Paper Title Improving Mathematical Problem Solving in an Introductory Engineering Course With the Testing Effect

Author(s) Jason W Morphew, University of Illinois at Urbana-Champaign; Mariana Silva, University of Illinois at Urbana-Champaign; Geoffrey Herman; Matt West, University of Illinois at Urbana-Champaign

Session Title Computational and Engineering Abilities in Higher Education and Adult Learners

Session Type Paper

Presentation Date 4/8/2019

Presentation Location Toronto, Canada

Descriptors Higher Education, Science Education, Testing

Methodology Quantitative

Unit Division C - Learning and Instruction

DOI 10.302/1431147

Improving Problem-Solving in an Introductory Engineering Course with the Testing Effect

Literature Review

Research and reform efforts to improve science, technology, engineering, and mathematics (STEM) education have largely focused on eschewing the traditional lecture in favor of active learning (e.g., Freeman, 2014). However, comparatively little attention has been paid toward transforming the traditional assessment paradigm used in introductory STEM courses. This is due, in part, because examinations are generally viewed as methods to measure learning rather than as a mechanism to enhance learning (Hartwig & Dunlosky, 2012). The lack of attention on assessment suggests a need for research into ways to apply findings from cognitive psychology about the testing effect as a critical means for improving students' learning and persistence in STEM.

Testing effect

Retrieval practice, often in the form of test taking, has been shown to produce better long-term retention in both clinical studies (Darley & Murdock, 1971; Roediger & Butler, 2011), as well as secondary and university classrooms (Bangert-Downs, et al., 1991; McDaniel, et al., 2013), compared with re-studying materials. For example, McDermott et al. (2014) utilized a within-subjects design with middle school students where the course material was randomly assigned to be either tested, restudied, or not tested or restudied. Students recalled facts at a higher rate for course material that was tested than for course material that was either restudied or not tested.

In the laboratory, much of the research concerning the testing effect has focused on memory tasks, while research in the classroom has utilized content focused on declarative memory tasks, such as word pairs in second language learning (Kang, et al., 2013), factual recall in psychology (McDaniel, et al., 2012), short answer questions in medical education (Larsen, et al., 2009), recalling facts from a lecture (Butler & Roediger, 2007), and multiple choice questions involving recalling or applying definitions (McDaniel, et al., 2011).

Although most research concerning the testing effect has used either identical, or very similar, questions for both practice and testing, a few studies have demonstrated improvements for rephrased questions (e.g., McDaniel, et al., 2007), analogical problem-solving tasks (Peterson & Wissman, 2018), and inferential and application questions covering previously tested material even when testing using questions involving factual recall (Butler, 2010; Thomas, et al., 2018). However, it is unclear whether testing can facilitate transfer to new, or untested material like would be found on final exams (e.g., Little, et al., 2012; Wooldrige, et al., 2014).

The benefit of testing on calculation-based problem-solving tasks, such as those found in introductory STEM courses, is less clear. Few laboratory studies have examined the benefits of testing in mathematical problem-solving contexts with most of them finding no evidence of a testing effect (e.g., Leahy, et al., 2015; van Gog, et al., 2011). For example, van Gog et al. (2015) found no advantage for testing over repeated studying for problem-solving tasks involving electrical circuits or probability distributions. In contrast, Blinded (2018) found that students completing multiple practice tests before an exam improved their exam scores relative to those who did not complete practice tests.

Assessment Frequency

The impact of increasing the number of assessments within a course is mixed. Laboratory studies have generally found that increasing the number of retrieval attempts leads to greater retention (e.g., Vaughn & Rawson, 2011). However, the evidence concerning the effect of increasing assessments frequency in classroom contexts is mixed (Bangert-Drowns, et al., 1991; Deck, 1998; Downs, 2015). The effect of increased assessment frequency on learning within courses that involve mathematical problem-solving is less clear. Two studies have found that weekly or daily tests lead to better performance on a final exam when compared to restudy or one midterm exam, but not when compared to three midterm exams (Pikunas & Mazzota, 1965; Townsend & Wheatley, 1975). However, several studies have found no effect of increased testing (Dineen, et al., 1989, Stephens, 1977; Ward, 1984).

Research suggests that testing facilitates deep processing of the material, strengthen pathways for correctly recalled information, weaken pathways for information which was incorrectly recalled, and increases the associations between the components of the target information (Bjork, 1975; Hopper & Huber, 2018; Kornell, et al., 2009). In addition, providing students with the opportunity and motivation to restudy the material on which they were tested appears to enhance the extent that learning occurs from retrieval practice (Pyc & Rawson, 2010; Rawson & Dunlosky, 2012). Enhanced learning after testing is thought to occur via test-potentiated learning, a term used to describe the beneficial effects of studying following retrieval practice, which seems to make studying more effective when it occurs after testing (Soderstrom & Bjork, 2014; Chan, et al., 2018).

Research in cognitive science suggests that instructors should engage students in more frequent testing that encourages students to reflect on and use feedback to enhance their learning. However, the effect of increased testing in classroom contexts utilizing mathematical problem-solving is unclear. In addition, there is a lack of empirical evidence concerning the effect of formative assessment on learning (Dunn & Mulvenson, 2009). The study described in this paper focuses on the following research question. To what extent does shorter and more frequent testing that incorporates aspects of formative assessment improve learning and retention in an engineering course as measured by the final exam?

The goal of this study was to examine the effects of testing frequency on problem-solving in an introductory engineering course, we employed a quasi-experimental design that prioritized the ecological validity and avoided ethical dilemmas arising from randomly using different assessment strategies for students taking the same course. Although random assignment was not used, the ecological validity and large sample size provides a controlled and well-powered study, such that the findings may be considered robust.

Methods

Participants

Four hundred eighty students enrolled in an introductory solid mechanics course at a large Midwestern university across two semesters were used for this study. Descriptive statistics are presented in Table 1.

Procedure

Students in the less-frequent assessment semester completed two exams while students enrolled in the course during the more-frequent assessment semester completed seven bi-weekly exams and five quizzes. In addition, during the more-frequent assessment semester, students could retake a different version of any exam the week after each of the seven bi-weekly exams.

The instructor, course content, final exam, and other course activities were kept constant except for homework. Fewer homework assignments were assigned during the more-frequent assessment semester and was graded for technical writing and group collaboration skills rather than content. The course, homework, and exam formats by semester are summarized in Table 1.

Results

Effect of Frequent Testing on Final Exam Performance

Students scored almost seven percentage points higher on the final exam when assessed more frequently (83.6%, SD = 13.0 compared to 76.7%, SD = 14.0). A Shapiro-Wilk test of normality indicated that the distribution of exam grades deviated from normality, and Levene's test indicated homogeneity of variance between semesters. However, a one-way ANOVA was conducted on the final exam score since ANOVA is robust to minor deviations from normality with large sample sizes. The ANOVA indicated that students scored higher on the final exam when assessed more frequently, $F(1, 478) = 31.14$, $p < .001$, with a medium effect size, $d = .51$.

The distribution of letter grades earned on the final exam are shown in Figure 1. An overall Chi-Square test of independence indicated that the two grade distributions differed, $\chi^2(3) = 33.21$, $p < .001$. Post-hoc Chi-Square tests indicated that c were 2.0 times more likely to receive an A, $\chi^2(1) = 20.03$, $p < .01$, $OR = 2.58$, and 2.4 times less likely to receive a failing grade (70% or lower), $\chi^2(1) = 23.58$, $p < .01$, $OR = 0.30$, on the final exam.

Final Exam Performance Controlling for Demographics and Differences in Ability

The students enrolled in both semesters had similar demographic profiles across, however a Chi-Square test of independence indicated that student grades in the prerequisite course differed between semesters, $\chi^2(4) = 12.4$, $p < .05$, with students in the more-frequent testing semester more likely to earn an A and less likely to earn a C. In addition, a Kruskal-Wallis test indicated that these students also had slightly higher ACT math scores, $\chi^2(1) = 11.16$, $p < .01$.

To control for demographic factors and student ability we conducted a hierarchical regression analysis. Because the data is skewed with a ceiling effect, beta regression models were fit with proc glimmix with SAS version 9.4 (see Table 2). Beta regression models values between 0 and 1, so the percentage on the final exam was expressed as a decimal and 0.01 was subtracting from all values to avoid scores at the boundaries (Smithson & Verkuilen, 2006). ACT math scores were centered to avoid issues of multicollinearity. The models indicate that students assessed with the more frequent exam schedule scored higher on the final exam even after controlling for demographics and differences in prior ability. In addition, the interactions between measures of student ability and semester were not significant, indicating that testing appears to benefit students of all ability levels.

Discussion

In this study, students who were assessed with the more frequent exam schedule scored higher on the final exam even when controlling for demographics and differences in ability, scoring about seven percentage points higher on average. More importantly, the percentage of students receiving A's was twice as large, and the percentage of students receiving failing grades was more than two times lower under the more frequent assessment schedule. This is an important threshold in many introductory courses because students do not earn course credit for earning either D’s or F’s, and are less likely to persist in the major if they earn low grades in introductory courses (King, 2015; Cromley, et al., 2016).

Little research has investigated individual differences in test-enhanced learning. For example, Carpenter et al. (2016) note that testing, compared to restudying, was better for high-performing students, but not for middle- and low-performing students. Conversely, Rawson, et al. (2013) found no differences in learning from using testing across ability groups. In the current study, the findings suggest that both low- and high-performing students benefitted from increased testing on the final exam.

Prior studies have found that assessment schedules incorporating more frequent exams often lead to better performance on those exams, however the improved performance is not always observed on a comprehensive final exam (e.g., Dineen, et al., 1989). In addition, much of the research concerning the testing effect has focused on memory tasks (e.g., Roediger & Karpicke, 2006b), multiple-choice application or inference questions (e.g., Thomas, et al., 2018), or reading comprehension (Karpicke & Aue, 2015). The present investigation extends this work by examining the testing effect in a semester-long engineering course that requires students to engage in mathematical problem-solving. Coursework in engineering requires students to be able to construct well-developed and connected conceptual understandings, in addition to building the strong retrieval pathways necessary for memory tasks, to successfully solve problems typically found in introductory engineering courses.

It is possible, that the combination of increased frequency of assessments with second-chance testing may provide additional benefits that explain why prior studies have not documented a testing effect for mathematical problem-solving tasks. Although we cannot disentangle the effects because we changed both aspects of the assessment schedule at the same time, we offer some theoretical reasons why the combination of increased frequency and second-chance testing may provide additional benefit for students’ learning.

The use of second-chance testing likely motivates students to engage in studying the exam material after receiving feedback on their performance. Evidence for this interpretation comes from the fact over 70% of the students elected to attempt the second try on each exam, including students who earned high (or even perfect) scores on their first attempt. Second chance testing may have led students to take advantage of test-potentiated learning more than under a traditional assessment schedule. Regardless of the exact causality, the combination of both frequent testing and second-chance testing appears to be a promising new assessment paradigm that could dramatically improve outcomes in STEM education and merits more research.

References

Bangert-Downs, R. L., Kulik, J. A., & Kulik, C-L., C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research, 85*, 89-99.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. Solso (Ed.), *Information Processing and Cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Erlbaum.

Butler, A. C. (2010). Repeated testing produced superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1118-1133. DOI: 10.1037/a0019902

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review, 24*, 369-378. DOI: 10.1007/s10648-012-9205-z

Carpenter, S. K., Lund, T. J., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review, 28*, 353-375. doi: 10.1007/s10648-015-9311-9

Chan, J. C. K., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective. *Journal of Memory and Language, 102*, 83-96. DOI: 10.1016/j.jml.2018.05.007

Cromley, J. G., Perez, T., & Kaplan. A. (2016). Undergraduate STEM achievement and retention: Cognitive, motivational, and institutional factors and solutions. *Policy Insights from the Behavioral and Brain Sciences, 3*, 4-11. DOI: 10.1177/2372732215622648

Darley, C. F., & Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology, 91*, 66-73. DOI: 10.1037/h0031836

Deck, D. W. (1998). *The Effects of Frequency of Testing on College Students in a Principles of Marketing Course*. (Doctoral dissertation). Retrieved from http://theses.lib.vt.edu/theses/available/etd-110298-195932/unrestricted/dis.pdf.

Dineen, P., Taylor, J., & Stephens, L. J. (1989). The effect of testing frequency upon the achievement of students in high school mathematics courses. *School Science and Mathematics, 89*, 197-200.

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

Dunn, K. E., & Mulvenson, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research, & Evaluation, 14*, 1-11.

Freeman, S., Eddy, S. L., McDnough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 8410-8415. DOI: 10.1073/pnas.1319030111

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin and Review, 19*, 126-134. DOI: 10.3758/s13423-011-0181-y

Hopper, W. J., & Huber, D. E. (2018). Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects. *Journal of Memory and Language, 102*, 1-15. DOI: 10.1016/j.jml.2018.04.005

Kang, S. H. K., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review, 20*, 1259-1265. DOI: 10.3758/s13423-013-0450-z

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27*, 317-326. DOI: 10.1007/s10648-015-9309-3

King, B. (2015). Changing college majors: Does it happen more in STEM and do grades matter? *Journal of College Science Teaching, 44*, 44-51. DOI: 10.2505/4/jcst15_044_03_44

Kornell, N., Hays, M. J., & Bjork, R. (2009). Unsuccessful Retrieval Attempts Enhance Subsequent Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 989-998. DOI: 10.1037/a0015729

Larsen, D. P., Butler, A. C., & Roediger, H. L. III. (2009). Repeated testing improves long term retention relative to repeated study: A randomised controlled study. *Medical Education, 43*, 1174-1181. DOI: 10.1111/j.1365-2923.2009.03518.x

Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review, 27*, 291-304. DOI:10.1007/s10648-015-9296-4.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23*, 1337-1344. DOI: 10.1177/0956797612443370

McDaniel, M. A., Agarwal, P. K., Huesler, B. J., McDermott, K. B., & Roediger, H. L. III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*, 399-414. DOI: 10.1037/a0021782

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*, 494-513. DOI: 10.1080/09541440701326154

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*, 360-372. DOI:10.1002/acp.2914

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition, 1*, 18-26. DOI: 10.1016/j.jarmac.2011.10.001

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L. III., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology, Applied, 20*, 3-21. DOI: 10.1037/xap0000004

Morphew, J.W., Mestre, J.P., Chang, H-H., Kang, H-A., & Fabry, G. (under review). Using computer adaptive testing to assess physics ability and improve exam performance in an introductory physics course.

Peterson, D. J., & Wissman, K. T. (2018). The testing effect and analogical problem-solving. *Memory*, DOI: 10.1080/09658211.2018.1491603

Pikunas, J., & Mazzota, D. (1965). The effects of weekly testing in the teaching of science. *Science Education, 49*, 373-376.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*, 335. DOI: 10.1126/science.1191465

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

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

Roediger, III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*, 20-27. DOI: 10.1016/j.tics.2010.09.003

Roediger, H. L. III., & Karpicke, J. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181-210. DOI: 10.1111/j.1745-6916.2006.00012.x

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological Methods, 11, 54-71. doi: 10.1037/1082-989X.11.1.54

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language, 73*, 99-115. DOI: 10.1016/j.jml.2014.03.003

Stephens, L. J. (1977) The effect of the class evaluation method on learning in certain mathematics courses, *International Journal of Mathematical Education in Science and Technology, 8*, 477-479, DOI: 10.1080/0020739770080414

Thomas, R. C., Weywadt, C. R., Anderson, J. L., Martinez-Papponi, B., & McDaniel, M. A. (2018). Testing encourages transfer between factual and application questions in an online learning environment. *Journal of Applied Research in Memory and Cognition*, *7*, 252-260. DOI: 10.1016/j.jarmac.2018.03.007

Townsend, N. R., & Wheatley, G. H. (1975). Analysis of frequency of tests and varying feedback delays in college mathematics achievement. *College Student Journal*, *9,* 32-36.

van Gog, T., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., & Verkoeijen, P. P. J. L. (2015). Testing after worked example study does not enhance delayed problem-solving performance compared to restudy. *Educational Psychology Review, 27*, 265-289. DOI: 10.1007/s10648-015-9297-3

van Gog, T., Kester, L., & Paas F. L. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology, 36*, 212-218. doi: 10.1016/j.cedpsych.2010.10.004

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22*, 1127-1131. DOI: 10.1177/0956797611417724

Wooldrige, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition, 3*, 214-221. DOI: 10.1016/j.jarmac.2014.07.001

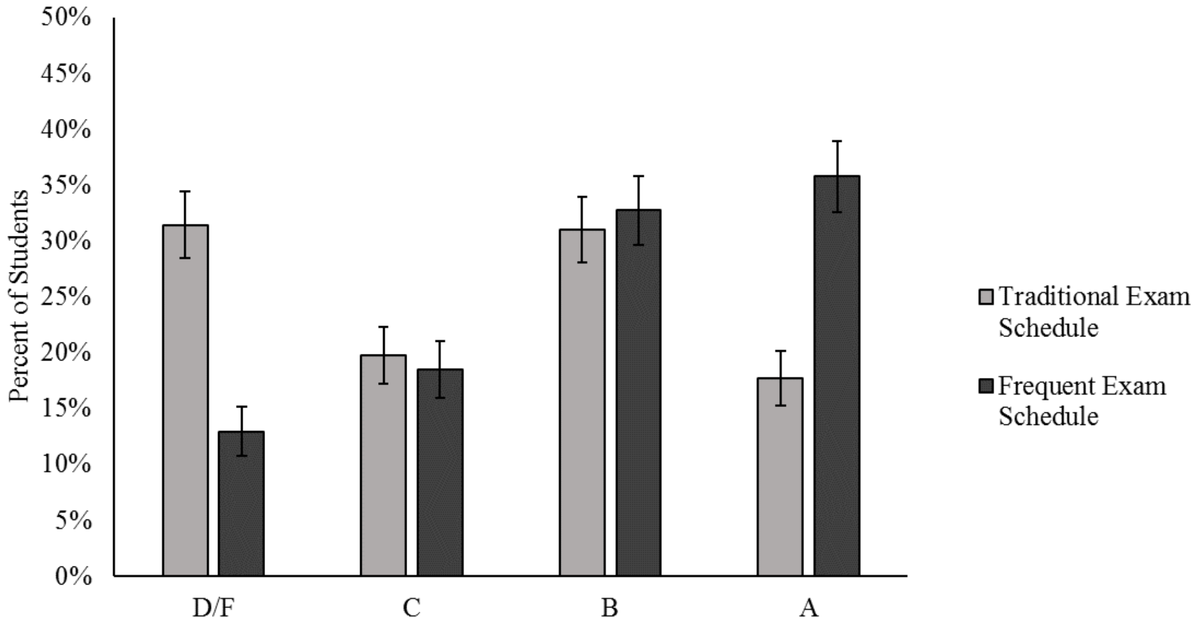

Figure 1: Letter grade distribution on final exam by semester

Table 1

Descriptive Statistics for Gender and Ethnicity, Course Format, Homework and Assessment Question Types by Semester

	Less Frequent Exams	More Frequent Exams
Demographics		
Number of students	248	232
Student gender	20% female, 80% male	21% female, 79% male
Student ethnicity	11% Asian American, 58% Caucasian, 20% International, 10% URM	15% Asian American, 47% Caucasian, 28% International, 9% URM
Prior Ability		
Prerequisite course grade	2.98; 95% CI: [2.87-3.09]	3.19; 95% CI: [3.06-3.32]
ACT math score	31.8; 95% CI: [31.3-32.2]	32.9; 95% CI: [32.4-33.3]
Exam Format		
Number of exams	2 required exams	7 required exams, 7 optional second-chance exams, 5 in-lecture short quizzes
Exam length	2 hours	50 minutes
Exam format	Pen-and-paper	Computerized (but pen-and-paper for in-lecture quizzes)
Exam Feedback	Per-question and total scores available after about 1 week	Immediate per-question and total scores provided during exam
Course Format		
Homework format	Weekly computer homework assignments and 9 written homework assignments	Weekly computer homework assignments and 4 written homework assignments
Lecture format	Three 50-minute lectures per week; same experienced instructor; same content; active learning lectures with clickers	Same Format
Recitation format	One 50-minute recitation per week; active collaborative problem solving in four-person teams; guided by graduate teaching assistants	Same Format

Note: URM – Under Represented Minority, Identical Final Exam were given

Table 2

Association between total exam score and exam schedule controlling for demographic variables and student ability.

	Model 1		Model 2		Model 3		Model 4		Model 5	
	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*
Gender	0.190 (0.08)	.03	0.127 (0.08)	.09	0.128 (0.08)	.09	0.090 (0.08)	.28	0.093 (0.08)	.26
International	0.148 (0.09)	.08	0.066 (0.08)	.39	0.067 (0.08)	.38	0.103 (0.17)	.54	0.096 (0.17)	.57
URM	-0.063 (0.12)	.59	-0.001 (0.11)	.99	0.004 (0.11)	.97	0.031 (0.10)	.76	0.032 (0.10)	.76
Prerequisite			0.517 (0.03)	< .001	0.531 (0.05)	< .001	0.518 (0.04)	< .001	0.519 (0.04)	< .001
ACT math							0.033 (0.01)	< .01	0.049 (0.02)	.02
Course	0.420 (0.07)	< .001	0.330 (0.06)	< .001	0.334 (0.07)	< .001	0.334 (0.07)	< .001	0.331 (0.07)	< .001
Prereq*Course					0.028 (0.07)	.68				
ACT math*Course									0.022 (0.02)	.39
Pseudo R^2	.07		.39		.39		.47		.47	
AIC	-679.55		-806.11		-804.28		-626.32		-625.13	
BIC	-654.51		-777.53		-771.62		-596.33		-591.38	
N	480		438		438		314		314	