Effect of Assessment Methods on Performance in Mechanics of Materials

Dr. Ron Averill, Michigan State University

Ron Averill joined the faculty at Michigan State University in 1992. He currently serves as the Associate Chair of Undergraduate Studies in the Department of Mechanical Engineering. His research focus is on design optimization of large and complex systems, analysis of composite materials and structures, and design for sustainable agriculture.

Dr. Geoffrey Recktenwald, Michigan State University

Dr. Recktenwald is a lecturer in Mechanical Engineering at Michigan State University where he teaches courses in in mechanics and mathematical methods. He completed his degree in Theoretical and Applied Mechanics at Cornell University in stability and parametric excitation. His active areas of research are dynamic stability, online assessment, and instructional pedagogy.

Dr. Sara Roccabianca, Michigan State University

Sara Roccabianca is an Assistant Professor in the Department of Mechanical Engineering at Michigan State University (MSU). She was born and raised in Verona, Italy and received her B.S. and M.S. in Civil Engineering from the University of Trento, Italy. She received her Ph.D. in Mechanical Engineering from the University of Trento in 2011. She then was a Postdoctoral Fellow at Yale University, in the Department of Biomedical Engineering, working on cardiovascular mechanics. Sara's research at MSU focuses on urinary bladder mechanics and growth and remodeling associated with bladder outlet obstruction (e.g., posterior urethral valves in newborn boys or prostate benign hyperplasia in men over the age of 60). Her goals are to (i) develop a micro-structurally motivated mechanical model to describe the non-linear elastic behavior of the urinary bladder wall, (ii) develop a stress-mediated model of urinary bladder adaptive response, and (iii) understand the fundamental mechanisms that correlate the mechanical environment and the biological process of remodeling in the presence of an outlet obstruction.

Effect of Assessment Methods on Performance in Mechanics of Materials

1. Introduction

The current study was motivated by the findings of numerous informal classroom experiments and hundreds of informal discussions with students, conducted during the previous four years. The goal of these efforts was to better understand students' current study habits, test taking strategies and attitudes towards learning. Among the findings, we discovered that the vast majority of students do want to learn, are quite capable of solving challenging engineering problems and feel like they are working very hard, yet their performance in quantified assessment is often below their expectations.

For many students, instead of using effective study strategies to fully understand key concepts and to master problem solving techniques, the goal of their current study and test taking strategy is to "maximize partial credit." The most common version of this strategy looks essentially like this.

1. Memorize problems from the homework, in-class examples, or previous exams.
2. Match each problem on the exam to one of the memorized problems that most closely resembles it.
3. Write down the memorized solution, making adjustments along the way so that the solution looks more relevant to the exam problem.

This strategy is often very effective at getting a passing grade or better, in large part because many instructors grade with a partial credit model based on the "correct approach," which is often not well-defined. If the memorized problems have similarities to the exam problems, it is difficult for a grader to truly distinguish whether or not a student used the correct approach. When partial credit is then awarded too generously, this flawed study approach is reinforced, and students do not feel the need to solve problems completely or correctly.

The above strategy is even more effective on open-book exams or exams that allow students to copy lots of solved example problems on a personal one or two-page formula sheet. In this case, there is no need to memorize the problems.

Despite being effective at "getting through" a class, this learning strategy has very little value in terms of creating knowledgeable and capable engineers. According to studies by cognitive scientists [1-4] (quote below from [4], page 156):

> "Example learners tend to memorize the examples rather than the underlying principles. When they encounter an unfamiliar case, they lack a grasp of the rules needed to classify or solve it, so they generalize from the nearest example they can remember, even if it is not particularly relevant to the new case."

Moreover, pre-test cramming of these example problems into short-term memory promotes almost no retention of the information [4,5].

Our informal student interviews also indicated that many students are copying most of their homework solutions from online resources. While collaboration on homework has probably occurred at some level since the concept of graded homework was introduced, the practice of purely copying homework without thinking about its substance is now widespread. This calls into question the value in grading and assigning course credit for homework. Additionally, the use of online products with randomized parameter values for homework problems does not help this situation. In this case, students can substitute new values of the parameters into the available online solutions or shared spreadsheets, which has very little value for learning.

In some cases, a passing grade in a class can now be obtained through a combination of copying online solutions to obtain a nearly perfect homework score and maximizing partial credit on exams by memorizing a few example problems. These approaches do not contribute in a positive way to the desired student outcomes.

The flawed student learning strategies described above are not necessarily new, but their increased magnitude and widespread usage are relatively recent. They are enabled in part by the internet and social media tools, but also by certain types of grading models and course structures. Based on discussions with colleagues at other institutions, these increasing trends are observable across universities and disciplines.

With these ideas in mind, we developed an experiment to investigate a different approach to course management, with the hypothesis that students will adapt their study habits to changing assessment strategies. This was implemented in a sophomore level Mechanics of Materials course (i.e., Mechanics of Deformable Solids, ME222 at XXX). We do not claim that the style of assessment used in the current experiments is unique, but it is uncommon in today's classrooms.

The main goal of these experiments was to determine if a change in assessment strategies would lead to better study strategies and thus improved student performance. The results strongly support this assertion.

2. Modified Course Design

The course design and methods used in this study were based on a series of informal discussions and interviews with students, which revealed the following behaviors:

1. Students tend to focus on maximizing partial credit as opposed to understanding.
2. Students focus more on memorization than developing or practicing a systematic problem solving process.
3. Students are subverting the benefits of homework by using online solutions.

The goal of this paper is not to discuss the validity of these observations, but rather the effects of a course design intended to address them.

We hypothesized that our current paradigm for teaching and grading was in part enabling these flawed approaches to learning, so a new course design was developed to encourage more productive learning behaviors. The course design had three primary features, one to address each of the above observed student behaviors. The primary design features were: Mastering, Variation, and No Graded Homework.

Mastering

The primary feature of the course design was to move toward a mastery model of grading. In this model, students receive credit only for correct solutions or solutions with minor errors, as described in Table 1. This means that any conceptual mistake made in the solution process results in no credit. The purpose of this grading method is to only give students credit for understanding how to solve the problem completely.

Table 1. Rubric used to grade each problem on exams.

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

In Level II scores described in Table 1, there are two necessary conditions for classifying an error as minor:

1. The mistake is a minor algebraic error, computational error, error in units or significant digits, or other human mistake such as misreading a value in the problem statement.
2. If the identified error had not been made, the final solution would have been correct.

When either of these conditions is not true, the error is assumed to be conceptual and the work does not demonstrate minimum competency. It thus receives no credit.

Grading appeals. In order to reduce the grading effort and increase the benefits associated with the rubric in Table 1, the initial round of exam grading only gave credit to correct solutions (Level I). With complete solutions in hand, students were encouraged to rework exam problems to locate their mistakes. If these mistakes fell into the category described in Level II of Table 1, then a written appeal could be submitted to obtain the partial credit defined in the rubric. More details about the rubric and the grading scheme are described in [6,7].

Multiple exam attempts. We assumed that students had many years of experience working within the paradigm of maximizing partial credit. It could be unreasonable, then, to insert these students into a course that requires mastery. Among other reasons, it is likely that they have not developed proper study habits or the skills necessary to review and correct their work during an examination. To account for this, three opportunities were provided on each of the four midterm exams. This meant that a total of twelve exams were administered during the semester, in addition to the final exam. For each of the four midterm exams, the final score was the sum of the best scores in each section (described below) from any of the three exam attempts. There was only one attempt on the final exam, which had a similar structure as the midterm exams.

Reasonable test difficulty and time. Instead of solving a small number of lengthy problems, as was common practice in previous versions of this course, in the current model the exam was divided into 4 sections.

> Section 1: Conceptual questions – Short questions that require little or no calculation but address a fundamental concept in the course.
>
> Section 2: Simple problems – Questions that require very little time to solve and involve only 1-2 computations. These problems may be very simple cases or address a specific step in the solution to a more complex problem.
>
> Section 3: Average problems – Medium length problems that contain no avoidable complications.
>
> Section 4: Challenge problem – One problem that requires a student to demonstrate a complete solution process for a problem that is more complex than those in Section 3.

Additionally, exams were written to be short, with a goal of taking only about 2/3 of the allotted time. The remaining 1/3 was left for students to review and correct their work. The total exam time was 90 minutes.

Variation

The second feature of the course design was variation. This means that each exam was created from scratch by the professors. Each one of the three versions of an exam covered the same concepts and solution processes, but with a reasonably broad variation in the specific problems that were used in the assessment. The purpose of this variation was to decrease the value of memorizing problems.

As the benefits of problem memorization were removed, students needed to be directed toward a more effective approach to learning. To this end, we created "The Compass". The Compass is a detailed, step-by-step, problem solving process for each type of problem in the course. Throughout the semester, we trained our students to use the Compass to map out their solutions. This is a key part of the current course design, as it gives students a way forward that is productive and helpful.

No Graded Homework

The third feature is that homework was not collected. While we did not collect homework, we did strongly emphasize practice. The goal here was to promote meaningful practice, not copying homework solutions for the sake of a homework grade. Students were provided with a list of suggested practice problems, many with fully worked solutions that followed the Compass. These problems were not collected, but students were told that success in the course depended on them using the practice problems correctly to prepare for exams.

We do not suggest that the details of the course design presented here are the ideal or only solutions to the observed student behaviors. Rather, we present them as an attempt to investigate what types of changes in course design might give rise to more beneficial student learning practices as well as improved overall performance.

3. The Current Study

In the fall semester of 2016, the authors conducted an experiment in three sections of Mechanics of Deformable Solids (course code ME222 at XXX), often referred to as Mechanics of Materials. The three sections were taught by different instructors. Two of the sections used the modified course design and assessment approach as described above, while the third section (control) used a course model and assessment approach that mirrors the current standard. Briefly, the standard course design used in Section 2 employed graded homework, two midterm exams, and the use of partial credit based on "correct approach" in the grading of every assignment.

A common final exam was administered across all three sections to determine the effects of the different assessment approaches used during the semester. The three course sections took the final exam at the same time, but in different rooms.

4. Results

In this section data are presented as mean $\pm$ standard deviation of the mean and were evaluated using one-way ANOVA (analysis of variance) when comparing more than two groups, and using a Student's t-test when comparing two groups.

Comparison of Student Groups Entering the Study

Student records were obtained and analyzed to determine if there were any significant differences among the students in each of the three sections as they entered the course. In particular, we measured differences among incoming cumulative GPA's and the grades in the prerequisite course Statics (course code CE221 at XXX). A summary of information for each of the three sections is provided in Table 2.

Table 2. Summary of information for each section of the course.

	Section 1	**Section 2**	**Section 3**
Number of Students	50	63	61
Median Cumulative GPA before FS16*	3.23	3.50	3.32
Median CE221 Grade*	3.00	3.50	3.00
Course Assessment Model	Modified	Control	Modified

* On a 4-point scale.

The data in Table 2 and in Figures 1 and 2 below show that, prior to the experiment, there were no significant variations in the student populations in the three sections, though the students in Section 2 have slightly higher median scores than those in Sections 1 and 3.

Figure 1 shows a comparison of the cumulative GPA (on a 4-point scale), calculated at the end of the summer semester 2016 (before the beginning of the experiment), of the students attending ME222 in the fall of 2016. The data are shown for Sections 1, 2, and 3. The mean value of the cumulative GPA of the students attending Section 2 (the control) was significantly higher when compared to the students attending Sections 1 and 3. The statistical comparison of the data show weak significance ($p<0.08$ calculated using one-way ANOVA).

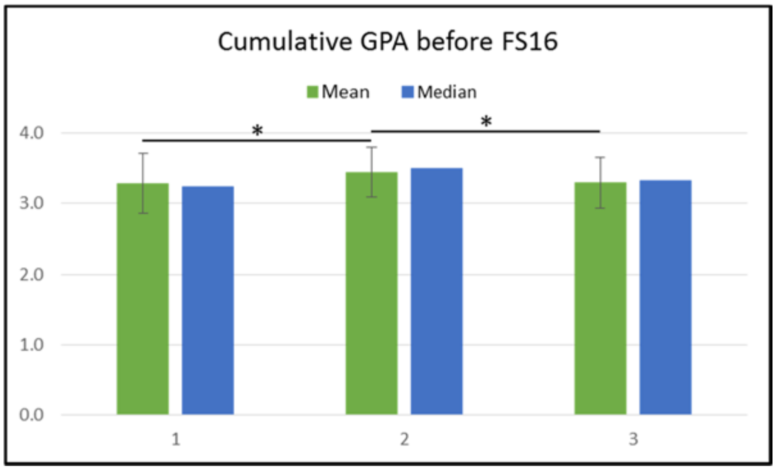

Figure 1. Cumulative GPA (on a 4-point scale) of the students attending ME222 in the fall semester of 2016. All data are presented as mean ± standard deviation of the mean (in green), and median (in blue). The asterisks represent statistically significant differences (* represents $p< 0.08$, weak significance calculated using one-way ANOVA).

Figure 2 shows the grade (on a 4-point scale) achieved by the same cohort of students in the course CE221: Statics, which is a prerequisite to ME222. No statistical difference was found between the three sections, though the mean and median grades are higher for Section 2. The number of credits enrolled in by the students in each Section was also compared and showed no statistical difference (data not shown).

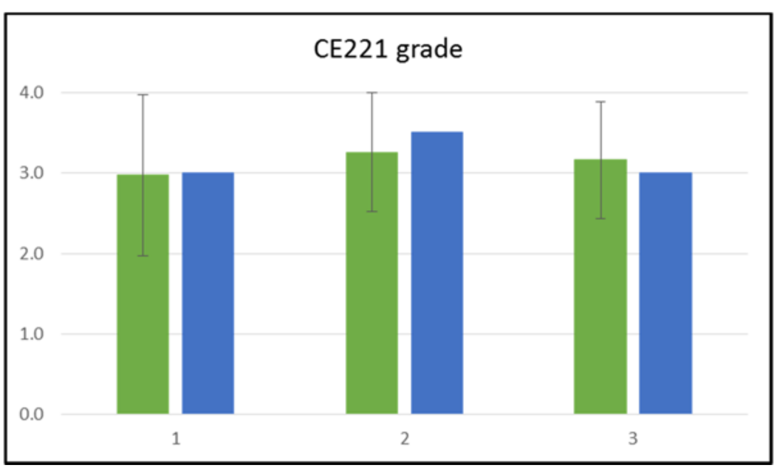

Figure 2. Grades in the prerequisite course CE221: Statics for the students attending ME222 in the fall semester of 2016. All data are presented as mean ± standard deviation of the mean (in green), and median (in blue).

Final Exam Results

The final exam was graded twice – once using the rubric in Table 1 and again using the control group's rubric based on partial credit.

Figure 3 shows the scores achieved in the final exam for the course ME222 in the fall semester of 2016, using the grading rubric presented in Table 1. The mean value of the grade achieved in the final exam by the students attending Section 2 (control) was significantly lower ($p<0.00001$ using one-way ANOVA) when compared to the Students attending Sections 1 and 3. No statistical difference was found between the Sections 1 and 3.

Among the two sections using the modified assessment approach across the semester, the final exam mean scores differed by 2.3 points (out of 100), and the median scores differed by 3 points. Compared to the control section, the final exam mean scores in the sections using the modified assessment approach were 30.2 points higher (out of 100), while the final exam median scores were 30 points higher. Under commonly used grading levels, this is a difference of three letter grades.

When the exams were graded using the control group's rubric based on partial credit (data not shown in figures), similar results were obtained. Among the two sections using the modified assessment approach across the semester, the final exam mean scores differed by 1.2 points (out of 100), and the median scores differed by 1 point. Compared to the control section, the final exam mean scores in the sections using the modified assessment approach were 26.1 points higher (out of 100), while the final exam median scores were 26 points higher. Under commonly used grading levels, this is a difference of two and one-half letter grades.

These results suggest a high impact of the modified course model and grading technique presented here. Further, the data show that equal benefits were obtained for two different instructors.

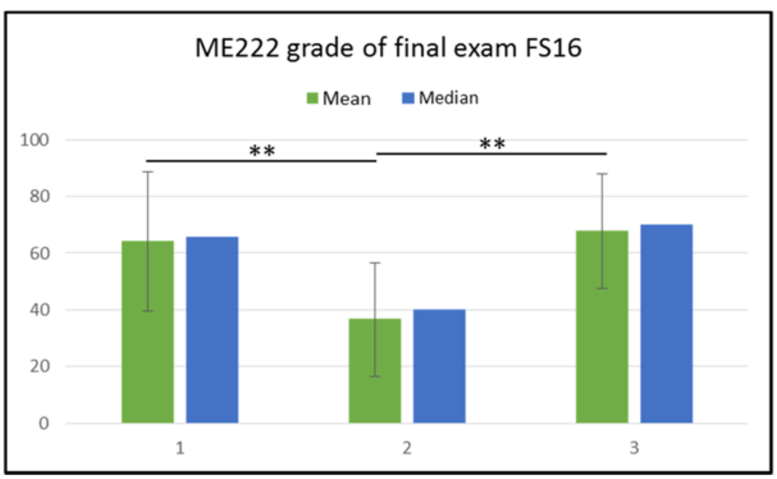

Figure 3. Grades for the final exam of the course ME222 in the fall of 2016. All data are presented as mean ± standard deviation of the mean (in green), and median (in blue). The asterisks represent statistically significant differences (represents p< 0.00001, strong significance calculated using one-way ANOVA).**

Repeatability of the Experiment

In the fall of 2017 a similar experiment was conducted. The same instructors that taught Section 1 and Section 2 in the original experiment during fall 2016 again taught Section 1 and Section 2 in fall of 2017. There was not a third section, so each of the sections had approximately 50% more students than in the previous year. Both instructors made small changes to the instructional model based on the results of the previous experiment, so the details of the course models are not exactly the same as in the original experiment. However, each section used essentially the same assessment approaches as they did in the previous experiments, so the results are still very relevant to the current discussion.

Figure 4 shows the grades achieved in the final exam for the course ME222 in the fall semester of 2017. Data are shown here for Sections 1 and 2, where Section 2 was again the control. Students attending Section 2 achieved significantly lower grades (p<0.00001 using Student's t-test) when compared to Section 1.

Compared to the control section, the final exam mean score in the section using the modified assessment approach (Table 1) was 15.8 points higher (out of 100), while the final exam median score was 21 points higher.

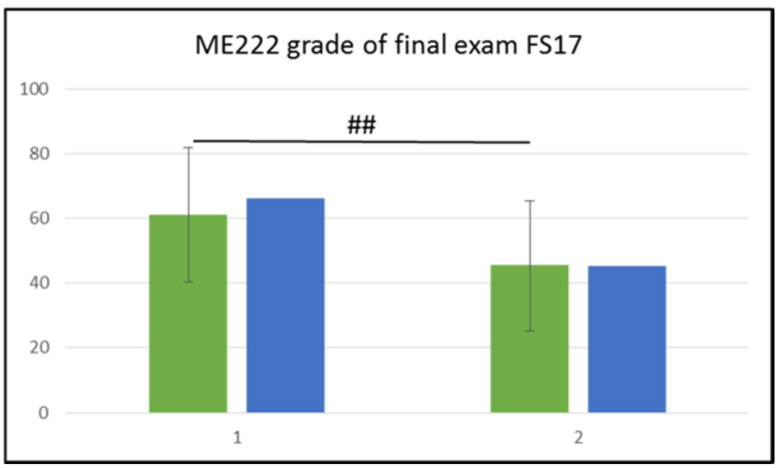

Figure 4. Grades for the final exam of the course ME222 in the fall of 2017. All data are presented as mean ± standard deviation of the mean (in green), and median (in blue). The asterisks represent statistically significant differences (## represents p< 0.00001, strong significance calculated using Student's t-test).

5. Additional Observations

When making significant changes to a course design, it is important to take a holistic approach. As the above modified course design was developed, we tried to anticipate issues that might result from any particular change and address them in a way that gave students the best chance of success in the new system. Nevertheless, we did observe the following unintended consequences of our course re-design.

Student stress. We presumed that students would panic upon learning that they would be required to obtain the correct answers on exams. We successfully mitigated that panic with a detailed discussion of why we were using this approach and how we were using multiple exams to minimize adverse effects to their GPA. However, having multiple exams meant that students took about one exam each week. By the end of the semester this created some fatigue and related stress in the student population. Additionally, many students were still trying to rely on memorization rather than follow a process based on the Compass, so their stress grew as their grades reflected that memorization would not work.

Forced practice and spaced repetition. Though it was not the primary intent, we realized midway through the semester that the weekly exams had become an intense forced practice session. During exams, students would sit uninterrupted for 90 minutes once a week to work on course problems. The repetition and spacing of these sessions likely contributed to improved learning due to increased levels of retrieval practice [4,5,8,9].

Interleaving of material. Common student mistakes were tracked for each exam. When a common mistake was observed in an average or challenge problem, the next exam would often have a simple question that was designed to address that specific concept. This naturally required us to include earlier concepts on exams, creating a constant interleaving of course material, a practice that has been shown to be very beneficial to learning [4,5,10-12]. Moreover, the exam schedule itself required us to test old material while teaching new material, adding to the amount of interleaving in the course.

Organization of solutions. An important observation we made at the end of the semester is that student solutions had become very orderly and followed a logical flow. This made grading easier and allowed us to see where a student's thinking was fuzzy. We believe that the consistent use of the Compass for instructor solutions to practice problems was mimicked by students trying to organize their own thoughts.

Lecture time. With 90 minutes of class time each week devoted to testing, lecture time was significantly reduced under the new model. This made it necessary to focus on the most important topics and example problems in class, and to assign readings and practice problems for out of class work. The measured improvements in performance occurred despite the reduction in lecture time compared to the control group.

6. Conclusions

The results of this study strongly support the hypothesis that students will adapt their study habits to changing assessment strategies. Compared to students in a control group that used conventional assessment methods (i.e., partial credit on exams based on "correct approach"), students in two course sections that used a mastery style assessment scored approximately three letter grades higher. These results also suggest that establishing higher performance standards in conjunction with increasing practice using a compass-guided solution process can lead to significantly higher student performance on exams.

7. References

[1] T. Pachur and H. Olsson, "Type of Learning Task Impacts Performance and Strategy Selection in Decision Making," *Cognitive Psychology*, Vol. 65, pp. 207-240, 2012.

[2] M.L. Gick and K.J. Holyoak, "Schema Induction and Analogical Transfer," *Cognitive Psychology*, Vol. 15, pp. 1-38, 1983.

[3] M.A. McDaniel, M.J. Cahill, M. Robbins and C. Wiener, "Individual Differences in Learning and Transfer: Stable Tendencies for Learning Exemplars versus Abstracting Rules," *Journal of Experimental Psychology: General*, Vol. 143, pp. 668-693, 2014.

[4] Peter C. Brown, Henry L. Roediger III and Mark A. McDaniel, *Make It Stick: The Science of Successful Learning*, Cambridge, MA: The Belknap Press of Harvard University Press, 2014.

[5] James M. Lang, *Small Teaching: Everyday Lessons from the Science of Teaching*, San Francisco, CA: Jossey-Bass, 2016.

[6] http://newleafedu.com/2018/02/27/grading/

[7] http://newleafedu.com/2018/03/13/rubrics/

[8] H.L. Roediger III and A.C. Butler, "Testing Improves Long-Term Retention in a Simulated Classroom Setting," *European Journal of Cognitive Psychology*, Vol. 19, pp. 514-527, 2007.

[9] B. Rogerson, "Effectiveness of a Daily Class Progress Assessment Technique in Introductory Chemistry," *Journal of Chemical Education*, Vol. 80, pp. 160-164, 2003.

[10] K.C. Bloom and T.J. Shuell, "Effects of Massed and Distributed Practice on the Learning and Retention of Second-Language Vocabulary," *Journal of Educational Research*, Vol. 74, pp. 245-248, 1981.

[11] D. Rohrer and K. Taylor, "The Shuffling of Mathematics Problems Improves Learning," *Instructional Science*, Vol. 35, pp. 481-498, 2007.

[12] S.K. Carpenter and F.E. Mueller, "The Effects of Interleaving versus Blocking on Foreign Language Pronunciation Learning," *Memory and Cognition*, Vol. 41, pp. 671-682.