

Second-chance Testing Course Policies and Student Behavior

Geoffrey Herman, Kavya Varghese and Craig Zilles

*Department of Computer Science
University of Illinois at Urbana-Champaign, USA*

[glherman, kmvargh2, zilles}@illinois.edu

Abstract—In this research category full paper, we present our findings on the effects of different course policies for second-chance testing on students’ studying and exam taking behavior. Second-chance testing, where students are allowed to take a second instance of an exam for some form of grade replacement, is a less expensive approximation of mastery-based learning that can be easily integrated into a broad range of college course structures. It encourages students to review course material after poor performance on an examination but limits the amount of resources instructors must invest in the creation of examinations or in grading them. There exists, however, a large space of potential course policies for integrating second-chance testing into a course and little prior research on how these policies affect student behavior.

This paper analyzes three different grading policies in use at Midwestern University. All of the policies attempt to encourage students to prepare adequately for the first-chance exam and review the material again before the second-chance exam, if they elect to take it. The first policy used a partial grade replacement policy with insurance: students’ grades could only improve by taking the second-chance exam but the first-chance exam always counts for at least one-third of a students’ grade on the examination. The second policy is identical, but required students to complete a zero-credit, online-homework assignment before being allowed to take the second-chance exam. The third policy implemented full grade replacement (even if the second score is lower) and capped the score that could be achieved on the second exam. By comparing these different course policies, we show that grading policies have a significant effect on whether students take second-chance exams.

We also performed a quasi-experimental study, adding second-chance exams to a course. We present data from students’ exam performance and from the course’s learning management system that suggest that adding a second-chance exam had no effect on student performance or study habits for the first-chance exam. However, the total amount of time that students studied did increase substantially as students who took the second-chance exam studied an additional 60% of their original effort.

Keywords—second-chance testing, assessment, STEM, higher education

I. INTRODUCTION

Efforts to improve STEM education have largely focused on eschewing the traditional lecture in favor of active learning [7], [11], but comparatively little attention has been paid to transforming the traditional assessment paradigm of “two midterms and a final” [16]. This lack of attention may, in part, be because exams are generally viewed as methods to measure learning rather than as a mechanism for learning in their own right [10].

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 [8]. Laboratory studies have robustly demonstrated that learning and retention of knowledge can be enhanced through retrieval practice that incorporates feedback [13], [22], increased use of formative assessment [5], and distributed practice [4], [21]. Efforts to translate these laboratory studies into the classroom, however, are sparse [1], [6], [17], [18].

One key impact of testing is that it primes students for future learning, providing critical metacognitive feedback to students about how well they have mastered material [19], [20]. The “one-shot” exams that are widely used in college education ignore this impact—they give students only one chance to demonstrate learning before the class moves inexorably forward, removing incentives to review material and making no use of the priming effect (see top of the illustrative Figure 1).

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 to the next [3], [14]. It has been shown consistently that mastery learning is more effective for learning than traditional instruction [9], [15]. In spite of its effectiveness, self-paced mastery learning is hard to 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 that is easier to integrate with a range of college course structures [2], [12]. 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 in grading them. Research on second-chance testing, while sparse, suggests that many students who retake an exam earn higher exam scores on the retake [12], [23], [24].

The past three years have seen rapid, sustained adoption of second-chance testing by instructors of ten large STEM courses—with a combined annual enrollment of over 6,000 students—at Midwestern University, in part due to an introduction of a computer-based testing facility [25], [26]. Second-

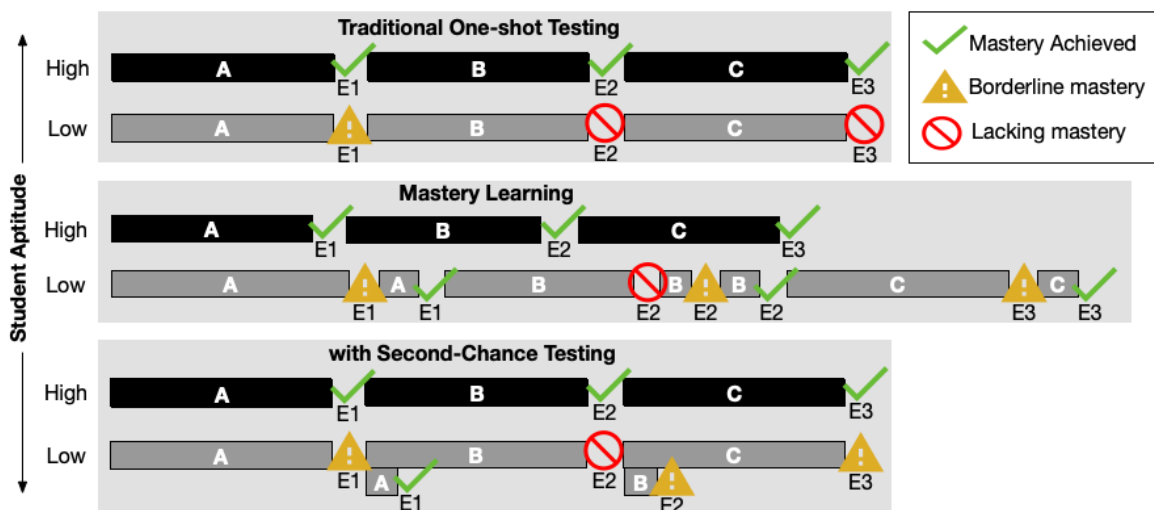


Fig. 1. An illustrative example comparing traditional (one-shot) exams, mastery learning, and second-chance testing with a hypothetical class with two mid-term exams (E1, E2) and a cumulative final (E3). Traditional one-shot testing works fine for students with high aptitude, but students with lower 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.

chance exams are typically offered one week after first-chance exams, giving students time to see their scores and to review material. This paper uses data collected in the courses to contribute to answering three research questions:

- 1) What impact does second-chance testing have on student learning and exam scores?
- 2) How does the choice of grade replacement policy influence which students elect to take second chance exams?
- 3) Does the introduction of second-chance testing influence students' preparation for the first-chance exam?

Our paper begins by discussing the importance and goals of a grade replacement policy and the diversity of policies used in courses studied in this paper (Section II). Then, we present our methods (Section III) and findings (Section IV) related to each of the above questions. We find that student learning improved in all courses (critically, that final exam failure rates significantly decreased), despite widely differing policies on how first-chance and second-chance scores were used to produce a final grade (Section IV-A). We show that these policies do significantly influence which students elect to take second chance exams (Section IV-B). Lastly, we provide initial data suggesting that use of an appropriate grade replacement policy prevents the introduction of second-chance testing from discouraging students from studying for the first-chance exam (Section IV-C). We conclude in Section V.

II. COURSES AND COURSE POLICIES

To be effective, an implementation of second-chance testing should be more than just "another roll of the dice" to see if a student can get a better score on a different assessment. Instead, it must encourage most students to engage in the study habits and test-taking behaviors that benefit them the most. Based on principles from test-potentiated learning and

spaced repetition, we want students to study and take each exam seriously and to re-study material with some spacing as needed to improve their mastery and retention of course material.

In addition, a grade replacement policy for second-chance testing needs to handle an inherent tension: we want students that will benefit from another exam to take it, but we want to discourage unnecessary test taking to manage both student and faculty/course staff workload. As a consequence, we have formulated the goals of a course policy for second-chance exams as follows:

- 1) As many students as possible demonstrate mastery of course material by the first-chance exam.
- 2) Students who demonstrate mastery of the course material do not return to take the second-chance exam.
- 3) Students who did not demonstrate mastery of the course material on the first-chance exam engage in spaced studying of course material after the first exam.
- 4) As many students as possible demonstrate mastery of course material by the second-chance exam.
- 5) Students see second-chance exams as a motivating opportunity to improve their mastery rather than a burdensome requirement or form of remediation.

The spread of the idea of second-chance testing throughout our institution has led to at least 10 courses permanently adopting the technique. Faculty members instructing each course have had the autonomy to implement second-chance testing as they see fit, including developing their own course policies for grade replacement. This autonomy has led faculty to explore many alternative grade replacement policies.

In a perfect world, we could offer second-chance exams with full grade replacement. In theory, if a student can demonstrate mastery on a second-chance exam, they should

TABLE I
LIST OF EXAMPLE COURSE 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 w/ 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.
Partial replacement	If student takes second-chance exam, student's final score is calculated based on both exam grades (e.g., one-third first exam plus two-thirds best exams <i>or</i> 10% worst exam score plus 90% best exam score).	Incentivizes students to do well on all exams while rewarding highest level of mastery.
Partial replacement w/ insurance	Same as above, except the final score is capped so that it can't be lower than the first-chance score.	Same as above, while reducing stress in the second-chance exam.
Partial replacement w/ insurance and extra homework	Same as above, except that students must complete an additional homework assignment to demonstrate that they have studied before being allowed to take the second-chance exam.	Same as above, but discourages students who have already demonstrated mastery from using more instructor time.

TABLE II
LIST OF STUDIED COURSES, THEIR TYPICAL ENROLLMENTS, AND THEIR SECOND-CHANCE POLICIES

Course Description	Level	Department	Typical Enrollment	Current course policy
Computer Organization	Soph.	Computer Science	~300	Full replacement w/ grade cap
Dynamics	Soph.	Mechanical Engineering	~400	Partial replacement w/ insurance
Solid Mechanics	Soph.	Mechanical Engineering	~300	Partial replacement w/ insurance
Intro. to Electronics (majors)	Fresh.	Electrical/Computer Engineering	~400	Partial w/ insurance + extra homework

receive full credit for that mastery, as, in general, we are less concerned with *when* students learn the material than *that* they do learn the material. We have, however, seen such a policy be universally problematic. Every time that an instructor has attempted such a policy, a non-trivial fraction of the students will forgo the opportunity to take the first chance exam or not sufficiently prepare for it, knowing that they have a chance for full grade replacement on the second chance. While this is likely rational time management for a few students, for most students it is better characterized as procrastination. This procrastination is understandable, but it weakens all of the benefits of offering second-chance exams (e.g., formative assessment, testing effect, test-potentiated learning, meta-cognition). As a result, instructors at our institution unanimously agree that the full grade replacement policy is not viable.

Instead, faculty have settled on a small collection of policies that seem to address the above goals, each emphasizing a different subset. A collection of these course policies are described in Table I, along with their justification.

III. METHODS

In this paper, we collected anonymized grade book information from four of the largest enrollment courses at Midwestern University: Computer Organization, Dynamics, Solid Mechanics, and Introduction to Electronics. For each course, we collected data about which students took each exam and their scores on first- and second-chance exams. Each of these course are required, gateway courses that students need to pass before progressing in their chosen majors. Additionally, later courses build on the knowledge that students should ideally learn in these courses.

We selected these courses because they are roughly equivalent in size (N = 300-400 students per semester) and have had stable course policies regarding second-chance testing for several semesters. These courses have also had stable exam design practices for several semesters (i.e., learning objectives assessed and exam modality have stayed constant), allowing for more even comparisons across semesters. These courses also provide a sort of natural experiment as the different courses implemented different course policies for how the second-chance test grades would be accounted. The courses, their average enrollment, and the policies they use are shown in Table II.

We used different statistical analysis procedures for each research question, so we describe our methods for each research question alongside the results for each research question.

IV. RESULTS

We present results relating to three aspects of second-chance testing. First, we present representative results demonstrating the impact of second-chance testing on student performance in our courses. These results are congruent with previous work on second-chance testing. Second, we present results on how the choice of course policy influences student behavior with respect to which students elect to take a second-chance exam. Third, we present results on an experiment to explore how the existence of a second-chance exam influences the study behavior of students.

A. Research Question 1: Improving Cognitive Outcomes

Across all of the offerings of second-chance testing at Midwestern University, instructors have always reported that the addition of second-chance exams led to better student

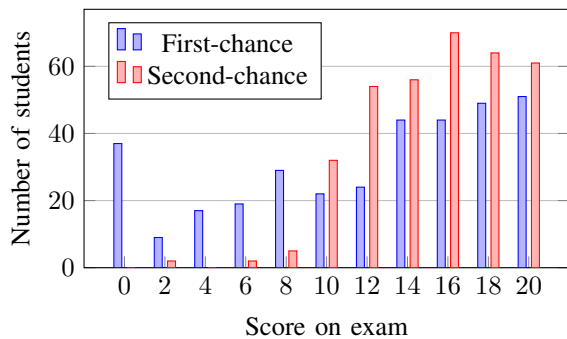


Fig. 2. Second-chance testing on a representative challenging learning objective from Computer Organization with $N = 345$ students. The percentage of students scoring below 50% on this objective fell from 32.2% of the class to just 2.6%.

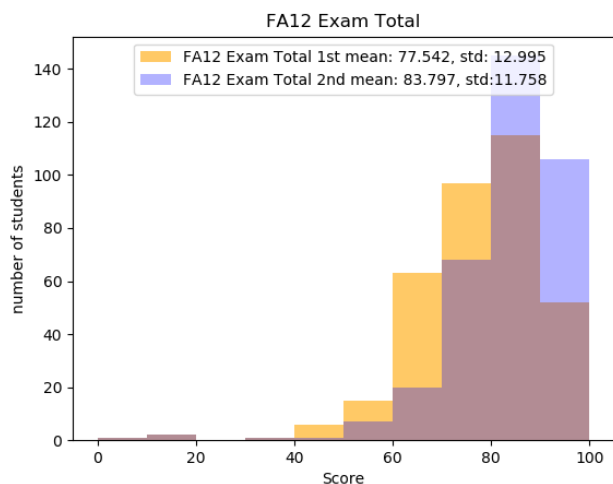


Fig. 3. Shift in the distribution of full-semester exam score totals in Computer Organization resulting from second-chance testing.

learning. These findings are consistent whether we look at aggregate performance as measured by student grades, reduced failing rates, or improved student learning measured per learning objective. In the case of courses that use a policy of partial replacement w/ insurance, increases in student grades are all but guaranteed as students cannot hurt their grade through a second-chance exam, so we will not present analysis of student grades for those courses. However, in a full replacement policy course, such as used in the sophomore-level computer organization course, improved student grades are not guaranteed. So we focused our initial analysis of student grade performance on this class.

To assess the affect of second-chance testing on students' grades in a full replacement policy, we compared student performance on both individual learning objectives and students' overall exam performance. Statistical reports from t-tests and Cohen's d effect sizes are reported.

In Figure 2, we show a representative distribution of scores for one particular learning objective tested on an exam. Students were given the opportunity to remediate this learning objective, being re-tested on a different problem based on the same concepts. With the addition of a second-chance exam,

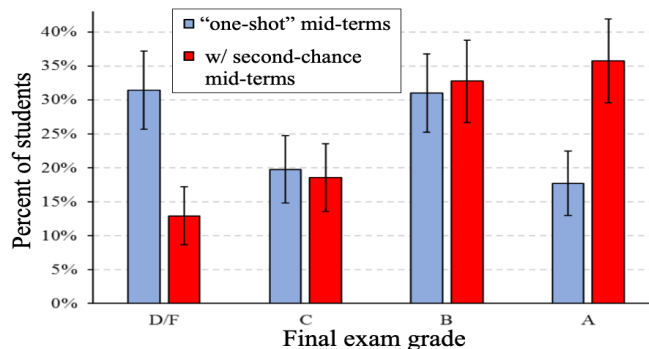


Fig. 4. Grades on an identical final exam in Solid Mechanics for two groups of students. High- and low-performing students had substantial gains from the introduction of more frequent exams with immediate feedback and second-chance exams relative to the group that had two "one-shot" mid-terms.

the mean student score on this learning objective increased statistically significantly ($p < 0.01$) with a large effect size (0.61). Additionally, the failure rate (students scoring below 60%) for this learning objective was more than halved from 36% to 11%. Notably, this improvement occurred even though students could score worse on the second-chance exam.

More generally from the same course, we see corresponding improvements in mean score and lower standard deviation in students' exam performance in general. Figure 3 shows the distribution of the students' exam total scores throughout the whole semester in the Computer Organization class. We have plotted what the students scores would be if we only counted the first-chance exams against the same class with the grade replacement from second-chance exams included. The inclusion of the second-chance exams statistically significantly ($p < 0.01$) improved the mean exam score by more than 8% (effect size: .48) and reduced the standard deviation of the score distribution by 10%.

To explore whether partial grade replacement policies similarly improved students' learning and to explore whether second-chance testing led to better retention of knowledge, we performed a quasi-experimental study in the sophomore-level introductory solid mechanics course, comparing students' performance on an identical "one-shot" final exam. The only differences between the offerings was the increased frequency of mid-term exams (5 1-hour exams vs. 2 2-hour exams) and the addition of second-chance mid-term exams. This analysis showed that the addition of second-chance exams statistically significantly ($p < 0.01$) improved students' performance on the final exam. As shown in Figure 4, these changes halved the number of D's and F's on the "one-shot" cumulative final exam and more than doubled the number of A's. While this experiment prevents us from discerning the individual contributions of more frequent testing vs. second-chance testing, both activities are motivated by test-potentiated learning and spaced repetition.

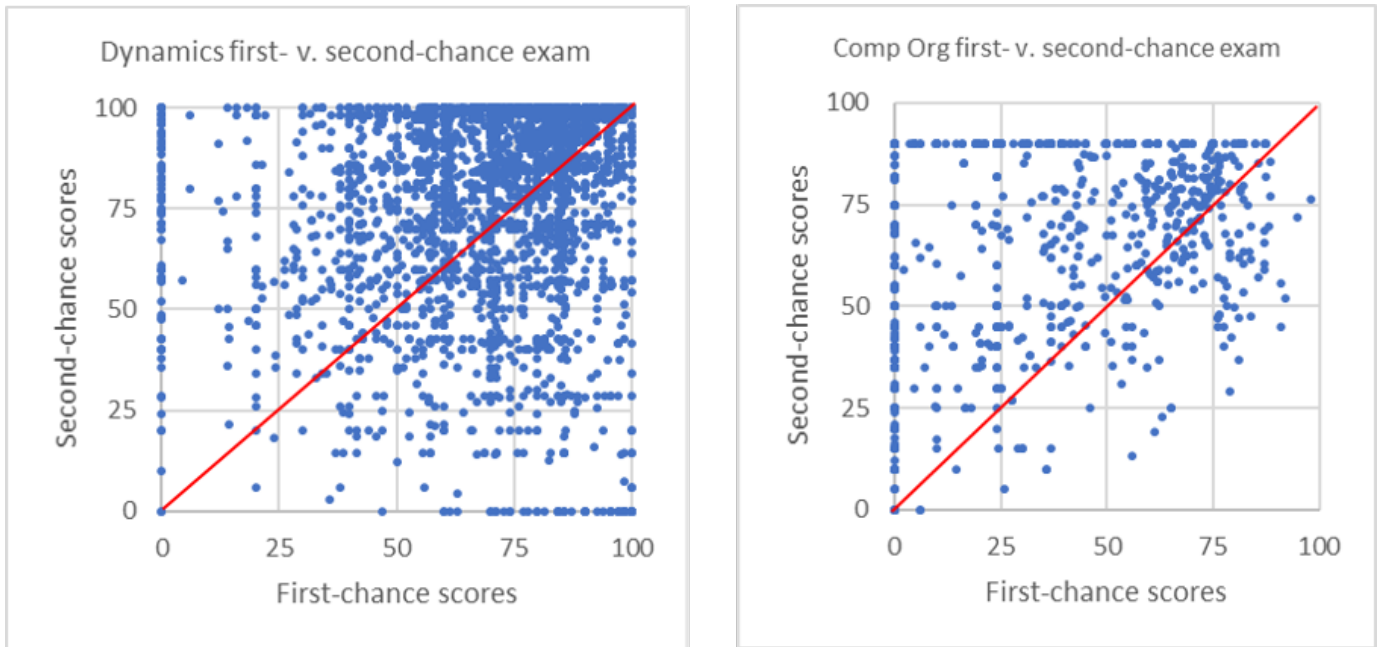


Fig. 5. Distribution of grades for students who took both a first and second chance exam. Left: Course (Dynamics) with partial grade replacement policy with insurance. Right: Course (Computer Organization) with full grade replacement policy with penalty.

B. Research Question 2: Course policies significantly influence who takes second-chance exams

In aggregate, we find that the choice of course policy can significantly affect how much of the class takes a second-chance exam. For example, Computer Organization had a lower average re-take rate (38% of students) than Dynamics (49% of students), in spite of the fact that average scores on first-chance Computer Organization exams were slightly lower (75% vs. 78%, significant at $p < 0.01$ with $N = 4400$ and $N = 5513$ exams, respectively). We believe the cause of this discrepancy is that Computer Organization uses a full replacement with a 90% grade cap policy (see Table I) that always takes the 2nd chance score if it is available, while Dynamics uses a partial grade replacement policy with insurance. We suspect that because the Dynamics second-chance exam can be taken risk free, more students take it.

Furthermore, we find that different populations of students elect to take second-chance exams in response to these different policies. Figure 5 plots students' first-chance scores versus their second-chance scores for Dynamics and Computer Organization students. It can be immediately seen that the correlations between first-chance and second-chance scores in these classes are different.

The first obvious aspect is the distinct lack of symmetry around the 45 degree line in the Computer Organization course. The data above the 45 degree line are students that performed better on the second-chance exam than the first-chance exam. In the Computer Organization course, students that scored below the 45 degree line are actually hurting their grade by taking the second-chance exam. From this graph, it appears that situation is more common than it is, as many

points stack on top of each other in the upper triangle. We find that 88% of the students perform better on their second chance exam in Computer Organization with a full replacement policy.

This rate was statistically significantly ($p < 0.01$) different than in the Dynamics class (70% of students performed better on the second-chance exam) where the partial grade replacement policy with insurance is used. This result is not surprising, as students can't hurt their grades by re-taking the exam. As such, we see students with high grades (e.g., 90%) taking the second-chance exam and scoring poorly (e.g., 0%) on the second chance exam. We imagine that these students go straight for the hardest problem on the second-chance exam and abandon the exam as soon they get enough answers wrong that the second chance cannot improve their grade. In fact, these plots reveal that a non-trivial number of students who earned perfect (or nearly perfect) scores returned to take the second-chance exam under the partial-grade replacement with insurance policy; we question whether this is a good use of student and instructor time.

In contrast, no students with perfect scores in Computer Organization are electing to take a second-chance exam. In fact, rationally, no student scoring above 90% should be taking the second-chance as doing so can only hurt their grade. Instead, a significant fraction of the data points are along the $x = 0$ line representing students that got no points on the first-chance exam and that got points on the second-chance exam and the $y = 90$ line representing students that likely got perfect scores on the second-chance after imperfect scores on the first chance. This part of the plot is more likely derived from the fact that many of the exams in Computer Organization are 1-hour exams consisting of a single programming problem where

TABLE III
STUDENT SCORES ON A FIRST-CHANCE EXAM DIDN'T SIGNIFICANTLY CHANGE WITH INTRODUCTION OF A SECOND-CHANCE EXAM.

Semester	second-chance available?	number of students, N	mean score on first chance, μ
Fall 2017	No	370	75.0
Spring 2018	No	267	75.0
Fall 2018	Yes	357	75.1

these behaviors result from having no passing tests on the first-chance exam and passing all of the tests on the second-chance exam, respectively.

Requiring an additional homework to earn the right to take a second-chance exam can ensure that students perform some preparation before a re-take and also discourage unnecessary exam taking (e.g., students with near perfect scores re-taking exams). Intro to Electronics instituted such an additional homework and saw a lower re-take rate (34% of students) relative to Dynamics (49% of students) even though their grading policies were otherwise identical and the average first-chance exam scores were not statistically different ($p = 0.11$, $N = 1368$ and $N = 5513$, respectively).

Nevertheless, the choice of policy is clearly influencing student behavior. We view these policies as different points on a continuum of how encouraging we should be about how many extra exams students should take. The partial replacement with insurance policy is particularly encouraging, so as to ensure that no potentially benefiting student is discouraged, but may unintentionally increase the exam burden on strong students (who may feel that they need to get every point) and unnecessarily increase exam administration and grading workload on faculty and their course staff. The full replacement policy with grade cap is less encouraging so as to manage grading workload and discourage unnecessary exam sitting, but may unproductively discourage students from taking a second-chance exam that may benefit them. At present we do not have evidence of the relative merits of these policies.

C. Research Question 3: Second-chance testing improves student study behavior, especially by producing more spaced repetition

One possible interpretation to the increase in exam scores due to second-chance exams shown in Figures 2 and 3 is that the students' knowledge of the existence of a second-chance exam leads them to study less for the first-chance exam, decreasing first-chance exam scores. In such a scenario, the benefit of second-chance testing would be less than is indicated by that data. To explore the extent of this effect, we would need a way to quantify students' study efforts to be able to compare them with and without second-chance exams.

In this section, we describe an initial experiment in this direction. While the majority of exams in the Computer Organization class offered second chances, this experiment focuses on a particular exam that was, prior to Fall 2018, offered as a one-shot exam. This exam is derived almost entirely from a large collection of short answer questions provided as homework through an online homework system. Practice exams are also provided through this online homework system.

The system allows students to practice the problems as much as they want, and no practice problems were provided outside of the system.

As such, we can expect that the amount of interaction a student has with this online system in the time leading up to the exam is a good proxy for their total amount of study for the exam. This online system logs every interaction that the student has with the system, so we have a detailed picture of when and how much each student studied.

Starting in Fall 2018, we offered a second chance for this exam and observed how the amount of student practice compared between the Fall 2018 semester and prior semesters. We found, somewhat surprisingly, that the existence of the second-chance exam had no statistically significant impact on students' exam performance on the first-chance exam, as shown in Table III. Furthermore, there was no significant difference in the amount of effort (either in time or number of problems solved) in student studying using the online homework system before the first/only exam. Students that elected to take the second-chance exam, however, did additional studying between the exams, corresponding to an additional 60% of their original effort.

V. CONCLUSION

Second-chance testing is a compelling technique both in theory and in practice. Its theoretical foundations are found in retrieval practice with feedback, increased use of formative assessment, distributed practice, and test-potentiated learning. Empirically, we find that it motivates lower performing students to engage in additional practice to remediate so that they can demonstrate higher levels of mastery of the material reducing failing grades.

There remains, however, a lot to be learned about how to best integrate second-chance testing into courses. In this paper, we presented results on the effect sizes we observe from second-chance testing on student grades and an initial experiment suggesting that the introduction of second-chance testing, provided a suitable grade replacement policy, doesn't impact first-chance exam scores. We discussed principles of designing grade replacement policies and showed that the choice of policies can have significant influence on which students elect to take second chance exams.

These studies, however, are limited, in part, because they rely on comparing policies across classes. We therefore cannot distinguish how much of the observed effects are due to differences in course content, student demographics, or instructor differences. As such, it is important future work to compare student behavior as a function of course policy in the same course.

REFERENCES

- [1] R. Averill, S. Roccabianca, and G. Recktenwald, "A multi-instructor study of assessment techniques in engineering mechanics courses," in *2019 ASEE Annual Conference & Exposition*, Tampa, Florida, June 2019.
- [2] R. L. Bangert-Drowns, J. A. Kulik, and C.-L. C. Kulik, "Effects of frequent classroom testing," *Journal of Educational Research*, vol. 85, pp. 1–12, 1991.
- [3] B. Bloom, "Learning for mastery," *Evaluation Comment*, vol. 1, no. 2, pp. 1–12, 1968.
- [4] S. K. Carpenter, N. J. Cepeda, D. Rohrer, S. H. K. Kang, and H. Pashler, "Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction," *Educational Psychology Review*, vol. 24, pp. 369–378, 2012.
- [5] I. Clark, "Formative assessment: Assessment is for self-regulated learning," *Educational Psychology Review*, vol. 24, pp. 205–249, 2012.
- [6] S. D. Downs, "Testing in the college classroom: Do testing and feedback influence grades throughout an entire semester?" *Scholarship of Teaching and Learning in Psychology*, vol. 1, pp. 172–181, 2015.
- [7] S. Freeman, S. L. Eddy, M. McDnough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, "Active learning increases student performance in science, engineering, and mathematics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, pp. 8410–8415, 2014.
- [8] G. Gibbs and C. Simpson, "Conditions under which assessment supports students' learning," *Learning and Teaching in Higher Education*, vol. 1, p. 3–31, 2005.
- [9] B. Gutmann, G. Gladding, M. Lundsgaard, and T. Stelzer, "Mastery-style homework exercises in introductory physics courses: Implementation matters," *Phys. Rev. Phys. Educ. Res.*, vol. 14, p. 010128, May 2018. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.14.010128>
- [10] M. K. Hartwig and J. Dunlosky, "Study strategies of college students: Are self-testing and scheduling related to achievement?" *Psychonomic Bulletin and Review*, vol. 19, pp. 126–134, 2012.
- [11] C. Henderson, A. Beach, and N. Finkelstein, "Facilitating change in undergraduate stem instructional practices: An analytic review of the literature," *Journal of Research in Science Teaching*, vol. 48, pp. 952–984, 2011.
- [12] S. M. Juhler, J. F. Rech, S. G. From, and M. M. Brogan, "The effect of optional retesting on college students' achievement in an individualized algebra course," *The Journal of Experimental Education*, vol. 66, no. 2, pp. 125–137, 1998.
- [13] S. H. K. Kang, K. B. McDermott, and H. L. I. Roediger, "Test format and corrective feedback modify the effect of testing on long-term retention," *European Journal of Cognitive Psychology*, vol. 19, pp. 528–558, 2007.
- [14] C.-L. C. Kulik and J. A. Kulik, "Mastery testing and student learning: A meta-analysis," *Journal of Educational Technology Systems*, vol. 15, no. 3, pp. 325–345, 1987. [Online]. Available: <https://doi.org/10.2190/FG7X-7Q9V-JX8M-RDJP>
- [15] C.-L. C. Kulik, J. A. Kulik, and R. L. Bangert-Drowns, "Effectiveness of mastery learning programs: A meta-analysis," *Review of Educational Research*, vol. 60, p. 265, 1990.
- [16] J. Laverty, S. Underwood, R. Matz, L. Posey, J. Carmel, M. Caballero, C. L. Fata-Hartley, D. Ebert-May, S. E. Jardeleza, and M. M. Cooper, "Characterizing college science assessments: The three-dimensional learning assessment protocol," *PLoS ONE*, vol. 11, no. 9, p. e0162333, 2016. [Online]. Available: <https://doi.org/10.1371/journal.pone.0162333>
- [17] M. A. McDaniel, J. L. Anderson, M. H. Derbish, and N. Morrisette, "Testing the testing effect in the classroom," *European Journal of Cognitive Psychology*, vol. 19, pp. 494–513, 2007.
- [18] M. A. McDaniel, R. C. Thomas, P. K. Agarwal, K. B. McDermott, and H. L. Roediger, "Quizzing in middle school science: Successful transfer performance on classroom exams," *Applied Cognitive Psychology*, vol. 27, pp. 360–372, 2013.
- [19] M. A. Pyc and K. A. Rawson, "Why testing improves memory: Mediator effectiveness hypothesis," *Science*, vol. 330, p. 335, 2010.
- [20] K. A. Rawson and J. Dunlosky, "When is practice testing most effective for improving the durability and efficiency of student learning?" *Educational Psychology Review*, vol. 24, pp. 419–435, 2012.
- [21] K. A. Rawson, J. Dunlosky, and S. M. Sciarтели, "The power of successive relearning: Improving performance on course exams and long-term retention," *Educational Psychology Review*, vol. 25, pp. 523–548, 2013.
- [22] H. L. Roediger, III and A. C. Butler, "The critical role of retrieval practice in long-term retention," *Trends in Cognitive Sciences*, vol. 15, pp. 20–27, 2011.
- [23] M. J. Roszkowski and S. Speat, "Retaking the sat may boost scores but this doesn't hurt validity," *Journal of the National College Testing Association*, vol. 2, pp. 1–16, 2016.
- [24] D. William B, W. J. House, and T. L. Boyd, "A test-retest policy for introductory psychology courses," *Teaching of Psychology*, vol. 11, no. 3, pp. 182–184, 1984.
- [25] C. Zilles, M. West, G. Herman, and T. Bretl, "Every university should have a computer-based testing facility," in *Proceedings of the 11th International Conference on Computer Supported Education, CSEU*, May 2019.
- [26] C. Zilles, M. West, D. Mussulman, and T. Bretl, "Making testing less trying: Lessons learned from operating a Computer-Based Testing Facility," in *2018 IEEE Frontiers in Education (FIE) Conference*, San Jose, California, 2018.