Comparison of Grade Replacement and Weighted Averages for Second-Chance Exams

Geoffrey L Herman*
glherman@illinois.edu
University of Illinois, Urbana-Champaign
Urbana, IL, USA

Zhouxiang Cai
zcai31@illinois.edu
University of Illinois, Urbana-Champaign
Urbana, IL, USA

Timothy Bretl
tbretl@illinois.edu
University of Illinois, Urbana-Champaign
Urbana, IL, USA

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

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

ABSTRACT

We explore how course policies affect students' studying and learning when a second-chance exam is offered. High-stakes, one-off exams remain a de facto standard for assessing student knowledge in STEM, despite compelling evidence that other assessment paradigms such as mastery learning can improve student learning. Unfortunately, mastery learning can be costly to implement. We explore the use of optional second-chance testing to sustainably reap the benefits of mastery-based learning at scale. Prior work has shown that course policies affect students' studying and learning but have not compared these effects within the same course context. We conducted a quasi-experimental study in a single course to compare the effect of two grading policies for second-chance exams and the effect of increasing the size of the range of dates for students taking asynchronous exams. The first grading policy, called 90-cap, allowed students to optionally take a second-chance exam that would fully replace their score on a first-chance exam except the second-chance exam would be capped at 90% credit. The second grading policy, called 90-10, combined students' first- and second-chance exam scores as a weighted average (90% max score + 10% min score). The 90-10 policy significantly increased the likelihood that marginally competent students would take the second-chance exam. Further, our data suggests that students learned more under the 90-10 policy, providing improved student learning outcomes at no cost to the instructor. Most students took exams on the last day an exam was available, regardless of how many days the exam was available.

*Corresponding Author

ICER '20, August 10–12, 2020, Virtual Event, New Zealand

ACM ISBN 978-1-4503-7092-9/20/08…$15.00
https://doi.org/10.1145/3372782.3406260

CCS CONCEPTS

• **Applied computing** → **Education**; *Computer-assisted instruction*; E-learning.

KEYWORDS

Computer education, assessment, second-chance testing, mastery, computer-based exams

ACM Reference Format:
Geoffrey L Herman, Zhouxiang Cai, Timothy Bretl, Craig Zilles, and Matthew West. 2020. Comparison of Grade Replacement and Weighted Averages for Second-Chance Exams. In *Proceedings of the 2020 International Computing Education Research Conference (ICER '20), August 10–12, 2020, Virtual Event, New Zealand*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3372782.3406260

1 INTRODUCTION

Efforts to improve student learning tend to focus on replacing lectures with active learning [11, 15] with little attention given to replacing the traditional assessment paradigm of "two midterms and a final" [23]. This lack of attention may be because faculty do not generally think of exams as a mechanism for improving students' learning [14]. This perception is unfortunate, as many studies indicate that how students are assessed may matter more than how they are taught: students decide what to learn based mostly on how they are assessed and whether they are given opportunities to respond to feedback from those assessments [12].

The traditional assessment paradigm of high-stakes, one-shot exams can be detrimental to students' learning because it provides few incentives for students to reflect on what they have learned (see the top of Figure 1). This metacognitive feedback is vital as it primes students for future learning [26, 27]. Additionally, if students' fail to learn prerequisite material, they are more likely to struggle to learn future information that builds on that material.

In contrast, self-paced mastery learning (see the middle of Figure 1) requires students to use the metacognitive feedback from testing, as they repeat exams to master each topic before moving on [2, 20]. It has been shown consistently that mastery learning is more effective for learning than traditional instruction [13, 21]. In spite of its effectiveness, self-paced mastery learning is hard to

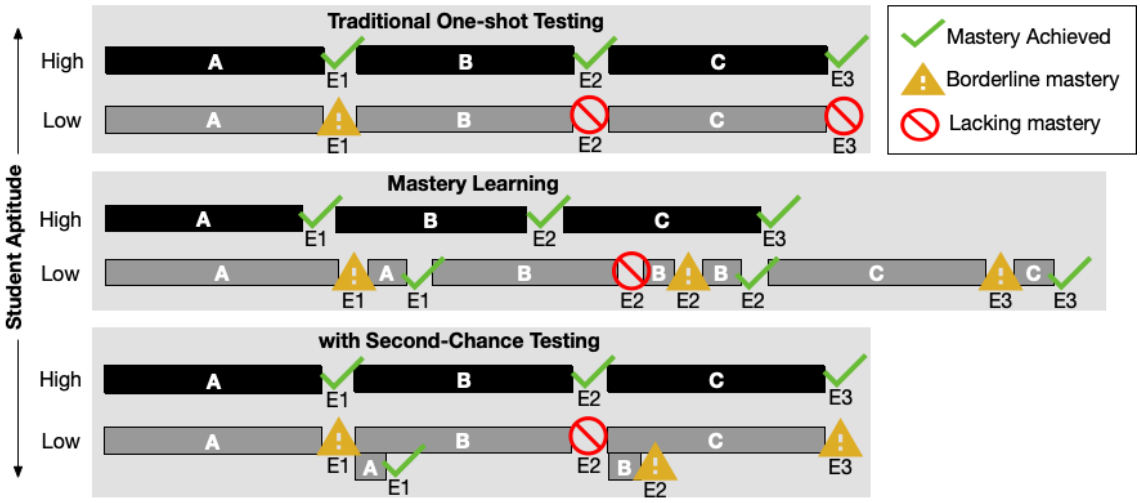

Figure 1: Comparing traditional (one-shot) exams, mastery learning, and second-chance testing with an illustrative class with two mid-term exams (E1, E2) and a cumulative final (E3). Traditional one-shot testing works fine for students with high aptitude, but students with lower aptitude don't learn the material sufficiently to demonstrate mastery on exams. In contrast, mastery learning gives students the flexibility to take assessments when they are ready for them and repeat them until mastery is achieved, but mastery learning is challenging to implement in most college environments. Second-chance testing provides students a chance to remediate after they receive feedback; furthermore, its test-potentiated learning helps students retain learned information longer to improve performance during the rest of the class.

adopt because it requires additional preparation by the instructor and because it conflicts with fixed-length semesters.

"Second-chance testing," where students can take a second instance of an exam to improve their grade (see the bottom of Figure 1), is an approximation to mastery learning that is less expensive and is easier to integrate with a range of college course structures [1, 17]. This model encourages students to review material after poor performance on an exam but limits the additional resources instructors have to invest in creating exams or grading them. Research on second-chance testing, while sparse, suggests that many students who retake an exam earn higher exam scores on the retake [16, 17, 30, 34].

Prior work has shown that different courses that used different grading policies had significantly different groups of students taking optional second-chance exams [16]. It could not be determined whether these differences were caused by the grading policies or other differences. Nor could it be determined whether the policies had any effect on students' learning. Understanding how course policies affect whether students take a second-chance exam can help faculty fine tune their policies to encourage students to increase their mastery of course material while managing the workload created by offering second-chance exams. We conducted a quasi-experimental study comparing different grading policies for the same course. We also experimented with increasing the size of the exam window (i.e., the range of dates that students can take an exam) for asynchronous, computer-based exams (similar to how students take tests like the SAT or GRE). In our experience, asynchronous, computer-based testing increases the long-term adoptability, scalability, and sustainability of second-chance testing. We explore the following research questions.

1) How does grading policy affect students' decisions to take second-chance exams?

2) How does grading policy affect students' exam preparation?

3) How does the change in grading policy affect students' performance and learning in the class?

4) How does the size of an asynchronous exam window affect when students opt to take exams?

2 BACKGROUND

Laboratory studies have documented a variety of ways by which exams and testing can be used to improve learning. Learning and retention of knowledge can be enhanced through retrieval practice that incorporates feedback, also known as the testing effect [18, 29]. The testing effect has been shown to be superior to other study strategies that students frequently employ such as re-reading course materials [28]. Learning can also be improved by increasing the use of formative assessment [9]. Unfortunately, one-shot exams generally do not serve as formative but rather only summative assessments. Finally, cramming or massed practice all at one time is less efficient for learning than distributing the same amount of studying over many smaller study sessions [4, 28]. Together, these results suggest that courses should focus on using many small assessments, distributed over a semester, with structures that encourage students to re-study and re-learn course material. Efforts to translate these laboratory studies into the classroom, however, are sparse [10, 24, 25].

Prior studies have generally documented that students generally improve their scores on a second-chance exam [16, 17, 30, 34]. Second-chance testing has even been shown to reduce DFW rates for courses. One critical challenge for second-chance testing is

Table 1: Example exam grading policies for second-chance exams

Policy	Description	Alignment with theory
Full Replacement (not viable)	If student takes second-chance exam, second-chance exam grade completely replaces the first-chance exam grade.	If students can demonstrate mastery, it shouldn't matter how long it took them.
Full Replacement with Grade Cap	Same as full replacement except student grade on second-chance exam is capped below 100% (e.g., 90%).	Same as above, except it incentivizes students on the first-chance exam.
Weighted Average	If student takes second-chance exam, first-chance and second-chance exams are averaged together using a pre-determined weighting scheme (e.g., 40% first-chance plus 60% second-chance [22].	Incentivize preparation for all exams.
Max Weighted Average	Same as weighted average, except weights are determined by the best and worst scores a student receives. (e.g., 10% worst exam score plus 90% best exam score).	Incentivizes students to do well on all exams while rewarding highest level of mastery.
Weighted Average with Insurance	Same as weighted average, except the final score is floored so that it can't be lower than the first-chance score.	Same as weighted average, while reducing stress in the second-chance exam.

motivating students to try their hardest on the first-chance exam. It was previously observed that fully replacing a first-chance exam score with the second-chance exam score consistently led a fraction of students to skip the first-chance exam entirely [16]. Instructors have employed a number of strategies to encourage students to take the first-chance exam seriously [16] (See Table 1 for a list of documented grading policies). In addition to the policies below, some instructors also require students to complete an additional homework assignment to prove that they studied before they would be allowed to take the second-chance exam [22].

Recent work collected data from several courses that used second-chance testing showed that courses that used a weighted average with insurance had more students taking second-chance exams, including large numbers of students who had earned A's or even perfect scores on the exam [16]. In contrast, full replacement with grade cap policies primarily encouraged lower performing students (C, D or F) to take the second-chance exam.

Literature on asynchronous exams where students can choose when to take their exam has reported that a large fraction of the students elect to take the exam at the end of the exam window [5, 6]. Furthermore, when students are allowed to pick their exam times, exam scores generally decrease throughout the exam period [3, 6]. A recent study suggests this decrease is largely due to weaker students, on average, taking the exam later in the exam period than stronger students [8]. When students take exams asynchronously, but are assigned their exam times, previous work has found that exam scores are stable within the exam period [19].

One potential motivation for weaker students to take exams later in the exam period is to cheat. One study of pencil-and-paper asynchronous exams where students selected their exam times found that exam items appeared easier in a Rasch model analysis after their first use on an exam [31]. Question randomization has been found to be an effective strategy for mitigating the potential for collaborative cheating (i.e., early exam takers passing information to late exam takers) on asynchronous exams [7].

3 COURSE CONTEXT

We performed a quasi-experimental study to determine the effect of grading policy on student test-taking behaviors and course performance. In this section, we describe the course and its policies in greater detail.

At the University of Illinois at Urbana-Champaign, Computer Architecture is a large-enrollment (300-400 students per term) required course for Computer Science majors offered every term. For the duration of the study, the course was taught using flipped lectures three days per week and a discussion/laboratory section once per week. Students were given a short, online homework assignment to complete before each lecture and completed weekly machine problem assignments (i.e., long homework coding problems).

Before lecture, students are provided with video lectures that introduce the course content so that lecture class time can be spent focusing on engaging students with multiple-choice questions (MCQs) to keep students engaged and to address common misconceptions. Students used clickers to respond to MCQs. Students earned attendance credit if they answered half of the MCQs during class.

The discussion/laboratory sections were taught by graduate teaching assistants and undergraduate course assistants. Students are encouraged to work in teams during these sections. These sections guided students through problem solving exercises to prepare them for each week's machine problem assignment. At the end of each discussion section, students are given a short quiz based on the content of the discussion section activity, to maintain individual accountability. The course is generally taken after students have taken at least two programming courses and a discrete mathematics course. The course begins by reviewing Boolean and propositional

logic before teaching students about datapaths, finite state machines, assembly language, pipelining, Caches, and basic concepts of parallel computing.

3.1 Online Learning System

Homework and exams were delivered using PrairieLearn [32, 33]. PrairieLearn is designed with the primary goal of providing students with an environment in which they can practice their knowledge of course content as much as possible [33]. Consequently, rather than writing individual homework problems, instructors are encouraged to write homework problem *generators*. In these generators, the instructor defines the parameters of interest to the problem and the bounds for those parameters. For example, rather than write a single question asking how many bits are needed to encode 50 pieces of information, the instructor would write a generator that asks how many bits are needed to encode N pieces of information and set N to randomly choose an integer between 33 and 255. By using randomization, students can generate practice problems on demand until they robustly learn the $\lceil log_2 N \rceil$ solution for this problem generator. These problem generators have access to the full capabilities of HTML5, permitting interactivity and allowing for the creation of questions in any format, including multiple-choice, free-response text submissions, and mathematical equations.

PrairieLearn also supports autograding of code submissions. For coding questions, students would receive a prompt, develop and test their code locally, and then upload their completed code to be graded. Coding questions are graded using a suite of test cases and points are awarded for the number of test cases that a student passed.

After each lecture, students were assigned to solve problems from approximately four problem generators. These assignments were intended to take no more than one hour for even a struggling student.

Most exams in the course were also administered using PrairieLearn. For exams, we created pools of question generators that covered the same learning objective. When a student took an exam, a set of question generators were randomly selected from the pool and random parameters were selected for that question generator. Consequently, every student would take a different exam but would be tested on the same learning objectives. Question generators provide a mechanism for exam security for asynchronous exams and a means by which students can generate as many practice problems as they want for studying.

3.2 Computer-based Exams

Students took seven midterm examinations and one final examination. All exams, except midterm 6, are taken in our campus's Computer-Based Testing Facility (CBTF) [35, 36]. The CBTF provides a secure, proctored testing environment that restricts students' internet access to only approved resources such as PrairieLearn [37].

For each exam, students are given a window during which they can schedule an exam at a time that is convenient for them. Consequently, all exams are asynchronous, necessitating the randomization features of PrairieLearn to ensure a level playing field for students regardless of the day that they take the exam. Because most exam questions were drawn from the same problem generators as the homework assignments, students perceive the exams to be fair even though they receive different exam variants. Exam windows are typically open for about three days before an exam closes.

Exams 2, 3, and 5 were strictly coding exams, presenting students with only a single exam prompt and giving students 50 minutes to write code that satisfied the prompt. For coding exams, students were allowed to use text editors, compilers, and debuggers to help them. Students were also given a small suite of example test cases to guide them in their development. Students were permitted to submit solutions multiple times. Some students failed to properly develop code for these exams, earning 0 points when graded solely by the autograder. Exams 1, 4, 7, and the Final were mixed format exams, containing coding questions, short-answer questions, multiple-choice questions, and computational questions.

Starting in the fourth week, an exam closes each week of the semester (See Table 2). Every other week, a required, first-chance exam on a new topic closes. In between those required exams, an optional, a second-chance exam covering the same topic as the previous week closed.

Table 2: Week each exam closed. Exams are numbered such that the first number indicates the exam number and the second number indicates whether the exam is a first- or second-chance exam (e.g., Exam 4.2 is Exam 4, second-chance). Exams with * are required exams.

Week	Exam	Content
4	1.1*	Boolean Logic Circuits
5	1.2	Boolean Logic Circuits
6	2.1*	Finite State Machines
7	2.2	Finite State Machines
8	3.1*	CPU datapath
9	3.2	CPU datapath
10	4.1*	Assembly programming (simple)
11	4.2	Assembly Programming (simple)
12	5.1*	Assembly programming (complex)
13	5.2	Assembly Programming (complex)
14	6.1*	Pipelines & Caches
15	7*	Pipelines & Caches
16	6.2	Pipelines & Caches
16	Final*	Comprehensive

3.3 Quasi-experimental conditions

For the three terms in this study, we tried to keep as many aspects of the course the same as possible, while primarily varying our exam policies. The same instructor taught all offerings of the course. The same topics were covered in the same order. Course policies for attendance were kept the same. Most homework problems were kept the same and most machine problems were kept the same. The computer-based exams for Exams 1, 2, 3, 4, 5, and 7 and the Final Exam used the same pool of questions.

To compare students' general level of preparation for the course, we used Exam 1 as a baseline measurement of students' incoming

ability levels (Table 3). A one-way ANOVA revealed no significant differences in students' performance on Exam 1 across terms [F = 2.59, p = 0.076]. Because no significant differences were found for any exam between FA18 and SP19 and these two terms had the same exam grading policies, we aggregate these two terms for all remaining analyses.

Table 3: Comparison of students' raw performance on Exam 1 as a baseline measure comparison of students' preparation for the course.

Term	N	Exam 1 $\mu(\sigma)$	Policy
FA18	356	82.85 (21.67)	90-cap
SP19	261	86.40 (20.06)	90-cap
FA19	353	85.63 (20.78)	90-10

Each term, 3 machine problems were modified to better align with the end-of-term design competition, which also changed each term. The format of these machine problems and their learning objectives were the same though: translate given C code into assembly code.

Most changes in the course focused on improving how we taught students about Caches. We provided students with more problem generators to study Caches. In FA19, we also converted most of the paper-based exam 6 into a computer-based exam using those new problem generators. Consequently, we exclude Exam 6 from any subsequent analysis of student grades.

The primary change that we analyze in this paper is the change of course exam policies. We made two primary changes: 1) how students' performance on first- and second-chance exams were aggregated to compute their final grade on exams and 2) the size of the window that students were given to take their exams.

All terms had optional second-chance exams. FA18 and SP19 used a Full Replacement with Grade Cap policy (See Table 1) where the cap for the second-chance exam was set to 90%. This policy will henceforth be called the *90-cap* policy. For example, suppose students A and B both scored a 50% on their first-chance exam. If student A scored a 100% on their second-chance exam, their final exam grade would be 90% (i.e., min(90%, 100%)). If student B scored a 70% on their second-chance exam, their final grade would be 70% (i.e., min(90%, 70%)). All midterm computer-based exams had a 3-day window. The final computer-based exam had a 7-day window.

FA19 used a Max Weighted Average policy (Table 1) with weights of 90% of their best score and 10% of their worst score. This policy will henceforth be called the *90-10* policy. Using the above scenarios, student A would earn a final score of $0.9 \times 100 + 0.1 \times 50 = 95$ and student B would earn a final score of $0.9 \times 70 + 0.1 \times 50 = 68$.

Exam 1 had a 3-day window (allowing for a baseline comparison) but all other midterm computer-based exams had variable length exam windows. The window for each exam was opened as soon as the course content relevant to that exam was covered but the exams closed with the same schedule used for FA18 and SP19 so that students never had to complete more than one exam per week. Table 4 shows the exam window lengths for each midterm exam. The computer-based final exam had a 7-day window.

Table 4: Size of first- and second-chance exam window for each midterm exam during FA19

Exam	1	2	3	4	5	7
Window (in days)	3	10	10	17	24	5

4 Q1: STUDENT TEST-TAKING DECISIONS

Research Question 1: How does the exam grading policy influence students' decisions to take second-chance exams?

4.1 Q1 Methods

To answer this research question, we examined which students decided to take second-chance exams based on their first-chance exam score. We provide two visualizations of this data to illustrate which students take the second-chance exam and how they performed on their second-chance exam relative to their first-chance exam. We test whether any differences in students' decisions are statistically significant using a logistic regression analysis. In the logistic regression, we use the student's score on the first-chance exam as a feature and the student's decision to take the second-chance exam as the dependent outcome. For example, if the student's grade on the first-chance exam is "C", then feature C is marked as 1, and A, B, D and F are marked as 0. Similarly, the outcome is 1 if the student took the second chance exam and 0 otherwise. There is one record for each student and each first-chance exam, and we fit a separate logistic model for the 90-cap and 90-10 policy exams. The logistic model is

$$\text{took-second-chance} \sim \text{logit}(A + B + C + D + F) \quad (1)$$

The logistic regression was fit using maximum likelihood with the `statsmodels` Python library. We compare the model coefficients to determine whether the change in grading policy affected students' decision to take second-chance exams.

Students were also asked ten questions using a 4-point Likert scale, to rate how likely they would be to take a second-chance exam for each letter grade for each grading policy. For example, for the 90-cap policy and the grade letter F, students were asked to respond to the following item, "If we used a policy of fully replacing your first-chance exam score with your second-chance exam score but your second-chance exam score would be capped at 90%, please indicate how likely you would be to take the second-chance exam if you earned a grade of F on the first-chance exam under this policy." Students rated their likelihood on the four-point scale: not at all likely, somewhat likely, most likely, definitely. To directly compare the results of this survey with students' observed behaviors, we re-categorized "definitely" responses as 1 and 0 otherwise, because definitely suggested the strongest inclination to take the second-chance exam, thereby providing a conservative estimate of students' actual behavior. We performed the same logistic regression comparison to compare students' surveyed responses as we did with students' actual behavior.

4.2 Q1 Results

In Figure 2, we plot the distribution of exam scores for students who took both the first- and second-chance exams. The size of each dot is proportional to the fraction of students who mapped to each

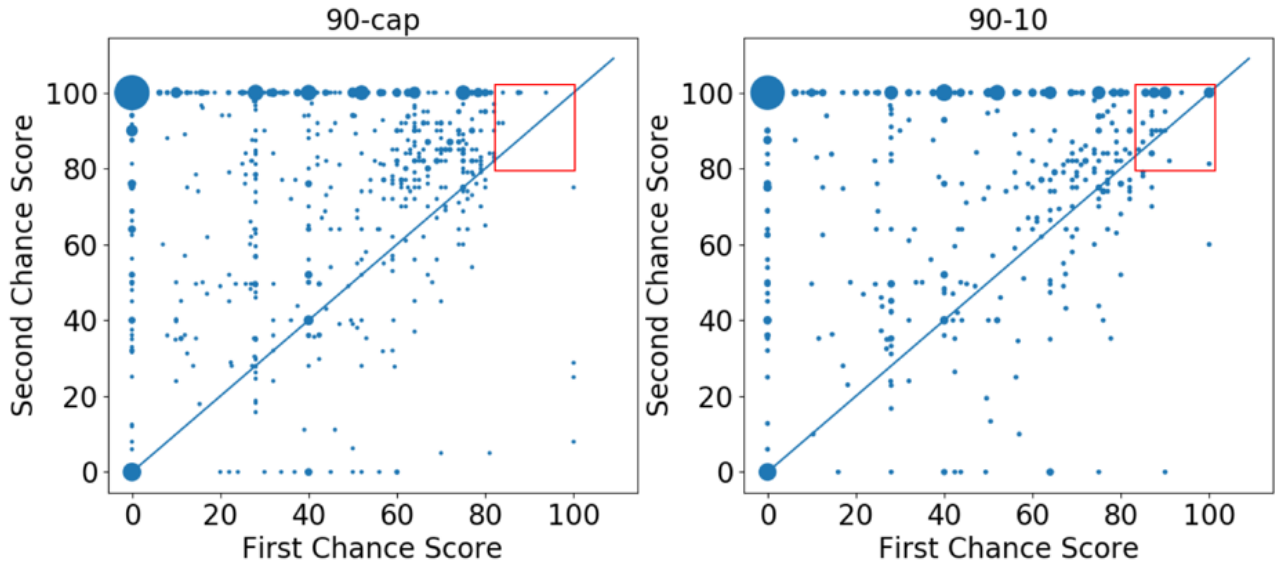

Figure 2: Distribution of exam scores for students who took both a first- and second-chance exam. Left: Students with the 90 Cap policy. Right: Students with the 90-10 policy.

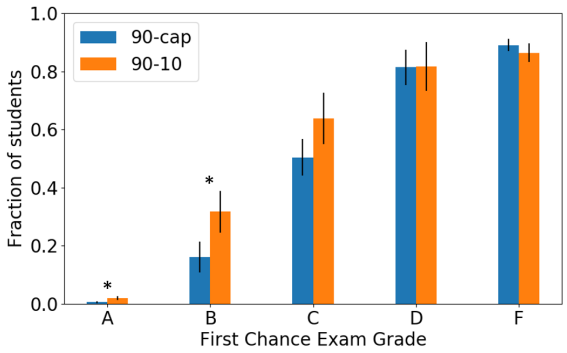

Figure 3: Fraction of students electing to take the second-chance exam, grouped by their first-chance exam grades. * indicates statistically significant (p < 0.05) differences.

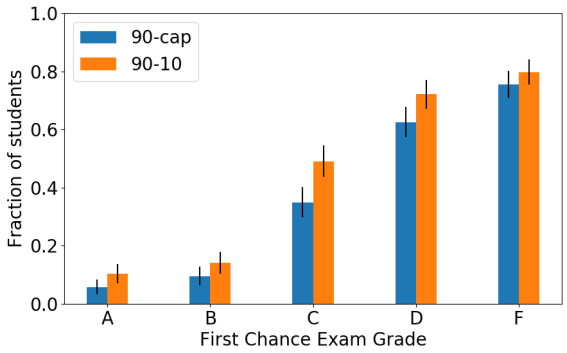

Figure 4: Students' surveyed decisions about whether they would consider taking the second-chance exam under each policy, provided they earned each grade on the exam.

dot. In Figure 3, we plot the percentage of students who elected to take a second-chance exam as a function of their letter grade on the first-chance exam (i.e., > 90% = A, 80–90% = B, etc).

Figure 2 reveals that the majority of students perform better on their second-chance exam than their first-chance exam, replicating prior findings [17, 22]. Figures 2 (denser dots in the top right) and 3 both suggest that the change in grading policy had little effect on students who earned a D or F on the first-chance exam, but that the 90-10 policy may have encouraged more A, B and C students to take the second-chance exam. The percentage of A students taking the second-chance exam rose from 0.6% to 2.0%, the percentage of B students taking the second-chance exam rose from 16% to 32% and the percentage of C students taking the second-chance exam rose from 50% to 64%.

Results from the logistic regression (See Table 5) reveal that students who score A or B grades are significantly more likely ($p < 0.001$ in both cases) to take the second-chance exam when using the 90-10 policy. Students who score C, D, or F grades do not have significantly different likelihoods of taking the second-chance exam under 90-cap and 90-10 policies. Figure 2 suggests that this is because students receiving lower grades on the first-chance exam are likely to take the second-chance exam under either policy, whereas the policy change has a large impact on the decisions of students with scores closer to 90% (A or B grades).

Results from a survey of students' hypothetical decisions (Figure 4), revealed no statistically significant differences between the two policies per grade letter but did reveal a significant difference when all responses were pooled ($p < 0.001$). The trends of the survey data align with students' actual behaviors. Under both policies, few students reported that they would definitely take a second-chance exam if they earned an A, but the percentage of students who said they would definitely take the second-chance exam increased as earned grades decreased. The actual percentage of B, C, and D students who took second-chance exams is higher than the conservative estimate of our survey analysis, suggesting that students' perception of their ability to improve on a second-chance exam or their willingness to take on risk in reality may be higher than their perceptions in theory.

Table 5: Logistic regression coefficients for model (1) for the 90-cap policy ($N = 3260$) and the 90-10 policy ($N = 2184$), and p-values for the coefficients differing by policy.

Coeff	90-cap (SE)	90-10 (SE)	p
A	−4.44 (0.25)	−3.15 (0.16)	$< 0.001^*$
B	−0.89 (0.18)	0.14 (0.16)	$< 0.001^*$
C	0.87 (0.13)	1.18 (0.21)	0.18
D	2.50 (0.25)	2.20 (0.28)	0.43
E	2.96 (0.13)	2.97 (0.17)	0.96

5 Q2: EXAM PREPARATION

Research Question 2: How does the exam grading policy influence students' exam preparation?

5.1 Q2 Methods

For each exam, we offered practice exams that drew most questions from the same pools of question generators as the actual exams and tested the same learning objectives. Practice exams could be taken as many times as a student felt they needed to prepare for an exam. PrairieLearn records each practice exam submission for every student. Consequently, these submission records can serve as a proxy for students' overall exam preparation. Figure 5 shows the average number of submissions each student made per day for Exam 1.

We binned all student submissions to estimate how much time students spent studying for each exam. All submissions during the first-chance exam window and four days before the exam window were binned as first-chance exam studying. All submissions after the close of the first-chance exam window and before the close of the second-chance exam window were binned as second-chance exam studying. We compared students' study habits using a one-way ANOVA for each exam using $\alpha = 0.05$ for significance testing.

5.2 Q2 Results

Across all exams, students in the 90-cap policy submitted an average of 2.34 practice exams ($\sigma = 2.01$) for first-chance studying and an average of 0.56 practice exams ($\sigma = 1.08$) for second-chance studying. Students in the 90-10 policy submitted an average of 2.13 practice exams for first-chance studying ($\sigma = 1.70$) and an average

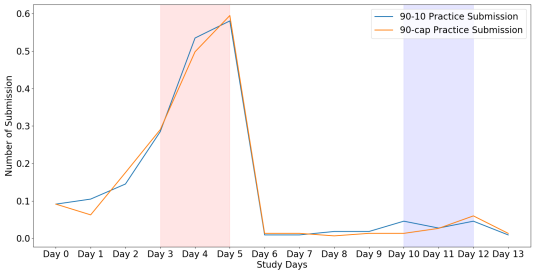

Figure 5: Average number of student submissions per day on practice exams and homework assignments as a proxy for overall student studying for exams. Red shadow area is the first-chance exam window and blue shadow is the second-chance exam window

of 0.32 practice exams ($\sigma = 0.60$) for second-chance studying. A one-way ANOVA ($F(3, 1978) = 240.04$, $p < 0.001$) revealed a significant difference in the amount of students' studying for exams. Tukey HSD post-hoc tests show significant differences between first- and second-chance exams but not across testing policies. For their second-chance studying, students studied only an additional 15–20% beyond their initial study efforts for the first-chance exam. Students studied about the same for the first-chance exam across both policies ($p = 0.14$) and about the same for the second-chance exam across both policies ($p = 0.08$).

6 Q3: STUDENT PERFORMANCE AND LEARNING

Research Question 3: How does the change in exam grading policy affect students' performance and learning in the class?

6.1 Q3 Methods

To measure the effect of the new testing regime on student performance, we compared student exam performance between the course offerings. For each exam in which a second-chance was offered (Exams 1–5), we quantified performance by taking the maximum of the raw first-chance exam score and second-chance exam score. For each exam in which a second-chance was not offered (Exam 7 and the Final Exam), we quantified performance by taking the raw first-chance exam score. We quantified performance in this way to eliminate effects caused by differing incentive structures. In particular, students have a stronger incentive to achieve higher scores on both exams with the 90-10 policy than with the 90-cap policy, whereas they have an equal incentive to achieve the highest maximum score with both policies. The way in which we quantified performance also mitigates effects caused by students' self-selection bias in taking second-chance exams and eliminates any skew in grades caused by the grading scheme.

We also collected DFW rates from the course as a whole because one core goal of using second-chance testing is to reduce the number of students failing courses for lack of opportunities.

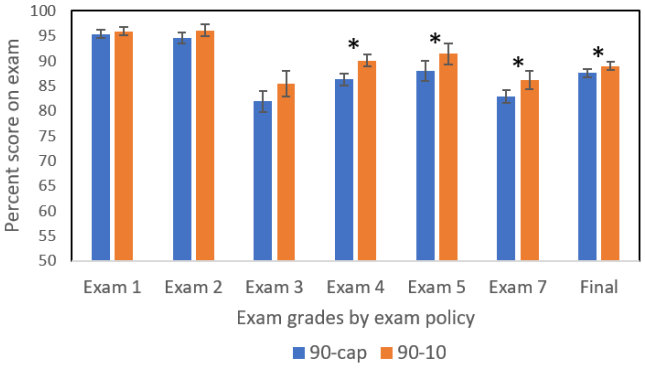

Figure 6: Student performance on exams across course policies. Significant differences marked with an *.

Grades and DFW rates were compared using unequal variances t-tests with $\alpha = 0.05$ as the threshold for significance. Effect sizes were measured with Cohen's d.

6.2 Q3 results

Table 6 reveals no significant differences in student performance on Exams 1-3, but reveals significant differences for all remaining exams with small effect sizes. These results suggest that students under both treatments began with similar baseline preparation but students in the 90-10 condition learned more as the course progressed. To make sure that this result was not simply caused by more students taking the second-chance exam and improving their score, we repeated the analysis using only students' first-chance exam scores and found the same results, except that the difference between exam 3 scores was now significant, favoring the 90-10 policy.

The 90-cap policy had a DFW rate of 11.9%. The 90-10 policy had a DFW rate of 10.1%. This difference was not significant ($p = 0.39$).

Table 6: Student performance on computer-based exams across course policies, as measured by the maximum raw score. Significant differences marked with an *. $N = 617$ for the 90 Cap policy. $N = 353$ for the 90-10 policy.

Exam	90 Cap $\mu(\sigma)$	90-10 $\mu(\sigma)$	t-test p	Cohen's d
1	95.4 (9.8)	95.9 (7.9)	0.397	0.05
2	94.6 (14.6)	96.1 (11.4)	0.067	0.11
3	81.9 (27.5)	85.2 (24.5)	0.051	0.13
4	86.3 (14.5)	90.1 (11.8)	$< 0.001^*$	0.28
5	88.0 (24.6)	91.4 (20.8)	0.023^*	0.15
7	82.9 (16.0)	86.2 (16.9)	0.003^*	0.20
Final	87.6 (9.9)	89.0 (8.6)	0.018^*	0.15

7 Q4: EXAM WINDOW SIZE

Research Question 4: How does the size of an asynchronous exam window affect when students opt to take exams?

7.1 Q4 Methods

Spaced repetition of studying is an effective and efficient method for improving student learning. Consequently, we originally forcibly spaced first- and second-chance exams to encourage spaced repetition by providing only 3-day exam windows each week. However, this restriction had an unintended consequence of creating too much space between when students learned course content and when they were tested on it. Our exam schedule has an exam (and its second chance) every two weeks, but some topics took less than two weeks to cover. In Fall 2019, we addressed this by keeping the exam closing dates the same but opening the exams as soon as the necessary material had been covered. By increasing the size of exam windows in this way, the exam window for the first- and second-chance exams could now overlap, potentially jeopardizing the goal of spacing students' studying.

To examine whether the increased exam window size jeopardized our goals of spaced repetition, we used students' exam submission data for Exams 1–5 (exams with second chances) to calculate how many days students took between taking their first- and second-chance exams for each exam window size. To better understand students' decision making under increasingly large window sizes, we also sought to understand whether students might have a preferred day to take exams (e.g., the last day of the exam window or Mondays). We explore this preference using two metrics: 1) the number of days before the exam window closed that students took the exam (henceforth *preferred day*) and 2) the number of unique days of the week that students took an exam. For the first metric, we calculated the average number of students who took an exam each day relative to the close of the exam window. To illustrate the second metric, if a student took all their exams on Friday, they had one *preferred day*. Likewise, a student who took two exams on Monday and three on Tuesday and another student who took four exams on Thursday and one on Wednesday would both have two *preferred days*. We calculated the percentage of students with each number of *preferred days*.

We also calculated linear regression coefficients to examine whether weaker students prefer to take exams later in the exam window as shown in previous studies [3, 8].

7.2 Q4 Results

Regardless of the exam window size, about 50% of students took the first- and second-chance offerings of the same exam exactly seven days apart. Additionally, about 97% of students had at least one day to study between the first- and second-chance offerings of the same exam. Few students took more than 14 days between their first- and second-chance exams when that was an option.

Figure 7 shows that most students took their exam on the closing day of the exam window. Figure 8 shows that most students (67%) take all of their exams on the same day of the week. Together these findings show that most students prefer to take their exams on Friday and that the 7-day spacing between first- and second-chance exams is a consequence of the one-week spacing of exam windows closing.

Figure 7 shows an additional unexpected pattern where it appears that there is a substantial increase in the number of students taking exams at the end of each exam week, creating a small ripple effect.

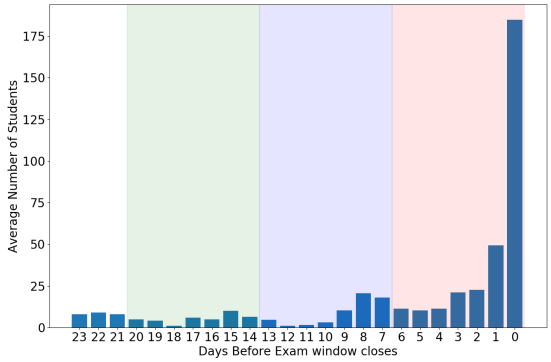

Figure 7: Average number of students who took an exam each day as a function of the number of days remaining in the exam window. Each panel is a week, ending on Friday. Most students took the exam on the last day of the exam window but there are also smaller spikes in tests taken at the end of each week.

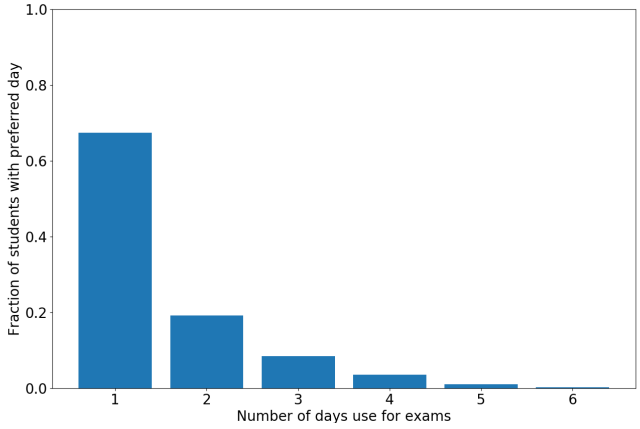

Figure 8: Stability in student exam day-of-week preferences. Over 60% of students took all of their exams on the same day of the week. Another 20% took all of their exams on only two days of the week.

This ripple and the stability of students' preferred day suggests that students may have also have a preferred day of the week to take an asynchronous exam for reasons separate from the closing day of the exam (e.g., course schedule).

Table 7 confirms that student performance generally decreased as the exam window progressed. Multiplying these slopes by the length of the exam window shows that student performance was fairly consistently a little more than one letter grade worse from the first day of the exam window to the closing day of the window.

Table 7: Slopes from linear regression comparing days before the exam window closed and students' average percentage score each day

	Exam1	Exam2	Exam3	Exam4	Exam5
Slope	−4.3	−1.3	−2.8	−0.8	−0.6

8 DISCUSSION

The change in exam grading policies primarily affected whether marginally competent (A or B) students decided to take second-chance exams. This difference makes intuitive sense when comparing the potential benefits and risks for students who decide to take the second-chance exam. Under the 90-cap policy, a student who earns an 85% grade can at most improve their grade by 5 percentage points (half a letter grade) but could potentially lose 85 percentage points (many letter grades). Under the 90-10 policy, the same student could improve their grade by 9.4 percentage points ($0.9 \times 100 + 0.1 \times 85 = 94.4$) and could lose at most 8.5 percentage points, both approximately one letter grade. The relatively equal amounts of risk and benefit for these students likely encouraged more of them to take the second-chance exam. Surprisingly, even a few students who earned close to 100% on their first-chance exam were willing to risk taking the second-chance exam.

Herman et al. found that the Weighted Average with Insurance led to many more students taking second-chance exams and many more students performing considerably worse on the second-chance exam [16]. One possible explanation for this effect is that the insurance policy may encourage students to take the second-chance exam as a roll of the dice, seeing if they might get lucky and improve their grade. We did not see a substantial increase in the number of students who performed worse on the second-chance exam using the 90-10 Max Weighted Average policy relative to the 90-cap policy. This lack of change suggests that although the risks of performing worse on the second-chance exam were not as dire for the 90-10 policy as for the 90-cap policy, the risk was still substantive enough that students did not waste their time and course resources taking second-chance exams that they did not prepare for. Consequently, it seems that the 90-10 Max Weighted Average is just as effective at encouraging students to seriously try during both of their attempts at the exam as the 90-cap Full Replacement policy, while also encouraging more students to study and learn more.

The change in exam grading policies had significant but small effects on students' learning and performance in the course. Although students began the semester with roughly the same amount of knowledge about course material, students in the 90-10 condition performed better as the semester continued. In contrast, the percentage of students failing or withdrawing from the course did not significantly improve. One possible explanation for these two changes is that marginally competent students had a greater incentive to develop a better mastery of early course content, better preparing them to learn later course content, creating the small but significant effect on students' performance on first-chance exams as a whole.

Exam 7 revealed a slightly larger effect size for students' performance on exams. This difference may be an outlier because

we extensively revised how we taught the topics of pipelines and Caches, potentially explaining some of this effect.

The change in policy did not have a significant effect on students' studying as measured by practice exam submissions. Consequently, increased time on task does not explain the change in students' grades well. We argue that the change in students' grades was likely caused mostly by the increased number of marginally competent students who took second-chance exams and further solidified their mastery of early content. This increased mastery better prepared these students for future learning. The change in policy did not appear to have an effect on students who lacked a general mastery of the course content (no change in DFW rate). These students were likely already taking second-chance exams under the old policy and thus were not improving their demonstrated mastery as a population.

8.1 Future work on grade replacement policies

Future studies could further explore the trade-offs between Max Weighted Average and other Weighted Average policies on students' test-taking behaviors and performance. Alternatively, other studies could examine whether less generous thresholds such as an 80% cap for full replacement policies or an 80-20 Max Weighted Average would have substantially different effects. The findings from this study and previous studies [16] suggest that students may indeed make their decisions to take second-chance exams primarily based on the balance of risks and rewards to their final grades. Additional quantitative studies could further explore the trade-offs between the various Weighted Average grading schemes reported in the literature. Future qualitative studies could seek to better understand the risk tolerance and decision-making that students use.

Additionally, in future work we are interested in studying the effects of second-chance testing on students' stress and anxiety and overall workloads in courses. For example, while it was not part of our research questions, we did ask students to provide their perceptions of how the different course policies affected their stress related to exams and which exam policies they generally preferred. Students expressed a moderately strong preference for the 90-10 policy overall and expressed that the 90-10 policy reduced their stress during the first-chance exam (Figure 9). The different policies had negligible effects on students' experiences of stress on the second-chance exam. Future studies could explore these effects more robustly, potentially connecting them to other sources of stress such as overall workload, stereotype threat, or identity beliefs.

8.2 Effect of exam window size

The closing date of the exam window appears to be the primary driver of when students take their exams. Because weaker students tended to take their exams later in the exam window [3, 8], the students who most needed to take the second-chance exam likely took most of their exams on the closing dates of each exam, creating the high rate of 7-day exam spacing. Unexpectedly, even students who took exams a week or more early tended to take their exams during the three days (Wednesday through Friday) that aligned with the three-day window of the first exam. It's unclear why this trend emerged but future studies could further explore whether this trend

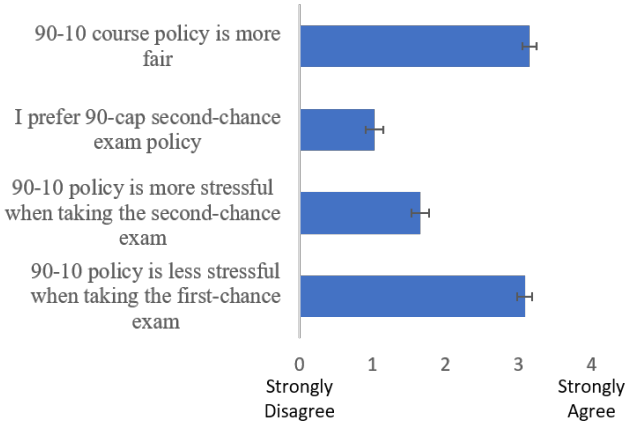

Figure 9: Students' expressed preferences on each grading policy.

is the result of a primacy effect (the first exam primed students to think that exams for the course should be taken Wednesday through Friday) or some other reason such as convenience. Future focus groups or interviews with students could help explore this dynamic.

9 CONCLUSION

This study suggests that the 90-10 Max Weighted Average exam grading policy does not harm any population of students relative to the 90-cap Full Replacement exam grading policy but may provide additional benefits for marginally competent students who could further solidify their understanding of course material if given a structure to study more. The change in policy unfortunately does not appear to provide any additional benefit for the weakest students. While the effect size of these changes is small, they are comparable to learning gains from more labor-intensive interventions and the level of effort needed to implement the better policy is zero. For instructors who are committed to second-chance exams and wish to encourage their students to focus and study for first- and second-chance exams while not offering trivial second-chance exams, the Max Weighted Average policy appears to be unilaterally better than the Full Replacement with Grade Cap policy. Future research on designing effective exam grading policies should focus on better understanding students' risk tolerance and how that affects their decision making and test anxiety.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. DUE 1915257. 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] Robert L. Bangert-Drowns, James A. Kulik, and Chen-Lin C. Kulik. 1991. Effects of Frequent Classroom Testing. *Journal of Educational Research* 85 (1991).
[2] B.S. Bloom. 1968. Learning for mastery. *Evaluation Comment* 1, 2 (1968), 1–12.

[3] E Robert Burns, Judith E Garrett, and Gwen V Childs. 2007. A study of student performance on self-scheduled, computer-based examinations in a medical histology course: is later better? *Medical Teacher* 29, 9-10 (2007), 990–992.

[4] S. K. Carpenter, N. J. Cepeda, D. Rohrer, S. H. K. Kang, and H. Pashler. 2012. Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review* 24 (2012), 369–378.

[5] Jacabo Carrasquel. 1985. Competency testing in introductory computer science: the mastery examination at Carnegie-Mellon University. In *ACM SIGCSE Bulletin*, Vol. 17. ACM, 240.

[6] Binglin Chen, Matthew West, and Craig Zilles. 2017. Do performance trends suggest wide-spread collaborative cheating on asynchronous exams?. In *Proceedings of the Fourth Annual ACM Conference on Learning at Scale*. Association for Computing Machinery.

[7] Binglin Chen, Matthew West, and Craig Zilles. 2018. How Much Randomization is Needed to Deter Collaborative Cheating on Asynchronous Exams?. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale (L@S '18)*. Association for Computing Machinery, New York, NY, USA, Article Article 62, 10 pages. https://doi.org/10.1145/3231644.3231664

[8] Binglin Chen, Matthew West, and Craig Zilles. 2019. Analyzing the decline of student scores over time in self-scheduled asynchronous exams. *Journal of Engineering Education* 108, 4 (2019), 574–594. https://doi.org/10.1002/jee.20292 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jee.20292

[9] I. Clark. 2012. Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review* 24 (2012), 205–249. https://doi.org/10.1007/s10648-011-9191-6

[10] S. D. Downs. 2015. Testing in the college classroom: Do testing and feedback influence grades throughout an entire semester? *Scholarship of Teaching and Learning in Psychology* 1 (2015), 172–181. https://doi.org/10.1037/st10000025

[11] S. Freeman, S. L. Eddy, M. McDnough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America* 111 (2014), 8410–8415.

[12] G. Gibbs and C. Simpson. 2005. Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education* 1 (2005), 3–31.

[13] Brianne Gutmann, Gary Gladding, Morten Lundsgaard, and Timothy Stelzer. 2018. Mastery-style homework exercises in introductory physics courses: Implementation matters. *Phys. Rev. Phys. Educ. Res.* 14 (May 2018), 010128. Issue 1. https://doi.org/10.1103/PhysRevPhysEducRes.14.010128

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

[15] C. Henderson, A. Beach, and N. Finkelstein. 2011. Facilitating Change in Undergraduate STEM Instructional Practices: An Analytic Review of the Literature. *Journal of Research in Science Teaching* 48 (2011), 952–984.

[16] G. L. Herman, K. Varghese, and C. Zilles. 2019. Second-chance testing course policies and student behavior. In *Proceedings of the 49th ASEE/IEEE Frontiers in Education Conference*. Cincinnati, OH, 7.

[17] Sandra M. Juhler, Janice F. Rech, Steven G. From, and Monica M. Brogan. 1998. The Effect of Optional Retesting on College Students' Achievement in an Individualized Algebra Course. *The Journal of Experimental Education* 66, 2 (1998), 125–137. https://doi.org/10.1080/00220979809601399

[18] S. H. K. Kang, K. B. McDermott, and H. L. III Roediger. 2007. Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology* 19 (2007), 528–558.

[19] C Kreiter, MW Peterson, K Ferguson, and S Elliott. 2003. The effects of testing in shifts on a clinical in-course computerized exam. *Medical Education* 37, 3 (2003), 202–204.

[20] Chen-Lin C. Kulik and James A. Kulik. 1987. Mastery Testing and Student Learning: A Meta-Analysis. *Journal of Educational Technology Systems* 15, 3 (1987), 325–345. https://doi.org/10.2190/FG7X-7Q9V-JX8M-RDJP arXiv:https://doi.org/10.2190/FG7X-7Q9V-JX8M-RDJP

[21] 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 (1990), 265.

[22] Richard Kunz, Monika Bubacz, and Jack Mahaney. 2011. Retests: A Rescue Plan for the Sophomore Slump. In *Proceedings of the 22011 ASEE Southeast Section Conference*.

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

[24] M. A. McDaniel, J. L. Anderson, M. H. Derbish, and N. Morrisette. 2007. Testing the testing effect in the classroom. *European Journal of Cognitive Psychology* 19 (2007), 494–513.

[25] M. A. McDaniel, R. C. Thomas, P. K. Agarwal, K. B. McDermott, and H. L. Roediger. 2013. Quizzing in middle school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology* 27 (2013), 360–372. https://doi.org/10.1002/acp.2914

[26] M. A. Pyc and K. A. Rawson. 2010. *Science* 330 (2010), 335.

[27] K. A. Rawson and J. Dunlosky. 2012. When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review* 24 (2012), 419–435. https://doi.org/10.1007/s10648-012-9203-1

[28] K. A. Rawson, J. Dunlosky, and S. M. Sciartelli. 2013. The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review* 25 (2013), 523–548. https://doi.org/10.1007/s10648-013-9240-4

[29] H. L. Roediger, III and A. C. Butler. 2011. The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences* 15 (2011), 20–27.

[30] M. J. Roszkowski and S. Speat. 2016. Retaking the SAT may boost scores but this doesn't hurt validity. *Journal of the National College Testing Association* 2 (2016), 1–16.

[31] Michaela Wagner-Menghin, Ingrid Preusche, and Michael Schmidts. 2013. The effects of reusing written test items: A study using the Rasch model. *ISRN Education* 2013 (2013).

[32] M. West. [n.d.]. PrairieLearn. https://github.com/PrairieLearn/PrairieLearn.

[33] 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.

[34] Davidson William B, William J. House, and Thomas L. Boyd. 1984. A Test-Retest Policy for Introductory Psychology Courses. *Teaching of Psychology* 11, 3 (1984), 182–184.

[35] 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*.

[36] 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)*.

[37] 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.