lower administrative overhead to faculty members and higher exam security, because the computers, networking, and filesystems are controlled by the institution.

This study aims to examine how different computer-based testing modalities affect student study behaviors and exam performance. We utilize a robust dataset from an auto-graded platform across multiple semesters at a large R1 university in the United States. This approach enables a detailed

analysis of student preparation strategies under various testing conditions and their correlation with performance outcomes.

2 METHODOLOGY

The study was conducted at the University of Illinois in a mandatory computer science course, Numerical Methods. This course employed a flipped classroom model, requiring students to complete pre-lecture assignments and participate in group activities during class sessions. Each topic was accompanied by a corresponding set of homework assignments, released weekly. The course's summative assessments consisted of six 50-minute exams featuring a variety of question types. All assessments were auto-graded and provided instant feedback through the PrairieLearn platform [4,5].

In PrairieLearn, questions are designed as “question generators” that use randomized variables to create unique question variants. This design allows students to engage in extensive practice by generating different versions of questions, while enabling instructors to reuse these question generators across different assessments and semesters.

To prepare for the exams, students had access to unlimited practice exam instances starting one week prior to each exam date. These practice exams were hosted on the same assessment platform and utilized the same question generators as those used in the actual exams. Both the official exams and practice exams were constructed as a series of “slots”, each associated with a pool of question generators of similar difficulty and concept coverage. Each time an exam (or practice exam) is generated, each slot receives a random draw from its pool. Typically, each exam comprised 9-12 slots, each containing 2-5 question generators, averaging about 40 question generators per exam. Given the unlimited practice opportunities, students could potentially encounter variants from all question generators included in the exam.

This study utilized course data spanning four semesters, from Fall 2021 to Spring 2023. During this period, students participated in exams under four different testing modalities: BYOD Unproctored, BYOD Zoom, BYOD In-person, and CBTF. The BYOD Unproctored modality allowed students to use their own devices to access the exam synchronously at a specified time without proctoring. In the BYOD Zoom modality, students were required to use a secondary device—such as a phone or tablet—to visually document their testing environment via Zoom. This device was used solely for proctoring purposes, while students completed the exams on their primary devices, which were not connected to Zoom. This setup maintained a proctoring ratio of about 40 students per proctor. For the BYOD In-person modality, students used their own computers in a classroom setting and were proctored by course staff at a ratio of approximately 20 students per proctor. In the CBTF modality, students took exams at a dedicated computer-based testing facility, selecting a convenient time within a designated 3-day period.

Over three semesters, students were grouped to alternate between different testing modalities in a crossover experimental design, as detailed in Table 1. This arrangement was designed to evaluate the impact of each testing modality on student performance [3,6]. The current study aims to understand how students use practice exams as a study resource, identify effective study strategies that correlate with higher exam performance, and explore whether these strategies vary based on the testing modality.

Table 1. Summary of testing modalities across the semesters included in this study.

Semester	***Testing Modality***
Fall 2021	BYOD Unproctored and BYOD Zoom (both synchronous)
Spring 2022	BYOD Zoom and BYOD In-person (both synchronous)
Fall 2023	CBTF (asynchronous)
Spring 2023	BYOD Zoom (synchronous) and CBTF (asynchronous)

Exam policies remained consistent across all semesters, regardless of the testing modality. Students were prohibited from communicating with others for assistance during the exams, including during the exam period in the asynchronous format. Although the use of online resources like Chegg or Stack Overflow was forbidden, students were encouraged to use and were provided access to all course materials, such as the online textbook and slides, in all testing modalities through the PrairieLearn platform.

3 RESULTS

The PrairieLearn platform comprehensively logs all student interactions with their assessments, capturing data such as the order of completed questions, the number of attempts, the duration, and score of each attempt. Our analysis aims to identify study patterns as students prepare for exams and to investigate which patterns correlate strongly with higher exam performance.

3.1 Time spent studying for exams

For this research, we define time spent studying for each exam as the summation of the time students spent actively engaging with practice exams. Note that we do not include time spent reading course materials, reviewing homework, or other known study activities. In PrairieLearn, a study session is defined as any period during which a student submits at least one answer per hour. A session concludes when an inactive period is detected, with the duration calculated up to the last submission. Subsequent submissions mark the beginning of a new session. Therefore, the total study time for an exam is the cumulative duration of all practice exam sessions.

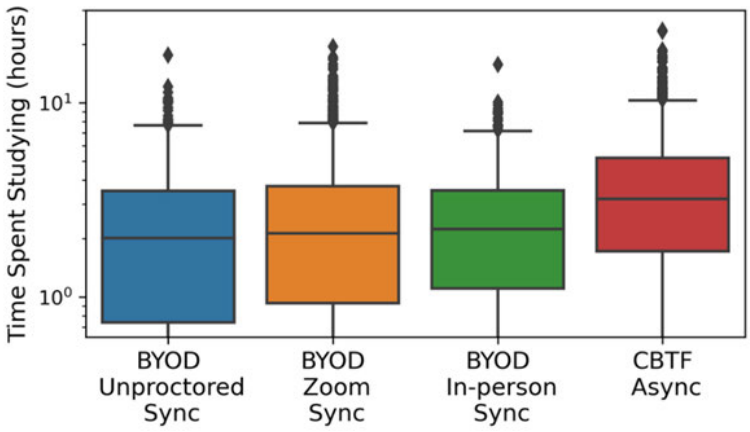


Figure 1. Distribution of study hours across the testing modalities. Each boxplot spans from the data 1st to the 3rd quartile, and the median is marked by the line inside the box. Log scale used for easier reading.

Figure 1 illustrates the distribution of study hours across different testing modalities, showing a clear trend: more stringent exam security protocols, such as those in the CBTF, are associated with increased study time. Conversely, the least restrictive modality, unproctored exams, correlates with the shortest study times. Table 2 provides a summary of average study times across these modalities. Tukey's range test shows that the average time spent studying for CBTF exams is statistically significantly different to the time spent on the other three BYOD modalities ($p < 0.001$), and that the average time between all the three BYOD modalities is not statistically significantly different to each other.

*Table 2. Average time spent studying for each testing modality. The **** indicates the time spent studying for CBTF exams is statistically significantly different to the time spent in the other modalities, with p<0.0001.*

Testing Modality	*Time (hours)*
BYOD Unproctored Sync	2.40
BYOD Zoom Sync	2.61
BYOD In-person Sync	2.66
CBTF Async	3.79 ****

To determine whether an increase in study time corresponds to better exam performance, we employed an ordinary least squares (OLS) regression model. This model is appropriate for quantifying the relationship between time spent studying and exam scores while controlling for confounding variables like GPA, which represents student ability:

$$s_{ij} = \beta \, log\,(t_{ij} + 1) + \alpha \, g_i + \gamma \quad (1)$$

Here s_{ij} is the predicted standardized exam score (z-score) that student i received in exam j, and t_{ij} is the time student i spent on the practice exam j, given in hours. We used the logarithm of the time due to its highly right-skewed distribution, as shown in Figure 1. g_i is the z-scored incoming GPA of student i. The parameters β and α are the regression coefficients estimating the change in exam score per unit of practice time and the coefficient corresponding to student ability, respectively. The intercept γ represents the baseline value of the dependent variable s_{ij} when all other variables are set to zero.

The regression results indicated that $\beta = 0.41$, $\alpha = 0.30$ and $\gamma = -0.49$, all statistically significant with $p < 0.0001$. This confirms that increased study time positively associates with higher exam scores.

To account for variations across different testing modalities, we replace β in Eq.(1) with the term $(\beta_o U_{ij} + \beta_1 BZ_{ij} + \beta_2 BI_{ij} + \beta_3 C_{ij})$, where U_{ij}, BZ_{ij}, BI_{ij}, and C_{ij} are indicator variables denoting whether student i took exam j in Unproctored, BYOD Zoom, BYOD In-person, and CBTF modalities, respectively (0 otherwise). The model results in $\beta_o = 0.54$, $\beta_1 = 0.46$, $\beta_2 = 0.41$, and $\beta_3 = 0.38$, all statistically significant $(p < 0.0001)$. Most of these β_k coefficients are statistically significant different from each other, except for β_2 and β_3 $(p = 0.31)$ and β_2 and β_1 $(p = 0.19)$. The relatively smaller coefficient for the CBTF modality suggests that because students take the CBTF exams more seriously, the impact of additional study time is somewhat diminished.

3.2 Defining the quality of students' study behavior

In the following two sections, we quantitatively explore the students' approach to studying. To facilitate this, we introduce four metrics derived from students' interactions with the question generators included in the practice exam instances.

Each practice exam generator consists of a number of slots (n), each of which is associated with a pool of question generators. Each question generator is identified by a unique question identifier (QID). Each time an exam is generated, one question is drawn from each pool, resulting in an n question exam. We define m as the number of distinct QIDs across the exam's n pools.

Metric 1 (Fraction attempted): This metric represents the proportion of all unique m QIDs that a student attempted. We consider a QID *attempted* if the student submitted an answer, independent of whether it was graded as correct or not. If a student generates enough practice exams, they can observe all m QIDs, resulting in a Metric 1 value of 1.0. This metric measures the breadth of a student's study efforts.

Metric 2 (Attempted correct fraction): This metric represents the proportion of attempted QIDs that a student got correct at some point. That is, if a student generated two practice exams and only attempted one question on each, but they happened to be the same question (same QID, hence "unique QIDs attempted" is 1), and they got it correct once and incorrect once ("unique QIDs correct" is 1), their Metric 2 value for this exam would be 1.0. This metric represents a student's diligence in mastering all of the questions that they attempt.

Metric 3 (Average correct per instance): This metric computes the average proportion of practice exam questions generated that a student gets correct. When practice exams are generated, one question is drawn from each pool and randomly parameterized and assigned a unique question *instance* identifier (QIID). This metric is computed by taking the ratio of QIIDs that a student gets correct (some of which could be for the same QID) and the number of QIIDs that the student created by generating practice exams. A student that attempts half of each practice exam that they generated and gets half of those correct would have a Metric 3 value of 0.25.

Metric 4 (Average attempted per instance): This metric is similar to the previous one, but isn't concerned with whether the student answered correctly, just whether they attempted it. It is computed by dividing the number of QIIDs a student attempted by the number of QIIDs the student generated. The student above, who attempted half of each practice exam generated, would have a Metric 4 value of 0.5. This ratio distinguishes students that are trying every question on a practice exam from those that are picking and choosing from the offered questions (independent of their ability).

Figure 2 shows the average values of these metrics for all students, grouped by testing modality. We observed significantly higher averages for Metric 1 in the most secure exam modality, the CBTF. This result is consistent with the results in Section 3.1; students who spend more time studying are likely to attempt more practice exams and, hence, observe a larger fraction of the exam questions.

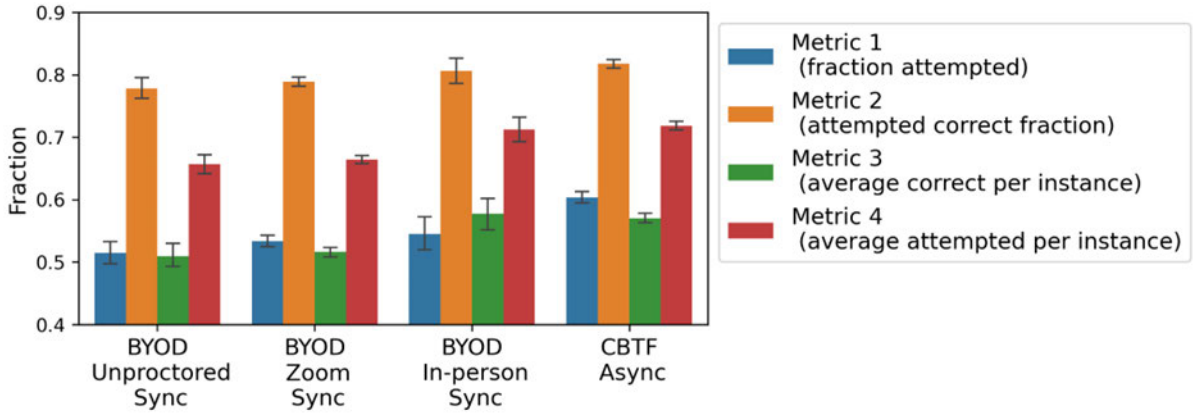


Figure 2: Average of each metric grouped by testing modality.

In our analysis, we wanted to determine which factor has a more significant impact on study performance: the amount of study, measured by the proportion of attempted questions (Metric 1), or the degree to which a student mastered the questions seen (Metric 2). To investigate this, we constructed a regression model incorporating these measures to assess their influence on performance.

$$s_{ij} = \kappa_1 M1_{ij} + \kappa_2 M2_{ij} + \alpha g_i + \gamma \quad (2)$$

In this equation, in addition to the parameters already defined in Eq.(1), $M1_{ij}$ is the proportion of distinct QIDs attempted by student i in exam j, and $M2_{ij}$ is the proportion of QIDs attempted by student i in exam j that they got correct at least once. The coefficients κ_1 and κ_2 represent these predictors, quantifying their influence on exam performance. The results of the regression analysis are shown in Table 3.

Table 3. Results from regression in Eq.(2)

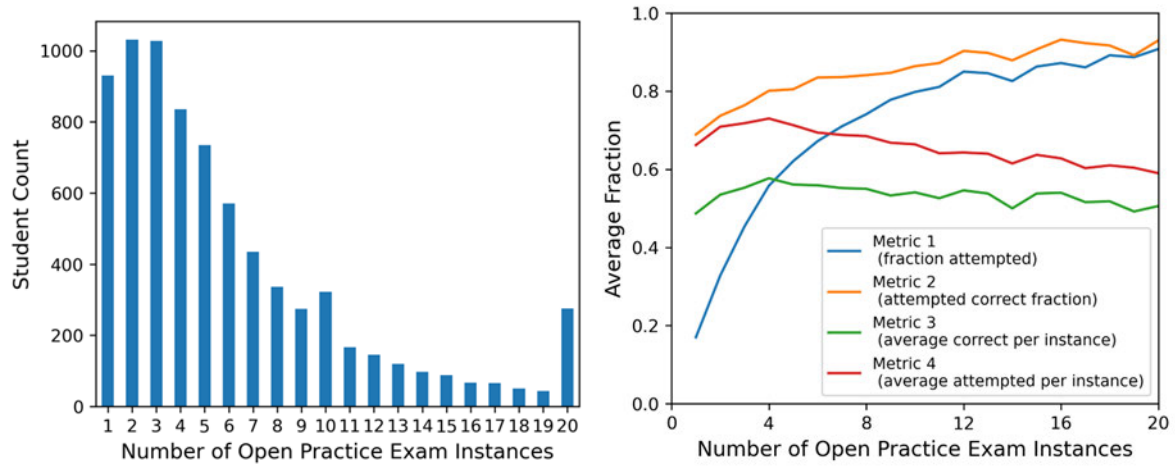
Regression parameters	***coefficient***	***p-values***
Metric 1 (fraction attempted)	$\kappa_1 = 0.55$	<0.0001
Metric 2 (attempted correct fraction)	$\kappa_2 = 2.08$	<0.0001
GPA	$\alpha = 0.19$	<0.0001
Intercept	$\gamma =$ -1.97	<0.0001

While the analysis in Section 3.1 indicated that more study time generally correlates with better performance, this regression reveals that student exam performance is better predicted by whether students can correctly answer the questions that they've attempted. Because these are correlations, it isn't clear in which direction the causality goes. Do high performing students get problems correct on both practice and actual exams, or do students that persist and master all of the questions that they encounter on practice exams do well on actual exams? Nevertheless, these results are consistent with expectations that the primary value in engaging with practice problems results from mastering them and not just seeing them before the exam.

When adding the terms $(\beta_o U_{ij} + \beta_1 BZ_{ij} + \beta_2 BI_{ij} + \beta_3 C_{ij})$ into the regression model in Eq.(2), the coefficients for Metrics 1 and 2 remain largely consistent, suggesting that our findings regarding the impact of study quantity and success on exam performance are robust across different testing modalities.

3.3 Determining students' study patterns

In this section, we explore whether the ability to open multiple practice exam instances influences study patterns and subsequently impacts exam performance. We calculated the *number of practice exam instances* (NPEI) each student created for each exam. Figure 3a shows the distribution of students (counted per exam rather than uniquely) as a function of NPEI. For instance, if a student opens one practice exam instance for the first two exams and three instances for the next four exams, they are counted twice under NPEI=1 and four times under NPEI=3. Figure 3a reveals that 50% of the student count opens 4 or fewer practice exam instances.



a) *b)*

Figure 3. (a) Student count per exam instance as a function of the number of open practice exam instances. Half the students open at most 4 practice exams. (b) average value of each metric varying with the number of open practice exams instances.

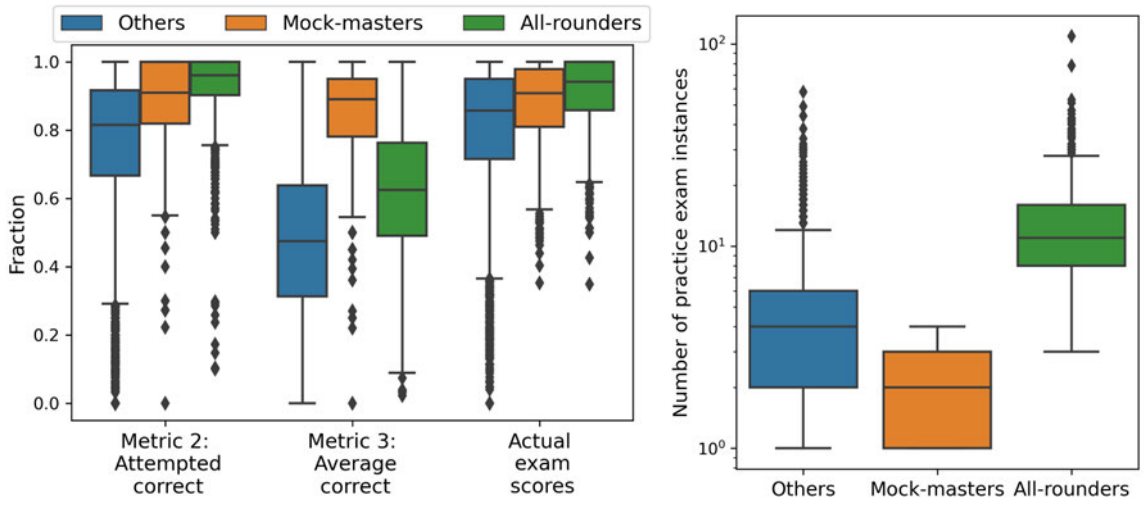
Figure 3b presents the average values of each metric as a function of the number of open practice exam instances. It highlights specific characteristics of these metrics. For instance, the fraction of attempted questions (Metric 1) is low when only a few practice exam instances are opened, as students are unable to encounter all available QIDs. However, this fraction increases as the number of instances grows, allowing students to attempt most of the available QIDs. As students attempt more questions, we can also observe an increase in the proportion of correctly attempted questions (Metric 2). Interestingly, the average correct fraction per exam instance (Metric 3) remains relatively constant regardless of the number of open practice exam instances. The average attempted per instance (Metric 4) decreases, suggesting that students who open many instances are mostly trying to see as many QIDs as possible, but are not necessarily making submissions to all questions.

Based on observations from Figure 3, we propose two study patterns based on the number of open practice exam instances and the metrics associated with the number of attempted QIDs (Metrics 1 and 4). Table 4 summarizes these categories and the criteria for identifying students within each one.

Table 4. Study pattern category descriptions and their corresponding criteria based on the number of open practice exam instances (NPEI) and metrics 1 and 4.

Category description	***Criteria***
Mock-masters: these students solve fewer practice exams and treat them like a real exam scenario, answering the majority of the questions.	NPEI <=4 and metric 4 > 0.9
All-rounders: this group of students wants to submit answers to as many QIDs as possible, aiming to see almost all questions that will appear in the exam.	metric 1 > 0.8 and not a Mock-master
Others: students who do not belong to the Mock-masters or All-rounders categories	None of the above

For our dataset, which includes 7590 entries from students taking an exam, we found 7 students who satisfy the criteria for both Mock-masters and All-rounder categories. Instead of considering this as a separate category, we included these students in the “Mock-masters” category. Students who don’t fall within these two categories are denoted as “Others”. It's important to note that a student may be identified in different categories on different exams; one might be a Mock-master while preparing for some exams and an All-rounder for others.



a) b)

Figure 4. (a) Metrics 2 and 3, and average exam scores for each category, revealing stronger performance from students in Mock-master and All-rounders when compared to others. (b) Number of practice exam instances opened by students in each category (log scale for better visualization), indicating that students in the all-rounders category open more practice exams than students in the "other" category.

Our dataset indicates that 10% of the students are Mock-masters and 20% are All-rounders. Figure 4a includes the distribution of metrics 2 and 3, which measure the accuracy of the submissions. It also shows the distribution of the actual exam scores for each category. Figure 4b shows the number of open practice exam instances.

Students in the All-rounders group, who by definition strive to attempt as many QIDs as possible, display the highest averages for Metric 2, indicating that they are not only aiming for content breath, but are also diligently mastering the content. Moreover, they open, on average, the highest number of practice exam instances, ensuring they have ample opportunities to correctly attempt most of the exam questions. Consequently, their scores on each practice exam instance are not necessarily high, as indicated by the low average for Metric 3.

In contrast, students in the Mock-masters group, who open fewer practice exam instances but take each one of them seriously by attempting the majority of the questions, display the highest average for Metric 3, indicating strong performance in these mock exams.

On average, Mock-masters' exam scores are 6.5 percentage points higher, and All-rounders' scores are 10.2 percentage points higher, compared to the other students, both statistically significant ($p < 0.001$). When examining these categories by testing modality, we find that a larger proportion of students prefer the All-rounder strategy during CBTF exams, as depicted in Figure 5.

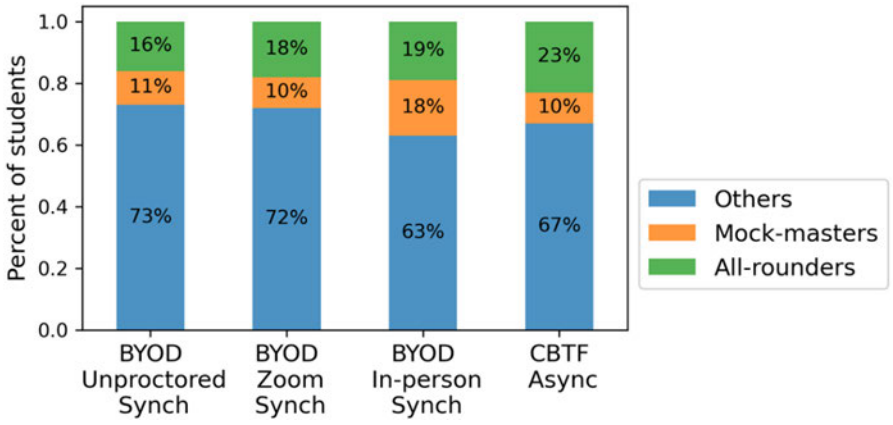


Figure 5: Distribution of the study pattern categories as a function of test modality, indicating that students tend to prefer an all-rounder approach when taking exams at the CBTF.

3.4 Investigating the impact of students' study timing on exam performance

We also wanted to understand when students engage with practice materials and how this timing correlates to exam performance. PrairieLearn records the time of students' submissions to each QIID during their practice sessions. Additionally, we have data on the exact times that students took their exams. For each practice exam submission, we calculate the variable "time elapsed", which measures the time in days between the submission and the corresponding exam date.

Our dataset includes 580,000 question submissions over the four semesters. Analysis reveals that 75% of these submissions occur within the 24-hour period immediately preceding the exam, 12% between 24 and 48 hours prior to the exam, and 13% more than 48 hours before the exam. Based on this, we proposed two metrics to characterize students' study behaviours:

- **Metric 5 (Cramming)**: this metric represents the proportion of submissions made within the 24-hour period leading up to the exam, indicating the intensity of last-minute study efforts.
- **Metric 6 (Early-start)**: this metric represents the proportion of submissions made more than 48 hours prior to the exam, reflecting early and potentially more distributed practice.

To investigate the effect of timing of study, we proposed the following regression model:

$$s_{ij} = \kappa_1 M5_{ij} + \kappa_2 M6_{ij} + \alpha g_i + \gamma \quad (3)$$

Here $M5_{ij}$ is the proportion of submissions made within the 24-hours period before exam j by student i and $M6_{ij}$ is the proportion of submissions that are completed more than 48 hours prior to the exam j by student i. This model controls for pre-existing academic ability with the GPA variable g_i. The coefficients from this regression are included in Table 5.

Table 5. Results from regression in Eq.(3)

Regression parameters	***coefficient***	***p-values***
Metric 5 (cramming)	$\kappa_1 = -0.29$	<0.0001
Metric 6 (early-start)	$\kappa_2 = 0.20$	<0.0001
GPA	$\alpha = 0.31$	<0.0001
Intercept	$\gamma = 0.22$	<0.0001

The coefficient for cramming is negative and statistically significant, indicating that students who emphasize last-minute study sessions do worse, while the coefficient for early practice is positive and also statistically significant. Again, we are not sure of the direction of causality for these results. Are students doing worse because their massed practice doesn't lead to durable learning, or are stronger students more organized and/or less prone to procrastination and engage in more distributed practice? Nevertheless, these findings are consistent with the benefits of distributed practice [7].

4 CONCLUSIONS

We identified several study behavior patterns based on metrics derived from students' interactions with the practice exams. In summary, we concluded that:

Time investment: Students generally spent more time studying for exams conducted with greater security measures. Proctored exams in institutional computer labs (CBTF) resulted in the highest study time, while unproctored remote exams saw the least.

Study quality: higher student performance was correlated more strongly with demonstrated mastery on practice exams (correctly answering a larger fraction of the question generators (QIDs) attempted) than from just attempting all of the different question generators. This finding is not impacted by the testing modality.

Study pattern: We observed two strategies that students engaged in that were correlated to better performance than the class as a whole: Mock-masters solved a few mock exams earnestly (attempting most questions) and All-rounders generated a large number of practice exams and selectively attempted

problems so as to see all distinct questions. A larger fraction of students behave as All-rounders (the best strategy) when taking exams in the most secure conditions (CBTF exams).

Study distribution: Taking practice exams more than 48 hours before the exam is positively correlated to exam performance. In contrast, doing the bulk of one's studying in the last 24-hours prior to the exam is negatively correlated with performance.

This study provides valuable insights into how computer-based testing modalities influence student study behaviors and performance. It suggests that increased exam security leads students to engage in more practice with and to use better practice strategies. In addition, our findings are consistent with existing literature finding that targeted practice and spacing study efforts over longer periods contribute to higher exam performance compared to cramming.

REFERENCES

[1] R. Ballagas, M. Rohs, J. G. Sheridan, and J. Borchers, "BYOD: Bring your own device," In Proceedings of the Workshop on Ubiquitous Display Environments, 2004.

[2] C. Zilles, M. West, G. Herman, and T. Bretl, “Every university should have a computer-based testing facility”, in *Proceedings of the 11th International Conference on Computer Supported Education (CSEDU 2019)*, 2019.

[3] R. Gulati, M. West, C. Zilles, M. Silva, “Comparing the Security of Three Proctoring Regimens for Bring-Your-Own-Device Exams”, in *Proceedings of the 55th ACM Technical Symposium on Computer Science Education*, SIGCSE 2024, 2024.

[4] M. West, G. L. Herman, C. Zilles, "Prairielearn: Mastery-based online problem solving with adaptive scoring and recommendations driven by machine learning," In *2015 ASEE Annual Conference & Exposition,* 26-1238, 2015.

[5] M. West, N. Walters, M. Silva, T. Bretl, C. Zilles, "Integrating diverse learning tools using the prairielearn platform," In *Seventh SPLICE Workshop at SIGCSE*, 2021.

[6] C. Emeka, M. West, C. Zilles, M. Silva, “A Comparison of Proctoring Regimens for Computer-Based Computer Science Exams”, in Proceedings of the 28th ACM Conference on Innovation and Te*chnology in Computer Science Education*, ITiCSE 2024, 2024.

[7] P. C. Brown, H. L. Roediger III, M. A. McDaniel. “*Make it stick: The science of successful learning”*. Boston, MA: Harvard University Press, 2014.