Effects of Mastery-Inspired Checkpoint Quizzes in a Large Introductory CS Course: A Mixed Methods Study

Sadia Sharmin
University of Toronto
Toronto, Canada
s.sharmin@utoronto.ca

Paul He
University of Toronto
Toronto, Canada
paulhe@cs.toronto.edu

Abstract

Students may learn at different paces due to differences in prior programming experience (PPE), physical or mental health challenges, and economic or lifestyle barriers (e.g., employment or caregiving responsibilities). Additionally, increasing use of AI tools for homework can contribute to inaccurate self-assessment and poor preparation for supervised tests. Motivated by these challenges, we introduced bi-weekly low-stakes checkpoint quizzes in a large (~500-student) introductory CS course. Inspired by alternative grading paradigms such as mastery learning (ML), each quiz could be attempted multiple times without penalty, offering students frequent feedback and opportunities for iterative improvement. Our mixed-methods study investigated the impact of these quizzes on performance, self-assessment, stress levels, and overall experience, using performance and survey data (N=456). Results showed that though retake opportunities allowed students to improve quiz performance, frequent retake attempts were associated with lower final exam outcomes, suggesting continued struggle on novel problems. Despite limited performance benefits, survey data revealed strong affective outcomes based on overwhelmingly positive student sentiment: students reported high Likert-scale ratings for learning/engagement and stress reduction value (though subgroup differences by gender, PPE, English fluency, and retake frequency suggest room to improve equity outcomes), and the majority of open-ended responses described the quizzes as helpful for improving self-assessment, reducing stress, and supporting meaningful learning. Overall, our implementation allowed students to experience some benefits of ML while retaining enough structure to prevent procrastination, illustrating how ML-inspired assessment can be incorporated into courses without a full course redesign.

CCS Concepts

• **Social and professional topics → CS1**; **Computer science education**.

Keywords

mastery, CS1, alternative grading, mixed methods, surveys

ACM Reference Format:
Sadia Sharmin and Paul He. 2026. Effects of Mastery-Inspired Checkpoint Quizzes in a Large Introductory CS Course: A Mixed Methods Study. In *Proceedings of the 31st ACM Conference on Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2026), July 10–15, 2026, Madrid, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3803400.3809375

ITiCSE 2026, Madrid, Spain

ACM ISBN 979-8-4007-2634-7/2026/07
https://doi.org/10.1145/3803400.3809375

1 Introduction

First-year CS students often learn at different paces, leading some students—often those from minoritized backgrounds or without prior programming experience—to struggle within courses [7, 16]. Others may develop a false sense of proficiency often due to the trending usage of AI tools for homework, resulting in inaccurate self-evaluation and subsequent inadequate preparation for major supervised assessments [11].

To address these concerns, we explored aspects of "mastery learning" (ML) – a pedagogical approach based on multiple re-assessment opportunities allowing students to demonstrate mastery of learning objectives at their own pace [2]. While this approach has been gaining popularity in education, implementations of it within large, undergraduate courses often face significant challenges such as issues with scalability, the potential to encourage procrastination, and increased instructional staff workload [1, 3, 6, 15].

In our large (~500 student) undergraduate introductory programming and proofs class, we wanted to allow students to experience the benefits of mastery learning, while mitigating some of the aforementioned challenges. Drawing from various existing implementations of alternative grading, consulting guidelines and recommendations for such assessments, and keeping the challenges of scale and effectiveness in mind, we developed a "mastery learning"-informed approach that aims to balance flexibility with structure: bi-weekly low-stakes checkpoint quizzes with retake opportunities. To prevent procrastination and scheduling conflicts, retake deadlines were strictly spaced one week apart.

As mentioned above, two key motivating factors for our implementation were: (1) students learning at different paces, often those from different backgrounds, and (2) students having difficulty assessing their own understanding. Our research questions were:

- **RQ1:** What percentage of students improved their quiz performance due to the retake opportunities? Did certain groups of students (e.g. based on gender, prior programming experience, and English language fluency) benefit more from the retakes than others?
- **RQ2:** Did the quizzes help students predict their term test and final exam performances?
- **RQ3:** How did the quizzes impact students' self-reported learning experience in the course, and their stress/anxiety levels within the course? Was there any difference based on gender, prior programming experience, English language fluency, and the total number of retakes they attempted?

Additionally, since this was our first time using these quizzes:

- **RQ4:** What suggestions and general thoughts did students have about the quizzes and ways to improve them?

2 Background

There has been a growing shift toward more equitable and mastery-based grading systems throughout education [1, 5, 14, 19]. Traditional grading models use a fixed time-frame to evaluate students, often leading to varied levels of achievement among individuals. Alternative grading approaches challenge this by shifting the focus from time-constrained assessment to evaluating progress at individual paces, aiming to ensure that all students have the opportunity to achieve sufficient understanding of each concept before advancing to more complex ones. A critical component of such models is the "feedback loop" involving multiple cycles of assessment, feedback, and reassessment [14]. This aims to shift the emphasis from earning points to meaningful learning through individualized feedback.

Mastery learning (ML), as initially conceptualized by Bloom [2], posits that every student can learn a concept if given sufficient time. Unlike traditional grading models where achievement varies while time remains constant, ML allows time to vary while aiming to hold achievement constant [18], based on the idea that fundamental concepts must be mastered before subsequent learning can occur. Various ML models exist [19], two key ones being Bloom's Learning for Mastery system [2] which involves instructor-paced learning and collaboration among students, and Keller's Personalized System of Instruction [9] which emphasizes independent, self-paced learning. Recent work by Szabo et al. [19] builds on this area of literature by presenting five distinct ML models, highlighting the adaptability and diversity in ML implementations, particularly focusing on computing education contexts.

The effectiveness of ML approaches has been well documented [1, 8, 10] being generally accepted as a promising pedagogical technique within the field of education, including Computer Science Education (CSE) [6]. In CS1 specifically, aspects of ML appear to have been incorporated with varying degrees of success. Robins' [17] Learning Edge Momentum theory suggests that students fall behind when they fail to grasp prerequisite concepts, reinforcing the importance of implementing mastery-based approaches in CS1. Their paper discussing their mastery-based CS1 course [12] reports that while student completion rates remained unchanged, students indicated improved learning of course material. However, students expressed concerns about time management and the course's perceived lack of organization. Additionally, the authors found that some students disengaged from the material after achieving a passing grade, instead of striving for mastery of all course content [12]. Ott et al. [15] also implemented a mastery learning-oriented CS1 course and found mixed results: an overall decrease in task completion and longer periods of student inactivity. They found interventions to increase student engagement with the course material to achieve only partial success. Another mastery learning based CS1 course [3] also faced issues with student procrastination, with students delaying coursework until the end of the term and beyond. They found no improvements in completion rates nor student performance, though students did report a deeper engagement with and understanding of the course material.

Despite the potential benefits of ML approaches, several logistical challenges (e.g. scalability, instructor workload, staff training, student engagement, and procrastination) have prevented a widespread adoption within CSE [1, 3, 6, 15, 19]. To encourage broader implementation in the future, this paper explores how mastery learning principles can be meaningfully integrated into an existing course without a full-scale course restructure.

3 Context and Methodology

3.1 Course Structure

Our course ran from September to December 2024, spanning 12 weeks of lectures, followed by a three-hour paper-based final exam. It covered both CS1 fundamentals and introductory proofs. Each week included six hours of lecture—two hours each on Mondays, Tuesdays, and Thursdays—and a two-hour tutorial session on Fridays. Initial enrollment for the course was 525 students, divided almost evenly among two lecture sections taught by different instructors. All core course materials and assessments were developed collaboratively and shared across both sections.

Each tutorial consisted of approximately 35 students and were used to administer checkpoint quizzes proctored by a Teaching Assistant (TA). Due to technical limitations and institutional regulations, the majority of CS courses at our institution use paper-based exams. The quizzes were thus also paper-based to help students practice for and evaluate their preparedness for larger assessments.

Each quiz focused on one chapter of the course textbook. As we typically covered two chapters per week, each tutorial session included two short quizzes. Each quiz contained 3-4 questions (question formats included short answer, proofs, coding questions, fill-in-the-blanks, multiple choice, etc.). While quizzes were designed to be completable within 30 minutes, students were allowed to use the full two-hour tutorial period to complete the two quizzes to reduce anxiety and accommodate different work paces.

3.1.1 Quiz Grading. Each chapter's quiz was worth 2% of the overall course grade. Quiz questions were graded using rubrics awarding partial marks for partially correct responses. Importantly, to reduce stress and encourage meaningful learning rather than a focus on earning points, all quizzes were graded so that any score above 77% would earn full marks. *(Note: The threshold of 77% was used as this is currently our department's threshold requirement for program entry).* For scores below 77%, the student kept the actual score earned (rather than being awarded the full 2/2), and opportunities for retakes were available. At the end of all retakes (if any), the highest score was kept as the final quiz score.

Quizzes were written during Friday tutorial sessions and graded by TAs the following week in a meeting coordinated by a lead TA to ensure consistency. Grades were typically returned by Tuesday night or Wednesday morning so the students would have enough time to prepare for a retake that Friday if needed.

3.1.2 Retake Formats and Scheduling. Each retake involved a new quiz version testing similar content. The first and second attempts were paper-based and held during tutorial sessions on alternating weeks (i.e. Week 1 tutorial = first attempt for Quiz 1+2, Week 2 tutorial = second attempt for Quiz 1+2, Week 3 tutorial = first attempt for Quiz 3+4, Week 4 tutorial = second attempt for Quiz 3+4, and so on). During retake weeks, we also gave students the option to complete their first attempt of the upcoming week's quizzes ahead of time. This was for two reasons: (1) to give students more flexibility in the quiz scheduling, allowing strong/keen students to work ahead (2) to reduce potential feelings of awkwardness or embarrassment among students attending retake sessions, who might otherwise be the only ones present in tutorial those weeks.

For those needing a third attempt, we sent an invitation link to sign up for a one-on-one verbal interview with an instructor or TA. These were 15-20 minute long meetings where students were asked 3-4 "technical interview"-style questions, testing similar content as the paper-based versions. Responses were evaluated during the interview using a standardized rubric, with partial marks awarded for partially correct responses. Scores were recorded during the interview and released to students once all interviews had been completed. A benefit of the third attempt format was allowing instructors to meet with and provide guidance to students who may be struggling with course material throughout the term.

3.2 Data Collection

3.2.1 Demographic Data. To explore whether our checkpoint quizzes were equitable and inclusive, and whether they were experienced differently based on student backgrounds, we collected data about students' gender, level of English fluency, and level of prior programming experience (PPE). The reason we wanted to look at gender and PPE is due to the longstanding finding that non-male students and students without PPE face unique challenges within CS1 courses. We considered English language fluency due to our particular class having several non-native English speakers within an English-speaking institution.

For gender, our survey options were: "Male", "Female", "Other" (with a field to specify) and "Prefer not to say". For English fluency, we asked "How fluent do you consider yourself in English?" and provided a scale from 1-5 with labels "Not fluent at all" (1) and "Very fluent" (5). For PPE, we asked "Did you have prior programming experience before this class?" with a 4-point scale labeled (from 1 to 4) as: "None", "Some experience (e.g., self-taught, coding tutorials, or minor exposure)", "Moderate experience (e.g., completed a high school or introductory programming course)" and "Extensive experience (e.g., advanced courses, internships, or projects)".

3.2.2 Performance Data. Student grades on checkpoint quizzes, two term tests and the final exam were included in our analysis. We also considered students' total number of retake attempts and improvement between the first and strongest attempts (i.e. the difference between the final quiz grade and first attempt grade). The average of these per-quiz improvements was considered the overall "mean learning gain".

3.2.3 Survey Data. To deepen our understanding of how the students perceived the quizzes, we also collected feedback using an online survey containing Likert-scale questions about learning benefits and stress regulation, and the following open-ended question: "What suggestions would you make to improve the checkpoint quizzes or how they are used, for future offerings of this course? Do you have any other thoughts you want to share?". The survey was administered to the 525 students enrolled in the class, once at the midterm point (~82% participation rate) and again at the end of the term (~87% participation rate).

3.2.4 Researcher Positionality and Ethics Approval. Both researchers served as instructors for the course during which the research was conducted. We recognize that our roles as both researchers and instructors could introduce potential biases and power dynamics. To mitigate these concerns, students were explicitly informed that their survey responses would have no impact on course grades and that instructors would not review any responses until after the course had concluded and final grades had been submitted. Additionally, data was anonymized using an automated script before analysis. All research procedures, including data collection and analysis, were approved by our institution's ethics review board.

4 Analysis and Results

4.1 Quantitative Analysis

4.1.1 Demographic Data. The distribution for pariticipants was: 71% male, 28% female, 1% "Prefer not to say"; majority of the class (88%) reported themselves to be highly fluent in English (values 4-5), and 12% of the class reported themselves to be less fluent (values 1-3); 66% of the class had moderate to extensive PPE, 22% of the class had some experience, and 11% of the class had no experience.

4.1.2 Performance and Learning Gains. Across the 10 checkpoint quizzes, an average of 75% of students achieved a full mark on their first attempt. This increased to over 90% through the use of retake opportunities. Students generally improved their performance with each retake: on average across the 10 quizzes, about 25% of all students needed a second attempt, and about 10% of all students needed a third. Among those who completed a third attempt, an average of 85% improved their quiz grade (70% achieved full marks).

A linear regression predicting mean learning gains (improvement from first attempt to final quiz marks) from gender, English fluency, and prior programming experience ($F(3, 448) = 26.0, R^2 = 0.149, p < 0.001$) revealed that, while gender and English fluency were not significant predictors, there was a significant negative association between programming experience and learning gains ($t = -8.34, p < 0.001$). That is, students with less prior experience experienced larger improvements due to retake opportunities.

To examine whether learning gains from quizzes were reflected in overall performance (as indicated by final exam scores), we conducted a hierarchal linear regression predicting exam scores. After controlling for gender, PPE and English fluency (Model 1: $R^2 = 0.123$), we added first-attempt quiz scores (Model 2: $R^2 = 0.473$), final quiz scores (Model 3: $R^2 = 0.487$), and lastly, total number of retakes across the term (Model 4: $R^2 = 0.546$). Model 1-2 comparisons indicate initial quiz performance strongly predicted exam outcomes ($b = 0.405, p < .001$ in Model 2). Adding final quiz scores added significant but minimal explanatory power ($\Delta R^2 = 0.014, p < .001$ between Model 2-3), suggesting that while students often improved their quiz performance through retakes, these improvements translated into only modest additional variance in exam performance beyond what was already predicted by initial quiz scores. Lastly, the number of total retakes across the term emerged as a significant negative predictor in Model 4 ($b = -2.319, p < .001$), with each additional retake attempt leading to a roughly 2.3-point decrease in exam score.

4.1.3 Factor Analysis. The survey included 15 Likert questions on a 5-point scale about students' experience with checkpoint quizzes related to their learning, engagement, and feelings of stress/anxiety about the course and its assessments. An exploratory factor analysis (EFA) was conducted on the final survey responses using maximum-likelihood factoring with oblimin rotation allowing factors to correlate. Our initial analysis based on parallel analysis and a scree test suggested four factors. However, two of the four factors had only 2-3 items each, suggesting a weak/unstable structure, and one item showed low loading ($< .3$) with high uniqueness (.9). We removed this item and reran the analysis following guidelines from Costello and Osborne [4] for factor retention and interpretability. The final model had two factors; all items loaded at or above .44, there were no item crossloadings, and the two factors were modestly correlated ($r = .26$) suggesting related but distinct constructs. Bartlett's test of sphericity was significant ($\chi^2(91) = 3577, p < .001$), and the KMO measure indicated excellent sampling adequacy ($KMO = .869$).

Factor 1 represented students' perceived learning/engagement value of checkpoint quizzes (involving questions such as "How helpful are the checkpoint quizzes in identifying which concepts need more of your attention?" and "How effective are checkpoint quizzes in helping you master the course content over time?"), and factor 2 represented stress reduction due to retake opportunities (with questions such as "How much does knowing you have the option to retake checkpoint quizzes reduce your stress?" and "How helpful do you find the retake opportunities are for managing your stress about mastering course material overall?"). Internal consistency was high for the perceived learning factor ($\alpha = .90$) and acceptable for the stress regulation factor ($\alpha = .75$).

To summarize students' perceptions, we computed two subscale scores based on the EFA results. Across both surveys, students reported high scores on both (Table 1).

Table 1: Factor Means Across Surveys (1-5 Likert scale)

	Midterm Survey		Final Survey	
Factor	M	(SD)	M	(SD)
Learning/Engagement Value	4.12	(0.64)	4.31	(0.58)
Stress Reduction Value	4.18	(0.84)	4.37	(0.74)

Note: Higher scores indicate stronger agreement.

Secondly, we calculated factor scores to analyze if students' demographic and background factors, and number of total retake attempts, affected survey responses. We conducted a multiple linear regression analysis for each factor score with gender, level of prior experience, English fluency and total number of retake attempts as predictors. For factor 1 (perceived learning/engagement value), the regression was statistically significant but explained a small proportion of variance ($F(4, 451) = 4.87, p < .001, R^2 = .041$) with self-reported English fluency being the only significant predictor ($t = 3.717, p < .001$), though the effect was quite small. For factor 2 (stress reduction value), the regression was statistically significant and explained a modest proportion of variance ($F(4, 451) = 11.8, p < .001, R^2 = .095$). All four predictors were significant. Greater prior programming experience ($t = 2.62, p = .009$) and higher English fluency ($t = 2.34, p = .02$) were associated with greater perceived stress reduction, and female students reported greater stress reduction benefits than male students ($t = -2.29, p = .022$). Lastly, students with more total retake attempts perceived less stress reduction ($t = -2.75, p = .006$).

4.1.4 Self-assessment. To analyze whether the quizzes were a good way for students to predict their readiness for the tests and exam, we considered: (1) Were quiz scores correlated with the term tests covering the same chapters and with the final exam? (2) Did students *feel* the quizzes helped their self-assessment?

Quizzes 1-6 tested the same content as test 1, and quizzes 7-10 as test 2. The final exam was cumulative and included content from the entire course (i.e. quizzes 1-10). We found that average quiz performance for quizzes 1-6 was significantly correlated with Test 1 performance ($r = 0.5, p < .001$). The average for quizzes 7-9 was also significantly correlated with Test 2 ($r = 0.8, p < .001$). Lastly, the average of all quizzes (1-10) was significantly correlated with the final exam scores ($r = 0.6, p < .001$). These moderate to strong correlations suggest the quizzes should have been fairly useful self-evaluation tools for students. Based on survey data, most students agreed with this statement. On the final survey, students were asked if the checkpoint quizzes had helped them "predict [their] performance/readiness" for test 1 and test 2 (the final exam had not yet occurred), with three response options. Student responses were distributed as follows:

- **Test 1 -** Yes, definitely (57%), Somewhat (35%), Not at all (8%);
- **Test 2 -** Yes, definitely (46%), Somewhat (41%), Not at all (13%)

4.2 Qualitative Analysis

Following grounded theory procedures outlined by Miles et al. [13], we conducted open coding on open-ended survey responses to identify recurring themes. Before analysis, we removed blank and meaningless inputs (e.g., "no suggestions," "N/A"), as well as any responses unrelated to checkpoint quizzes (e.g. "I love CS", "Thank you for teaching us"). After filtering, 174 responses from the midterm survey and 176 from the final survey were analyzed.

Our analysis followed a three-stage process. First, both researchers reviewed all responses and developed an initial set of codes independently; the codes were purely descriptive (e.g., "mentions stress reduction," "suggests more retakes"), rather than interpretive, to limit researcher bias within coding. No existing literature or theoretical frameworks were consulted to ensure the codes emerged directly from the data. Next, the independently generated codes were compared, and any discrepancies were resolved through discussion until consensus was reached. As with the initial coding, rather than relying on predefined categories, we allowed themes to emerge organically from the data. Through axial coding, similar codes were grouped together, revealing key themes that reflected students' perspectives. Both researchers collaboratively refined and finalized these themes, which are discussed in the next section. It should be noted that though we analyzed the midterm and final surveys separately, the same themes were prevalent in both.

4.2.1 Praise for Checkpoint Quizzes. Though our question specifically asked students for improvement suggestions, the most prevalent theme – corresponding to 108 of 174 responses (62%) in the midterm survey and 98 of 176 (56%) in the final survey – expressed praise and appreciation for the quizzes. As these were not addressing our specific question, we grouped "praises" into one theme.

Many students described the quizzes as "good", "great", "amazing", or "perfect". Specifically, students highlighted that the quizzes were helpful for self-assessment, reduced stress, helped them stay on track, reinforced their understanding of course material, built their confidence for larger course assessments, and improved their learning. Several responses in both the midterm survey and final survey expressed a desire for other courses to implement checkpoint quizzes and mentioned the positive impact the flexibility of the quizzes had on their mental health. Notably, many students expressed that they felt cared for due to the course structure, e.g. "I feel that it centers around our success and education as students." and "This really makes me feel like the instructors want us to succeed by giving us an opportunity to demonstrate our mastery of certain knowledge, even if we don't get it the first time."

Lastly, the following feedback from one student was a great indication that we are meeting one of our key goals of inclusivity: "As a neurodivergent student, I'd just like to say I really love the flexible structure of the course. [...] Out of all the assessment systems I've encountered so far in my courses, I believe the checkpoint quiz system has been the most effective in motivating me to stay on track, understand, and retain the course material. The flexibility and low stakes are particularly valuable to me as a neurodiverse individual who also struggles with anxiety."

4.2.2 Quiz Content. The second most prevalent theme across both surveys was a desire for more difficult questions that better align the quiz difficulty to the tests/exam. To preserve the "low-stress" nature of the current quizzes however, some students suggested adding an *optional* challenge question, e.g.: "I felt as though the midterms had significantly more difficult questions, and having a graded optional (or even mandatory though that might defeat the purpose of checkpoint quizzes) questions of the same difficulty would be helpful."

4.2.3 Coverage. Several students suggested more comprehensive checkpoint quizzes that incorporate material from previous units, rather than each quiz focusing on one chapter. Some students mentioned this would motivate them to revisit earlier topics more frequently, thus helping them retain knowledge over time. Other students mentioned cumulative quizzes would provide a better "overall depiction of [their] abilities". Lastly, some suggested increasing quiz length for more thorough coverage of the content.

4.2.4 Grading. A suggestion for longer quizzes also frequently showed up in responses discussing the general scoring of the checkpoint quizzes. Since the quizzes were meant to be completed within 30 minutes, they were quite short in length (about 3-4 questions long, where each question should not take more than 5-10 minutes). The final grade for each quiz was generally out of a total of 12-15 marks. Some students expressed frustration about minor deductions in the quiz grade leading to significant reductions to the overall percentage for that quiz. Students suggested that longer quizzes with more questions and a larger spread of points would help mitigate this issue by reducing the weight of each individual mistake and providing a more balanced assessment.

4.2.5 Feedback and Solutions. Several students requested that solutions be posted shortly after grading. We had chosen not to post solutions and instead encouraged students to work through the solutions on their own or seek personalized feedback from instructors. Another common theme was students requesting more detailed feedback on their incorrect quiz responses, e.g.: "I would have preferred more thorough feedback because sometimes I would get the question wrong but there would be no elaboration on what specifically I should have done differently."

5 Discussion

5.1 RQ1: Performance

As discussed in Section 4.1.2, the retake opportunities allowed over 90% of students to receive the full marks on each quiz, compared to the average (across all quizzes) of 75% of students achieving full marks on the first attempt. These results suggest that retake opportunities substantially supported student improvement from the first attempt, consistent with prior mastery learning literature. A linear regression analysis suggested this was especially beneficial for students with less PPE. However, despite improvements within quizzes, the impact of retakes on overall performance gains (as indicated through final exam scores) was minimal. Notably, the total number of retakes across the term emerged as a significant negative predictor of exam performance, which likely suggests students who required more attempts continued to struggle on larger assessments. In the future, retake frequency could thus be used as an early diagnostic tool to identify students in need of additional support/resources supplementing retake opportunities.

5.2 RQ2: Self-assessment

One of the motivations for introducing the quizzes was to give students more frequent feedback about their standing in the course. In past offerings of the course, the written proctored assessments included only the term tests and final exam; anecdotally, many students who underperformed on the first term test expressed shock about the results. Since written supervised assessments significantly differ from computer-based homework assignments, it was understandably difficult for many students to judge their own preparedness. By introducing bi-weekly, paper-based quizzes with questions of a similar format to those on term tests and the exam, we hoped to improve students' ability to self-evaluate their readiness. According to the correlational and survey data reported in Section 4.1.4, the average quiz performance was significantly correlated with larger assessments testing the same material and the vast majority of students found the quizzes to be at least somewhat useful for predicting their performance/readiness for those assessments. As mentioned in Section 4.2.1, the open-ended responses also included students describing the quizzes as helpful for self-assessment and building confidence for larger assessments.

5.3 RQ3: Learning Experience and Stress Levels

Based on both the quantitative and qualitative analysis, students generally found the quizzes to have positive effects on their overall learning and stress levels. As shown in Table 1, Likert-scale scores regarding perceived learning/engagement value and stress reduction benefits of the checkpoint quizzes was very high throughout the term. Consistent with these quantitative findings, students' open-ended responses also frequently described the quizzes as being helpful for reducing stress, keeping up with course material and reinforcing understanding of core concepts (Section 4.2.1).

Linear regression analysis indicated that perceived learning and engagement benefits of checkpoint quizzes were largely consistent across student gender, prior programming experience and English fluency. This means, although students with less PPE showed greater learning gains (as per Section 4.1.2), those with more experience still perceived similar value in the quizzes for supporting their learning. Additionally, the number of retakes a student completed did not significantly affect the perceived learning/engagement value, suggesting that the quizzes were perceived as similarly beneficial regardless of how often retake opportunities were used. Overall, these findings suggest that checkpoint quizzes were generally seen as beneficial for learning/engagement across students with varying backgrounds and quiz performance (as suggested by total retakes).

As for stress reduction benefits, we observed mixed outcomes in relation to our equity goals. Female students (a traditionally disadvantaged and minoritized group in CS) reported greater stress reduction benefits than males, suggesting a potential for improving gender inclusivity. However, the remaining significant findings were less aligned with our equity goals. Students with more PPE and higher English fluency reported greater stress reduction benefits, suggesting that the quizzes may not have equitably supported less advantaged students (i.e., those with less prior experience or lower English fluency). Lastly, students with more retake attempts reported diminished stress reduction, suggesting the retakes may have been less reassuring to those who were struggling the most. These findings motivate further exploration into the experiences of disadvantaged groups and frequent retakers.

5.4 RQ4: Suggestions for Improvement

To address this RQ, we consider our qualitative analysis of student responses to the survey questions asking for improvement suggestions and feedback. As mentioned in Section 4.2.1, student responses were largely positive, with most students praising the current implementation rather than, or in addition to, offering suggestions for improvement, indicating the quizzes were well-liked.

In terms of improvements, students expressed a desire for increased quiz difficulty, coverage, and feedback quality. While students appreciated the quizzes' low-stress nature, some suggested the inclusion of optional challenge questions, offering students a more accurate gauge of their readiness for larger tests. Secondly, the grading structure, particularly the impact of minor errors on the overall score, was a source of frustration for some students. We are thus considering alternate grading schemes such as a pass/fail model where "mostly correct" answers could be deemed sufficient. This could alleviate the pressure caused by small mistakes, while still encouraging students to demonstrate a solid understanding of the material. Lastly, several students wished for more detailed feedback about their mistakes. We could potentially achieve this by integrating TA-led weekly review sessions into the course.

6 Future Plans

Quantitative analysis revealed limited performance benefits and equity outcomes. Students with more retake attempts had significantly lower exam scores, suggesting that frequent retakers could benefit from targeted interventions, such as TA review sessions or one-on-one support, to help transfer learning gains from quizzes to larger assessments. We also plan to conduct an interview study with frequent retakers and other potentially disadvantaged or minoritized subgroups (e.g., female students, students with less prior programming experience, and students with lower English fluency) to better understand their experiences and identify more effective ways to support them.

Secondly, based on qualitative feedback, we will consider the following improvements for future iterations: optional challenge questions to better align quiz difficulty with the tests/exam, revised grading criteria to reduce the impact of minor mistakes, and more detailed feedback through TA-led pre-retake review sessions.

Additionally, one challenge with our current implementation was the high workload for course staff due to quizzes being paper-based (requiring significant overhead in terms of printing, distributing, grading, etc). In the future, we plan to conduct proctored, computer-based quizzes instead, which would reduce the workload of paper quizzes as well as provide immediate feedback to students via automated testing results. We will collect data to examine the impacts of this change, including effects on the self-assessment aspect since our larger assessments will continue to be paper-based due to technical and institutional limitations.

7 Conclusion

This paper explores the pedagogical benefits and potential improvements of a "mastery learning"-informed implementation, confirming that several benefits of ML persist even without a full course redesign. Our approach balanced flexibility with structure, giving students opportunities to revisit and refine their understanding while maintaining steady progress. Survey data showed that checkpoint quizzes were largely well-received; students cited them as helpful for self-assessment, stress reduction, and overall learning. However, learning gains across quiz retakes were not reflected in final exam performance, and students needing frequent retake attempts had lower exam outcomes (despite improvements on quiz scores across retakes). Across subgroups, perceived learning and engagement values were similar, but stress reduction value differed significantly, suggesting mixed equity outcomes. Despite these limitations, the affective benefits were clear: even when asked to suggest improvements, several students expressed enthusiastic support for the current design, noting how it supported their challenges and made them feel cared for. This suggests checkpoint quizzes are a valuable tool for supporting affective aspects of the student experience, but further revisions are needed in future iterations to better address performance, equity, and scalability.

References

[1] Claudio Álvarez, Nickolas Falkner, Päivi Kinnunen, Jaromir Savelka, and Lisa Zhang. 2025. Show Me the Mastery Learning! Obstacles to Adoption and Opportunities for New Solutions. In *Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education V. 1* (Nijmegen, Netherlands) *(ITiCSE 2025)*. Association for Computing Machinery, New York, NY, USA, 639–645. doi:10.1145/3724363.3729104

[2] Benjamin S Bloom. 1968. Learning for Mastery. Instruction and Curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. *Evaluation comment* 1, 2 (1968), n2.

[3] Jennifer Campbell, Andrew Petersen, and Jacqueline Smith. 2019. Self-paced Mastery Learning CS1. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (Minneapolis, MN, USA) *(SIGCSE '19)*. Association for Computing Machinery, New York, NY, USA, 955–961. doi:10.1145/3287324.3287481

[4] Anna B Costello and Jason Osborne. 2005. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation* 10, 1 (2005), 1–9.

[5] Joe Feldman. 2023. *Grading for equity: What it is, why it matters, and how it can transform schools and classrooms.* Corwin Press, Thousand Oaks, CA, USA.

[6] James Garner, Paul Denny, and Andrew Luxton-Reilly. 2019. Mastery Learning in Computer Science Education. In *Proceedings of the Twenty-First Australasian Computing Education Conference* (Sydney, NSW, Australia) *(ACE '19)*. Association for Computing Machinery, New York, NY, USA, 37–46. doi:10.1145/3286960.3286965

[7] Michail N Giannakos, Ilias O Pappas, Letizia Jaccheri, and Demetrios G Sampson. 2017. Understanding student retention in computer science education: The role of environment, gains, barriers and usefulness. *Education and Information Technologies* 22 (2017), 2365–2382.

[8] Thomas R Guskey and Therese D Pigott. 1988. Research on group-based mastery learning programs: A meta-analysis. *The Journal of Educational Research* 81, 4 (1988), 197–216.

[9] Fred S Keller. 1968. Good-bye, teacher... *Journal of applied behavior analysis* 1, 1 (1968), 79.

[10] Chen-Lin C Kulik, James A Kulik, and Robert L Bangert-Drowns. 1990. Effectiveness of mastery learning programs: A meta-analysis. *Review of educational research* 60, 2 (1990), 265–299.

[11] Rabab Marouf, Iouri Kotorov, Yuliya Krasylnykova, and Hamna Aslam. 2024. Evidence of Learning in the Presence of AI Tools in Computer Science Higher Education. In *KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*. Springer, Singapore, 331–341.

[12] Brendan McCane, Claudia Ott, Nick Meek, and Anthony Robins. 2017. Mastery Learning in Introductory Programming. In *Proceedings of the Nineteenth Australasian Computing Education Conference* (Geelong, VIC, Australia) *(ACE '17)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3013499.3013501

[13] Matthew B Miles and A Michael Huberman. 1994. *Qualitative data analysis: An expanded sourcebook.* Sage, Thousand Oaks, CA, USA.

[14] Linda B Nilson, David Clark, and Robert Talbert. 2023. *Grading for growth: A guide to alternative grading practices that promote authentic learning and student engagement in higher education.* Routledge, New York, NY, USA.

[15] Claudia Ott, Brendan McCane, and Nick Meek. 2021. Mastery Learning in CS1 - An Invitation to Procrastinate?: Reflecting on Six Years of Mastery Learning. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1* (Virtual Event, Germany) *(ITiCSE '21)*. Association for Computing Machinery, New York, NY, USA, 18–24. doi:10.1145/3430665.3456321

[16] Eric Roberts. 2000. Strategies for encouraging individual achievement in introductory computer science courses. In *Proceedings of the Thirty-First SIGCSE Technical Symposium on Computer Science Education* (Austin, Texas, USA) *(SIGCSE '00)*. Association for Computing Machinery, New York, NY, USA, 295–299. doi:10.1145/330908.331873

[17] Anthony Robins. 2010. Learning edge momentum: A new account of outcomes in CS1. *Computer Science Education* 20, 1 (2010), 37–71.

[18] M. A. Robinson. 1992. Mastery learning in public schools: Some areas of restructuring. *Education* 113, 1 (1992), 121–126.

[19] Claudia Szabo, Miranda C. Parker, Michelle Friend, Johan Jeuring, Tobias Kohn, Lauri Malmi, and Judithe Sheard. 2025. Models of Mastery Learning for Computing Education. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1* (Pittsburgh, PA, USA) *(SIGCSETS 2025)*. Association for Computing Machinery, New York, NY, USA, 1092–1098. doi:10.1145/3641554.3701868