A Method to the Midterms: The Impact of a Second Midterm on Students' Learning Outcomes

Kelly Keus, Jamie Grunwald, and Neil Haave*

University of Alberta, Augustana Campus

4901 – 46 Avenue, Camrose, AB, CANADA, T4V 2R3

*Primary contact: nhaave@ualberta.ca

Abstract

Midterm exams are a multi-use tool, providing evaluation of students for professors but also acting as a learning tool for students. Midterms may improve learning outcomes by contributing to the testing effect: the phenomenon in which retrieval of learned material (i.e., testing) produces improvements in long-term retention beyond those produced through additional rehearsal or re-exposure (i.e., studying or re-reading). Additionally, increased frequency of testing may impact student behaviors and attitudes (e.g., spaced practice, self-efficacy), increase the testing effect, or impact both, which ultimately improves learning outcomes. This study considered the differential impact of one versus two midterm exams on students' exam difference scores (final exam score minus first midterm exam score). We also considered whether two midterm exams differentially impacted low- and high-achieving students. Results suggest that two midterm exams benefit freshmen but not junior students.

Keywords: testing effect, frequency effect, midterm exam, student learning outcomes

Introduction

Midterm and final exams are common forms of assessment implemented in undergraduate university courses to determine the degree of students' mastery of course material. However, midterms can act as a multi-use tool, providing evaluation of students for professors but also acting as a learning tool for the students. Usually, courses will have one or more midterm exams spaced throughout the semester in addition to a final exam; these midterm exams may or may not be cumulative (Myers & Myers, 2007). Although there are anecdotal preferences for the number of midterms a course should have, there is limited research on the benefits of one versus two midterm exams on the outcome of students' final exam scores. Our study was designed to fill this gap in the research by considering whether a second midterm could improve student learning outcomes. Studies supporting midterm exams as a learning tool cover two broad areas of research: testing effects and frequency effects.

Testing Effects

Interest in the testing effect has generated significant research both in labs and classroom settings. The testing effect occurs when retrieval of learned material (i.e., testing) produces improvements in long-term retention beyond those produced through additional rehearsal or re-exposure (i.e., studying or re-reading) (Brame & Biel, 2015; Carpenter, 2012; Roediger & Butler, 2011). Early laboratory research on the testing effect was predictably structured (Carpenter, 2012). A learning phase allowed participants to encode the material. This was followed by a testing phase or re-study (control) phase allowing participants to either retrieve or re-read the material. Finally, a second test phase was used to determine retention of the material. The positive impact of testing in early work implied that testing should be introduced into educational settings to improve achievement (Spitzer, 1939; Wheeler & Roediger, 1992). However, laboratory conditions do not adequately mirror educational settings, therefore, substantial work has now been done to ensure that the testing effect holds true in classroom settings.

A plethora of classroom research suggests that the testing effect is robust. The testing effect occurs despite differences in test materials (e.g., words, prose, pictures, spatial locations), test formats (e.g., multiple choice, short answer, free recall, quiz), and timing (e.g., minutes versus weeks between testing phases) (Bae et al., 2018; Carpenter, 2012; Carpenter & Kelly, 2012; McDaniel et al., 2007; Rowland, 2014). Additionally, the testing effect has been duplicated across multiple disciplines (e.g., psychology, biology, chemistry) (Bailey et al., 2017; Pyburn et al., 2014; Schwieren et al., 2017) and different populations (e.g., primary school, university) (McDaniel et al., 2007; Roediger & Butler, 2011; Spitzer, 1939). Furthermore, the testing effect is not limited to retention of learned material (i.e., rote memory); the testing effect has been shown to improve application of material, improve

knowledge-based inferences, promote transfer of rules to novel contexts or knowledge to a different knowledge domain, and facilitate learning of new material (Brame & Biel, 2015; Carpenter, 2012). Finally, the testing effect can be increased when tests are combined with feedback (Bailey et al., 2017; Brame & Biel, 2015; Foss & Pirozzolo, 2017; Roediger & Butler, 2011; Schwieren et al., 2017) and when multiple tests are offered (i.e. three or more) (Bailey et al., 2017; Foss & Pirozzolo, 2017; Roediger & Karpicke, 2006; Wheeler & Roediger, 1992).

Two recent meta-analyses provide strong evidence for the testing effect based on laboratory research (Rowland, 2014) and classroom research (Schwieren et al., 2017). Rowland (2014) suggested two theoretical frameworks that may explain the testing effect: retrieval effort theories and the bifurcation model. Retrieval effort theories suggest that the difficulty and effort during the initial testing phase impact the intensity and depth of processing leading to a testing effect (Rowland, 2014). Whether difficulty increases retrieval routes, supports specific types of processing (i.e., item-specific processing), or allows for elaboration of memory traces remains unclear. The bifurcation model suggests that tests produce non-normal distributions of memory strength over time (Kornell et al., 2011; Rowland, 2014). Specifically, successfully tested (i.e., retrieved) items receive a large boost in memory strength, un-retrieved items receive no boost, and re-studied material receives a small boost. Thus, testing does not reduce the speed of forgetting, but increases memory strength for successfully tested items and makes them more likely to remain above a recall threshold during the final testing phase, thereby bifurcating the distribution.

Despite significant research, there has been limited consideration of whether the testing effect is equally powerful in various student subpopulations. Pyburn et al. (2014) argued that learning tools do not affect all students equally and specific attention should be focused on whether the testing effect as a phenomenon is equally apparent in disadvantaged populations. They examined whether a pre-test differentially influenced low- and high-skilled English language comprehenders. They found that a multiple-choice pre-test was more beneficial to low-skilled English comprehenders; additionally, the pre-test closed the achievement gap between these two groups. There is also a small selection of research suggesting that a negative testing effect (i.e., when a testing phase causes a decline in learning outcomes) is due in part to the cognitive ability of the participants. Mulligan et al. (2018) suggested differences in encoding might explain why there are only a few inconsistent instances of a negative testing effect. Briefly, the negative testing effect is potentially tied to the type of processing that occurs during the testing phase versus the requirements of the final test. Item-specific processing during the testing phase reduces a participant's ability for inter-item processing (and vice versa). Item-specific information helps distinguish one target from another and improves the odds of retrieval (e.g., the ground finch Geospiza conirostris can eat cactus-flowers). Inter-item relational information is categorical or grouping information; that is, common features of targets (e.g., all ground finches are seed-eaters). Inter-item relational information is tied to successful free recall. Therefore, when the testing phase forces one type of processing but success on the final test requires the other type of processing a negative testing effect may result. For example, if the testing phase includes a multiple-choice question asking a student which finch eats cactus flowers, inter-item processing leads to the answer Geospiza conirostris. However, in the re-study condition, a student may recognize that the given list of finches all eat seeds and are therefore ground finches. If the final test is a free recall test in which students are asked to list ground finches, inter-item processing is more useful to access the categorical information that all ground finches are seed eaters than the specific exception that can also eat cactus flowers. More importantly, Mulligan et al. (2018) found that manipulating the type of processing interacted with the cognitive ability of the student, particularly in the re-study control condition. A student's cognitive ability limits their ability to recognize and process categorical information during the re-study phase (i.e., the fact that the list of birds given in the re-study condition are all seed eaters and thus ground finches). Therefore, high-achieving students in the re-study condition could outperform low-achieving students in the testing condition when the test forces them to encode item-specific details and miss inter-item details that are more useful for a final exam that requires categorical knowledge. The testing effect research supports the use of a midterm as a useful learning tool, and limited research on frequency also suggests two midterms may be more beneficial than one (Bailey et al., 2017; Foss & Pirozzolo, 2017; Roediger & Karpicke, 2006; Wheeler & Roediger, 1992). Additionally, research on the negative testing effect and disadvantaged student subpopulations suggests that the number of midterms may differentially impact low and high achievers (Mulligan et al., 2018; Pyburn et al., 2014).

Frequency Effects

It is difficult to separate a phenomenon like the testing effect from other aspects of testing, such as frequency because a single test can potentially impact students across various theoretical frameworks. As already noted, increasing frequency has been shown to

increase testing effects (Bailey et al., 2017; Foss & Pirozzolo, 2017; Roediger & Karpicke, 2006; Wheeler & Roediger, 1992). However, frequency research makes novel predictions regarding subpopulations and potential limits on the impact of frequency. The frequency research suggests different underlying causes for the impact of increased frequency; for example, spaced or distributed practice, improved self-efficacy, reduced procrastination, or student-instructor relations (Bailey et al., 2017; Myers & Myers, 2007). Increasing test frequency has been shown to improve individual test scores as well as final exam scores (Bailey et al., 2017; Myers & Myers, 2007). Unfortunately, each of these studies used multiple cumulative exams (6-10 midterms); therefore, whether educators will see an increase in performance using a second non-cumulative midterm remains unclear. There is some suggestion that the expectation of a cumulative exam is enough in itself to increase student performance (Lawrence, 2013). Lawrence (2013) specifically tested differential impacts of cumulative exams on low and high achievers. While all students benefited from cumulative exams (versus non-cumulative exams), she found that the benefits were greater for low-achieving students. Due to the limited research on student subpopulations, Lawrence's work supports considering low- and high-achieving students separately in the present study, even though our second midterm exam is non-cumulative.

When considering what level of frequency is necessary to create improvements, a meta-analysis by Bangert-Drowns et al. (1991) suggests that extremes are unnecessary. Frequency varies substantially and while they concluded that increasing frequency of tests improved student achievement on final exams, they also noted that students are only at a serious disadvantage when they receive no tests at all. Furthermore, they determined that improvements in student learning diminish as test frequency increases: having one midterm exam benefits student learning more than no exams but having four exams will not produce a four-fold improvement in final exam results. These findings suggest that a second midterm may be a sufficient increase in frequency to produce a positive impact on student achievement.

Our project had two objectives: to determine if changing the frequency of midterm exams from one to two improves student learning outcomes and to consider whether testing influences low- and high-achieving students differently. We hypothesized that students in courses with two midterm exams would show greater improvement on their final exam score relative to their first midterm exam score than students in courses with a single midterm exam. Additionally, we predicted that low-achieving students would disproportionately benefit from two midterms.

Methods

Courses analyzed in our study were selected from the courses taught by one of the co-authors (NH) between 1990 and 2018, and syllabi were compared for their assignment breakdown and the number of midterm exams. The courses included in our study were selected based on whether the types of assessments and year of implementation were similar, except for the number of midterm exams. In total, four iterations of freshman cell biology and two iterations each of junior cellular biology and junior biochemistry I and II were selected for analysis. Freshman cell biology courses selected for inclusion in this study were offered in fall 2000 (1 midterm), 2003 (1 midterm), 2001 (2 midterms), and 2002 (2 midterms). Selected junior cell biology courses were offered in fall 1992 (2 midterms) and 1993 (1 midterm), junior biochemistry I courses were taught in winter 2010 (1 midterm) and fall 2010 (2 midterms), and the junior biochemistry II courses were from winter 2013 (1 midterm) and 2011 (2 midterms).

The one- and two-midterm cohorts for freshman cell biology and junior biochemistry I and II were similar in course structure: lab component (30-40%), quizzes (5-10%), midterm (20-30%), and cumulative final exam (35%). The one- and two-midterm cohorts for junior cell biology both had a lab component (40%), term paper (15%), and similar weighting for the midterm exams (one midterm = 20%; two midterms = 15% + 10%) and final exam (one midterm = 30%; two midterms = 35%). In all courses, the second midterm exam in the two-midterm condition was not cumulative, but each would contribute to the material on a cumulative final exam. All lectures were taught by the same instructor (author NH) and so were taught in a similar style. While course structure was similar, individual course elements occasionally differed from year to year (e.g., different textbooks or lab manual editions, different lab instructors, fresh quiz and exam questions). Therefore, the potential exists for confounding variables because the classes were not absolutely identical. The freshman biology courses used the same syllabus, and each of the junior cell biology, biochemistry I, and biochemistry II courses used the same syllabus for the same course. But clearly, the syllabi differed between courses (the syllabi were different for each of freshman biology, junior cell biology, junior biochemistry I, and junior biochemistry II). Student marks and class demographics from the selected courses were collected from the instructor's grade books, and students' identities were anonymized with a study ID before data analysis. Students who did not fulfill the assessment requirements of the study (i.e., did not complete one of the midterm exams or the final exam)

were removed from the dataset before analysis. This study was approved by the University of Alberta Research Ethics Board (Project #82145).

Our study had a 2 (midterm: one or two) x 2 (achievement level: high or low) x 2 (course level: freshman or junior) between-subjects factorial design. To assess improvements in final exam scores we chose to compare difference scores (i.e., final exam score minus midterm one exam score) rather than raw scores. Difference scores are better able to tell us how each students' performance changed across the semester and act as our dependent variable. To determine if there were differential impacts on weaker students, students were split into high- versus low-achieving cohorts based on whether they fell in the upper or lower 50% of the course, as determined by the median score of the first midterm exam. Finally, because we collected data from courses aimed at two different year levels, freshman and junior, course level became an additional factor. Rather than compare individual classes (e.g., cell biology vs biochemistry), we combined students into a single freshman cohort (N = 118) and a single junior cohort (N = 84). There were no significant differences between the first midterm scores of the freshman one- and two-midterm cohorts and between the junior one- and two-midterm cohorts indicating that students in the one- and two-midterm cohorts started out academically similar.

Results

The 2 x 2 x 2 analysis of variance (ANOVA) showed a main effect for achievement, $F(1,188) = 5.555$, $p = .019$. High-achieving students (mean exam score difference = -4.635, SEM = 1.135) had significantly different mean difference scores than low-achieving students (mean exam score difference = -.761, SEM = 1.188). There was no main effect for midterm exam score or course level.

There was an interaction effect for midterm exams and course level, $F(1,188) = 4.137$, $p = .043$, in which freshman students were impacted by the number of midterms while junior students were not (Figure 1). Specifically, freshmen who received one midterm performed significantly poorer on their final relative to their midterm exam (mean exam score difference = -5.885, SEM = 1.566) than freshmen who received two midterms (mean exam score difference = -4.635, SEM = 1.135).

There was no interaction effect between the number of midterm exams and achievement level: low-achieving students did not differentially benefit from a second midterm exam relative to high-achieving students.

Discussion

Our primary goal was to consider whether increasing midterms from one to two exams would improve learning outcomes in undergraduate biology courses. Within the testing effect research, there is a strong consensus that retrieval practice leads to better long-term retention than re-study alone (Rowland, 2014; Schwieren et al., 2017). There is also evidence to suggest that increasing the frequency of testing will lead to greater improvements in learning outcomes (Bailey et al., 2017; Bangert-Drowns et al., 1991; Foss & Pirozzolo, 2017; Myers & Myers, 2007; Roediger & Karpicke, 2006). Whether frequency improves the testing effect, alters student attitudes and behaviors (e.g., spaced studying), or impacts both, remains unclear. Regardless of the mechanism, we expected that two midterm exams would result in improved final exam scores relative to their first midterm exam score. Our results partially support this prediction. An ANOVA found a significant interaction effect between course level and number of midterms indicating that freshman students were positively impacted by a second midterm while junior students were not. This is similar to the impact that an e-portfolio assignment can have on student learning (Haave, 2016). Freshmen who received a second midterm exam did not perform as poorly on their final exam relative to their first midterm exam compared to those who completed only one midterm exam: a second midterm exam rescued freshman students from a significantly poorer final exam result. Freshmen are a unique student population as they are transitioning from high school to university while learning to become self-directed learners. Having freshmen practice retrieving their learning in the classroom (something they typically do not incorporate into their own study regime, Brown et al., 2014) is beneficial in the short-term, but may also benefit their ongoing development as learners. In contrast, juniors may be sufficiently self-directed learners that there is no additional impact from a second midterm. Therefore, junior students may require other kinds of learning interventions to continue their development as self-directed learners.

We were also interested in considering the subpopulation of low achievers. We believed that low achievers would see a greater benefit from two midterms than high achievers, but our results do not support this prediction. While we saw a main effect for achievement (i.e., there was a difference in how high- versus low-achieving students performed on their final vs their first midterm exam), we found no interaction effect to suggest that low or high achievers benefited from the second midterm in a unique way. Both low and high achievers did worse on the final compared to the midterm. Low achievers had a significantly smaller difference score, meaning their midterm and final

marks remained more similar than those of high achievers. This result is contradictory to other research on disadvantaged populations. For instance, Pyburn et al. (2014) found that a multiple-choice pre-test led to improved exam performance, but low-skilled English comprehenders benefited more than high-skilled English comprehenders. It appears that initial learning

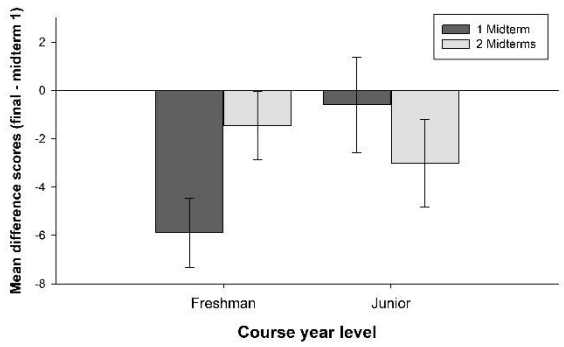

Fig. 1. The impact of course level and number of midterms on difference scores (final minus the first midterm exam score). ANOVA results indicate a significant interaction effect between course level and number of midterm exams, $F(1,188) = 4.137$, $p = .043$. Error bars represent standard error of the mean.

ability may not impact the influence of a second midterm exam. This result is unexpected as it could be argued that freshmen are not as experienced learners as juniors which is why freshmen benefit from a second midterm exam whereas juniors do not. Clearly, initial achievement level and learning ability/experience have a more complicated relationship than we anticipated.

Conclusion

Our results suggest that a second midterm exam may improve learning outcomes for students enrolled in a freshman but not a junior biology course. Additionally, a second midterm exam did not differentially improve the final exam scores relative to the midterm exam scores for low-achieving students. A primary limitation to our study is that it only analyzes biology courses. In addition, we were able to match only a handful of course iterations for analysis which limited our sample size. The small sample negatively impacted the effect size and power of the statistical test. Furthermore, while differences in course structure were minimized by using courses offered close in year and with similar course structuring external to the additional midterm exam, we were not able to account for all variations, such as students' prior GPA, relying instead on the first midterm exam score as an indicator of academic ability or preparation. A possible confounding factor is that the junior cell biology course had a term paper rather than in-class quizzes which our statistical analysis could not address. More robust conclusions will require future research with access to a larger campus population as well as additional disciplines. Future testing of sophomores and seniors may also provide additional information about the impact of course level. One obvious question is whether sophomores and seniors will show a similar pattern; that is, will additional midterm exams impact sophomores but not seniors? Finally, we cannot make any claims regarding the mechanism by which two midterm exams improved student learning outcomes. One future direction for research is to attempt to make distinctions between the testing and frequency effects. Distinguishing between these two mechanisms remains problematic. However, in terms of useful interventions, it is sufficient to recognize that regardless of why, testing in the classroom acts as a beneficial learning tool, not simply a necessity for program assessment purposes.

Acknowledgements

The University of Alberta supported this research. JG received funding from NH's research stipend as Associate Director of our Centre for Teaching and Learning. KK was supported by a grant held by NH from our Teaching and Learning Enhancement Fund. Our thanks to Paula Marentette for her statistical advice.

References

BAE, C.L., THERRIAULT, D.J. AND J.L. REDIFER. 2018. Investigating the testing effect: Retrieval as a characteristic of effective study strategies. Learn. and Instr 60: 206-214.

BAILEY, E.G., JENSEN, J., NELSON, J., WIBERG, H.K. AND J.D. BELL. 2017. Weekly formative exams and creative grading enhance student learning in an introductory biology course. CBE—Life Sci. Educ. 16(1): 1-9.

BANGERT-DROWNS, R.L., KULIK, J.A. AND C. L.C., KULIK. 1991. Effects of frequent classroom testing. J Educ. Res. 85(2): 89–99.

BRAME, C.J. AND R. BIEL. 2015. Test-enhanced learning: The potential for testing to promote greater learning in undergraduate science courses. CBE- Life Sci. Educ. 14(2): 1-12.

BROWN, P.C., ROEDIGER, H.L. AND M.A. MCDANIEL. 2014. To learn, retrieve. Pp. 23–45. Make it stick: The science of successful learning. The Belknap Press of Harvard University Press, Cambridge, MA.

CARPENTER, S.K. 2012. Testing enhances the transfer of learning. Current Directions in Psychol. Sci. 21(5): 279–283.

CARPENTER, S.K. AND J.W. KELLY. 2012. Tests enhance retention and transfer of spatial learning. Psychonomic Bulletin and Review 19(3): 443–448.

FOSS, D.J. AND J.W. PIROZZOLO. 2017. Four semesters investigating frequency of testing, the testing effect, and transfer of training. J Educ. Psycho.l 109(8): 1067–1083.

HAAVE, N. 2016. E-portfolios rescue biology students from a poorer final exam result: Promoting student metacognition. Bioscene: J. Coll. Biol. Teach.. 42(1): 8–15.

KORNELL, N., BJORK, R.A. AND M.A. GARCIA. 2011. Why tests appear to prevent forgetting: A distribution-based bifurcation model. J Mem. Lang. 65: 85–97.

LAWRENCE, N.K. 2013. Cumulative exams in the introductory psychology course. Teach. Psychol. 40(1): 15–19.

MCDANIEL, M.A., ANDERSON, J.L., DERBISH, M.H. AND N. MORRISETTE. 2007. Testing the testing effect in the classroom. Eur. J Cogn. Psychol. 19(4–5): 494–513.

MULLIGAN, N.W., RAWSON, K.A., PETERSON, D.J., AND K.T. WISSMAN. 2018. The replicability of the negative testing effect: Differences across participant populations. J Exp. Psychol. Learn. 44(5): 752–763.

MYERS, C.B. AND S.M. MYERS. 2007. Assessing assessment: The effects of two exam formats on course achievement and evaluation. Innov. High. Educ. 31(4): 227–236.

PYBURN, D.T., PAZICNI, S., BENASSI, V.A. AND E.M. TAPPIN. 2014. The testing effect: An intervention on behalf of low-skilled comprehenders in general chemistry. J Chem. Educ. 91(12): 2045 2057.

ROEDIGER, H.L. AND A.C. BUTLER. 2011. The critical role of retrieval practice in long-term retention. Trends Cogn. Sci. 15(1): 20–27.

ROEDIGER, H.L. AND J.D. KARPICKE. 2006. Test enhanced learning: Taking memory tests improves long-term retention. Psychol. Sci. 17(3): 249–255.

ROWLAND, C.A. 2014. The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. Psychol. Bull. 140(6): 1432-1463.

SCHWIEREN, J., BARENBERG, J. AND S. DUTKE. 2017. The testing effect in the psychology classroom: A meta-analytic perspective. Psychol. Learn. Teach. 6(2): 179–196.

SPITZER, H.F. 1939. Studies in retention. J Educ. Psychol. 30(9): 641–656.

WHEELER, M.A. AND H.L. ROEDIGER. 1992. Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. Psychol. Sci. 3(4): 240–2.