Test-Enhanced Learning in Post-Secondary Biology Courses: The Effect of Cues and Incentives on High-Level Learning

Bryn Ellen St. Clair
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Life Sciences Commons

BYU ScholarsArchive Citation
https://scholarsarchive.byu.edu/etd/8920

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
Test-Enhanced Learning in Post-Secondary Biology Courses:

The Effect of Cues and Incentives
on High-Level Learning

Bryn Ellen St. Clair

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Jamie L. Jensen, Chair
Harold L. Miller Jr.
Ross Larsen
John Bell

Department of Biology
Brigham Young University

Copyright © 2021 Bryn Ellen St. Clair
All Rights Reserved
ABSTRACT

Test-Enhanced Learning in Post-Secondary Biology Courses: The Effect of Cues and Incentives on High-Level Learning

Bryn Ellen St. Clair
Department of Biology, BYU
Doctor of Philosophy

Cognitive scientists and psychology researchers have given growing attention to evidence of the testing effect, that is, the improvement of students’ recall through memory-retrieval practice in the form of quizzes and exams. While laboratory experiments consistently show dramatic positive effects on learning through the testing effect, discipline-specific education researchers have sought to generalize these findings in real, instead of simulated classrooms. The objective of this dissertation was threefold: (1) To survey current literature on the testing-effect as it applies to learning biology at the post-secondary level. In this review, I consider how further research on the testing effect may be useful for instructors’ decisions regarding its use. (2) To describe findings from a quasi-experimental design in a post-secondary biology class 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. (3) To describe findings from a study that compared student learning in conditions where cued exams were in place versus conditions in which they were absent. Student learning improved in the former condition relative to the latter. I discuss the implications of the results in all of these studies for further research and application.

Keywords: test-enhanced learning, post-secondary, biology education, STEM, testing effect
ACKNOWLEDGEMENTS

I thank my advisor, Dr. Jamie Jensen, for her tireless effort to develop and execute this project. From the moment I heard her first talk about it, I knew it was for me. My committee members, Drs. Miller, Ross, and Bell, were all incredibly helpful in analyzing, interpreting, and writing this work. I’m grateful for their patience and time. I will remember the conversations and principles they taught me forever. I admire and revere them all.

This dissertation is based on research conducted on students at Brigham Young University. I am grateful for their willing participation. Additionally, Max Putnam was an active participant in the data analysis; thank you. Danny Ferguson was great to hear and bounce off ideas from me in the lab.

I am grateful to the institution at BYU and all those who work to make it function, including those who supplement the funds donated.

Finally, I would like to acknowledge with gratitude the support and love of my family — my parents, Allen and Lori; my parents-in-law, Larry and Rieta; my children—Isaac, Seth, Eli, Grace, and Daniel; and my husband, Sam. They all kept me going, and this dissertation would not have been possible without their support, the endless effort of time, attention, and mercy when there was a lack of attention from me.

I realize that words cannot compensate or adequately describe the amount of support given. Understanding that, I will still try—thank you.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE PAGE</td>
<td>i</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>Chapter 1: Modulators of Test-Enhanced Learning in Post-Secondary Biology</td>
<td>1</td>
</tr>
<tr>
<td>Abstract</td>
<td>2</td>
</tr>
<tr>
<td>Introduction</td>
<td>2</td>
</tr>
<tr>
<td>How does the classroom structure of assessment influence the testing effect in biology education?</td>
<td>4</td>
</tr>
<tr>
<td>Assessment Frequency</td>
<td>5</td>
</tr>
<tr>
<td>Assessment Incentives</td>
<td>6</td>
</tr>
<tr>
<td>Assessment Feedback</td>
<td>7</td>
</tr>
<tr>
<td>How does assessment format and cognitive skill influence the testing effect?</td>
<td>8</td>
</tr>
<tr>
<td>Assessment Format</td>
<td>8</td>
</tr>
<tr>
<td>Assessment and cognitive skills</td>
<td>8</td>
</tr>
<tr>
<td>What student characteristics influence the testing effect in biology?</td>
<td>9</td>
</tr>
<tr>
<td>Test Anxiety</td>
<td>9</td>
</tr>
<tr>
<td>Individual Student Differences in Academic Performance</td>
<td>10</td>
</tr>
<tr>
<td>Conclusion</td>
<td>11</td>
</tr>
<tr>
<td>References</td>
<td>14</td>
</tr>
<tr>
<td>Chapter 2: Test-Enhanced Learning and Incentives in Biology Education</td>
<td>20</td>
</tr>
<tr>
<td>Abstract</td>
<td>21</td>
</tr>
<tr>
<td>Introduction</td>
<td>21</td>
</tr>
<tr>
<td>Methods</td>
<td>23</td>
</tr>
<tr>
<td>Subjects</td>
<td>23</td>
</tr>
<tr>
<td>Study design</td>
<td>24</td>
</tr>
<tr>
<td>Outcome measure and Independent Variables of Interest</td>
<td>26</td>
</tr>
<tr>
<td>Covariates</td>
<td>28</td>
</tr>
<tr>
<td>Analysis</td>
<td>29</td>
</tr>
<tr>
<td>Results</td>
<td>30</td>
</tr>
<tr>
<td>Model 1</td>
<td>30</td>
</tr>
</tbody>
</table>
Chapter 3: Test-Enhanced Learning: Cues Enable Scientific Skill Development

Abstract ...................................................................................................................................... 40
Introduction ............................................................................................................................. 40
Method ....................................................................................................................................... 42
  Participants ............................................................................................................................. 42
  Study Design .......................................................................................................................... 42
  Covariates ............................................................................................................................... 45
Analysis ..................................................................................................................................... 46
Results ....................................................................................................................................... 46
Discussion .................................................................................................................................. 47
Acknowledgments ..................................................................................................................... 49
References ................................................................................................................................. 50
LIST OF TABLES

Table 2.1: Sample Intended Learning Outcomes with Coordinated Unit and Final Exam Items 27

Table 2.2: Generalized Test Anxiety Survey Questions............................................................... 29

Table 2.3: Summary of Multiple Regression Analysis Model 1 .................................................. 31

Table 2.4: Summary of Multiple Regression Analysis Model 2 .................................................. 32

Table 3.1: Sample Learning Outcomes......................................................................................... 44

Table 3.2: Hierarchical Multiple Regression of Final-Exam Scores ............................................. 47
LIST OF FIGURES

Figure 2.1: Description of the Study Design .......................................................... 25

Figure 3.1: Study Design ...................................................................................... 45
Chapter 1: Modulators of Test-Enhanced Learning in Post-Secondary Biology

Bryn Ellen St. Clair*
Department of Biology
Brigham Young University
Provo, UT 84602

Jamie L. Jensen
Department of Biology
Brigham Young University
Provo, UT 84602

*Corresponding Author. Department of Biology, 4102 LSB, Brigham Young University, Provo, UT 84602, bes@byu.edu

I hereby confirm that the use of this article is compliant with all publishing agreements.
Abstract

Cognitive scientists and psychology researchers have given growing attention to the testing effect, that is, the improvement of students’ recall through memory-retrieval practice in the form of quizzes and exams. While laboratory experiments consistently show dramatic positive effects on learning through the testing effect, discipline-specific education researchers have sought to generalize these findings in real, instead of simulated classrooms. Our objective in this review was to survey recent findings on the testing effect in post-secondary biology education. We found that: (a) Increased exam frequency increases the testing effect; (b) Corrective feedback on exams may enhance the testing effect; (c) Incentives, such as points, may decrease the positive outcome of learning through the testing effect, though little research in actual classrooms on this widely used practice is found; (d) Individual differences in student achievement and preparation may moderate the effect. We consider how further research on the testing effect may be useful for instructors’ decisions regarding its use.

Introduction

In recent decades, science education reform has involved stronger collaboration between cognitive and education researchers interested in applying their findings from laboratory research to subject-specific disciplines. Discipline-Based Education Research (DBER) is central to the effort at the post-secondary level. Several principles of learning from cognitive-science laboratory studies have hypothetical applications for instructional practice in STEM classrooms. In this article, we review the recent findings on one such principle of learning, for application in post-secondary biology education —the testing effect.
The testing effect is the improvement of students’ learning through classroom testing. It is also known as retrieval practice, practice quizzing, or test-enhanced learning (for a review see Roediger & Butler, 2011). For example, information that appears on a quiz will more likely be recalled later than information not tested. Terms such as test, exam, and quiz are most commonly associated with assessment in education. In simple language, assessment is the measurement of how much students have learned of what the teacher has taught. In this way, assessment informs teachers of the effectiveness of their instruction in terms of student learning (Black & Wiliam, 1998). Even though an exam on previously presented content is sometimes considered, in itself, a neutral learning event, researchers in cognitive psychology, education and neurobiology have reported that learning may be enhanced when it is retrieved through testing (Roediger & Karpicke, 2006). Specifically, retrieving information from memory may enhance the cues for future retrieval.

Early research on the testing effect occurred in the classroom (Gates, 1917; Jones, 1923; Spitzer, 1939). A recent resurgence of research on this topic has been largely conducted in cognitive-science laboratories in controlled studies (Roediger & Butler, 2011) using varied material, including word pairs (Carrier & Pashler, 1992; Karpicke & Roediger, 2008), general facts (Butler, Karpicke, & Roediger, 2008), trivia (McDaniel & Fisher, 1991), and textual passages (Kang, McDermott, & Roediger, 2007). Researchers have also examined the differential effects of the testing effect using varied assessment formats, including multiple-choice items (Marsh, Roediger, Bjork, & Bjork, 2007), open- and closed-book items (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008), and inference items (Karpicke, 2012). The results of a multitude of studies on the testing effect have been featured in confirmatory meta-analyses (Bangert-Drowns, Kulik, & Kulik, 1991; Phelps, 2012; Rowland, 2014; Schwieren,
Barenberg, & Dutke, 2017). Officials at the Institute of Education Sciences, sponsored by the US Department of Education, have recommended the adoption of testing and retrieval practice at all levels of education, including the post-secondary level (Pashler et al., 2007).

The extent to which research on the testing effect applies to learning biology at the post-secondary level is critical in making decisions in its use (Daniel 2012). Given both the consistent findings on the testing effect in laboratory studies, and researchers’ confidence in recommending the method as a means of improving student learning, there is a growing motivation to understand the mechanistic boundaries that may influence the testing effect as it applies to learning in discipline-specific classrooms. Perhaps due to the demand for STEM education reform, the body of testing effect literature in recent years has expanded to include information pertinent to the application of the testing effect in post-secondary biology classrooms. A summary of the testing effect in biology education at the post-secondary level is not found in the current literature. Here we present an overview of the testing effect specifically in post-secondary biology education and discuss the implications for further research and application of the testing effect. The key questions being addressed are:

(1) How does the classroom structure of assessment influence learning through the testing effect in biology education?

(2) How does assessment format and content influence learning through the testing effect in biology education?

(3) What student characteristics influence learning through the testing effect in biology education?

How does the classroom structure of assessment influence the testing effect in biology education?
Assessment Frequency

Researchers have found a positive relationship between the higher frequency of classroom assessments and academic achievement (Phelps 2012). The importance of testing frequency has also been shown more specifically in biology education research at the university-level. For example, frequent quizzing improved learning outcomes in post-secondary biology over standard unit exams (Bailey, Jensen, Nelson, Wiberg, & Bell, 2017). Haak, HilleRisLambers, Pitre, & Freeman (2011) also implemented a highly structured course design in an introductory biology course based on daily and weekly assessments in problem-solving, data analysis, and other higher-order cognitive skills. The design was associated with improved performance in all students enrolled in the course and reduced the performance gap between socioculturally disadvantaged and non-disadvantaged students. In a final example, Pape-Lindstrom, Eddy, and Freeman (2018) measured an increase in student performance in a community college biology course when they implemented frequent pre-class online and open-book reading assessments.

Overall, increased frequency of testing experiences appears to improve learning. The increased frequency of assessment has been reported as positively correlated with lower course-failure rates, higher course point totals, and higher scores on midterm assessments (Freeman et al, 2007) as well as increased academic motivation (Healy, Jones, Lalchandani, & Tack, 2017). Importantly, Leeming (2002) surveyed University of Memphis students and reported their greater satisfaction with courses that included more frequent assessments. Students also indicated that they learned more as a result. By simply increasing the frequency and number of exam experiences instructors can enhance the testing effect in the biology classroom for students.
Assessment Incentives

Incentives in terms of assessment scores and points are commonly used to motivate students and may have an influence on the testing effect. However, researchers typically have not treated classroom incentives per se as an experimental variable with regard to the testing effect. Hinze and Rapp (2014) awarded monetary compensation to lab study participants based on performance. Subjects scored relatively lower on high-stakes biology exams than low-stakes biology exams. The current application of incentives in the biology education classroom studies on the testing effect is varied and a clear understanding of the implications with regard to the testing effect is not well-defined in the literature. While researchers intuitively recommend low-stakes quizzing as an important safeguard against student test-anxiety in post-secondary biology, little classroom research has addressed this idea experimentally. In one exception, St. Clair, Putnam, and Jensen (in press) demonstrated no difference between students’ performance on exams with high- and low-incentive levels (21% vs 10% of the course points total on exams) in an introductory college-level biology course. Further research on classroom incentives and learning through the testing effect will be important to understand the interaction between more extreme levels of course incentives and the testing effect.

A course instructor may remove the incentives from an assessment and make the experience voluntary. Student self-reported voluntary use of self-testing is correlated with increased student achievement (Hartwig & Dunlosky, 2012). Specifically, in university-level biology, Carpenter et al. (2017) reported that students who opted for quizzes as a review tool in a general-biology course scored higher on the initial examination in the course than those students who selected reading-based review instead. Subsequently, researchers promoted quiz completion using a classroom presentation of the differential outcome on examination performance.
Increasingly more students participated in voluntary assessment practice prior to each subsequent examination, producing higher mean scores on the examinations. Others have modeled effective optional learning strategies (Rodriguez, Rivas, Matsumura, Warschauer, & Sato, 2018) and offered voluntary workshops promoting the testing effect as a learning strategy in large post-secondary biology courses (Stanger-Hall, Shockley, & Wilson, 2011). Both strategies led to improved student learning.

Assessment Feedback

Informing students of the assessment items they missed and the correct answer to those items is generally referred to as feedback and is an influential factor in the testing effect (Kang, McDermott, & Roediger, 2007). Retrieval practice is effective without feedback, but feedback enhances learning with the testing effect (Pashler, Cepeda, Wixted, & Rohrer, 2005; Lavigne, & Risko, 2018). Jacoby, Wahlheim, and Coane (2010) found that feedback enabled bird classification with fewer exam experiences. In another study, researchers displayed feedback to subjects following initial fill-in-the-blank items with process-based biology concepts (e.g., stages of mitosis) and found an increase in performance on final fill-in-the-blank questions when feedback was given (Pan, Hutter, D’Andrea, Unwalla, & Rickard 2018).

Oliver, Renken, & Williams (2018) assessed students on common biology misconceptions. Those students prompted to self-explain misconceptions after assessment feedback were more likely to overcome the misconceptions on a final criterion test. The majority of studies on the testing effect and biology learning use feedback as a consistent part of testing, and the empirical evidence from laboratory studies seem to support this practice, yet relatively few studies exist that examine the effects of feedback on the testing effect specifically in the
biology classroom. More work in this area is needed to understand the role of feedback on the testing effect in biology education.

How does assessment format and cognitive skill influence the testing effect?

Assessment Format

The initial test format may influence the final test success (Kang et al., 2007). According to Glover (1989), short answer, and fill-in-the-blank item formats both increased the testing effect over multiple-choice and true-or-false formats. However, Little and Bjork (2015) found that multiple-choice items were more effective when they contained strong distractor options and feedback. More specifically, Pagliarulo reported that multiple-choice and short answer assessment formats could be useful on complex biology content (2011). Hinze (2010) assessed post-secondary students on biology content in laboratory experiments and found that cued-recall assessment format (a sentence that includes pertinent content preceding the assessment item) improved performance on memory items while removing the cues from recall items made retrieval practice more difficult and less effective. Presumably, generating information on one's own, if successful, could increase the effectiveness of free-recall assessment items over cued-recall assessment items (Carpenter & DeLosh, 2006), yet there is an inherent balance between increasing effortful processing and overloading a student's ability to successfully do the task (see Pyc & Rawson, 2009). More research needs to be done specifically in post-secondary biology classrooms with regard to item type and the effectiveness of the testing effect to bring about student learning.

Assessment and cognitive skills
Although many instructors seek to develop high-level cognitive thinking in their students, most assessment items are specific to memory retrieval of subject content rather than application, analysis, evaluation, and creativity (Momsen, Long, Wyse, & Ebert-May, 2010). As such, quiz and exam questions that are related in content subject-material but do not focus on the same specific learning outcome or concept may not show a testing effect (Nguyen & McDaniel, 2015). In biology, it appears that the standard procedure of using quiz questions from the test bank provided by ancillary sources (e.g., textbook companies) may not benefit student performance unless the summative exam questions are closely tied to the targeted learning outcomes created for the course and taught by the instructor (Wooldridge, Bugg, McDaniel, & Liu, 2014). Instructors should pay specific attention to coordinating intended learning outcomes with assessment items to enable learning through the testing effect.

Researchers argue for the strength of the testing effect with complex material (Karpicke & Aue; 2015; Rawson, 2015; Burns, 2010). Jensen, McDaniel, Woodard, and Kummer (2014) found routine quizzing requiring application, analysis, and evaluation of biology material could be useful in promoting both conceptual and higher-order skills performance on the final exam in a biology class. Agarwal (2011) reported that a match in initial and final cognitive processing on assessment items (e.g., quizzed and tested on a specific skill) benefits long-term higher-order skills in learning biology. Further research is needed on the testing effect using complex material learning in biology in the post-secondary level classroom including valuable reasoning skills used in scientific discovery and problem solving.

What student characteristics influence the testing effect in biology?

Test Anxiety
Test anxiety is common among undergraduate students. In a survey, Gerwing (2015) found that 38.5% of student respondents reported test anxiety. High test anxiety typically is associated with poorer test performance, test avoidance, loss of motivation, decreased memory retrieval, and impaired attention (Wolf & Smith, 1995; Zeidner, 2005). In the laboratory, Tse and Pu (2012) replicated the testing effect using word pairs while also measuring attention to relevant detail and test anxiety. They found that students who scored lower on attention to relevant detail but higher in test anxiety made more errors on average on the final assessment. England, Brigati, and Schussler (2017) surveyed learners in undergraduate biology courses that featured active learning pedagogy including in-class clickers. High-test anxiety accompanied lower self-reported GPA and a weaker intention to persist in the biology major.

By contrast, moderate test anxiety may enhance assessment performance (Keeley, Zayac, & Correia, 2008). A majority of students report decreased test anxiety when they use retrieval practice (Agarwal, D’Antonio, Roediger, McDermott, & McDaniel, 2014) and low-stakes in-class quizzing (Khanna, 2015) to prepare for a summative course assessment. A clear picture of the relationship between individual learners test anxiety, the testing effect, and biology material in a college classroom is weakly defined in the literature, partially due to the variable nature of individual students and their reaction to test experiences. Clearly isolating variables in the ecologically complex classroom is challenging yet needed to further clarify the mechanisms surrounding these commonly experienced pedagogical tools.

Individual Student Differences in Academic Performance

Researchers have begun to study individual student performance differences and the testing effect. While some researchers have demonstrated the benefit of quizzing in biology to
students of diverse academic abilities (Orr & Foster, 2013; Pape-Lindstrom, Eddy, & Hogan, 2014), Hubbard and Couch (2018) found that the use of in-class clickers benefited high-performing students more than low-performing students. Carpenter et al. (2016) found that among undergraduate biology learners, all benefited from the use of frequent assessment, but high-performing students benefited more from it than mid- and low-performing students. Bailey, Jensen, Nelson, Wiberg, and Bell (2017) studied increased quiz frequency and categorized students into learning history. They found that mid- and late-learners (those who did not show mastery until the second half of the course) comprised 24% of the class and specifically benefited more from the increased assessment frequency. Butler (2010) found that repeated testing produced improved average success on assessment items with biology inference questions if prior learning of individuals was included in the model. Individual differences in student prior learning and academic ability may impact the outcome of learning biology through the testing effect.

Conclusion

The application of research findings from cognitive psychology to post-secondary classrooms may yield significant benefits in STEM education reform. The evidence supporting the testing effect in particular may enable learning for students in post-secondary biology classrooms. Increased test frequency influences the testing effect in the biology classroom and is the most obvious recommendation for immediate classroom application of the testing effect. Exam feedback has effectively been shown to be an influential moderator of the testing effect in the literature, though no specific study in this search has experimentally applied feedback to a biology classroom.
Course incentives, such as points or stakes also may affect the result of the testing effect in the biology classroom. Current application of incentives in biology education research is varied and there is not a clear understanding of the interaction with the testing effect and points. Hypothetically, researchers recommend low-stakes quizzing as an important preventive for student test-anxiety in post-secondary biology, though little classroom research has addressed this idea experimentally. Further empirical work on classroom incentives, such as points and learning through the testing effect will be important to understand the interaction.

The influence of individual learner achievement on the testing effect in biology is also a consideration in the success of students in STEM and its application to the postsecondary classroom overall. Students experience test-anxiety, but the influence of anxiety on the testing effect in post-secondary biology is not thoroughly demonstrated. Early work shows that most students can benefit from testing, though it seems that learners embody characteristics that enable learning biology through the testing effect differently from others. Continued experimental application of varied classroom structure on the testing effect would illuminate mechanism boundaries as well as an understanding of the population to which those effects apply.

Early work on the testing effect in biology education indicates that the relationship between the initial and final exam questions may influence the testing effect. Biology instructors can accentuate exam experiences by ensuring that the required cognitive skills in both formative and summative exam items align with the designated learning outcomes in the course. The connection between content, cognitive process and coordinated learning outcome on the initial and final exam may impact the testing effect. Practice retrieval, and in many cases, practice processing, on an exam, if successful, could increase the effectiveness of learning through tests,
yet there is an inherent balance between increasing effortful processing and overloading a student's ability to successfully do the task. Further classroom research is necessary to illuminate the development of desired skills required for deep application of biology material through the testing effect on assessment experiences.

Research efforts will continue to illuminate and support reforms in STEM education. The testing effect is a promising principle of learning that has the potential to aid post-secondary biology teaching. Effective instruction needed for deep application and conceptual knowledge in biology education will require further understanding of the mechanistic boundaries of the testing effect as they apply to the biology classroom.
References


Chapter 2: Test-Enhanced Learning and Incentives in Biology Education

Bryn Ellen St. Clair*
Department of Biology
Brigham Young University
Provo, UT 84602

Paul Putnam
Department of Biology
Brigham Young University
Provo, UT 84602

Harold L. Miller Jr.
Department of Psychology
Brigham Young University
Provo, UT 84604

Ross Larsen
Department of Instructional Psychology and Technology
Brigham Young University
Provo, UT 84602

Jamie L. Jensen
Department of Biology
Brigham Young University
Provo, UT 84602

*Corresponding Author. Department of Biology, 4102 LSB, Brigham Young University, Provo, UT 84602, bes@byu.edu

I hereby confirm that the use of this article is compliant with all publishing agreements.
Abstract

Based on results from laboratory studies, 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 post-secondary 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 post-secondary biology classrooms.

Introduction

Biology educators seek effective instructional methods to increase students’ ability to think scientifically. In recent decades, numerous results from 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, Anderson, Derbish, & Morrisey, 2007; Karpicke & Roediger, 2008). Practice testing, or test-enhanced learning, has received extensive support across different types of learning materials (for reviews see Pan & Rickard, 2018; Roediger & Butler, 2011). Test-enhanced learning includes the testing effect, in which learners’ recall of information that appeared in tests was greater than that of information not tested. Cognitive researchers and education policymakers suggest applying the testing effect to real-world
educational settings including those in post-secondary biology courses to improve student thinking and learning (Carpenter, Lund, & Coffman, 2017; Pagliarulo, 2011; Pashler et al., 2007).

Cognitive scientists have typically studied the testing effect in the laboratory where the use of artificial stimuli in controlled conditions may not capture the complex interactions that occur in real classroom conditions (McDaniel, Roediger, & McDermott, 2007). For example, although practice retrieval from memory might be helpful to student learning and retention in a laboratory study (Carrier & Pashler, 1992; Pan & Rickard, 2018, Roediger & Karpicke, 2006), an instructor may vary the application of the testing effect from the laboratory conditions by adjusting the type of retrieval used (e.g. recall vs. recognition), the frequency and duration of the classroom tests, the feedback format, or even the incentives structure of those exams. (Hogan & Kintsch, 1971; Pan et al., 2019).

Incentives, in the form of grades or points, are a common practice whether in low- or high- stakes settings. Researchers commonly recommend quizzing at a low-stakes or low-incentives level (e.g. Brame & Biel, 2015; Roediger, Agarwal, McDaniel, & McDermott, 2011). However, researchers typically have not treated classroom incentives per se as an experimental variable. 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 increased, the testing effect decreased. They suggested that the reduction in learning was the result of an increased demand on attentional processes. Tse & 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, Salehi, & Cotner, 2017).
The objective of our study was to assess the outcome of incentives on the testing effect on a series of unit exams in a post-secondary biology course. We focused on the following research questions: 1) Does the testing effect improve student learning in post-secondary 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 post-secondary biology course; and we 2) assessed the role of incentives during unit exams on the subsequent retention of course content on a final comprehensive exam in a post-secondary 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. This university’s total undergraduate enrollment is 31,233 students and the 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 that 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 of 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 5 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 five-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 & Krathwohl, 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 2.1.

![Figure 2.1: Description of the Study Design](image)

We divided the course content and applied an experimental treatment of higher incentive to two sections in Fall 2018 and lower incentives to two sections in Spring 2019. We measured student learning in both sections on both contents on a cumulative final exam.

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.

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 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 2.1.
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. Higher student scores on tested final assessment items would indicate that students receive a learning benefit through the testing effect from an exam experience.

We then assessed student learning with incentives through the testing effect with the variable course points treatment. As mentioned above, we divided the course content and split
the treatment between the sections. 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 had content in both areas measured. The remaining final exam questions were items that were completely untested (as mentioned above) and those items associated with learning outcomes 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 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; 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, Kummer, & Godoy, 2015) as it is highly correlated with performance in science classes (e.g., Johnson & 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 from here on out 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 7-point Likert scale. We standardized the data by taking a percentage of the total. For an example of survey question see table 2.2.

Table 2.2: Generalized Test Anxiety Survey Questions

<table>
<thead>
<tr>
<th>Generalized test-anxiety survey questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>I feel anxiety during in-class quizzes</td>
</tr>
<tr>
<td>I feel anxiety during tests in the testing center</td>
</tr>
<tr>
<td>The anxiety I feel during class quizzes prevents me from demonstrating my learning.</td>
</tr>
<tr>
<td>The anxiety I feel during tests in the testing center prevents me from demonstrating my learning.</td>
</tr>
</tbody>
</table>

Analysis

We first tested that students are learning through the testing effect by comparing the exam content that was tested to the exam content that was untested in an independent sample t-test. To address our second 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 (student scientific reasoning (LCTSR), course content, student generalized test-anxiety) in multiple regression. 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 & Rubin, 2019). We did all analyses in SPSS version 25 for the diagnostic plots and used MPlus version 8.3 for the multiple regression. We measured the equivalence of groups of those who did and did not answer the anxiety student survey through an independent sample t-test.

Results

Our initial analysis showed that there was a difference between mean student performance on the tested content (m = .71, SD = .14) versus untested content (m = .64, SD = .18), p < .001, n = 514.

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), p = .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 covariates (LCTSR, incentives treatment, 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 < .001); the third, incentive level, was not statistically significant (p = .305). Regression coefficients and standard errors can be found in Table 2.3.
Table 2.3: Summary of Multiple Regression Analysis Model 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SE_{B}</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incentives treatment</td>
<td>-0.349</td>
<td>0.340</td>
<td>-0.040</td>
</tr>
<tr>
<td>LCTSR</td>
<td>0.402</td>
<td>0.040</td>
<td>0.394*</td>
</tr>
<tr>
<td>Content A</td>
<td>2.436</td>
<td>0.303</td>
<td>0.308*</td>
</tr>
</tbody>
</table>

Note. * p < .05; B = unstandardized regression coefficient; SE_{B} = Standard error of the coefficient; β = standardized coefficient.

Content A 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 standard deviations, which can be considered a small effect. LCTSR scores had an unstandardized beta of 0.402 indicating that for every one unit increase in LCTSR score the final exam score increased by 0.402. The standardized beta for this independent variable was 0.394 indicating that for every one standard deviation increase in LCTSR the predicted final exam score increased by 0.394 standard deviations, a moderate effect size.

Model 2

For our second model, we ran multiple regression on final student exam scores from LCTSR, incentives treatment, 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 < .001); the fourth, incentives, was not statistically significant (p =
Regression coefficients and standard errors can be found in Table 2.4. 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 the first model.

Table 2.4: Summary of Multiple Regression Analysis Model 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SE&lt;sub&gt;B&lt;/sub&gt;</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incentives treatment</td>
<td>-.537</td>
<td>.341</td>
<td>-.061</td>
</tr>
<tr>
<td>LCTSR</td>
<td>.341</td>
<td>.043</td>
<td>.335*</td>
</tr>
<tr>
<td>Content</td>
<td>2.519</td>
<td>.302</td>
<td>.319*</td>
</tr>
<tr>
<td>Trait-anxiety</td>
<td>-4.184</td>
<td>1.179</td>
<td>-.181*</td>
</tr>
</tbody>
</table>

Note. * p < .05; B = unstandardized regression coefficient; SE<sub>B</sub> = Standard error of the coefficient; β = standardized coefficient

Interaction model.

We ran an interaction model that included generalized test-anxiety and incentives. This was done to see if 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 > .05). Thus, the 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 (i.e. Carpenter et al., 2017; Hubbard & Couch, 2018). Still others have hypothesized and recommended enhancement of the testing effect using incentives with low-stakes on exams in classrooms (Brame & Biel, 2015; Roediger et al., 2011).

Our results did not find a difference in the performance of those students given unit exams at 10% of the total course points vs. those students given unit exams at 21% of the course points. The apparent contradiction in our findings and cognitive researchers’ recommendation demonstrates the challenge in the translation from laboratory to cross-disciplinary application of principles of learning (Talanquer, 2014). Most post-secondary biology courses offer incentive structures different from typical laboratory techniques, which include monetary compensation based on performance (Hinze & Rapp, 2014) or exemption from further study duties (Clark, Crandall, & Robinson, 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 with those produced by systematic experimentation in an actual classroom setting.

In this study, we measured student generalized test-anxiety in a pre-course 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, Jensen, Nelson, Wiberg, & Bell, 2017). While high test-anxiety typically is associated with poorer test performance (Zeidner, Matthews, Elliot, & Dweck, 2005) and weaker intention to persist in a biology major (England, Brigati, & Schussler, 2017) moderate test-anxiety can enhance assessment performance (Keeley, Zayac, & Correia, 2008). Our data support the previously demonstrated idea that anxiety reduces student performance on biology exams. Continued experimental separation of student test-anxiety and incentive levels during biology exams will clarify the differences between these two variables.

Further research in post-secondary biology classrooms is needed to direct the effective application of the testing effect. Future research avenues include an even more extreme application of incentives on the testing effect in even more diverse classroom settings. It is common for biology instructors to apply even greater incentive levels than we did in this study. Furthermore, low-stakes assessment of student outcomes may be modified by the loss of student motivation (Wise & DeMars, 2005). 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.
Acknowledgements

The authors would like to 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


Chapter 3: Test-Enhanced Learning: Cues Enable Scientific Skill Development

Bryn Ellen St. Clair*
Department of Biology
Brigham Young University
Provo, UT 84602

Paul Putnam
Department of Biology
Brigham Young University
Provo, UT 84602

Harold L. Miller Jr.
Department of Psychology
Brigham Young University
Provo, UT 84604

Ross Larsen
Department of Instructional Psychology and Technology
Brigham Young University
Provo, UT 84602

Jamie L. Jensen
Department of Biology
Brigham Young University
Provo, UT 84602

*Corresponding Author.
Department of Biology, 4102 LSB, Brigham Young University, Provo, UT 84602, bes@byu.edu

I hereby confirm that the use of this article is compliant with all publishing agreements.
Abstract

Cognitive scientists have recommended the use of test-enhanced learning in science classrooms, including post-secondary biology. Instructors use different styles of assessment in real-world learning environments, including cued testing that may include a "cheat sheet" or a "study guide." Cued exams include essential content required to assist students in answering test items. We designed a classroom study in an introductory college biology course and compared student learning in conditions where cued exams were in place versus conditions in which they were absent. Student learning improved in the former condition relative to the latter. We discussed the implications of the results for further research and application.

Introduction

In recent decades, cognitive scientists and neurobiologists have drawn attention to the testing effect, which is the improvement of students' learning through quizzes and exams (e.g., Carpenter et al., 2017; van den Broek et al., 2016). Also known as practice quizzing, test-enhanced learning, and retrieval practice, the testing effect has been demonstrated across different types of learning materials and subject matter (see reviews by Pan & Rickard, 2018 and Roediger & Butler, 2011). Given the empirical evidence, officials at the Institute of Education Sciences sponsored by the US Department of Education recommended the application of test-enhanced learning at all levels of public education, including post-secondary science, technology, engineering, and math (STEM; Pashler et al., 2007; also see Daniel, 2012).

Cognitive scientists have studied the testing effect with educationally appropriate materials in laboratory settings and have reported enhanced student learning (Roediger & Karpicke 2006), but these results may not reflect complex interactions that occur in real
classroom conditions (McDaniel, Roediger, & McDermott, 2007). For example, when an instructor in real-life settings may vary the test format or the test's complexity (Greving & Richter, 2018) and produce variable results through the testing effect in their students' learning. In fact, a common criticism in the testing-effect literature is the relative simplicity of the items used, that is, that they require only low-level cognitive skills (see, e.g., Carpenter, 2012). In accordance with calls to increase students' scientific thinking (see AAAS, 2010), instructors seek to develop targeted skills in their students, such as content application, analysis, and evaluation instead of memory for scientific content per se (Momsen et al., 2010). The testing effect robustly improves student skills with scientific content application, analysis, and evaluation (Jensen et al. 2020). Researchers have encouraged that research on the testing effect occurs in an actual discipline-specific classroom setting to enhance its ecological validity (St. Clair et al., 2020).

A common pedagogical practice that may modulate the testing effect is cued exams. Sometimes called a "cheat sheet" or a "study guide," cued exams include essential content applicable to assessment items. As the testing effect is based on the retrieval of self-generated content (Roediger & Karpicke, 2006), providing specifically arranged retrieval cues may systematically constrain that content to the test taker's advantage. Theoretically, providing content on exams may reduce a student's cue-target strengthening through retrieval practice. To date, post-secondary biology-education researchers have not treated cued exams as an experimental variable in classroom studies.

We designed our study to explore the effects of the testing effect using cued exams with complex, high-level educational material in a real-world post-secondary biology course. We focused on the following research questions: 1) Does the testing effect improve student learning?
Moreover, 2) Does the use of cued exams with high-level items impair the testing effect in such courses?

**Method**

During the fall semester of 2018, we 1) compared students' scores on final exams composed of items related to previously tested and untested content material in a post-secondary biology course; and 2) assessed the impact of cued exams on subsequent retention of content in the same course.

**Participants**

We performed this study at a private university in the western United States. The institutional review board at the institution approved the proposed research (IRB #17219). The university's total undergraduate enrollment at the time of the study was 31,233 students. Its admissions process was highly selective, with a new admissions grade-point average of 3.86 and a mean American College Testing score of 28. It is a private religious institution with a largely culturally homogenous student body. The introductory biology course was a requirement in the university's general-education curriculum. Participants ranged from freshmen students to seniors and came from various academic disciplines outside of the life sciences. We recruited 371 students for our study. All provided written consent to participate.

**Study Design**

The same instructor taught students in two introductory biology sections to control for variance between experimental groups. Enrollment was self-selected. The course sections were
scheduled back-to-back in the same classroom and used the same required textbook and other course materials. We organized the course into five units according to topic. Each student received the same comprehensive list of learning outcomes for each unit. After each unit, the instructor asked students to complete a cued exam and a non-cued exam representing the relevant learning outcomes. The instructor administered unit exams at the university testing center during a five-day window. The university testing center proctored the final exam as well. Exam items had a multiple-choice format and were composed according to the high-level criteria of the Bloom cognitive hierarchy (Anderson & Krathwohl, 2001). Our course included active learning pedagogy and included learning activities beyond the assessments, such as homework application, formative quizzes, and class participation.

To assess the testing effect on student learning, we divided the course content in half (Content A and Content B). Students in each section took five cued exams on each designated half (Content A in section 1 and Content B in section 2) and five non-cued exams on the other half (Content B in section 1 and Content A in section 2) for a total of 10 exams per section (see Figure 1). Each exam contained between 30-33 items. We counterbalanced the course content on the exams to control for the content difficulty in our statistical model. The cued exams' total points accounted for 21% of the overall course points, as did the total points for the non-cued exams. In other words, 42% of a student's total points in the course were accounted for by the ten exams.

Exam items were associated with specific learning outcomes. Each cued exam included a summary of relevant biology content for that unit. We listed terms and definitions in paragraph form for each intended learning outcome. The cues consisted of terms and definitions and were
developed by one researcher and validated by another researcher. Sample items appear in table 3.1.

Table 3.1: Sample Learning Outcomes

| Learning outcome: Evaluate the most likely reproductive isolation mechanism in a given scenario. |
| Associated cue: |
| Reproductive isolation occurs when two species do not mate because of certain isolation factors. The biological species concept is primarily based on this idea. Postzygotic barriers include the inability to reproduce fertile or viable zygotes. Prezygotic barriers include things such as different mating rituals or signals (behavioral isolation), when two species do not come into contact with each other because they occupy different habitats (habitat isolation), when two species are physically unable to mate (mechanical isolation), and sperm and egg cannot combine (gametic isolation). |

Coordinated unit and final exam items

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 ______ I _______ ; whereas, Black-crested mangabeys and Johnston’s mangabeys produce offspring that are sterile, making their speciation due to: _______ II _______.

A. Behavioral isolation (I), Habitat isolation (II)
B. Gametic isolation (I), Post zygotic barriers (II)
C. Mechanical isolation (I), Gametic isolation (II)
D. Habitat isolation (I), Post zygotic barriers (II)

Associated final exam item

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

Note. Correct answers appear in italics.
We measured student learning on the final exam common to both sections of the course. To detect a testing effect, we compared scores on exam items previously seen on an exam (tested) to scores on items that had not been previously seen on any assessment (untested). The final exam included 49 tested items and 13 untested items. Of the items identified with a specific learning objective, there were 27 items associated with Content A and 22 items associated with Content B. Each specific learning objective on the unit exams (both cued and non-cued) had a coordinated summative assessment item on the final. The final assessment did not include identically worded exam items as the unit exams; rather, New items measured the same specific learning objectives. See table 3.1.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Introductory Biology Section 1</th>
<th>Introductory Biology Section 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 Unit Exams with Cues</td>
<td>5 Unit Exams without Cues</td>
</tr>
<tr>
<td>Exam Content</td>
<td>Content A</td>
<td>Content B</td>
</tr>
<tr>
<td>Final Exam</td>
<td>Both sections take the same final: content A and B</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* We divided the course content into two halves and administered cued exams for one half and non-cued exams for the other half in the two sections of the course. We measured overall student learning in each of the sections using a cumulative final exam.

Figure 3.1: Study Design

Covariates

We measured students' level of scientific reasoning ability using Lawson's Classroom Test of Scientific Reasoning (LCTSR; Lawson et al., 2000). The LCTSR is a content-independent test of formal reasoning skills, including correlational, probabilistic, proportional, and hypothetico-deductive reasoning. Others have used the LCTSR as a covariate (Jensen,
Kummer, & Godoy, 2015; Johnson & Lawson, 1998). Researchers have established validity and reliability for the test (Bao, Xiao, Koenig, & Han, 2018; Lawson et al., 2000). We counterbalanced the course content (Content A v, Content B) to control the content difficulty in our statistical model.

Analysis

We compared two groups of students' performances on two types of exam items: those previously tested with specific learning objectives and those for which it had not. We used repeated-measures analysis of covariance (ANCOVA). Following this analysis, we examined the relationship between student performance on the final exam, cued exams, student scientific reasoning (LCTSR), and exam content (A or B). We used hierarchical multiple regression to determine whether LCTSR scores and content (A or B) were significantly related to the final exam score. We checked for multiple regression assumptions, including linearity, independence of residuals, homoscedasticity, multicollinearity, and data normality. We did all analyses in SPSS v. 25 for the diagnostic plots and used M plus v. 8.3 for the analysis.

Results

A repeated-measures ANCOVA, using LCTSR as a covariate, revealed a significant difference between mean student performance on the tested content (M=0.73, SD=.13) versus untested content (M=0.65, SD=0.18), F(1, 375)=28.09, p<0.001, n=371). Our hierarchical regression analysis predicted the final exam score using two covariates (LCTSR and exam content). The independent variable of interest was the cued exam treatment with cued exams coded as one and the control of no cued exams coded as 0. Data were linear, and linear
regression assumptions were assessed and met through visual inspection of histograms and residual plots produced in SPSS. The multiple regression model significantly predicted the final student exam score. In each case, the covariates improved the prediction of the final score. (R² = .222, p < .001; see Table 3.2).

Table 3.2: Hierarchical Multiple Regression of Final-Exam Scores

| Variable | Model 1 | | Model 2 | | Model 3 | |
|----------|---------|--------|---------|--------|--------|
|          | B       | β      | B       | β      | B      | β      |
| Intercept| 70.234** | 4.734** | 39.458** | 2.66** | 38.363** | 2.586** |
| Cues     | 3.448** | 0.116** | 3.466** | 0.117** | 1.415* | 0.048* |
| LCTSR    |         |        | 1.697** | 0.439** | 1.695** | 0.438** |
| Content  |         |        |         |        | 4.308** | 0.145** |

R² | .014* | .206** | .222*

Note: N=371 *p ≤ 0.05 **p ≤ 0.001

Discussion

In our study, we observed the testing effect in an experiment involving previously tested exam items and untested items in a semester-long course in post-secondary introductory biology. Researchers had previously reported the testing effect in similar courses (Jensen et al., 2014 ηp² = .039). Some researchers have argued that the testing effect declines with high-level cognitive content compared to rote-memory performance (Van Gog & Sweller, 2015); others (Agarwal et
al., 2007, Jensen et al., 2014; Karpicke & Aue, 2015) have asserted the efficacy of the testing effect for high-level cognitive content, which is consistent with our view. Instructors can use exams to enable high-level cognitive science learning at the post-secondary level.

We further examined the testing effect with cued unit exams and uncued unit exams throughout the semester-long course and found improved student learning on high-level cognitive content using cued exams. Smith, Blunt, Whitten, and Karpicke (2016) suggested that the use of prompts on assessment items specific to high-level cognitive content may draw students' attention to more relevant or previously known information and thereby promote subsequent retrieval. Another potential explanation for our findings is that exposure to cued items facilitated students' application of subject-specific content (see Jensen et al., 2020), thereby increasing the amount of successfully executed knowledge construction activities at which the respective learning tasks are targeted. Furthermore, the application of relevant content on increasingly complex tests may enable further learning (Roelle & Berthold, 2017; Wise et al., 1989). These considerations yield the prediction that the net benefit of incorporating directed, cued content into learning tasks should increase as learning tasks become more complex. Future researchers may give focused attention to this putative relationship.

Retrieval practice can be applied to more than encoding rote memory through exam experiences. The assumption implied through the terminology of the testing effect is that students must participate in an exam to improve learning by retrieving low-level facts. Researchers have tried to encourage the extension of the testing effect principle to include retrieval practice through informal settings, such as self-imposed recall, sometimes termed practice testing. This study's results highlight the idea that practice testing should be applied to cognitive processes beyond simple recall. Our results demonstrate that even when students are
given exam cues in the form of rote memory details, alleviating the need to recall content solely, practice testing improves learners' application of objective skills. While these results may be counter-intuitive if considered through a traditional lens of the testing effect, they make sense when considered in the light of retrieval practice—that practicing skills improve skill acquisition, be it memory, or other.

The present experiment provides evidence supporting cued exams' role in developing higher-order scientific thinking skills in post-secondary biology students. Regardless of the mechanistic explanation, it is clear that providing cues on complex application, evaluation, and analysis exam items in post-secondary biology courses can improve learning by enhancing subsequence encoding.

Acknowledgments

We thank our undergraduate research assistants for their help with the research and data collection. We did not receive grants from funding agencies in the public, commercial, or not-for-profit sectors.
References


