A Quantitative Analysis of When Students Choose to Grade Questions on Computerized Exams with Multiple Attempts

Ashank Verma, Timothy Bretl, Matthew West, Craig Zilles
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{ashankv2, tbretl, mwest, zilles}@illinois.edu

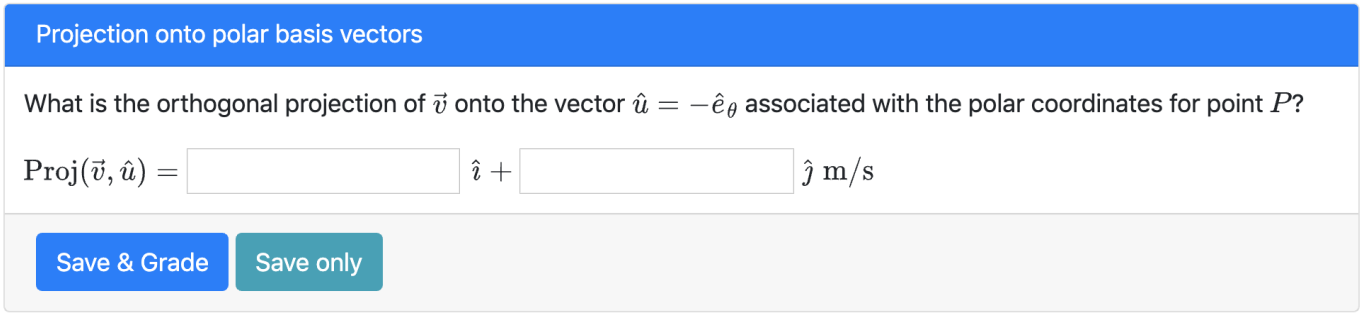

Figure 1. The PrairieLearn system presents students with two options on how to progress through an exam. Students can submit their answers for immediate feedback (Save & Grade) or store them for bulk grading later during the exam (Save only). This paper analyzes student choices.

ABSTRACT

In this paper, we study a computerized exam system that allows students to attempt the same question multiple times. This system permits students either to receive feedback on their submitted answer immediately or to defer the feedback and grade questions in bulk. An analysis of student behavior in three courses across two semesters found similar student behaviors across courses and student groups. We found that only a small minority of students used the deferred feedback option. A clustering analysis that considered both when students chose to receive feedback and either to immediately retry incorrect problems or to attempt other unfinished problems identified four main student strategies. These strategies were correlated to statistically significant differences in exam scores, but it was not clear if some strategies improved outcomes or if stronger students tended to prefer certain strategies.

Author Keywords

assessment; computerized exams; multiple attempts; agency; computer-based testing

INTRODUCTION

Computerized exams enable the auto-grading of a broad range of question types, including numeric, symbolic, and programming questions. For these auto-graded questions, exam designers can give students multiple attempts at a question, potentially for reduced credit. Doing so can partly mitigate the need for manually-assigned partial credit. For example, a student that makes a small computational error may identify the error when their first submitted answer is marked wrong.

Allowing students to have multiple attempts, though, necessitates giving them feedback during their exam. There has been little research on best practices for how feedback should be given during an exam, as it isn't possible on traditional manually-graded pencil-and-paper exams, outside of multiple-choice exams.

Furthermore, it isn't obvious when the "best" time is to give students feedback about their answers. Having answers scored immediately upon submission is potentially the most efficient approach for students, because then they can retry the problem while the details are fresh in their memory. Immediate negative feedback, however, could be anxiety provoking in some students, which could negatively impact their ability to do other exam problems. For these students, submitting answers to all questions before receiving feedback (as is done on a tradition pencil-and-paper exam) could be the best strategy, and then they could use their remaining time to re-work any incorrect problems.

In this paper, we consider a scenario where students are given the agency to decide when their exam answers should be graded and report on the student behavior with respect to this choice. Specifically, our study considers a collection of course offerings that ran computerized exams using PrairieLearn [8]

L@S '20 August 12–14, 2020, Virtual Event, USA

 ISBN 978-1-4503-7951-9/20/08…$15.00

DOI: https://doi.org/10.1145/3386527.3406740

a web-based assessment platform used for computer-based exams, which lets student have submitted answers graded immediately or stored for later grading, as shown in Figure 1.

RELATED WORK

The research literature focused on giving students multiple attempts on exams is focused almost entirely in the context of multiple-choice exams where students are allowed to "answer until correct" (AUC). A number of strategies have been developed to enable AUC for pencil and paper exams, most recently using the Immediate Feedback Assessment Technique (IFAT) which are multiple-choice bubble sheets that use scratch off material and students receive credit proportional to the number of answer bubbles not scratched off. Frary provides a survey of early work on the reliability validity of AUC in relation to traditional multiple-choice tests [5].

One of the motivations for AUC testing is to enhance the learning that occurs during the exam through providing immediate feedback process [2]. Furthermore, permitting students to have multiple attempts provides psychometric advantages by boosting both the mean item discrimination and overall test-score reliability, when compared to tests scored dichotomously (correct/incorrect) based on the initial response [7]. In addition, researchers find a strong correlation between students' initial-response successes and the likelihood that they obtain partial credit when they make incorrect initial responses suggesting that partial credit is being granted based on partial knowledge that remains latent in traditional multiple-choice tests [7].

Using an instrument similar to the State-Trait Anxiety Inventory (STAI), Attali and Powers found that student anxiety was lower after completing exam sections that provided immediate feedback [1]. Aggregate statistics regarding anxiety, however, could be misleading. DiBattista and Gosse found that using IFAT reduced test anxiety for a majority of students, but that nineteen percent of students self-reported that immediate feedback interfered with their test performance [4]. Interestingly, however, students reporting such interference expressed desire to use IFAT on future exams in similar (high) rates as non-impacted students. Richmond also provides an anecdotal estimate that 10% of students have higher test anxiety as a result of IFAT [6].

None of this work provides insight into how students approach taking their exam when multiple attempts at the same question are available.

METHODS

Exam data was collected for two semesters (Fall 2017, Spring 2018) in three courses at the University of Illinois at Urbana-Champaign: Introductory Dynamics, Introductory Solid Mechanics, and Introduction to Electronics. In these courses, the computerized exams were run using PrairieLearn in our campus's Computer-Based Testing Facility (CBTF) [9, 10]. The CBTF is a proctored computer lab that allows students to schedule their exams at times convenient to them during an allotted range of days. The computers' networking and filesystems are controlled to prevent unwanted communication or web browsing [11]. Facilities similar to the CBTF have been developed at other universities [3].

Our data set consists of every answer submission from students for the exams in the previously mentioned course-semesters. Each record includes an anonymized student ID, the submission type (`Save & Grade` or `Save only`), the exam name, the question ID, the question score, and submission time. In addition, we had the overall exam scores for each anonymized student. In total, across all 6 different course-semesters, our data set included 77 distinct exam offerings, 1,928 students, 16,054 student-exam pairs, and 298,005 answer submissions. Note that some exams were optional re-takes offered to students, so not all students in a specific course-semester took all of the exams offered that semester.

We characterized students exam-taking strategies as follows. By sorting each student's exam submissions by submission time, we can observe the order the student attempted answering the questions and the actions that they took. We distinguished seven actions which are described as follows:

1. Save-only a question, then save-only the same question
2. Save-only a question, then save-and-grade the same question
3. Save-only a question, then move on to a different question
4. Save-and-grade a question correctly, then move on to a different question. *(Note that, since this action is the only one possible after a question is marked correct, it reveals no information about the student's strategy.)*
5. Save-and-grade a question incorrectly, then save-only the same question
6. Save-and-grade a question incorrectly, then save-and-grade the same question
7. Save-and-grade a question incorrectly, then move on to a different question

For each student-exam pair, we counted the instances of each type of transition and constructed a Markov model to represent the student's behavior. The frequencies associated with the edges in the Markov model were then collected into a vector. We ran K-Means clustering from the Python SciKit library on the vectors (one per student-exam) independently for each course-semester. We ran the clustering algorithm several times with a target number of clusters ranging from 1 to 10.

RESULTS

Students use the save-and-grade button more than the save-only button. Save-only's represented only 4.07% of all answer submissions and only 23% of student-exams included even a single save-only. A distribution of the ratio of save-only to save-and-grade submissions across all student-exams is shown in Figure 2. This suggests that a majority of the students preferred receiving immediate feedback to saving and working through other questions.

Table 1. Six observed test taking behaviors.

Cluster	Name	Description	Count
A	Acers	The data points in this cluster are of those students who always get the questions right.	2614
B	Retriers	No one saves in this cluster. If they get the question wrong, the majority retry the same question.	5479
C	Save and Move	Students who save in this cluster mostly move on to a different question. If they get the question wrong, they mostly try retrying the question.	3456
D	Save and Retry	The data points in this cluster are of those students who save and then submit the same question.	2414
E	Split Retriers	No one saves in this cluster. If they get the question wrong, they are split between moving on to a different question and trying the same question again.	1513
F	Failers	The data points in this cluster are mainly of those students who get all the questions wrong.	578

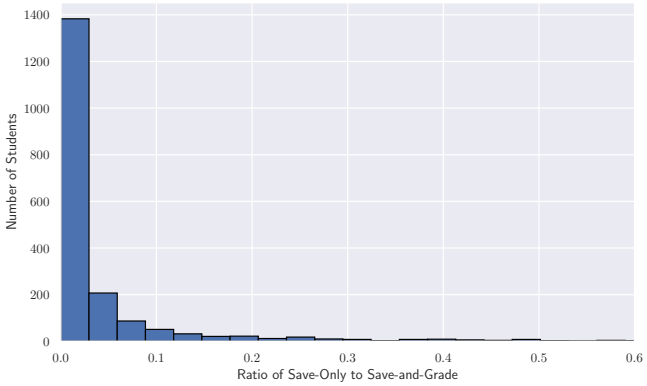

Figure 2. Distribution of save-only's to save-and-grade's across all student-exams in our data set. The majority (72%) of students rarely use save-only at all (ratio of less than 0.03). Less than 1.6% student-exams have ratios greater than 0.6.

We found that using $k = 5$ clusters produced good clustering in each of the course-semesters.[1] Comparing the clusters across the six course-semesters, we found six distinguishable behaviors, which we summarize in Table 1. Note that these clusters are ordered from highest to lowest in exam score averages.

Four of the clusters (A, B, C, and D) were present in all of the course-semesters. Cluster F was only present in both semesters of one of the courses and, we suspect it is due to the nature of the exams in the class. There are usually one or two long questions in the (50-minute) exams for this class, making it more common to receive a 0%. In the other two courses, however, Cluster F was replaced with Cluster E, which consists of students who are split between retrying the same question and moving on to a different question. Figures 3 and 4 display statistics for all 6 clusters with data from all 6 course-semesters, ordered by best exam performance.

Two of the clusters—the Acers (Cluster A) and the Failers (Cluster F)—arose because the students' strong and weak performances, respectively, resulted in some edges of the Markov model having near 0 frequencies for these students, which caused the clustering algorithm to distinguish these two populations apart from the other four. We don't consider these clusters as indicative of any particular strategy, but include them for completeness in our analysis.

[1] After plotting the number of clusters versus the inertia for the clusters, there was an apparent "elbow" in the generated curve at 5 clusters in all of the course-semesters.

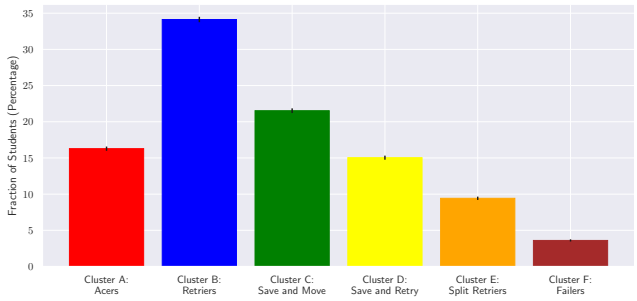

Figure 3. Proportion of students in each cluster.

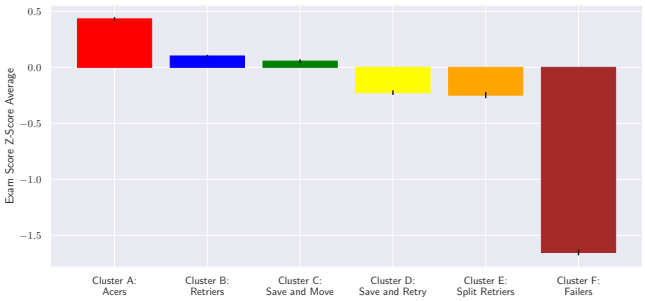

Figure 4. Mean z-score of exam performance for each cluster with 95% confidence intervals.

We found that the most common strategy among students was completing exam questions in order, retrying those that they got wrong until they ran out of attempts or arrived at the correct answer (Cluster B). Figure 3 shows the proportion of students in each cluster to the total number of students. The second most common behavior is more cautious. Students in Cluster C prefer to save their questions, move on, and come back to it in the future rather than working on that question until it is completed. The smallest clusters were the Failers (Cluster F) and Split Retriers (Cluster E). This is unsurprising, as Cluster F had students from only 2 course-semesters and Cluster B had students from 4 course-semesters, while the rest of the clusters had students from all 6 course-semesters.

We find that the clusters had statistically different average normalized exam scores. Figure 4 plots the average z-scores of exam percentages. The exam scores were normalized per exam per class for each cluster. The bars represent the averages of those normalized exam scores for each specific cluster. Unsurprisingly, Clusters A (Acers) and F (Failers) had the highest and lowest exam score averages, respectively. Clusters B and C had above average normalized exam scores, while Clusters D and E had lower than average normalized exam scores. We ran 15 t-tests with all possible cluster pairs to check for statistical significance between cluster scores. All cluster-pair average normalized exam scores were statistically different from each other ($p < 0.05$) except Clusters D and E.

DISCUSSION AND CONCLUSIONS

We used k-means clustering on 298,005 student submissions to categorize student grading-choice behavior on computerized exams in a situation where students can choose to grade questions at any time and retry them for reduced credit if incorrect. We found four clusters that were stable across multiple offerings of the three classes from which we had data, and a further two clusters that were each present in a subset of the classes. Examination of these six clusters showed that they were each associated with distinct student behaviors on the exams.

In these six clusters, we observe four different student strategies relating to grading choices (clusters B to E in our categorization); clusters A and F are primarily distinguished by the students almost always answering questions correctly or incorrectly, respectively, which shed little insight into their choices. The most popular behavior was to immediately grade a question and retry it if the answer was incorrect, and this was also associated with the highest average exam scores. This is consistent with students wanting to retry the question while it is still fresh in their memory. However, saving answers and moving on to a different question was also quite popular, and was associated with almost as high average exam scores, suggesting that deferred grading is also a viable strategy. The least popular and lowest-scoring strategies were those where the student saved an answer but then chose to grade it immediately, or where the student did not immediately reattempt an incorrect answer. Both of these strategies seem consistent with students who are unsure about how to solve a question.

From these results we see that question grading choice behavior is correlated with different outcomes, as measured by total exam score. However, we are unable to tell whether the different grading behaviors are causing the exam score differences, or whether students of different abilities are self-selecting into different strategies. This is an area of potential future research via an experimental manipulation that restricts students to particular strategies, for example by removing the option to save-only. It is also important to consider the impact of different grading strategies on student stress levels during the exam, and observational or experimental studies to probe this would be valuable.

ACKNOWLEDGMENTS

This work was partially supported by NSF DUE-1915257 and the College of Engineering at the University of Illinois at Urbana-Champaign under the Strategic Instructional Initiatives Program (SIIP). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Yigal Attali and Don Powers. 2010. Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement* 70, 1 (2010), 22–35.

[2] Richard O Beeson. 1973. Immediate knowledge of results and test performance. *The Journal of Educational Research* 66, 5 (1973), 225–226.

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

[4] David Dibattista and Leanne Gosse. 2006. Test anxiety and the immediate feedback assessment technique. *The Journal of Experimental Education* 74, 4 (2006), 311–328.

[5] Robert B Frary. 1989. Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education* 2, 1 (1989), 79–96.

[6] Aaron S Richmond. 2017. Scratch and win or scratch and lose? Immediate Feedback Assessment Technique. (2017).

[7] Aaron D Slepkov and Alan TK Godfrey. 2019. Partial Credit in Answer-Until-Correct Multiple-Choice Tests Deployed in a Classroom Setting. *Applied Measurement in Education* 32, 2 (2019), 138–150.

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

[9] C. Zilles, R. T. Deloatch, J. Bailey, B. B. Khattar, W. Fagen, C. Heeren, D Mussulman, and M. West. 2015. Computerized Testing: A Vision and Initial Experiences. In *American Society for Engineering Education (ASEE) Annual Conference*.

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

[11] Craig Zilles, Matthew West, David Mussulman, and Timothy Bretl. 2018. Making testing less trying: Lessons learned from operating a Computer-Based Testing Facility. In *2018 IEEE Frontiers in Education (FIE) Conference*. San Jose, California.