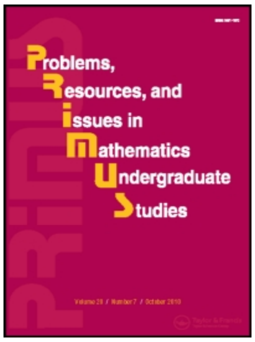

PRIMUS

Problems, Resources, and Issues in Mathematics Undergraduate Studies

ISSN: 1051-1970 (Print) 1935-4053 (Online) Journal homepage: https://www.tandfonline.com/loi/upri20

The Effect of Cumulative Tests on the Final Exam

Jonathan E. Beagley & Mindy Capaldi

To cite this article: Jonathan E. Beagley & Mindy Capaldi (2016) The Effect of Cumulative Tests on the Final Exam, PRIMUS, 26:9, 878-888, DOI: 10.1080/10511970.2016.1194343

To link to this article: https://doi.org/10.1080/10511970.2016.1194343

Accepted author version posted online: 01 Jul 2016.
Published online: 13 Sep 2016.

PRIMUS, 26(9): 878–888, 2016

ISSN: 1051-1970 print / 1935-4053 online
DOI: 10.1080/10511970.2016.1194343

The Effect of Cumulative Tests on the Final Exam

Jonathan E. Beagley and Mindy Capaldi

Abstract: Mathematics teachers often give cumulative final exams, but little research has been done on the effects that cumulative exams given throughout the semester have on student grades and content knowledge. This study went beyond a cumulative final exam, and investigated the benefits of cumulative versus non-cumulative semester exams on students' final exam scores. Comparing control and experimental sections of an introductory mathematics course showed a positive effect of cumulative exams on final exam scores. Additionally, low-scoring students in the experimental sections performed better on the final exam compared with the average of their semester exams than low-scoring students in the control section. This interaction was not statistically significant, however.

Keywords: Cumulative, testing effect, spacing, exams

1. INTRODUCTION

Many mathematics teachers require students to take cumulative final exams. Memory science research supports both this practice and the general use of frequent and spaced testing. In laboratory word list experiments, expectation of a cumulative final test increased memory processes and resulted in stronger recitation of material [16]. Even in the classroom, cumulative final exams led to higher content exam scores from students, both immediately following the course and 18 months later [7]. Additionally, Lawrence [9] found additional evidence of the benefits of cumulative exams for introductory psychology students. Students who were given cumulative exams throughout the semester had significantly higher scores (p = 0.035) than students who only had a cumulative final exam [9, p. 17]. The aim of this study is to extend the results of Lawrence to a mathematics course. Throughout this paper, the term

Address correspondence to Jonathan E. Beagley, Department of Mathematics and Statistics, 1900 Chapel Dr., Valparaiso University, Valparaiso, IN 46383, USA. E-mail: jon.beagley@valpo.edu

"semester exams" will be used to distinguish tests taken throughout the semester from the final exam.

2. LITERATURE REVIEW

Often, teachers use exams simply as a method of assessing learning. However, the positive effects of testing on recall, or the *testing effect*, are well documented in the memory science literature. The very act of giving students tests can result in significant learning gains, as active retrieval of information leads to better retention [1, 5, 12]. Tests have proven more advantageous than additional opportunities to study in terms of boosting memory of material [5]. This gain holds even when feedback is not provided to the student [3]. As Pyc et al. [12, p. 78] state, "The benefits of practice tests have been documented hundreds of times and the benefits of tests may last up to nine months after information has been initially learned." Mathematics teachers, in particular, often promote the need to practice problems instead of merely reading notes or previously worked homework. This can be done by including both old and new material on exams, effectively making each exam cumulative. The theory is that repeated retrieval of information leads to more effective learning.

If we accept that the ability to recall information benefits from repeated retrieval, does it matter how that repetition is scheduled? Human cognition research indicates that it does [2, 6]. The *spacing effect* describes the learning advantage that occurs when repeated exposure to information is presented in intervals, called spacing gaps, rather than concentrated in one lesson or test. A typical spacing effect study would have experimental subjects initially learning information, after a gap of time reviewing it again, and then after a final gap (*testing delay*), they would be tested over the material. The control group would either study the information twice in a row or have a smaller spacing gap, and then be tested after the same final testing delay [2]. In both experimental and classroom settings, the spacing effect increases long-term retention of information [2, 6, 13, 15]. Combining testing with spaced intervals results in more gains [5]. Requiring additional cumulative exams throughout the semester increases the number of exam-initiated retrievals, which was shown to be beneficial by Lawrence [9].

Lawrence used one section of an introductory psychology course as a control group, which took three non-cumulative semester exams and one cumulative final exam. An experimental section of the class had cumulative portions on the second and third semester exams, specifically 10 out of 40 multiple choice questions. The experimental section performed significantly better on the cumulative part of the final exam. Additionally, students in each section were split into two groups, low-scorers and high-scorers, based on the median of the first exam. Low-scorers in the cumulative section had significantly ($p < 0.001$) higher course grades, meaning an average difference of a

C+ to a B, than their peers in the non-cumulative section. Results for high-scorers were not statistically significant [9, p. 17]. Differences in the effects of pedagogical practices on struggling students versus strong students has appeared in other educational research as well [8, 14].

Little empirical research has been done on using cumulative exams throughout a course. In a 2011 literature search, Lawrence [10] found no relevant results. No similar cumulative exam studies were found as of the current 2015 project in this paper. However, the testing effect was studied in an online college course where quizzes, with feedback, resulted in a stronger performance on subsequent exams than simply reading additional material [11]. Another interesting finding by McDaniel et al. [11] was that short-answer quizzes produced stronger benefits than those that were multiple-choice. Another study, Rohrer and Taylor [13], applied the spacing effect to college mathematics students, and found that distributing the practice of abstract mathematics problems across two sessions instead of one doubled the scores of students. Although this study aimed to extend the results of Lawrence [9] to a mathematics course, not all of the methodology was identical. The mathematics exams, for example, were short answer instead of multiple choice. Similarities between the two studies include examining differences on the final exam performance between cumulative versus non-cumulative course sections, as well as differences between low-achieving and high-achieving students.

3. METHODS

The study participants were students in an introductory course in mathematics for non-mathematics majors at a midsized liberal arts school in the Midwest. Students primarily take this course to fulfill a quantitative reasoning requirement for graduation. The course requires only algebra as a prerequisite, and covers topics including lines, systems of linear equations, linear optimization and the simplex method, probability, Markov chains, and game theory. The sample consists of three sections from the spring semester of 2015 (2015 A, 2015 B, 2015 C), as well as two sections from the fall semester of 2014 (2014 A, 2014 B). The sections 2015 A and 2015 B were taught by one instructor, the first author of this paper, and 2015 C, 2014 A, and 2014 B were taught by another instructor, the second author. Both instructors have taught this course a minimum of four times prior to the fall semester of 2014.

The course used three in-class semester exams and a cumulative in-class final exam. The cumulative final exam is the same for both semesters with a common rubric. All exam questions were short-answer type. Each of the semester exams in the control sections (2015 A, 2014 A, 2014 B) contained only problems from the newer material covered since the previous exam. There were 80 students in these three sections. In the two experimental sections, 2015 B and 2015 C, a topic from the first exam appeared on the second exam, and a topic from the second exam appeared on the third exam. There were 44 students in these two sections. A few students who withdrew from the course or did not take the final exam were not included in these numbers or the analysis of scores. Students did not know which older topic would be chosen for the cumulative question. The cumulative topic and question were agreed upon by both instructors, based on what material we knew would be important to later topics in the course. If students in the experimental sections reviewed the material from the first exam for the later three exams in the course, then there would be three spacing gaps and two additional testings for these topics. Similarly, studying the material that was new to the second exam for the third exam and final exam would result in two spacing gaps and one more testing compared with the control sections.

To keep the exams the same length between the experimental sections and the control sections, one question was replaced between the copies of the exams and was worth one tenth of the exam's points. Students in the two experimental sections were told that they would have cumulative semester tests on the first day of class, along with some explanation of the benefits. The 2015 A section was not told that other sections of the course were being given cumulative semester tests. Regardless of section, each student took the same final exam.

The dependent measures in this study were the scores on the cumulative final exam. We predicted that students in the experimental sections would outperform the control students on the final exam, which would be consistent with the results of Lawrence [9].

We took no formal measure of student attitudes about cumulative semester exams, though anecdotal evidence from students implied that the cumulative exams were helpful when studying for the final exam.

4. RESULTS AND DISCUSSION

Experience has shown that exam scores by different sections of students can vary dramatically. Therefore, the authors thought it best to compare students' performance on the semester exams with their final exam scores, as opposed to just comparing final exams across sections. This way, we would be able to account for students' previous knowledge about topics in the course. We made the distinction of a result being significant if a two-sample t-test, assuming equal variances, produced a $p < 0.05$. See Table 1 for a summary of the data.

4.1. Performance on Final Exam versus First Exam

Students' first semester exam scores were used to develop a baseline for their prior knowledge. The first exam did not have a cumulative question, as there was no older material, and thus was the same for experimental and control

Table 1. Summary of results

	Mean (M)	Standard deviation (SD)
First exam (experimental sections)	78.11	16.88
First exam (control sections)	81.99	13.45
Semester exam average (experimental sections)	75.75	15.33
Semester exam average (control sections)	80.06	11.38
Final exam scores (experimental sections)	65.05	19.10
Final exam scores (control sections)	63.96	15.49
Final exam - first exam (experimental sections)	−13.07	14.29
Final exam - first exam (control sections)	−18.03	14.39
Final exam - average semester exam (experimental sections)	−10.70	10.46
Final exam - average semester exam (control sections)	−16.10	11.96

students within each instructor's sections. When comparing the performance on the first exam for students in the experimental sections, we see that they did worse on that exam ($M = 78.11, SD = 16.88$) compared with the students in the control sections ($M = 81.99, SD = 13.45$). This was nearly a significant difference with $p = 0.08$.

The scores for the experimental sections were better on the final exam ($M = 65.05, SD = 19.10$) than the students in the control sections ($M = 63.96, SD = 15.49$). This was not a significant difference with $p = 0.37$. By taking the difference between the score on the final exam and the score on the first exam, we can see the impact of the intervention. In general (for all instructors at the university who teach the course), students' scores in this course drop from the semester tests to the final exam. The cumulative sections did better by this measure, with a smaller drop ($M = -13.07, SD = 14.29$), than the students in the control sections ($M = -18.03, SD = 14.39$). This was a significant difference with $p = 0.03$. The experimental, or cumulative, sections started out with more low-level students. However, these sections ended the semester relatively strongly.

4.2. Performance on Final Exam versus Semester Exam Average

Furthermore, if we want to compare the performance of the students over the entire semester, we could measure their learning by the means of their three semester exam scores. As with the first exam, students in the experimental sections scored lower when looking at their average semester exam grades ($M = 75.75, SD = 15.33$) than the students in the control sections ($M = 80.06, SD = 11.38$). This was a significant difference with $p = 0.04$.

We believe, as was seen on the first exam scores, that this is because the students in the experimental sections started out at a lower-level in terms of prior mathematics knowledge.

If we compare the difference between the score on the final exam with the average score on a semester exam, we again see the impact of cumulative semester exams. The students in the experimental sections had a smaller grade loss on the final $(M = -10.70, SD = 10.46)$ than the students in the control sections $(M = -16.10, SD = 11.96)$. This was a significant difference with $p < 0.01$.

4.3. Performance of High Scorers versus Low Scorers

Since pedagogical interventions can affect students differently, depending on their level, we also investigated this interaction. As we did not have access to any data about the students in this study, other than their scores in this course, one way to measure baseline knowledge was the first semester exam in the course. This first exam covered the same material regardless of section, whereas the second and third semester exams differed due to the addition of the cumulative question and the removal of another question. Therefore, as in Lawrence [9], we use the median score on the first exam as a method to distinguish between high-scoring students and low-scoring students.

To analyze this data, we used a 2 × 2 factorial ANOVA, using high scoring and low scoring as one factor, and experimental and control as the other factor.

Studying comparisons with the first exam did not reveal interesting results. If we use the measure of the difference between the score on the final exam and the first exam, there was no significant difference between high-scoring students in the experimental sections $(M = -16.19, SD = 13.77)$ and the high-scoring students in the control sections $(M = -20.99, SD = 13.73)$ with $p = 0.57$. Similarly, there was no significant difference between low-scoring students in the experimental sections $(M = -9.94, SD = 14.42)$ and the low-scoring students in the control sections $(M = -14.76, SD = 14.57)$ with $p = 0.58$. There was no interaction between the two factors with $p = 0.996$; an interaction plot for this is shown in Figure 1.

Consider, then, the average of the semester exams. For high-scoring students, using the measure of the difference between the score on the final exam and the average semester exam score, the experimental sections $(M = -11.19, SD = 11.43)$ performed better than the students in the control sections $(M = -15.13, SD = 10.94)$, but this was not a significant difference with $p = 0.56$. However, considering the low-scoring students using the same measure, we see that the students in the experimental sections $(M = -10.21, SD = 9.65)$ did better than the students in the control sections $(M = -17.17, SD = 13.06)$. This result was nearly significant with $p = 0.11$. There was some interaction between the factors with $p = 0.49$, but not a significant interaction as seen in Figure 2. One reason we did not see a statistically significant result in this factor analysis is that the sample sizes are small.

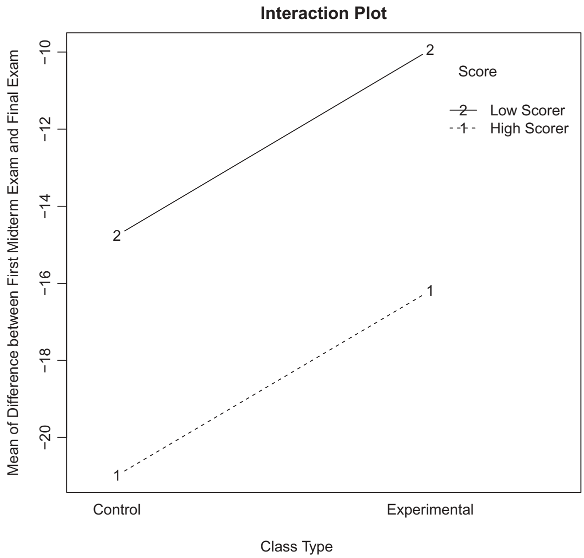

Figure 1. Interaction between difference of first exam score and final exam score.

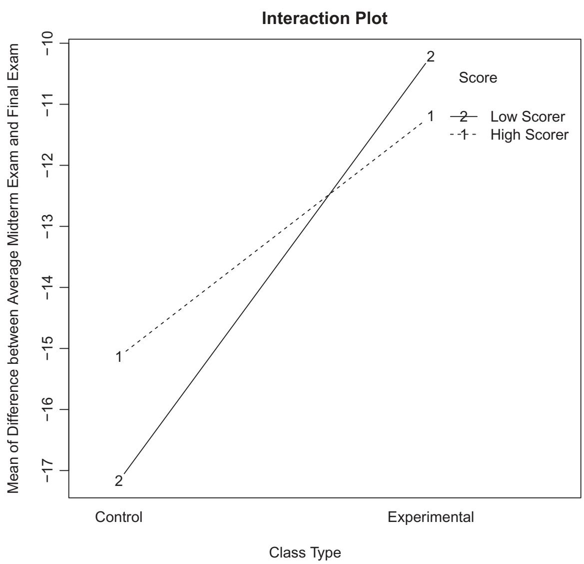

Figure 2. Interaction between difference of average semester exam score and final exam score.

4.4. Performance on Specific Questions on the Final Exam

Of particular interest are the two cumulative questions that were added to the second and third semester exams; topics which were then later covered on the final exam. Consider a question on duality of the simplex method. Out of a possible 17 points on the final exam, students in the experimental sections scored approximately 2.5 points higher than control section students ($M = 9.67, SD = 5.37$ compared with $M = 7.13, SD = 4.24$). This showed a significant difference of $p = 0.02$. This problem was the cumulative exam question on the third exam.

The other question, also out of a possible 17 points, covered finding the matrix inverse of a 2×2 matrix. Students in the experimental sections did better on this question ($M = 9.39, SD = 6.48$) than students in the control sections ($M = 9.36, SD = 5.88$). However, this was not a significant difference with $p = 0.49$.

Recall that on the semester exams the experimental sections had significantly lower scores, on average, than the control sections. This was true even for the first exam alone. For the experimental students to get scores on these two cumulative questions at the same level of, or significantly higher than, the control students indicates a positive testing effect.

5. CONCLUSIONS

The experiment in this study was analyzing the effect of cumulative semester tests on the cumulative final exam in an introductory mathematics course. The testing effect was applied by requiring students to retrieve material more often than just through studying for the cumulative final exam. Students were also given feedback, which was previously shown to increase testing effect benefits. These retrievals were scheduled with spacing gaps, which were the weeks-long time periods between exams. Across two instructors, two semesters, and multiple sections of the class, statistically significant results indicate that this intervention benefited students, in general. Low-scoring students had a non-significant advantage over high-scorers.

We saw specific positive effects on final exam scores for the topics chosen to be cumulative portions of the semester exams. However, only one of these topic questions had results that were significant. The stronger findings for the duality question could be due to the fact that this cumulative problem was on the third, more recent, exam. One finding in the research on the spacing effect is that the length of the spacing gap, relative to the testing delay, affects the amount of advantage that results to the learner [4]. The laboratory results of Cepeda et al. [4] indicate that the optimal spacing gap is 10–20% of the length of the testing delay. Realistically, in a classroom setting where cumulative semester exams are being used, those tests must be given with fixed gaps between them. Another approach could be to administer quizzes at beneficial points between exams.

It is possible that differences between the fall and spring semesters, or the instructors, could have affected students' performance. However, the authors feel that extending the study past one semester or a single instructor strengthens the argument that cumulative semester tests benefit students. Both instructors individually saw a smaller drop from the semester test average to the final exam score in their experimental sections compared with their control sections. Since one instructor had students only in the spring, and the other had fall and spring, the advantages of cumulative tests seem to hold regardless of semester.

In general, semester exam averages were lower for the experimental sections. Since this was a statistically significant difference on the first exam, it could just be a difference in prior knowledge or ability level. Another factor contributing to lower scores on the second and third exams might be the inclusion of a cumulative question, worth a tenth of the exam points. Since that portion of the exam was older material it could have proved more challenging, even for students who reviewed previous exam topics.

Future iterations of this experiment might include more, perhaps shorter, cumulative problems on the second and third tests. The authors would like to see further evidence of the effect of cumulative tests, especially in mathematics courses. Since significant benefits of cumulative tests have been found for psychology and mathematics students, it would also be useful to extend the results to other disciplines. Future research should aim to replicate results, perhaps with larger sample sizes or in higher-level courses. Neither Lawrence [9] nor this study experimented in courses above the introductory level. Additional studies of the long-term benefits of cumulative semester exams in mathematics would also be interesting; specifically, retesting students after the semester concludes. Including the effects of other types of "tests," such as quizzes, could reveal insightful conclusions about the effects of testing and spacing, as well.

ACKNOWLEDGEMENTS

The authors wish to thank the other members of the faculty learning community on applying memory science to the classroom: Nora Belzowski, Kieth Carlson, Melissa Desjarlais, Randa Duvick, Laurie Eberhardt, and Tiffany Kolba.

REFERENCES

1. Butler, A. C. 2010. Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 36: 1118–1133.

2. Carpenter, S. K. 2014. Spacing and interleaving of study and practice. In V. Benassi, C. Overson, and C. Hakala (Eds), *Applying Science of Learning in Education* (pp. 274–278). American Psychological Association. Retrieved from the Society for the Teaching of Psychology web site: http://teachpsych.org/ebooks/asle2014/index.php.
3. Carpenter, S. K., H. Pashler, J. T. Wisted, and E. Vul. 2008. The effects of tests on learning and forgetting. *Memory & Cognition*. 36(2): 438–448.
4. Cepeda, N. J., E. Vul, D. Rohrer, J. T. Wixted, and H. Pashler. 2008. Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*. 19: 1095–1102.
5. Cull, W. L. 2000. Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*. 4: 215–235.
6. Delaney, P. F., P. P. J. L. Verkoeijen, and A. Spirgel. 2010. Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *The Psychology of Learning & Motivation: Advances in Research & Theory*. 53: 63–147.
7. Khanna, M. M., A. S. Brack, and L. L. Finken. 2013. Short- and long-term effects of cumulative finals on student learning. *Teaching of Psychology*. 40(3): 175–182.
8. Kogan, M. and S. Laursen. 2014. Assessing long-term effects of inquiry-based learning: A case study from college mathematics. *Innovative Higher Education*. 39(3): 183–199.
9. Lawrence, N. K. 2013. Cumulative exams in the introductory psychology course. *Teaching of Psychology*. 40(1): 15–19.
10. Lawrence, N. K. 2014. Examining the benefits of cumulative exams. In V. Benassi, C. Overson, and C. Hakala (Eds), *Applying Science of Learning in Education* (pp. 274–278). American Psychological Association. Retrieved from the Society for the Teaching of Psychology web site: http://teachpsych.org/ebooks/asle2014/index.php
11. McDaniel, M. A., J. L. Anderson, M. H. Derbish, and N. Morrisette. 2007. Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*. 19: 494–513.
12. Pyc, M. A., P. K. Agarwal, and H. L. Roediger. 2014. Test-enhanced learning. In V. Benassi, C. Overson, and C. Hakala (Eds), *Applying Science of Learning in Education* (pp. 274–278). American Psychological Association. Retrieved from the Society for the Teaching of Psychology web site: http://teachpsych.org/ebooks/asle2014/index.php
13. Rohrer, D. and K. Taylor. 2006. The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology*. 20: 1209–1224.
14. Saville, B. K., D. Pope, J. Truelove, and J. Williams. 2012. The relation between GPA and exam performance during interteaching and lecture. *The Behavior Analyst Today*. 13(3): 27–31.

15. Sobel, H. S., N. J. Cepeda, and I. V. Kapler. 2011. Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*. 25(5): 763–767.
16. Szpunar, K. K., K. B. McDermott, and H. L. Roediger. 2007. Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition*. 35: 1007–1013.

BIOGRAPHICAL SKETCHES

Jonathan E. Beagley received his B.S. from Illinois Institute of Technology in Chicago, IL and, M.S. and Ph.D. at George Mason University in Fairfax, VA. His dissertation work was over discrete and convex geometry, but has broad interests in combinatorics and the scholarship of teaching and learning. He enjoys spending time with his wife, two cats, and dog.

Mindy Capaldi attended Georgetown College in Georgetown, KY before moving on to graduate school at North Carolina State University in Raleigh, NC. Although she studied algebraic topology for her dissertation, Mindy's interests now include mathematics education and the scholarship of teaching and learning. She likes to think about how to be a better teacher and bounce those ideas off of her husband, Professor Alex Capaldi. When not contemplating work, Mindy is usually playing with son Henry, reading, or cuddling with her cats.