Computerized exam reviews: in-person and individualized feedback to students after a computerized exam

Wayne L Chang, University of Illinois at Urbana - Champaign

Wayne Chang is a Lecturer in the Mechanical Science and Engineering Department at the University of Illinois at Urbana-Champaign. He received his BS, MS, and Ph.D. in Mechanical and Aerospace Engineering from the University of California, Irvine. He also serves as the course coordinator for introductory engineering mechanics courses in the Grainger College of Engineering.

Prof. Matthew West, University of Illinois at Urbana - Champaign

Matthew West is an Associate Professor in the Department of Mechanical Science and Engineering at the University of Illinois at Urbana-Champaign. Prior to joining Illinois he was on the faculties of the Department of Aeronautics and Astronautics at Stanford University and the Department of Mathematics at the University of California, Davis. Prof. West holds a Ph.D. in Control and Dynamical Systems from the California Institute of Technology and a B.Sc. in Pure and Applied Mathematics from the University of Western Australia. His research is in the field of scientific computing and numerical analysis, where he works on computational algorithms for simulating complex stochastic systems such as atmospheric aerosols and feedback control. Prof. West is the recipient of the NSF CAREER award and is a University of Illinois Distinguished Teacher-Scholar and College of Engineering Education Innovation Fellow.

Prof. Craig Zilles, University of Illinois at Urbana - Champaign

Craig Zilles is an Associate Professor in the Computer Science department at the University of Illinois at Urbana-Champaign. His research focuses on computer science education and computer architecture. His research has been recognized by two best paper awards from ASPLOS (2010 and 2013) and by selection for inclusion in the IEEE Micro Top Picks from the 2007 Computer Architecture Conferences. He received the IEEE Education Society's Mac Van Valkenburg Early Career Teaching Award (2010), campus-wide Excellence in Undergraduate Teaching (2018) and Illinois Student Senate Teaching Excellence (2013) awards, the NSF CAREER award, and the Univerisity of Illinois College of Engineering's Rose Award and Everitt Award for Teaching Excellence. He also developed the first algorithm that allowed rendering arbitrary three-dimensional polygonal shapes for haptic interfaces (force-feedback human-computer interfaces). He holds 6 patents.

David Mussulman, University of Illinois at Urbana - Champaign

Dave is an Instructional Technology Facilitator with the University of Illinois at Urbana-Champaign's Engineering IT Shared Services. He helps instructors select and integrate technologies into their courses to enhance student learning and improve course administration.

Carleen Sacris, University of Illinois at Urbana-Champaign

Carleen Sacris manages the day-to-day operations for the Computer-Based Testing Facility (CBTF) including the hiring and training of all full-time, graduate, and undergraduate proctors. She sets lab policies, engages with faculty regarding academic integrity concerns, and assists students on scheduling and disability accommodations. Before joining the Grainger College of Engineering, she was a Special Education Area faculty member at the University of the Philippines. She moved to the US in 2006 and received her master's in Teaching English as a Second Language at the University of Illinois at Urbana-Champaign. She thoroughly enjoys working with students and educators in all the different contexts she has found herself in throughout the years.

Computerized exam reviews: in-person and individualized feedback to students after a computerized exam

Abstract

Computerized testing centers are a promising new technology for running exams in large (200+ students) courses. They eliminate many of the logistical problems of pencil-and-paper exams; no conflict exams need to be scheduled, exams are graded efficiently and consistently, and timely feedback is provided to students. Computerized testing can be used to dramatically shorten the feedback cycle between student learning and feedback from assessment and enables the use of frequent testing and second-chance testing in large courses, which has been shown to lead to significant improvements in learning outcomes.

However, in some courses involving mathematical problem solving, an important student dissatisfaction with computerized testing is that numerical-answer questions are typically graded solely on the correctness of the final answer. The two major concerns reported by students are: (1) limited access to the assessment and corresponding learning opportunities post-assessment, and (2) the lack of partial credit given for correct solution procedures with incorrect final answers.

To address these concerns from students, a large public Midwestern university has developed a new exam-review service to provide in-person feedback to students after the completion of computerized exams, with the option of human-assigned partial credit for a correct solution procedure. These review sessions are hosted in the computerized testing facility to ensure the integrity of exam problems for future use. In these review sessions, students are able to go through their scratch work collected at the end of the exam, and any program code they wrote to solve problems, under the guidance of a course staff member. The format of the session is student guided in nature, where course staff are present to assist with the identification of conceptual errors.

In this paper, we present the design of the review system in a large-scale computerized testing facility, including the scheduling logistics, software support, course staff training, and guidance to students. Detailed data from student usage is reported, including survey data of student desire for exam review and the degree to which our system addresses this desire. We find that usage of exam review sessions to depend on three factors: the difficulty of the exam, whether exam regrading was offered during the session, and if a retry exam was available the following week.

1. Introduction

In engineering education, exams are commonly viewed only as a means for measuring learning rather than as an important part of the learning process[13]. To the contrary, a number of studies find that how students are assessed can 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[11]. Laboratory studies have shown that learning and retention of knowledge is enhanced through retrieval practice that incorporates feedback[16;23], increased use of formative assessment[7], and distributed practice[6;21].

Computerized testing centers were developed to reduce the overhead of running exams even in large (200+ student) classes, to make it easier to perform assessment in a fashion that positively impacts education. In the three courses discussed in this paper (Intro. Statics, Intro. Dynamics, and Intro. Solid Mechanics), shorter (50-minute) computerized exams are run every two weeks, students are given immediate feedback on their exam performance, and optional second-chance exams are offered in the off weeks. Frequent, low-stakes exams along with second-chance exams encourage students to keep up with and master the course material, leading to better final exam outcomes[20].

Three characteristics of our computerized testing center are key to improving assessment for both students and faculty. First, by running the exams on computers, we can write complex, authentic (e.g., numeric, programming, graphical, design) questions that are auto-gradable, enabling an immediate feedback and a reduction in grading load. Second, rather than write individual questions, we write question generators—small pieces of code that use randomness to produce a collection of problems—allowing us to give each student different question instances and permitting the question generators to be used semester after semester. Third, because each student has a unique exam, we allow students to schedule their exams at a time convenient to them within a specified day range, providing flexibility to students, avoiding the need to manage conflict exams, and allowing large classes to be tested in a relatively small facility.

While the introduction of computer-based testing into this engineering course sequence has led to positive educational outcomes, some students have expressed concerns about the testing format. In particular, students are accustomed to two aspects of pencil and paper exams: 1) the exams are generally returned to the students, and 2) the exams are manually graded with some amount of partial credit given to students for showing their work. As our computerized exams are graded interactively, before leaving the testing center, students know which problems they got correct and what the correct answers are for the problems they got wrong. As this is sufficient feedback for many students, these courses have elected to not return the exams to students to permit re-use of exam problems in future semesters. Partial credit for shown work is not provided on our computerized exams because it is technically difficult to do so via autograding. Recent findings suggest that awarding partial credit for shown work can create perverse incentives for students[4], but even in circumstances where we choose on principle to not provide such partial credit, we need to address students' desire for such partial credit.

To address these students' concerns, we instituted *exam review sessions* for computerized exams. These exam review sessions permit the students to return to our computerized testing center at times when it isn't being used for exams and review their tests in the presence of teaching assistants from the course. These exam review sessions have addressed a need of the students in these courses. In this paper, we make three contributions.

1. We describe the operation of these exam reviews, including the scheduling logistics, software support, course staff training, and guidance to students.

2. We provide student survey data that demonstrates that the availability of these exam

reviews have had statistically significant impact on students perceptions of their ability to learn from the mistakes on their exams.

3. We analyze the factors that influence attendence of exam reviews and identify three factors that increase attendence: lower average exam scores, opportunities to get exam regrades, and the availability of a retry exam in the following week.

This paper is organized as follows: We first review related work in Section 2. Then, in Section 3, we describe the operation of our exam review sessions, the data that we collected, and our analysis methods. Results are presented in Section 4. Finally, we discuss the results and conclude in Section 5.

2. Related Work

This work was inpired by the review sessions held by graduate assistants (GA) at the Evaluation and Proficiency Center (EPC) at the University of Central Florida. The EPC includes a computer testing center similar to the one described in this paper, providing asynchronous, proctored, auto-graded computer-based exams for engineering courses[9]. In addition, the EPC has a separate 30 seat tutoring area where students can come after the testing window has closed to review their exams and discuss them with the onsite GA[10]. Students' exam scratch paper is scanned after their exams and the scans are available during this review. Their papers describe the review sessions and compute the effort (and cost) savings of this approach. This paper complements that work by presenting data about students perceptions about the exam reviews and their utilization as a function of average exam score and regrading policy.

Previous work suggests that post-exam review can be an important contributor to student learning. Addison performed a correlational study that found that students in two psychology classes that attended all of the lectures that presented detailed reviews of exams had statistically significantly better overall exam performance than students that missed one or more of the exam reviews, even though each group had statistically equivalent performance on the first exam (i.e., before the first exam review)[1]. Surveys of students find mixed opinions about whether post-exam review is a good use of lecture time and that these views influence their attendence[19].

Levant et al. find that providing explanations beyond the correct answer on a first test can have statistically significant positive impact on student scores on a re-test[18]. Their medical school study compared providing just the correct multiple choice response (control treatment) with providing explanations of the question's rationale, why the correct answer was correct, and why each incorrect answer was not correct (experimental treatment). Providing explanations increased the score improvements from test to re-test by a larger degree ($3.92 \pm 7.12\%$, $n = 197$) than correctness-only feedback ($2.29 \pm 6.83\%$, $n = 192$).

Because many students don't review their exams beyond checking the score[14], some faculty have introduced post-exam assignments with the intention of helping students learn from their mistakes and develop metacognition. Generally, this process is found to be somewhat resource intensive but effective[2;3;5;22]. Faculty perceive that the technique leads to improved conceptual understanding and students believe that it helps them learn[14]. In the courses discussed in this paper, second-chance exams are offered as a motivation for students to review their exams and a means to demonstrate their revised understanding.

3. Methods

3.1. Implementation of the exam review system

Exam reviews were implemented within our computerized testing center (CTC). Exams in the CTC are conducted over 3 to 5 day windows, and the review sessions are available in a 2 to 3 day window after the end of the exam window. This ensures that all exams have been completed before the first student attends a review session.

Students sign up and make reservations for an exam review time slot using the same scheduling software used for the computerized exams themselves. Each exam review slot is 25 minutes long, and there are normally about 8 to 12 different times available over 2 to 3 days to permit students to find a slot that works with their schedule. Students may only sign up for one review session per exam.

During exam review sessions the CTC is blocked off from normal exams to eliminate the potential for distracting students taking exams. Using the CTC in a dedicated review mode allows for multiple courses to conduct reviews during the same time slots, improving the utilization of both the physical space and proctors.

For courses using the review system, student scratch paper is collected by proctors at the end of each exam. Scratch paper is not useful for reviews if it cannot be understood by course staff, so examples of good and bad scratch work are shown to students prior to taking the exam and they are encouraged to write neatly and show all work.

During review sessions, students have access to their hand-written scratch paper, as well as all the answer submissions and associated files from their computerized exam. Teaching assistants (TAs) for the course are staffed at every review session to answer questions from students. The TA-to-student ratio is 1:5. The goal of the review session is for the students to have a personalized experience in identifying the particular step in their solution procedure that was incorrect, which is different for every student. Students are encouraged to first try to work out their errors on their own, and they are frequently successful in resolving their issues without help. In cases where students cannot resolve their errors alone, TAs can step in and help. TAs are instructed not to explicitly show students how to solve an exam question from beginning to end, but rather to focus on identifying the conceptual or procedural error. In cases where students students struggle to even get started with a problem, they are redirected to office hours outside of the review sessions for assistance on understanding course concepts.

In one course (Dynamics), review sessions allow students to submit regrade requests for partial credit. In this course, students may submit a regrade request for one problem per exam and the maximum partial credit that can be awarded is 60%. The regrade request consists of a written justification for how their scratch paper from the exam shows a mainly-correct solution procedure and an identification of the error that they made. This request is submitted to the TA in the review session, who will award partial credit based on a regrade rubric established by the course instructor. Importantly, the regrade decision is only made after the review session is over, to avoid subjecting the TA to direct pressure from students.

In order to ensure the integrity of the exam problems for future course offerings, CTC proctors are present during review sessions and exam security protocols are followed during reviews. This

Course	Students	Exams	Retry exams	Regrades?	Review attendances
Statics	600	5	5	No	402
Dynamics	200	6	5	Yes	346
Solids	153	6	5	No	164

Table 1: Courses included in the study. Each review attendance is one student attending a review session for one exam.

includes no electronics being brought into reviews, no notes being taken out, pre-assigned seating, and entry to reviews only with a reservation and identify verification. TAs are trained in CTC policies before reviews.

3.2. Course details

Three courses implemented the new exam review system in one semester, as shown in Table 1. These courses were all introductory mechanics courses that are normally taken at the Freshman or Sophomore level. There were a total of 953 student enrollments, 17 exams, and 15 optional retry exams.

All three courses used an exam pattern of a regular exam every other week, with optional retry exams for partial score replacement in the intervening weeks. The retry scoring policy[15] was "Weighted Average with Insurance". That is, given an exam score E and a retry exam score R, the total score T was computed as

$$T = \max\left(E, \frac{1}{3}E + \frac{2}{3}R\right). \tag{1}$$

This means that a student could replace up to 2/3 of the points they lost in the regular exam with points from the retry exam. Between 15% and 65% of students elected to take the retry exam in any given week.

3.3. Student surveys

We conducted anonymous end-of-semester surveys of students in Statics in the semester before the review system was implemented ("pre-intervention", $N = 83$) and in the semester with the review system ("post-intervention", $N = 330$). We used five-point Likert scales and analyzed them using parametric t-tests to see whether pre- and post-intervention results were significantly different[8;12]. Additionally, we focused on the least satisfied students (those answering "Strongly disagree" on the Likert scale) and computed relative risk as a measure of the change in proportion of students in this group.

3.4. Review attendance data collection and analysis

In all three courses that used the exam review system (see Table 1) we collected the number of students who took each exam, the average score on the exam, and the number of students who attended the corresponding review session. To understand whether the courses had different review attendance rates we computed average rates and used a one-way ANOVA and pair-wise t-tests. To analyze whether review attendance was different for regular and retry exams, we fitted an OLS (Ordinary Least Squares) model of the form

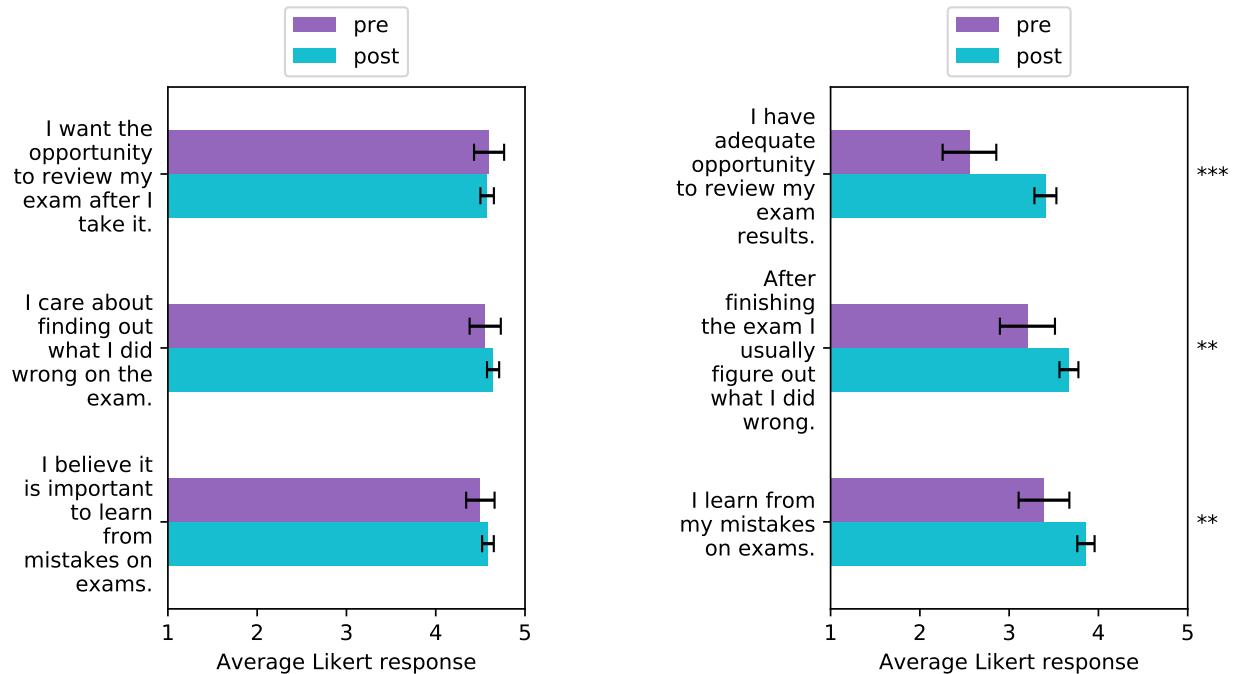$$\text{review-rate} \sim \text{course} + \text{retry-exam}, \tag{2}$$

Figure 1: Average responses to survey questions. The scale ranges from 1 ("Strongly disagree") to 5 ("Strongly agree"). Error bars are 95% confidence intervals. Asterisks on the right side indicate statistically significant differences between pre- and post-intervention results.

where "review-rate" is the exam review attendance rate, "course" is a categorical variable, and "retry-exam" is 0/1 indicator.

To investigate whether average exam score affected the review attendance rate we only considered regular exams (not retry exams), because the retry-exam scores are not reliable indicators of student knowledge or concern about the exam. This is due to the fact that the grading policy (see Eq. (1)) meant that the retry exam score could not lower the total score and so many students who took the retry exam abandoned it halfway through or otherwise failed to take the exam seriously. To test for a score effect, we fitted an OLS model of the form

$$\text{review-rate} \sim \text{course} + \text{score}, \tag{3}$$

where "course" is a categorical variable and "score" is the average percentage score on the exam.

4. Results

4.1. Pre/post student survey results

Figure 1 shows the survey results. Questions in the left panel asked students about how much they agree that it is important to review exams, find out what they did wrong, and learn from their mistakes. Students in both pre- and post-intervention groups report high agreement with these statements (average about 4.5 on a 5-point scale). There were no statistically significant differences between pre- and post-intervention responses to these questions ($p = 0.82, 0.35$, and 0.33, respectively).
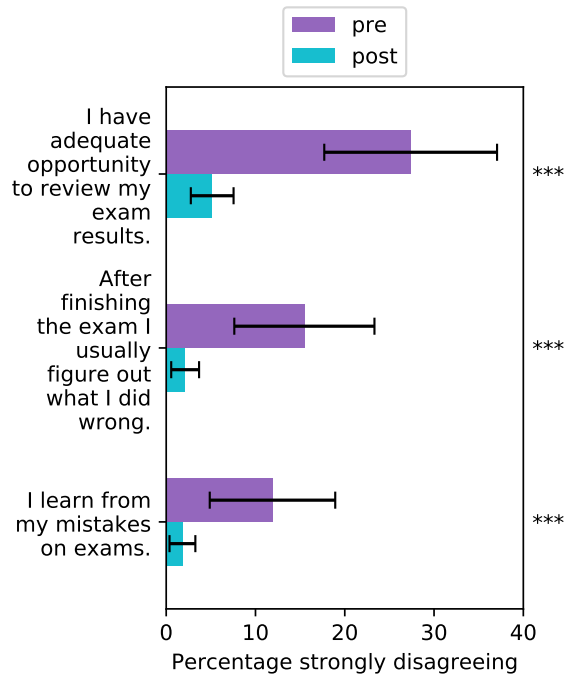
Figure 2: Percentage of students strongly disagreeing with the survey question statements. The three questions are the same as in the right panel of Figure 1. Error bars are 95% confidence intervals. Asterisks on the right side indicate statistically significant differences between pre- and post-intervention results.

Questions in the right panel of Figure 1 asked students how much they agreed that they could do each of the items in the left panel (review exams, find out what they did wrong, learn from their mistakes). On all three items there was a statistically significant increase in agreement from pre- to post-intervention ($p < 0.001, 0.006, 0.002$, respectively). The effect sizes were $d = 0.07$ (medium[17]), 0.04 (small), and 0.04 (small), respectively. This indicates that the review system made a statistically significant and moderate-sized difference to students' average ability to not only review their exams, but also to learn from them, at least judged by this self-report measure.

On all three items the post-intervention group using the new exam review system reported a positive level of agreement with statements about their ability to review exams and learn from them. The average responses were 3.41 (95% CI [3.28, 3.53]) for opportunity to review, 3.67 (95% CI [3.56, 3.78]) for finding out what they did wrong, and 3.86 (95% CI [3.76, 3.96]) for learning from mistakes.

Figure 2 and Table 2 show the percentage of students strongly disagreeing with the three statements about their ability to do each of the items (review exams, find out what they did wrong, learn from their mistakes). This shows dramatic reductions after the introduction of exam reviews. Students were 5.5 times less likely to strongly disagree that they could review their exams and even less likely to disagree that they could figure out what they did wrong and learn from their mistakes.

Statement	Pre [95% CI]	Post [95% CI]	RR [95% CI]	p
I have adequate opportunity to review my exam results.	27.4% [17.7, 37.1]	5.2% [2.8, 7.6]	5.5 [3.1, 9.8]	< 0.001
After finishing the exam I usually figure out what I did wrong.	15.5% [7.6, 23.3]	2.1% [0.6, 3.7]	7.4 [3.0, 17.9]	< 0.001
I learn from my mistakes on exams.	11.9% [4.9, 18.9]	1.8% [0.4, 3.3]	6.7 [2.5, 17.9]	< 0.001

Table 2: Percentage of students strongly disagreeing with survey statements, together with relative risk (RR) between pre- and post-intervention values and the p-value for pre- and post-intervention values being different.
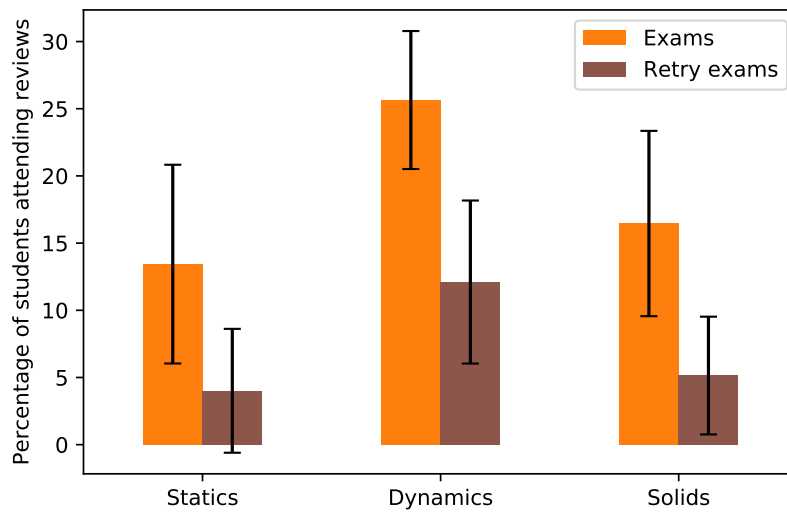

Figure 3: Average exam review attendance rates by course, for both regular exams and retry exams. Error bars are 95% confidence intervals.

	Average exam review attendance rate			
Course	Exams	[95% CI]	Retry exams	[95% CI]
Statics	13.4%	[6.0, 20.8]	4.0%	[-0.6, 8.6]
Dynamics	25.6%	[20.5, 30.8]	12.1%	[6.0, 18.2]
Solids	16.5%	[9.6, 23.4]	5.1%	[0.8, 9.5]

Table 3: Average exam review attendance rate for each course, broken out by regular exams and retry exams.
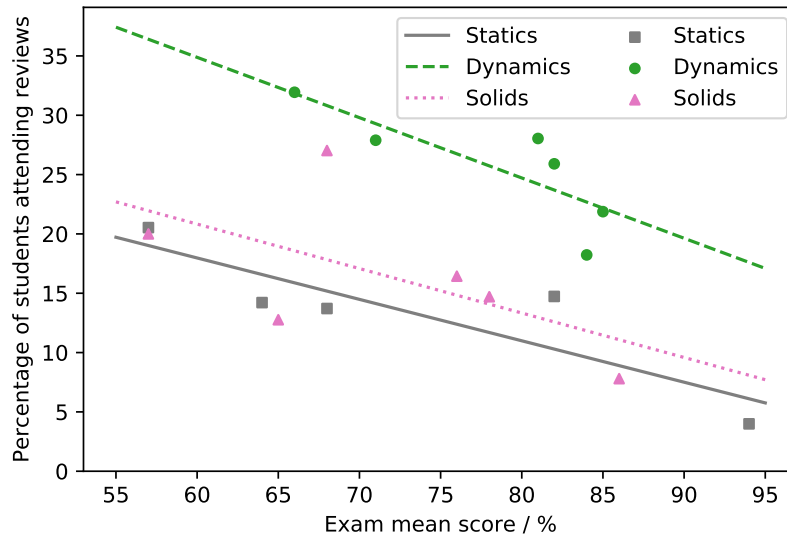
Figure 4: Exam review attendance rates versus exam scores for regular exams. Lines are linear least-square fits for each course.

4.2. Exam review attendance rates

The average attendance rates of exam reviews for the different courses are shown in Figure 3 and Table 3. The average rate over all courses was 13.3% (95% CI [10.1, 16.5]). However, Dynamics had approximately twice the rate of review attendance as Statics or Solids. This difference was statistically significant ($p = 0.003$ from a one-way ANOVA). The review attendance rates for Statics and Solids were not significantly different to each other ($p = 0.43$ from a two-sided t-test).

As can be see from Table 3, the attendance rate of reviews for retry exams was about 60% lower than for regular exams. In the OLS model (2) the retry-exam coefficient was -11.5 (95% CI $[-15.1, -7.9]$) percentage points, which is significantly negative ($p < 0.001$). This indicates that students were significantly less likely to attend exam reviews for retry exams.

Figure 4 shows the exam review attendance rates plotted against the mean score for each exam. In the OLS model (3) the score coefficient was -0.38 (95% CI $[-0.60, -0.17]$) percentage points per percentage point, which is significantly negative ($p = 0.002$). This shows that students were much less likely to attend exam reviews when the average exam score was high.

5. Discussion and Conclusions

We described the successful implementation of an exam review system at scale for exams in a computerized testing facility. This system dramatically reduced the number of the most unhappy students (more than a five-fold reduction) with a relatively small investment of effort.

The exam reviews were each 25 minutes long and 912 individual exam reviews were conducted in the first semester of operation. All reviews were conducted under proctored conditions in our computerized testing facility with the aid of course teaching assistants.

Students consistently agreed that it is important to review exams, find out what they did wrong, and learn from their mistakes. These opinions were the same between the pre-intervention

semester without a review system and the post-intervention semester with reviews.

Survey data indicates that the exam review system moderately increased the average of students' perceived opportunity to review their exams, find out what they did wrong, and learn from their mistakes. These effects were all statistically significant.

More significant than the average increase in student perceptions of review efficacy was the dramatic reduction in the number of students who strongly disagreed that they had adequate opportunity to review their exams to learn from their mistakes. This was the most unhappy group of students before the exam reviews were instituted, and this group massively declined in size (from 27% of students to just 5% of students, a 5.5-times reduction).

Overall, 13.3% of students attended any given exam review session. However, there were three key factors which affected the attendance rate. First, Dynamics had twice the review attendance rate of Statics and Solids. This is presumably due to the regrading policy in Dynamics, which offered partial credit for one question for each student attending the review.

Second, retry exams had less than half the review attendance rates of regular exams. This might be because there were no more retries offered. This would suggest that students are highly motivated to review their exams when they are about to take a retry, but without this immediate incentive they are much less likely to attend. This would suggest that reviews would be much less used in courses that only have a few exams and do not offer retries.

Third, students were much more likely to attend an exam review if average scores on the exam were low. For example, in courses without regrading in reviews attendance rates were around 20% when the average score was 60%, but fell to just 4% of students when the average score was 94%. This effect explained much of the week-to-week variability in review attendance rates.

While the review system described in this paper was able to efficiently scale to large courses and improve students' perceptions of their ability to learn from mistakes, we have not quantified whether there are actual learning improvements from the reviews. We leave this question for future work.

References

[1] William E. Addison. Consequences of missing postexam review sessions. *Teaching of Psychology*, 22(2):121–123, 1995. doi: 10.1207/s15328023top2202_7.

[2] Thomas Ammirati. Targeted test revision–another approach to science testing. *Journal of College Science Teaching*, 28(2):117, 1998.

[3] Gillian Andaya, Victoria D Hrabak, Sarah T Reyes, Rafael E Diaz, and Kelly K McDonald. Examining the effectiveness of a postexam review activity to promote self-regulation in introductory biology students. *Journal of College Science Teaching*, 46(4), 2017.

[4] Ronald C. Averill, Geoffrey Recktenwald, and Sara Roccabianca. Effect of assessment methods on performance in mechanics of materials. In *Proceedings of the 2018 American Society of Engineering Education (ASEE 2018)*, 2018.

[5] Cynthia Bolt-Lee and Sheila D Foster. Examination retakes in accounting: increasing learning by writing after the exam. *Language and Learning Across the Disciplines*, 4(2): 40–46, 2000.

[6] 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*, 24:369–378, 2012.

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

[8] J. C. F. de Winter and D. Dodou. Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research, and Evaluation*, 15(11), 2010.

[9] Ronald F. DeMara, Navid Khoshavi, Steven D. Pyle, John Edison, Richard Hartshorne, Baiyun Chen, and Michael Georgiopoulos. Redesigning computer engineering gateway courses using a novel remediation hierarchy. In *2016 ASEE Annual Conference & Exposition*, New Orleans, Louisiana, June 2016. ASEE Conferences. https://peer.asee.org/26063.

[10] Ronald F. DeMara, Tian Tian, Soheil Salehi, Navid Khoshavi, and Steven D. Pyle. Scalable delivery and remediation of engineering assessments using computer-based testing. In *Proceedings of IEEE Integrated STEM Education Conference (ISEC)*, March 2019.

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

[12] S. E. Harpe. How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, 7(6):836–850, 2015. doi: 10.1016/j.cptl.2015.08.001.

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

[14] Charles Henderson and Kathleen A Harper. Quiz corrections: Improving learning by encouraging students to reflect on their mistakes. *The physics teacher*, 47(9):581–586, 2009.

[15] G. Herman, K. Varghese, and C. Zilles. Second-chance testing course policies and student behavior. In *Proceedings of the 2019 Frontiers in Education (FIE 2019)*, 2019.

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

[17] M. A. Kraft. Interpreting effect sizes of education interventions. *Educational Researcher*, 2019. (preprint, Annenberg Working Paper, Brown University).

[18] Beth Levant, Wolfram Zückert, and Anthony Paolo. Post-exam feedback with question rationales improves re-test performance of medical students on a multiple-choice exam. *Advances in Health Sciences Education*, 23(5):995–1003, 2018.

[19] Trent W Maurer, Laura Frost, Diana Sturges, Simone Charles, Deborah Allen, J Michelle

Cawthorn, and Cherry C Brewton. Faculty and student perceptions of influences on post-exam attendance. *Journal of the Scholarship of Teaching and Learning*, 9(3):38, 2009.

[20] J. W. Morphew, M. Silva, G. Herman, and M. West. Frequent mastery testing with second-chance exams leads to enhanced student learning in undergraduate engineering. *Applied Cognitive Psychology*, 34(1):168–181, 2020. doi: 10.1002/acp.3605.

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

[22] John M Risley. Reworking exams to teach chemistry content and reinforce student learning. *Journal of Chemical Education*, 84(9):1445, 2007.

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