Effect of Mastery-graded Exams on Student Outcomes in Statics and Mechanics of Solids Course

Dr. Hadas Ritz, Cornell University

Hadas Ritz is a senior lecturer in Mechanical and Aerospace Engineering, and a Faculty Teaching Fellow at the James McCormick Family Teaching Excellence Institute (MTEI) at Cornell University, where she received her PhD in Mechanical Engineering in 2008. Since then she has taught required and elective courses covering a wide range of topics in the undergraduate Mechanical Engineering curriculum. In her work with MTEI she co-leads teaching workshops for new faculty and assists with other teaching excellence initiatives. Her main teaching interests include solid mechanics and engineering mathematics.

Dr. Kathryn Dimiduk, Cornell University

Kathryn Dimiduk is the Director of the Teaching Excellence Institute in the College of Engineering at Cornell University. She received her B.A. in Physics from Cornell University and her Ph.D. in Applied Physics from Stanford University. Her current research interests are in engineering education.

Dr. Andrew van Paridon

Effect of Mastery-graded Exams on Student Outcomes in Statics and Mechanics of Solids Course

Abstract

We piloted a mastery-style assessment method in one section of a sophomore-level Statics and Mechanics of Materials course at Cornell University. The main goal of this approach is to move student and course staff effort away from rote completion and grading of homework problems and towards building problem solving skills, as demonstrated with successful work on exams. Our work is inspired by the work of Averill, Roccabianca, and Rechtenwald at Michigan State University (MSU) [1, 2], but our implementation differs from theirs in several key aspects.

During Fall 2019, we taught two lecture sections of the course. The control section was graded in the traditional manner with a portion of the course grade due to graded homework and all exams graded with generous partial credit, while the experimental section did not have their written homework collected or graded and exams were graded with a constrained partial credit, or mastery-graded rubric. For the mastery-graded exams, students received full marks for a correct solution method leading to the correct answer, and received no credit if either the method or the answer was not correct. Some partial credit was possible if students could show that their method was conceptually correct and their wrong answer due to a calculator error or similar. Students in the experimental section had the opportunity to retake parts of each exam up to two additional times to improve their scores. The two sections were given the same final exam.

Key differences between our implementation and that of Averill, et. al. include: students in both of our sections were required to submit online homework, which was auto-graded by the learning management system; we had only two midterm exams compared to their five midterm exams; due to time constraints (50 minute lecture periods versus 90 minute evening midterm time slots), students in the experimental section could retake only parts, not all, of the exams.

Students were surveyed midway through the semester and again at the end, and asked about their achievement of course outcomes and experiences with the novel grading scheme. The learning outcomes and experiences of the two sections are compared and suggestions for future implementation included. The general applicability of this method to other classes at other universities is also discussed.

Introduction

A common set of frustrations for engineering faculty surrounds homework in large-enrollment required introductory textbook-based courses. From the student point of view, textbook-based homework assignments can often feel rote and repetitive, and therefore not a good use of their time. Solution manuals to most textbooks are available to students online and it is very common for students to copy their homework answers from those solutions. From the course staff point of view, it can be very time consuming to carefully grade weekly homework. Since only a small number of students ever look at their graded homework assignments, it can feel like wasted effort to give helpful feedback on student work. Knowing the ubiquity of available solutions also disincentivizes spending time and effort on grading weekly assignments.

If students are neither putting forth significant effort nor getting helpful expert feedback on a weekly basis, it's unlikely they will learn the problem solving skills necessary for content mastery. However, the course grading rubric in many classes allows students in that situation to get a good grade in the course through significant partial credit on high-stakes exams even without clear understanding of the concepts or ability to solve problems completely and correctly. A typical exam problem grading rubric awards points for recognizing that certain types of equations will be necessary, even if the equations themselves are formed incorrectly. For example, a student might earn significant partial credit on a statically indeterminate torsion problem even if they never recognize it as a statically indeterminate problem, simply by writing some equilibrium equations and some angular deflection equations. As faculty, it's frustrating to see students leave our courses without ever solving a core problem correctly.

An ideal class structure might put significant course staff effort towards providing timely expert feedback on student problem solving attempts and significant student effort on deliberate practice to the point of being able to solve new problems correctly from start to finish. Minimal student time should be spent practicing skills they have already mastered and no course staff time should be spent writing feedback that students never see. Averill, et. al. at Michigan State University (MSU) [1, 2] have implemented an approach to homework and exams in their mechanics of materials course that aims to meet these goals. Sharing their same frustrations, here at Cornell University we have piloted an intervention based on their work.

In this paper we describe our piloted intervention and point out some key differences between our work and that at MSU. Our method used one control section and one experimental section. After different homework and exam rules for the sections throughout the semester we compared their performance on the final exam and a concept test. We also surveyed students twice during the semester and held focus groups to collect information about the student experience in the experimental section compared to the control section, asking questions about stress, study habits, and confidence in course material, for example. As results and discussion sections show, we did not achieve the significant learning gains that the MSU studies did. We conclude with thoughts on the general applicability of a mastery graded approach to other schools.

Methodology

While inspired by and based on the course design in [1], there were several key differences in our implementation versus theirs. Here we describe the attributes of our course and then highlight important differences from the work at MSU.

Course overview

Students enrolled in ENGRD 2020 attended three 50-minute lectures and one 50-minute TA-lead discussion section each week. Students were registered for either the experimental or the control lecture section, taught by different faculty members, but students in both lecture sections were intermingled in the discussion sections and had frequent course-related interactions. This research was conducted under IRB protocol number 1907008977 which required students to opt-in to participation in the study, so a detailed informed consent document was introduced during the first class. Because of the opt-in requirement and the interconnectedness of students in the two lectures, details of the differences between the experimental and control sections were made explicitly clear to the students during the first week of the semester. Students then had the opportunity to switch sections if their class schedules allowed. Regardless of which lecture section they were enrolled in, all students had the opportunity to opt in or out of participation in the study. Note that study participation had no effect on their activities during the semester or the grading scheme they were judged by, but only determined whether this study considers their data in our analysis. Approximately 25 students (from a total enrollment of approximately 240) opted out of participation.

Students enrolled in the control section were graded on:

- weekly written homework (primarily textbook problems as well as some additional problems related to in-class activities);
- approximately bi-weekly online homework (OLHW) administered through the university's learning management software (LMS) and primarily focused on "flipped" topics in the course (topics covered in posted videos and the textbook but not covered during in-class lecture time);
- attendance at lectures (noted through classroom polling software) and discussion sections; and
- two midterms and a final exam.

Students in the experimental section were graded on all of the same things except they did not submit weekly written homework. Additionally, midterm exam administration and grading was significantly different between the two sections, as described below.

Homework and OLHW

As mentioned in the introduction, some of the primary issues with homework include the ease with which students can copy solutions to textbook problems from online sources and the excessive course staff time and effort spent grading homework, when any feedback is rarely read by students. For those reasons, while both sections were assigned weekly homework, the experimental section did not submit their work. Solutions were posted after the submission deadline each week for all students to use while studying. However, all students did submit OLHW which was administered as "quizzes" in the LMS. Since this was automatically graded in the LMS, no course staff time was necessary to grade and provide individual feedback. Also, since the problems are not numbered problems taken from the textbook used for the course, it's difficult for students to find solutions online. The amount of effort necessary to cheat the system would be greater than the effort to complete the assignment. Since the OLHW primarily covered flipped topics or preparation for in-class activities, it was important that all students participate in those assignments.

Exams

At Cornell University, large-enrollment courses typically have 90-minute evening midterm exams, outside of regular class time. These evening exam schedules are determined approximately a year in advance, so the number of midterm exams was pre-set at two. Additionally, since the students were all in the same course (same discussion sections and assignments, regardless of which lecture section), it was important to have the same number of exams and cover course content at the same pace in both lecture sections. The mastery-graded approach introduced in [1] emphasizes the need for multiple exam attempts for mastery-graded students. Balancing all of these goals and constraints lead to the following course exam design.

All students took version A of each midterm together in the prearranged 90-minute evening exam slot, which was a Tuesday evening for both exams. The exam was divided into three sections: section 1 had conceptual problems or simple one-step calculations and was worth 40 points; section 2 had homework-type problems, but on the easier side, and was worth 30 points; section 3 was a challenge problem, similar to a harder homework-type problem, and was worth 30 points. Students in the control section were graded using typical partial credit rubrics that reward using key concepts and making a reasonable attempt, even with significant errors. That grading took approximately one week.

Students in the experimental section had a much stricter exam grading rubric but could choose to take version B and/or version C of the exam (covering the same content but with new problems) according to the following rules. All of their exams were graded according to the mastery-graded rubric set forth in [3] and summarized in Table 1. That grading took place the night of the exam and grades were available for review immediately[1]. This initial grading was strictly on correctness (for each problem, students received 0% or 100%) which allowed one professor plus one TA to grade over 100 students' exams in less than two hours. Students were then responsible for reviewing their graded exams and the posted solutions and identifying whether their mistakes were minor enough to be worth partial credit (for example, a calculator error or misreading a

[1]The online grading platform Gradescope was used for all sections of the course. This was a crucial aspect of implementing the experimental design because it allowed very rapid grading of the mastery-graded exams, made the graded exams available to the students instantaneously, and organized all regrade requests in an easy to access and easy to review way.

Table 1: Rubric used for exam grading for mastery-graded exams, following [3].

Competency	Score	Description
Meets Minimum Competency	100%	Correct answer fully supported by a complete, rational and easy to follow solution process, including required diagrams and figures
	80%	Incorrect answer due to one or two mechanical errors but supported by a correct solution process as described above
Does Not Meet Minimum Competency	0%	Incorrect answer due to conceptual or procedural error(s)

Table 2: Timeline for exam grading, regrading, and retakes for the midterm exams for the experimental section.

	Mon	Tue	Wed	Thu	Fri
Week I		Version A (evening)	Grades posted	Regrade requests due	Regrade results posted; Version B (in class)
Week II	Grades posted	Regrade requests due	Regrade results posted; Version C (in class)		Grades posted
Week III		Regrade requests due			

given value). If so, they submitted regrade requests which were evaluated by Friday morning, three days after version A. Version B of the exam was administered during class time on Friday. The grading and regrade request process was repeated and version C of the exam was administered during class time the following week on Wednesday. Table 2 summarizes the examination timeline for each midterm in the experimental section. The primary reason for such a condensed timeline during each exam (three versions of the exam administered within just over a week) was to prevent the experimental and control sections from being off-track with each other for an extended period of time. Since the students were all in discussion sections together each week the content had to move forward at the same rate.

Since the evening exam time was 90 minutes but the class period during which versions B and C were administered was only 50 minutes long, it was impossible to have students retake a full version of the exam. Therefore, section 1 of the exam (conceptual and simple problems) was only on version A. Versions B and C had retake opportunities only for sections 2 and 3 (homework-type and challenge problems). Students in the experimental section could optionally take exam versions B and/or C, in which case their overall exam score was found using "super scoring" of the three exam attempts: the best score on each section added together.

For the final exam all students in both sections took just one version of the final exam. For that exam, there was no regrade cycle. The exams were all graded according to the partial credit scheme and afterwards (and only for the purposes of this study, not for course grades) the rubric items were mapped onto the 0%-80%-100% grading scheme.

Key differences from MSU

Several key differences between the implementation in this study and the implementation in [1, 2] are noted here:

- This course is Statics and Mechanics of Materials, whereas the MSU course is Mechanics of Materials with a prerequisite of Statics. That may require moving through course content at a more rapid clip in this implementation since there may be more content to cover during the semester.
- Class periods at MSU during the time of their implementation included 80- or 90-minute class periods, which allowed full length in-class exams, whereas here the longest class period is 50 minutes. Therefore our retake exams were limited to only sections 2 and 3 instead of full coverage of exam sections.
- At MSU the exam retakes happened at least one week apart whereas in our implementation three version of each exam were administered over an 8 day period. This may have had an impact on student stress as discussed later in this paper.
- At MSU the first version of their implementation had four midterm exams (two retakes available on each) and the revised version of their implementation had five midterm exams with one retake available on each. Their exams start during the second or third week of the semester. Here we have only two midterm exams and the first doesn't take place until the sixth week.

The main aspects we were able to transfer from MSU to our implementation were the mastery-graded exams, the (partial) retake opportunities with super scoring, and no collection or grading of textbook-based weekly homework.

Data collection

The hypothesis of this study is that students who are graded throughout the semester on the ability to correctly solve problems will achieve better mastery of the course material than students who are graded on exams according to a generous partial credit policy. To determine whether mastery was achieved, we considered final exam scores graded both using partial credit and mastery-graded approaches, and student performance on the Statics Concept Inventory (SCI) [4] which was administered as both a pre-test at the beginning of the semester and post-test during the last week of classes (before the final exam).

Because students were aware of the pedagogical experiment and had some flexibility to self-select into one section or the other (subject to their own scheduling constraints) it was important to control for confounding variables. We collected demographic information (including gender, ethnicity and under-represented minority (URM) status) from the registrar, as well as GPA.

Student reaction to the mastery-graded approach was explored using a mid-semester survey after the first midterm exam and additional questions on the end of semester course evaluations, as well

Table 3: Population information for the experimental and control sections. GPA is out of 4.00 (4.33 for A+) and is reported as mean±standard deviation.

	Experimental	Control
Count	93	114
% female	53	61
% URM	26	30
GPA:		
all	3.35±0.90	3.36±0.60
male	3.24±1.09	3.27±0.69
female	3.44±0.63	3.42±0.53
non-URM	3.41±0.93	3.43±0.65
URM	3.16±0.79	3.20±0.46

as focus groups with students in the experimental section conducted after the second midterm exam. Questions on those surveys are discussed below.

Results

Statistical analysis

Comparison between performance of the lecture sections is meaningless without some understanding of the population in each section. There are limits to what information can be gleaned from registrar's data - for example, we didn't measure students' interest in the subject matter, or tendency towards test anxiety - but based on the information available, the control and experimental section were relatively well matched. Table 3 contains a demographic summary of the two sections. There are no statistically significant differences in GPA between the two sections and they have roughly equivalent demographics. Note the unusually high percentage of female students in the control section. There are several possible explanations, though we cannot say definitively. Since we can only consider data for students who opted in, it's possible that male students disproportionately opted out. This class is required by several majors, some of which have higher percentages of female students. Perhaps students from those majors, either due to class scheduling conflicts or interest in the material or some other reason, disproportionately enrolled in the control section.

Looking at final exam scores graded both ways (with partial credit, and mastery-graded), as response variables, we did Students' t-tests with lecture section, sex, and URM-status as predictor variables. If lecture section was a significant predictor variable, it would support our hypothesis, but we also wanted to see whether sex or URM-status were primary predictors of success without regard to lecture section. Table 4 shows the results for these different groups. The only predictor variable with statistical significance at the $p < 0.05$ level is URM-status for the partial credit graded final exam. Some items were close to statistic significance with sex at $p < 0.0635$ for the partial credit grading and lecture section at $p < 0.0508$ for the mastery-graded exam. A 1-way ANOVA analysis with GPA as a predictor variable unsurprisingly shows very strong significance

Table 4: Mean final exam scores for different subgroups, as well as p-values for t-tests for lecture section, sex, and URM status.

Group	Partial credit grading	Mastery-graded
Experimental	67.5	51.5
Control	63.9	46.0
Lecture section p-value	$p < 0.1282$	$p < 0.0508$
Male	68.0	51.6
Female	63.7	46.2
Sex p-value	$p < 0.0635$	$p < 0.0581$
non-URM	67.0	49.9
URM	61.4	44.6
URM status p-value	$p < 0.0378$	$p < 0.0975$

($p < 0.0001$) for both grading methods. The primary hypothesis, that students subjected to mastery-graded exams throughout the semester will perform better at the mastery-graded final exam, is hinted at but is not borne out to the level of statistical significance.

Some interesting trends appear when further dividing the groups. For example, in the experimental section there was almost no difference in exam scores by gender (67.8 for men vs 67.2 for women with partial credit, and 52.2 vs. 51.0 without) but in the control section there was a 7-8 point difference. On the other hand, in the control section there was a very small difference by URM status (64.4 non-URM vs. 62.8 URM with partial credit, and 46.5 vs. 44.9 without) but in the experimental section there was a 9-11 point difference by URM-status. The groups who seemed to have benefitted most from the experimental section methodology are females and non-URM students. However, none of these trends rise to the level of statistical significance.

The version of the SCI administered in this course used 8 questions, each of them randomly selected for each student from a bank of three equivalent options[2]. The average gain (that is, (post-test score)$-$(pre-test score)) is between 1.9-2.9 for all subgroups, and we found no significant effects based on section.

Midsemester surveys

After the first midterm (and all of its retakes) the students in both sections were given an anonymous midsemester survey. We used that to learn about students' efforts and experience in the course. Students in the experimental section reported completing 70% of the homework each week, on average. They self-reported spending an average of 2.6 hours (standard deviation 1.9 hours) versus 4.3$\pm$1.8 hours in the control section. Students were also asked how many hours they spend on the course outside of class time besides time spent on homework. In the experimental section they reported 2.3$\pm$2.1 hours and in the control section 1.7$\pm$1.2 hours.

[2]The complete SCI includes three equivalent questions covering each of 9 concepts, but sliding friction was not covered in this class so we removed it from the concept test.

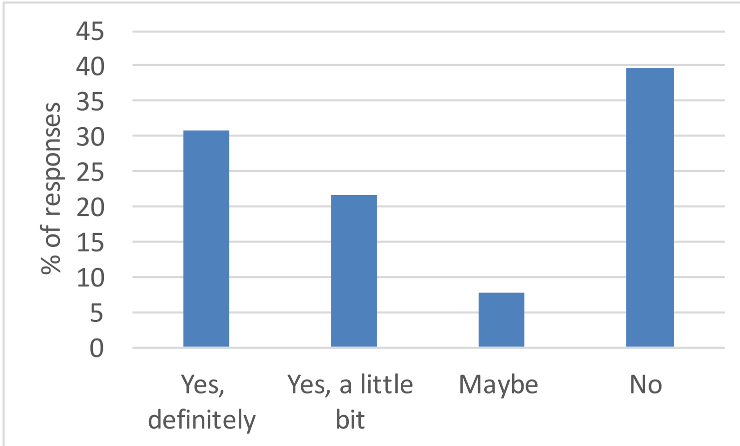

Figure 1: Experimental section responses to the midesemester survey question, "Does the exam structure with conceptual questions and problem solving questions where the problem solving questions component can be taken again with new problems, but [limited] partial credit is given, change how you study for this course compared to other courses?"

Students in the experimental section were asked whether the novel grading scheme changed how they studied in this course compared to other similar courses, and their responses are shown in Figure 1. When given the opportunity to elaborate in an open ended question, typical comments include things like:

"I still try to cover all the material, but especially focus on understanding the basic ideas, since I know I can't retake that section."

"I still study the same way, but it creates a lot more stress while studying."

"It really will change the way I study for [midterm] 2 and changed the way I studied for the retakes after I experienced [midterm] 1A. I will study conceptual things rather than just trying to figure out the math. I also actually look at solutions and try to figure out what I did wrong. Now that I'm trying to get full points, I try to fully complete a problem and study every aspect of it."

"I am scared so I study more."

Students in both sections were asked to report on their stress in this course compared to others; the results are shown in Figure 2. Students in the experimental section were more likely to say this class is more stressful (total between "much more" and "somewhat more" stressful) than other classes, but were also more likely to say the class is less stressful than other classes. This suggests that the experimental course design was divisive, appealing to some subset of students but stressful (and, according to some comments, frightening) to a larger group.

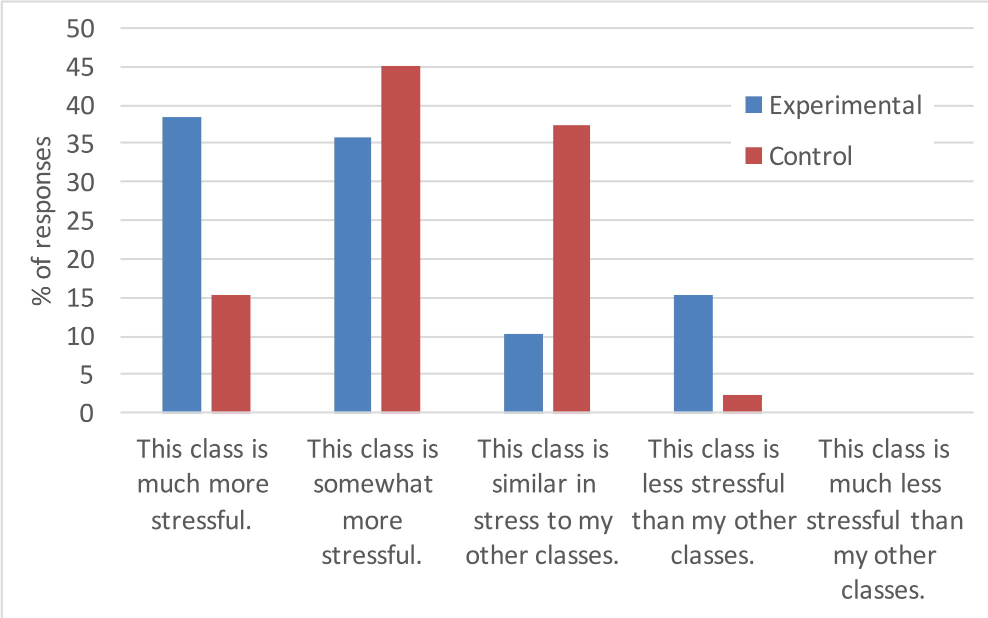

Figure 2: Responses from both sections to the midesemester survey question, “How does the stress from this class compare with that of other similar classes you have taken?”

End of term surveys

Students were asked survey questions at the end of the course prior to completing their final exam. Some observations from their responses are reported here. The students in the experimental section reported higher levels of confidence in their ability to use the material from this course in later courses, as shown in Figure 3. Some students responded to this question by evaluating the necessity of this course to support later courses in their specific major, rather than considering their own ability. These responses were ignored. Students were also asked the question; “When you see a new problem, how do you approach solving it?” In the experimental section, 18 out of 38 responses (47%) remarked they would form some sort of plan using the identified knowns, unknowns, and relevant equations. This compared with 15 out of 43 responses (35%) from the control group. The control group had a higher number of responses that suggested trying to use relevant equations without mentioning any form of planning (14 versus 8 responses). The other responses were either focused primarily on Free Body Diagrams or looking for unknowns only, or simply using a step-by-step brute force approach.

Students in the experimental section were given the opportunity to report how their study practices changed due to the exams, homework and grading for this course. Overall the responses generally split students across a few divisions. Some students found that they could strategize how they approach exams, giving them more time to target their study habits and learning from the retakes, while other students found that the added exams simply added excess stress to a busy exam period. Similarly, from the perspective of homework, some students enjoyed the flexibility to work on the homework in their own time and enjoyed having the solutions available, while others found the lack of rigid deadlines detracted from their motivation to keep up. The split appears to be along lines of self-motivation, and appears to be about evenly split based on the

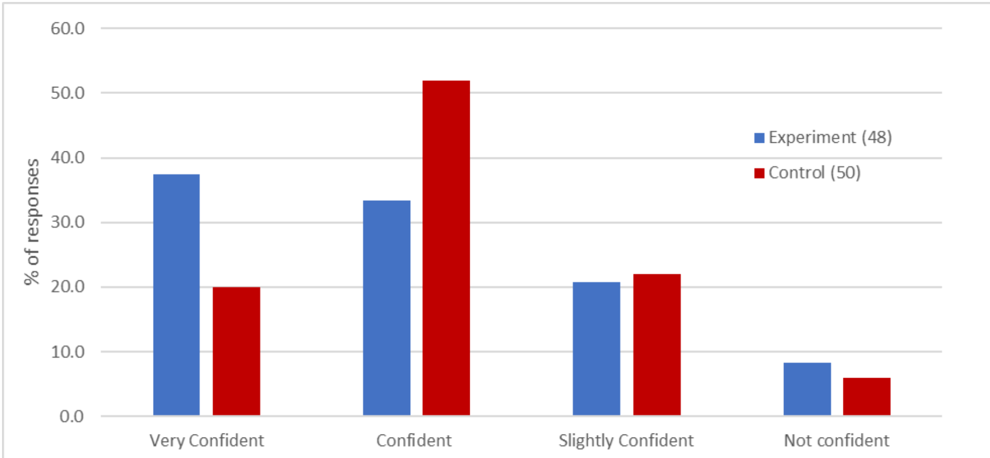

Figure 3: Responses from both sections to the end of term survey question, "How confident do you feel that you could use the material you learned in this course in a later course?"

surveys. Five students specifically reported that they felt compelled to finish practice questions to accurate full solutions whereas they might have previously left problems nearly but not totally complete before checking solutions. Some students reported that they spent more time on practice problems, and others reported they spent more time ensuring they understood the conceptual elements of the course. Representative comments are included here:

"Given the retake option, I found that I actually studied less up front - only enough to know I could get through section one - then continued to study for the next week until retakes were over, which honestly in the end had me studying more than I normally would."

". . . instead of making summarized notes like other classes, I found that annotating my problem sets was much more efficient for understanding course material [thoroughly]."

"The retake [midterm] system has benefitted my learning like no other."

"I have [been] focusing on the concepts [more] than I usually do with other courses."

"we weren't required to do the problem sets, [so] they would fall to the very back of my priorities."

Finally, students also provided some feedback on how they felt about the experiment in general.

"I believe this experiment placed unnecessary extra stress on students."

"Being in the experimental section put less stress on me homework wise. I was even more motivated to complete the homework even though it wasn't graded because I genuinely wanted to learn how to solve the problems."

"I found that this really took a toll on my self-esteem and self-confidence, which further led to poor exam performance simply because of nerves . . . I think the experience did make me a better studier, but I think it gave me much more stress than necessary"

"I think the experiment had a major flaw in it as the two classes were given incredibly different resources when it came to skill of lecturer."

Focus groups

Student response to the experimental approach to exams and homework was also explored during two focus groups following the second mid-term and its make-ups. The first point the students wanted to make clear was that they did not feel the experiment was valid because the experimental class and the control class had different instructors. The experimental section was taught by an experienced instructor who had taught the course several times already, while the control section was taught by a novice instructor. The students felt this was a dominant effect even though the instructors worked closely together, the lecture slides were mostly the same, the homework problems, textbook, activities, discussion sections, and exam coverage was the same. The final exam was exactly the same between the sections. The students clearly thought the lecture delivery had a big impact on their learning. Note that the final exam results and pre-post gains on the SCI do not support that impression of the students.

Two different modes of doing the homework were described. Some students did the homework regularly each week, but did appreciate being able to slip it a day or two occasionally as it wasn't collected. They looked at answer sheets after doing the homework. The majority of students in the first focus group (and, they reported, in the class), however viewed themselves as too busy to really do the homework until the week before the exam. They then did all the homework in the week leading up to the exam. Their approach was to do the homework with the solutions open. There was a consensus amongst the first focus group students that it would be better to have to turn something in, even if it was just existence graded to keep them more current in their studying. There was considerable frustration around the exam and its makeups. Students didn't change their studying from other courses and just viewed that they had to take three exams which they found added stress. Many of them felt the process didn't really help their grades or understanding and didn't like the constrained partial credit. They weren't sure it was quite fair that a few students they knew had focused on the first part of the exam the first time and then the later parts for the make-ups to the benefit of their grade. From the discussion it was clear that the focus group included students ranging from having done poorly even with the makeups through a student who had done very well. The general attitude was skeptical of the whole experiment.

The second focus group involved a different set of students. In this group students still weren't sure the experiment was valid, but there was a much stronger sense that the new approach was helping them learn. There was still a mix in how homework was done, but there was less of leaving it till the week of the exam. One student indicated that she was doing much better than she usually did in this course because she could see what she didn't know on the first exam and then focus on learning it. She felt much more confident in her ability to use the material in a later course. Several students said that instead of just getting an exam back and feeling badly, they would look to see what they had missed and try and learn it for the next exam. They liked that feeling and way of looking at the first exam. However, they still found the exam plus two makeups more stressful than a single exam as most of the stress was just due to taking an exam that week and the retakes felt like another exam. This was even more stressful when the makeups overlapped their studying for an exam in another class. These students also thought homework should be at least existence graded. They suggested maybe calling the exams quizzes since there were so many of them and that would reduce the stress level. The students were less frustrated with the new approach and could see value in it for their learning, but not all of them were convinced it was worth the extra exams. They wondered what would happen on the final and how their grades would come out overall.

Discussion and conclusions

Our implementation of a mastery-graded approach in our sophomore level Statics and Mechanics of Materials course did not lead to significant learning gains over a more typical homework and exam approach for the overall student population. For many students it caused an increase in the stress they felt related to the course, though some students, particularly stronger students, preferred this novel approach. For the students who preferred this approach, they cited such things as having flexibility to study and complete homework on their own schedule, getting to strategize in their exam time usage (focussing only on the part of the exam they couldn't retake the first time around, then focussing on one section at a time in subsequent retakes), and having to study their mistakes on an exam to earn partial credit.

For the students experiencing higher stress in this approach, they noted having to take three versions of an exam within just over a week as being excessive, being anxious about the lack of leniency in grading, and not being forced to work on the homework every week (and therefore not keeping up with the work or studying).

Even though the approach did not lead to significantly better learning outcomes in general, some groups may have experienced some benefit. In particular, the gender gap in final exam scores which appears in the control section does not appear in the experimental section. However, the instructor of the experimental section was female and of the control section male, and some research suggests instructor gender may affect achievement of students by gender [5, 6].

Some aspects of the course were very appealing to the instructors and may be implemented in future courses. The requirement for students to look at their incorrect exam problems and identify and correct their mistakes before receiving partial credit is an excellent way to force deliberate practice: focussing effort on difficult aspects of the work [7]. The partial credit must be more generous than the mastery-graded approach requires, however, if the additional exam opportunities aren't available. Also, not grading weekly homework significantly cuts down on course staff time. Since students rarely look carefully at their graded homework, any time spent grading it is essentially wasted. In the future grading homework only on existence might be enough incentive for the students to work on it each week but low enough staff time requirement to be justified.

The main likely reason for the difference in outcomes in this study compared to the group at MSU is the smaller number of exams. At MSU they were able to have an exam two out of every three weeks which was spread out enough to prevent serious fatigue but frequent enough to give students early and significant practice with the mastery-graded approach to exams. One purpose of this study was to assess how transferable the scheme may be from MSU to other courses or universities. Long class periods are necessary to allow the required frequency of exams, which may be difficult to come by in many courses. Our main conclusion is that in the absence of all aspects of the MSU implementation, the learning outcome gains may not be achievable.

Acknowledgements

This research was conducted under local IRB protocol number 1907008977. Dr. Erika Mudrak from Cornell Statistical Consulting Unit provided valuable assistance with the statistical analysis.

References

[1] R. Averill, G. Recktenwald, and S. Roccabianca, "Effect of Assessment Methods on Performance in Mechanics of Materials," In *ASEE Annual Conference & Exposition*, 2018.

[2] R. Averill, S. Roccabianca, and G. Recktenwald, "A Multi-Instructor Study of Assessment Techniques in Engineering Mechanics Courses," In *ASEE Annual Conference & Exposition*, 2019.

[3] R. Averill, "Rubrics," New Leaf Education, Blog, March 13, 2018, http://newleafedu.com/2018/03/13/rubrics/

[4] P. S. Steif, and J. A. Dantzler, "A statics concept inventory: Development and psychometric analysis," *Journal of Engineering Education*, vol. 94, no. 4, pp. 363-371, 2005.

[5] S. E. Carrell, M. E. Page, and J. E. West, "Sex and science: How professor gender perpetuates the gender gap," *The Quarterly Journal of Economics*, vol. 125, no. 3, pp. 1101-1144, 2010.

[6] F. Hoffmann and P. Oreopoulos, "A Professor Like Me: The Influence of Instructor Gender on College Achievement," *Journal of Human Resources,* University of Wisconsin Press, vol. 44, no. 2, 2009.

[7] D. L. Schwartz, J. M. Tsang, and K. P. Blair, *The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them*. pp. 39-51. WW Norton & Company, 2016.