Test-Enhanced Learning and Incentives in Biology Education

Bryn St. Clair,[†]* Paul Putnam,[†] Harold L. Miller,[‡] Ross Larsen,[§] and Jamie L. Jensen[†]
[†]Department of Biology, [‡]Department of Psychology, and [§]Department of Instructional Psychology and Technology, Brigham Young University, Provo, UT 84602

ABSTRACT

Cognitive scientists have recommended the use of test-enhanced learning in science classrooms. Test-enhanced learning includes the testing effect, in which learners' recall of information encountered in testing exceeds that of information not tested. The influence of incentives (e.g., points received) on learners who experience the testing effect in classrooms is less understood. The objective of our study was to examine the effects of incentives in a postsecondary biology course. We administered exams in the course using a quasi-experimental design with low and high point incentives and measured student learning. Although exposure to exams predicted better learning, incentive level did not moderate this effect, an outcome that contradicted recent laboratory findings that higher incentives decreased student recall. We discuss possible explanations of the disparate outcomes as well as the implications for further research on the testing effect in postsecondary biology classrooms.

INTRODUCTION

Biology educators seek effective instructional methods to increase students' ability to think scientifically. In recent decades, results from many cognitive science studies have shed light on this goal. In particular, cognitive scientists have shown that taking exams enhances thinking and learning (e.g., McDaniel *et al.*, 2007; Karpicke and Roediger, 2008). This phenomenon is called *practice testing*, *test-enhanced learning*, *retrieval practice*, and the *testing effect* and has received extensive support across different types of learning materials (for reviews, see Roediger and Butler, 2011; Pan and Rickard, 2018). Cognitive researchers and education policy makers suggest applying the testing effect to real-world educational settings, including those in postsecondary biology courses, to improve student thinking and learning (Pashler *et al.*, 2007; Pagliarulo, 2011; Carpenter *et al.*, 2017).

Research in cognitive science on learning and testing with incentives has illuminated mechanisms surrounding test-enhanced learning and inspired discipline-based education research (DBER) questions surrounding the use of tests in classrooms. As cognitive scientists attempt to experimentally isolate variables to attribute causality, both in the laboratory and in actual classroom settings, researchers in the growing field of DBER focus on describing mechanisms of learning within a more ecologically variable discipline-specific environment, such as the application of learning principles in a postsecondary biology classroom. For example, cognitive researchers have recommended the use of low-stakes quizzing to improve student learning based on a multitude of studies that show improved student learning through the testing effect (e.g., Roediger *et al.*, 2011), yet little classroom research has been done to define the parameters of low-stakes quizzing in biology classrooms. Increasingly, cognitive psychologists and discipline-based researchers are collaborating to extend the principles of the testing effect through experimentation from the cognitive research laboratory to a classroom application to define the mechanisms and parameters of test-enhanced learning in biology education (Jensen *et al.*, 2014; Talanquer, 2014).

Ido Davidesco, *Monitoring Editor*
Submitted Nov 08, 2019; Revised Apr 23, 2020; Accepted Apr 24, 2020
CBE Life Sci Educ September 1, 2020 19:ar40
DOI:10.1187/cbe.19-11-0226
*Address correspondence to: Bryn St. Clair (bes@byu.edu).

Researchers have explored the mechanisms of the testing effect on biology learning with authentic classroom variables. Hinze and Rapp (2014) studied the effects of incentives as a source of performance pressure on biology tests in a laboratory setting and found that, as pressure increases, the testing effect decreases. They suggested that the reduction in learning was the result of an increased demand on attentional processes. Tse and Pu (2012) had previously suggested that attention was divided by anxiety associated with increased performance pressure. More recently, research has shown that students with high trait anxiety perform worse on biology exams than those with low trait anxiety (Ballen *et al.*, 2017). Based on the variable impact of student anxiety on learning, researchers recommend quizzing at low-stakes or low-incentives levels (e.g., Roediger *et al.*, 2011; Brame and Biel, 2015), yet the mechanisms and boundaries of student performance with regard to incentives on assessments in a classroom have not been well defined.

Incentives, in the form of grades or points, are a common practice, whether in low- or high- stakes settings. However, DBER researchers typically have not treated classroom incentives as an experimental variable. The objective of our study was to assess the outcome of incentives on the testing effect on a series of unit exams in a postsecondary biology course. We focused on the following research questions: 1) Does the testing effect improve student learning in postsecondary biology? 2) Do incentives affect learning via the testing effect? 3) Do results from a real-world classroom study of the testing effect support those obtained in laboratory settings?

METHODS

In this study, during Fall semester 2018 and Spring 2019, we 1) compared student exam scores on tested and untested material to measure a testing effect in a postsecondary biology course and 2) assessed the role of incentives during unit exams on the subsequent retention of course content on a final comprehensive exam in a postsecondary biology course.

Subjects

We performed this study at a private university in the western United States. The institutional review board at our institution approved this research and granted permission for this study (IRB no. 17219). This university's total undergraduate enrollment is 31,233 students, and admissions are highly selective, with an incoming student average grade point average of 3.86 and American College Testing score of 28. It is a private religious institution with students who are relatively religious and culturally homogenous. The introductory biology course is a general education requirement for the university. The course enrollment is a representative sample of the university student body. Participants ranged from freshmen to seniors and came from a variety of disciplines outside the life sciences. We recruited 514 students. There were 142 students in the high-incentives treatment during the first semester and 372 students in the low-incentives treatment during the second semester. All participants granted written consent.

Study Design

We made significant effort to ensure as much group equivalence as possible, that is, the same instructor taught all sections of introductory biology during two consecutive semesters (Fall 2018, Spring 2019). During each semester, the course sections were taught back-to-back at the same time of day in the same classroom, with the same textbook and course materials. We organized the course into five units divided by subject. The students received a list of all of the intended learning outcomes for each unit. At the end of each unit of instruction, students were given an exam. The exam items were coordinated with the intended learning outcomes from the course. Students took the five unit exams throughout the semester in the university testing center facility. Students completed each unit exam within a 5-day window. Exam items were primarily application-, analysis-, and evaluation-type multiple-choice items, in other words, high Bloom's-level multiple-choice questions (Anderson *et al.*, 2001).

To assess student learning with incentives through the testing effect, we applied a variable course points treatment in a quasi-experimental design. We divided the course content in half (content A and content B). Students in Fall semester section 1 were treated with high-incentive exams on half of the course content (content A), while students in Fall semester section 2 were treated with high-incentive exams on the other half of the course content (content B). Students in Spring semester section 1 were treated with low-incentive exams on half of the course content (content B), while students in Spring semester section 2 were treated with low-incentive exams on the other half of the course content (content A). In addition, students in both sections were also given low-incentive quizzes on the opposite content (e.g., section 1 students were given low-incentive quizzes on content B and high-incentive exams on content A); see Figure 1. Each unit included content A and content B. The students did not know which content would be on the quiz or exam. Only the content that was assessed on the unit exams (either A or B) was included in the analysis. The quizzed content was not included in the analysis.

The point equivalence for the unit exams was 10% of the overall course point structure in the low-incentive treatment group and 21% of the overall course point structure in the high-incentive treatment group. We redistributed the extra points from the low-incentive course exams equally between the other areas of the course, including equal points to homework, surveys, attendance, and the final exam, in order to reduce the extra variable of student study attention based on point emphasis. Our course used an active-learning pedagogy that included learning activities beyond the assessments (e.g., homework application, formative quizzes, and class participation). Learning activities were incentivized heavily to encourage active participation. Thus, overall incentives available for the summative assessments were limited. As such, a doubling of assessment incentive was considered a substantial increase in points. However, we acknowledge that it is a relatively modest difference when compared with more traditional didactic classrooms, where the majority of points may be assigned to assessments.

Outcome Measure and Independent Variables of Interest

We measured student learning as a final comprehensive course exam. We administered an identical exam to all sections. The exam consisted of 90 multiple-choice questions. Students took the final assessment in the university testing center facility. Each learning outcome tested on the unit exams had a

	Fall 2018		Spring 2019	
	Introductory Biology Section 1	Introductory Biology Section 2	Introductory Biology Section 1	Introductory Biology Section 2
Experimental Treatment	5 Unit Exams with higher-incentives	5 Unit Exams with higher-incentives	5 Unit Exams with lower-incentives	5 Unit Exams with lower-incentives
Tested Content	Content A	Content B	Content B	Content A
Final measurement	Both sections take the same final, content A and B			

FIGURE 1. Graphical illustration of study design. Students in Fall semester received high-incentive exams (dark gray) on half of the content and low-incentive quizzes (white) on the other half of the content. Students in Spring semester received low-incentive exams (light gray) on half of the content section and low-incentive quizzes (white) on the other half of the content.

coordinated summative assessment item on the final. Coordinated unit exam items were not identical to the final assessment items; rather, new questions were designed to assess the same intended learning outcomes. For a sample unit exam item and final exam item, see Table 1.

To detect a testing effect, we compared student success on final assessment items designed to measure intended learning outcomes that were previously seen on an exam or a quiz (*tested*) with intended learning outcomes that were not previously seen on any exam or quiz (*untested*). The final summative assessment included 49 items that were tested and 13 items that were untested. Untested items on the final exam were coordinated with intended learning outcomes presented to students through in-class and out-of-class application activities. These learning outcomes were not seen on any previous quiz or exam. Higher student scores on tested final assessment items would indicate that students receive a learning benefit through the testing effect from an exam experience.

To detect an effect of incentives, we compared student scores on tested items on the final between those who had taken high-incentive exams with those who had taken low-incentive exams. For one section in each semester, content A was the tested content; by comparing these two sections, we assessed the difference between low and high incentives on content A. Likewise, in the other section for each semester, content B was the tested content; by comparing these two sections, we assessed the difference between low and high incentives on content B. There were 27 items designated as content A on the final exam and 22 items designated as content B on the final exam. Both semesters took both content A and content B on the exam along with the untested items (as mentioned earlier) to make a balanced and complete final exam. Differences in student scores on final assessment items between those that were previously tested at a low-incentive level and those previously tested at a high-incentive level would indicate that students receive differential learning benefits from the testing effect based on the incentive structure.

Covariates

In estimating the effect of testing and incentives, we controlled for student scientific reasoning ability, trait anxiety, and content difficulty. We measured students' scientific reasoning ability using Lawson's Classroom Test of Scientific Reasoning (LCTSR;

TABLE 1. Sample intended learning outcome, with coordinated unit and final exam items (correct answers shown in italics)

Sample intended learning outcome
Evaluate the most likely reproductive isolation mechanism in a given scenario.
Coordinated unit exam item and final exam item
There are about six different species of mangabeys in the genus *Lophocebus*. Osman Hill's mangabeys are found only in Cameroon, and Uganda mangabeys are restricted to just Uganda, making their speciation due to _____; whereas, black-crested mangabeys and Johnston's mangabeys produce offspring that are sterile, making their speciation due to _____. A. Behavioral isolation (I), Habitat isolation (II) B. Gametic isolation (I), Postzygotic barriers (II) C. Mechanical isolation (I), Gametic isolation (II) D. *Habitat isolation (I), Postzygotic barriers (II)*
There are nine different species of the baobab tree, six are native to Madagascar, two are native to mainland Africa, and one is native to Australia. Identify the most likely reproductive isolating mechanism keeping species separate for each situation: A. Behavioral isolation B. Gametic isolation C. Mechanical isolation D. *Geographic isolation*

TABLE 2. Generalized test-anxiety survey questions

I feel anxiety during in-class quizzes.
I feel anxiety during tests in the testing center.
The anxiety I feel during class quizzes prevents me from demonstrating my learning.
The anxiety I feel during tests in the testing center prevents me from demonstrating my learning.

Lawson *et al.*, 2000). The LCTSR is a content-independent test of basic formal reasoning skills including correlational, probabilistic, proportional, and hypothetico-deductive reasoning. Others have used the LCTSR as a covariate to control for student reasoning ability (e.g., Jensen *et al.*, 2015), as it is highly correlated with performance in science classes (e.g., Johnson and Lawson, 1998). Validity and reliability are well established on this measure (Lawson *et al.*, 2000). We controlled for differences in course content difficulty by adding content as a dummy variable into our model to account for any variation in results that were determined by differences based on course content selection.

We measured and controlled for student self-reported trait anxiety. This is the level of anxiety that students generally feel toward testing situations, not the anxiety they specifically felt during our test administrations. For ease of exposition, we will refer to this measure hereafter as "generalized test anxiety." We administered a voluntary survey given at the beginning of the course. Students responded to four questions on a five-point Likert scale. Due to a clerical error, one version of the survey provided had a seven-point Likert scale. We standardized the data by taking a percentage of the total. For an example of a survey question see Table 2.

ANALYSIS

We established evidence for a simple testing effect first by comparing the exam content that was tested with the exam content that was untested (not found in content A or B) in a repeated-measures analysis of covariance (ANCOVA). Following this analysis, to address our research question of interest, we examined the relationship between student performance on the final exam and the incentive treatment and used a variety of controls, including student scientific reasoning (LCTSR), course content, and student generalized test anxiety in multiple regression. A multiple regression analysis subsumes ANCOVA and has the added benefit of providing beta coefficient values (a measure of total effect of the predictor variable). An additional reason we chose a multiple regression analysis over an ANCOVA was to accommodate missing data points in our analysis. We checked for all assumptions of multiple regression, including linearity, independence of residuals, homoscedasticity, multicollinearity, and data normality. Due to a response rate of 57% on the voluntary anxiety survey, we used the full-information maximum-likelihood (FIML) method for missing data. FIML has been shown to outperform traditional missing-data techniques, such as listwise deletion or mean imputation (Little and Rubin, 2019). We did all analyses in SPSS v. 25 for the diagnostic plots and used M *plus* v. 8.3 for the multiple regression. We measured the equivalence of groups of those who did and did not answer the student anxiety survey through an independent-samples *t* test.

RESULTS

A repeated-measures ANCOVA, using the LCTSR as a covariate, showed that there was a significant difference between mean student performance on the tested content (M = 0.73, SD = 0.13) versus untested content (M = 0.65, SD = 0.18), $F(1, 481) = 28.09$, $p < 0.001$, $n = 483$.

Due to the low response rate on the voluntary anxiety survey, we ran an analysis to compare groups. There was no difference in mean student performance between those who answered the generalized test-anxiety survey (M = 17.63 SD = 4.15) and those who did not (M = 17.06, SD = 3.83), $t(512) = -1.582$, $p = 0.233$, $n = 514$. Assured of group equivalence, we proceeded to our multiple regression analysis of the variable of interest, incentive level.

Model 1

Our first model predicted the final student exam score using two covariates (LCTSR and exam content). The independent variable of interest was the high-incentive treatment, with high-incentive coded as 1 and the control coded as 0. Data were linear, and all other assumptions of multiple regression were assessed and met through visual inspection of histograms and residual plots produced in SPSS. The multiple regression model predicted the final student exam score. Two of the three variables added statistical significance to the model ($p < 0.001$); the third, incentive level, was not statistically significant ($p = 0.305$). Regression coefficients and standard errors can be found in Table 3. Content had an unstandardized beta of 2.436, indicating that content A material had a 2.436-point increase over content B material. The standardized beta for that independent variable was 0.308, indicating that the difference between content is 0.308 SD, which can be considered a small effect. LCTSR scores had an unstandardized beta of 0.402, indicating that, for every one-unit increase in the LCTSR score, the final exam score increased by 0.402. The standardized beta for this independent variable was 0.394, indicating that for every 1 SD increase in LCTSR, the predicted final exam score increased by 0.394 SD, a moderate effect size.

Model 2

For our second model, we ran multiple regression on final student exam scores with three covariates (LCTSR, exam content, and generalized test anxiety). The variable of interest again was the high-incentive treatment, with high-incentive labeled as 1. The data were linear, and all other assumptions of multiple regression were met. The multiple regression model statistically predicted the final student exam score. Three of the four variables added statistical significance to the model ($p < 0.001$); the fourth, incentives, was not statistically significant ($p = 0.11$).

TABLE 3. Summary of multiple regression analysis model 1 of the following variables predicting final exam score ($n = 514$ students)[a]

Variable	B	SE_B	β
Incentives treatment	–0.349	0.340	–0.040
LCTSR	0.402	0.040	0.394*
Content	2.436	0.303	0.308*

[a]B, unstandardized regression coefficient; SE_B, standard error of the coefficient; β, standardized coefficient.
*$p < 0.05$.

TABLE 4. Summary of multiple regression analysis model 2 of the following variables predicting final exam score (n = 514 students)[a]

Variable	B	SE_B	β
Incentives treatment	–0.537	0.341	–0.061
LCTSR	0.341	0.043	0.335*
Content	2.519	0.302	0.319*
Generalized test anxiety	–4.184	1.179	–0.181*

[a]B, unstandardized regression coefficient; SE_B, standard error of the coefficient; β, standardized coefficient.
*$p < 0.05$.

Regression coefficients and standard errors can be found in Table 4. Interestingly, the high-incentive treatment was still not statistically significant, even in the presence of generalized test anxiety. The pattern of results of the other independent variables with the final exam score remained the same as in the first model.

Interaction Model

We ran an interaction model that included generalized test anxiety and incentives and all covariates. This was done to see whether the effect of incentives was conditional on the level of anxiety of the student. We did not find any significance in the interaction term ($p > 0.05$). Thus, these results are not shown.

DISCUSSION

In this study, we applied variable incentives when testing undergraduate biology students during unit exams in a semester-long course and measured performance on a final comprehensive exam. Although we found enhanced performance on the final exam in content tested on unit exams, incentive level (high vs. low) did not change that performance. Other researchers have reported the testing effect in undergraduate biology (e.g., Carpenter *et al.*, 2016; Hubbard and Couch, 2018). Still others have hypothesized and recommended enhancement of the testing effect using incentives with low-stakes on exams in classrooms (Roediger *et al.*, 2011; Brame and Biel, 2015). Our results did not find a difference in the performance of those students given unit exams at 10% of the total course points versus those students given unit exams at 21% of the course points (see Figure 2).

This study provides a bridge from evidence found in laboratory research in cognitive psychology to a more applied subject-specific understanding of the principle of test-enhanced learning in biology and also demonstrates the challenge in the translation from laboratory to cross-disciplinary application of principles of learning (Talanquer, 2014). Most postsecondary biology courses offer incentive structures different from typical laboratory techniques, which include monetary compensation based on performance (Hinze and Rapp, 2014) or exemption from further study duties (Clark *et al.*, 2018). As noted by Hinze and Rapp (2014), "Laboratory-based manipulation of performance pressure … may not align perfectly to the kinds of real-world pressure experienced during classroom or standardized tests" (p. 605). Although the cognitive research perspective is useful in forcing attention on educationally relevant cognitive processes, the ecology of the real-world classroom presents competing systems that may moderate findings found in a more streamlined laboratory setting. This does not mean that the laboratory findings do not apply to the mechanisms in isolation, but rather that the classroom environment creates variables that may change the outcome of theoretical models. In our view, laboratory findings should be supplemented by those produced by systematic experimentation in an actual classroom setting.

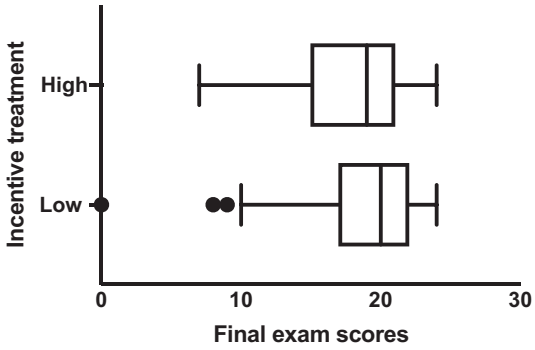

FIGURE 2. Box-and-whiskers plot mean comparison of high-incentive exam treatment items on final and low-incentive exam treatment items on final. The box indicates the interquartile range. The line indicates the median. The dots outside the line indicate outliers. Error bars represent mean standard errors.

In this study, we measured student generalized test anxiety in a precourse survey and used it as a covariate in our study. We had predicted that high generalized test anxiety would decrease student performance on the final exam and that that effect would further decrease with a high-incentive level. We found no effect of the latter. Throughout the course, multiple unit tests may have produced the testing effect regardless of student generalized test anxiety. Researchers have shown that frequent testing episodes decrease self-reported test anxiety (Agarwal *et al.*, 2014; Khanna, 2015) and increase learning in biology (Bailey *et al.*, 2017). While high test anxiety typically is associated with poorer test performance (Zeidner *et al.*, 2005) and weaker intention to persist in a biology major (England *et al.*, 2017), moderate test anxiety can enhance assessment performance (Keeley *et al.*, 2008). Continued experimental separation of student test anxiety and incentive levels during biology exams will clarify the differences between these two variables.

Further research in postsecondary biology classrooms is needed to direct the effective application of the testing effect. Our course included active-learning pedagogy, and as such, included learning activities beyond the assessments, including homework application, formative quizzes, and class participation, that were heavily incentivized to encourage active participation. As such, our study had a limited number of points available for assessment and applied only 10% of the overall course point structure in the low-incentive treatment group and 21% of the overall course point structure in the high-incentive treatment group (2% and 4.2% on each unit exam, respectively) to exams. It is possible that this difference was not large enough to prompt differences in student behavior. However, in this particular classroom structure, the difference between low-incentive and high-incentive treatments represents a doubling in points. As a consequence, the application of these findings is limited and may change in a course that applies even more extreme point values to course incentives, often found in a traditional didactic classroom. Additionally, the quasi-experimental nature

of our study design prevented us from randomly assigning students to course sections, so generalization of these findings should be done with caution. Future research avenues include an even more extreme application of incentive differences on the testing effect in even more diverse classroom settings. Additional research is also needed to understand the relationship between the number of exams administered to students in the course of the semester and the level of test anxiety and the testing effect as well as student preparation and attention to studying due to point differences.

ACKNOWLEDGMENTS

The authors would like to thank the undergraduate researchers for help in research implementation and data collection. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

Agarwal, P. K., D'Antonio, L., Roediger, H. L., McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition*, *3*(3), 131–139.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., ... & Raths, J. (2001). In Anderson, L. W., & Krathwohl, D. R. (Eds.), *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete edition). New York: Longman.

Bailey, E. G., Jensen, J., Nelson, J., Wiberg, H. K., & Bell, J. D. (2017). Weekly formative exams and creative grading enhance student learning in an introductory biology course. *CBE—Life Sciences Education*, *16*(1), ar2.

Ballen, C. J., Salehi, S., & Cotner, S. (2017). Exams disadvantage women in introductory biology. *PLoS ONE*, *12*(10), e0186419.

Brame, C. J., & Biel, R. (2015). Test-enhanced learning: The potential for testing to promote greater learning in undergraduate science courses. *CBE—Life Sciences Education*, *14*(2), es4.

Carpenter, S. K., Lund, T. J. S., 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.

Carpenter, S. K., Rahman, S., Lund, T. J. S., Armstrong, P. I., Lamm, M. H., Reason, R. D., & Coffman, C. R. (2017). Students' use of optional online reviews and its relationship to summative assessment outcomes in introductory biology. *CBE—Life Sciences Education*, *16*(2), ar23.

Clark, D. A., Crandall, J. R., & Robinson, D. H. (2018). Incentives and test anxiety may moderate the effect of retrieval on learning. *Learning and Individual Differences*, *63*, 70–77.

England, B. J., Brigati, J. R., & Schussler, E. E. (2017). Student anxiety in introductory biology classrooms: Perceptions about active learning and persistence in the major. *PLoS ONE*, *12*(8), e0182506.

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology*, *28*(4), 597–606.

Hubbard, J. K., & Couch, B. A. (2018). The positive effect of in-class clicker questions on later exams depends on initial student performance level but not question format. *Computers & Education*, *120*, 1–12.

Jensen, J. L., Kummer, T. A., & Godoy, P. D. D. M. (2015). Improvements from a flipped classroom may simply be the fruits of active learning. *CBE—Life Sciences Education*, *14*(1), ar5.

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test...or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, *26*, 307–329.

Johnson, M. A., & Lawson, A. E. (1998). What are the relative effects of reasoning ability and prior knowledge on biology achievement in expository and inquiry classes? *Journal of Research in Science Teaching*, *35*(1), 89–103.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966.

Keeley, J., Zayac, R., & Correia, C. (2008). Curvilinear relationships between statistics anxiety and performance among undergraduate students: Evidence for optimal anxiety. *Statistics Education Research Journal*, *7*(1), 4–15.

Khanna, M. M. (2015). Ungraded pop quizzes: Test-enhanced learning without all the anxiety. *Teaching of Psychology*, *42*(2), 174–178. https://doi.org/10.1177/0098628315573144

Lawson, A. E., Alkhoury, S., Benford, R., Clark, B. R., & Falconer, K. A. (2000). What kinds of scientific concepts exist? Concept construction and intellectual development in college biology. *Journal of Research in Science Teaching*, *37*(9), 996–1018.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.) (Wiley series in probability and statistics 793). Hoboken, NJ: Wiley. https://www.wiley.com/en-us/Statistical+Analysis+with+Missing+Data%2C+3rd+Edition-p-9781118595695

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*(4–5), 494–513.

Pagliarulo, C. L. (2011). *Testing effect and complex comprehension in a large introductory undergraduate biology course (Doctoral dissertation)*. Retrieved November 1, 2019, from https://repository.arizona.edu/handle/10150/202773

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning: IES practice guide (NCER 2007-2004)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved May 30, 2020, from https://files.eric.ed.gov/fulltext/ED498555.pdf

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology. Applied*, *17*(4), 382–395.

Roediger, H. L. 3rd, & Butler, A. C. (2011). *The critical role of retrieval practice in long-term retention*. *Trends in Cognitive Sciences*, *15*(1), 20–27.

Talanquer, V. (2014). DBER and STEM education reform: Are we up to the challenge? *Journal of Research in Science Teaching*, *51*, 809–819.

Tse, C.-S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *Journal of Experimental Psychology: Applied*, *18*(3), 253.

Zeidner, Moshe; Matthews, Gerald; Elliot, A. J., & Dweck, C. S. (2005). Evaluation anxiety. In Elliot, A. J., & Dweck, C. S. (Eds.), *Handbook of competence and motivation*. New York: Guilford.