Testing Effect on High-Level Cognitive Skills

Jamie L. Jensen,[†]* Mark A. McDaniel,[‡] Tyler A. Kummer,[†] Patricia D. D. M. Godoy,[§] and Bryn St. Clair[§]

[†]Department of Biology, Brigham Young University, Provo, UT 84602; [‡]Department of Psychological and Brain Sciences, Washington University, St. Louis, MO 63130-4899; [§]Department of Biology, Saint Joseph Catholic Schools, Ogden, UT 84403

ABSTRACT

The testing effect is one of the strongest learning techniques documented to date. Although the effects of testing on high-level learning are promising, fewer studies on this have been done. In this classroom application of the testing effect, we aimed to 1) determine whether a testing effect exists on high-level testing; 2) determine whether higher-level testing has an effect on low-level content retention; and 3) determine whether content knowledge, cognitive skill, or additional components are responsible for this effect. Through a series of two experiments, we confirmed a testing effect on high-level items. However, improved content retention due to testing was not observed. We suggest that this high-level testing effect is due to a better ability to apply specific skills to specific content when this application process has appeared on a previous exam.

INTRODUCTION

Recent calls for improving undergraduate science education (e.g., Association of American Universities, 2013) have motivated evidenced-based education research in science, technology, engineering, and math and have led researchers to seek collaborative partnerships in this effort (Talanquer, 2014). For example, as cognitive psychology produces information regarding principles of learning, the application of psychological principles in educational contexts is left to translational researchers to explore through discipline-specific classroom studies (Daniel, 2012). Many studies from cognitive psychology have highlighted the finding that assessments can be learning tools themselves and not merely a measurement of the success of curriculum in building student understanding. Researchers highlight assessments as a means to motivate student study behaviors and strengthen cues to understanding during the test-taking experience. Often referred to as the *testing effect*, test-enhanced learning is a promising principle demonstrated as a strengthening of retention of information that was previously tested or retrieved (for recent reviews, see Roediger *et al.*, 2011; Dunlosky and Thiede, 2013; Rowland, 2014).

The Testing Effect in the Laboratory

The majority of work on the testing effect has been focused on low-level memory tasks (Carrier and Pashler, 1992; Carpenter and Pashler, 2007; Carpenter and DeLosh, 2006; Carpenter *et al.*, 2008; Carpenter, 2009; Chan and McDermott, 2007; McDaniel *et al.*, 2007; Johnson and Mayer, 2009; Rohrer *et al.*, 2010). Researchers typically compare the final learning of two groups of participants: subjects learning new materials in a condition in which materials are studied and then re-studied are compared with subjects learning new materials in a condition in which materials are studied and then recalled. When effect sizes are reported, these testing effects qualify as large-sized effects ($\eta_p^2 > 0.14$). With the exception of a few studies reviewed next, research has not examined the testing effect with regard to outcomes reflecting high-level cognitive processes and deep conceptual understanding.

Jennifer Momsen, *Monitoring Editor*

Submitted Oct 8, 2019; Revised Apr 20, 2020; Accepted May 1, 2020

CBE Life Sci Educ September 1, 2020 19:ar39

DOI:10.1187/cbe.19-10-0193

*Address correspondence to: Jamie L. Jensen (Jamie.Jensen@byu.edu).

Several research studies conducted in the laboratory have suggested that the testing effect may play a role in higher-order thinking. It has been shown that, as a learner practices the application of subject content to a high-level cue or cognitive process, there is the potential for improved high-level process retention, and even transfer to other nontested items, with high effect ($\eta_p^2 > 0.14$; Jacoby *et al.*, 2010). Chan *et al.* (2006) found that testing on one set of content facilitated the retrieval of related (rather than the same) content on a later exam, suggesting that the testing effect may cross knowledge domains. In addition, Kang *et al.* (2011) found that, when they tested participants on a set of stimuli that reflected a functional rule (e.g., a mathematical function), the participants demonstrated better acquisition and transfer of the rule to novel stimuli relative to participants to whom they presented stimuli without testing. Similarly, Jacoby *et al.* (2010) found that testing students on bird classification schemes allowed them to better classify birds they had never seen before on a final test. Both of these studies suggest that underlying rules or principles may be subject to the testing effect as well. In addition, both Butler (2010) and Karpicke and Blunt (2011) recently showed that testing students on information allowed them to better make inferences from the information than if they had simply reread the information.

The Testing Effect in a Classroom Setting

While the testing effect has been consistently demonstrated in the laboratory, the outcomes in the classroom have only begun to be explored (Dobson and Linderholm, 2015) and are less clear (Nguyen and McDaniel, 2014; Rowland, 2014). Some researchers have described the outcome of the testing effect in applied classroom studies and have shown positive outcomes on content retention, with large effect sizes ($\eta_p^2 > 0.14$; e.g., Larsen *et al.*, 2009; McDaniel *et al.*, 2012; also Orr and Foster, 2013, but effect sizes not reported). For example, in a classroom experiment, quizzing students with application-level questions (Bloom's taxonomy; Bloom, 1984) produced better transfer of the target content to new application questions than when the content was not quizzed (McDaniel *et al.*, 2013). The benefit extended to quiz questions on the basic content of these questions as well. Jensen and others (2014) also showed an effect of testing in a biology classroom, although effect size was small ($\eta_p^2 < 0.06$; Jensen *et al.*, 2014). Some studies, however, have shown mixed or less conclusive results in the classroom (for a review, see Nguyen and McDaniel, 2014).

Recent work has focused on examining possible boundaries of the testing effect. Leahy *et al.* (2015) and Hanham *et al.* (2017) have suggested that the testing effect may not be obtainable using items with "high element interactivity"; that is, test items that require a student to process multiple related content items at the same time. Others have found that the testing effect may not extend to performance on summative test questions that are not similar to questions used for initial retrieval practice (Wooldridge *et al.*, 2014), thereby potentially limiting broad classroom application. Successful pedagogical application of the testing effect requires further discipline-based education research to understand and describe mechanisms that apply.

Bloom's Taxonomy

Appropriate classroom application of testing effect research may hinge on instructor alignment of their proposed learning outcomes and their assessment designs. This alignment of learning outcomes and assessments is based on categorization within the revised Bloom's taxonomy (Anderson *et al.*, 2001). In the present study, we refer to test items that target the first two levels of Bloom's as *low-level* items, a common convention (lower-order cognitive skills, or LOCS; see Zoller, 1993). We grouped together test items that reflected the three upper levels of Bloom's taxonomy (application, analysis, and evaluation) and refer to these as *high-level* items (higher-order cognitive skills, or HOCS; i.e., as opposed to the low-level items that required retention and basic comprehension only). This is consistent with recent practical recommendations for writing and considering multiple-choice exam items that reflect different levels of learning (Zimmaro, 2016, p. 26). Higher-level skills, sometimes referred to as scientific reasoning skills, are highly correlated with biology achievement at higher levels of Bloom's taxonomy (Lawson *et al.*, 2000a) and are closely associated with science process skills (e.g., controlling variables, interpreting data, and drawing conclusions). Presumably, in addition to content knowledge, these skills are necessary to perform on higher-level items.

The taxonomy was originally intended to be hierarchical (Anderson *et al.*, 2001; Krathwohl, 2002). In other words, to solve an evaluation problem, students would need to be familiar with the basic content of the question and then be able to apply the appropriate higher-order skill to successfully complete the problem. However, research on the hierarchical nature of Bloom's is mixed (e.g., Author [JLJ, MAM, and TAK] Jensen *et al.*, 2014; Kropp and Kropp, 1966; Madaus *et al.*, 1973; Seddon, 1978; Hill and McGaw, 1981).

To accommodate this view, Anderson and others (2001) created a revised taxonomy that adds an additional dimension to the original taxonomy, including both a subject matter content aspect and a cognitive process aspect (Anderson *et al.*, 2001). The content, or knowledge domain, includes factual knowledge, conceptual knowledge, procedural knowledge, and metacognitive knowledge; while the cognitive process domain includes the six levels of cognitive skills discussed earlier. Thus, any item can be classified by the content it requires *and* the process(es) applied to solve the item. In this respect, low-level items, as presently defined, require the student to only remember or understand the different knowledge domains. A high-level item, according to this definition, requires the student to be able to apply a HOCS to one or more of the knowledge domains. The revised taxonomy can be used to classify specific learning objectives by both the knowledge and cognitive process(es) involved. From this description, we can imply that a high-level item includes both a content component and a skill component.

As indicated at the outset, the fact that testing on low-level items enhances content retention is well established (for an extensive review, see Roediger and Karpicke, 2006); however, as noted, the research is scarce on the testing effect when the initial tests focus on high-level items (for ease of exposition, we label this *high-level testing*). The current study aims to investigate the effect of high-level testing on performance for high-level items on a criterion test and tries to parse out the possible factors identified in the preceding theoretical analysis mediating this effect; that is, an increase in the target content, cognitive skill, or a combination of both.

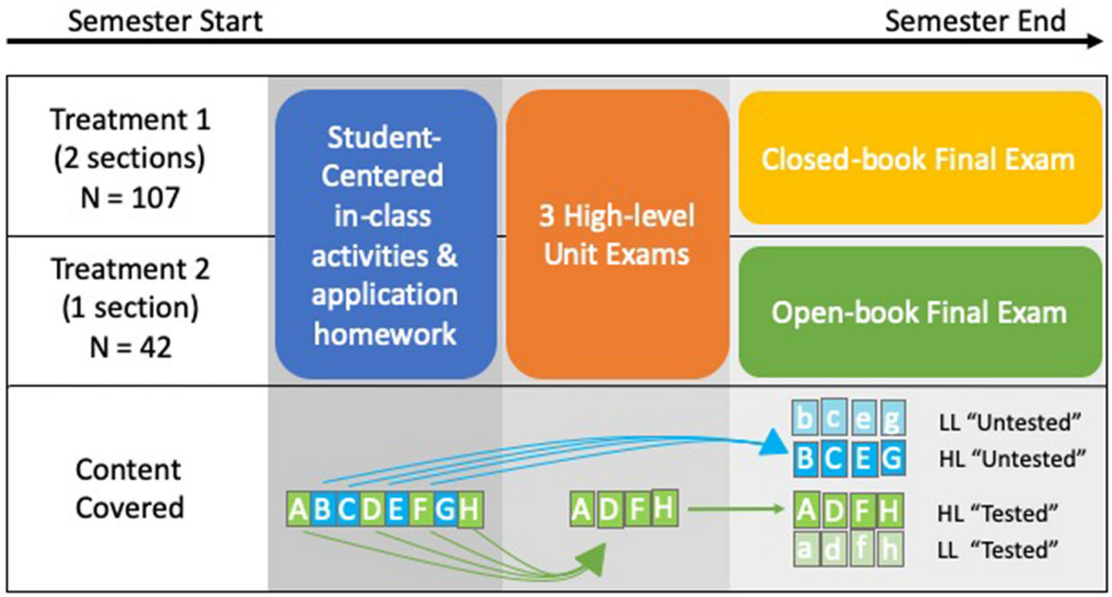

FIGURE 1. Experimental setup. Treatments 1 and 2 received identical student-centered in-class activities and application homework covering all content of the course. Both treatments were administered unit exams covering half of the content. Both treatments were administered a final exam containing tested and untested content at both low (LL) and high levels (HL) of Bloom's taxonomy. Treatment 2 was given access to their books and notes for the final exam.

OVERVIEW OF THE CURRENT EXPERIMENT

We report a classroom experiment to examine whether high-level testing (unit exams) enhances content acquisition and also extends to the enhancement of performance on final exam questions that require analysis and evaluation. We also attempt to identify the mechanisms supporting the testing effect on high-level questions. Finally, we describe a novel framework that we adopted to characterize high-level cognitive performance.

We conducted the experiment in the context of a university introductory biology course to address these issues. We examined the effect of high-level testing on final exam items. The final exam included both high- and low-level items that covered content previously tested on unit exams (using different items; referred to as *tested content*) and content covered in the class but never tested on a unit exam (referred to as *untested content*). We reasoned that the comparison of items covering tested content with items covering untested content would reveal the presence of a testing effect on high-level items and associated content retention on low-level items. Second, we further analyzed the high-level testing effect to determine whether content, cognitive skills, or a combination of both was responsible for the effect (on high-level final exam items). To do so, we set up two final exam testing conditions. In treatment 1, students took high-level exams throughout the semester and then a closed-book final exam consisting of both low- and high-level items that covered both tested and untested content. Treatment 2 was identical to the treatment 1 but with an open-book final exam allowing students ready access to content and thus eliminating content knowledge as a limiting factor in performance and somewhat isolating the contribution of cognitive skills to exam performance. Figure 1 graphically depicts the experimental setup.

The goal of treatment 1 was to examine the effect of testing in an authentic classroom context on final closed-book exam performance on both high- and low-level items. Higher performance on high-level summative test items covering tested content relative to high-level items covering untested content would indicate a high-level testing effect. Higher scores on final low-level items associated with tested content over untested content would support an increase in basic content retention from high-level testing.

The goal of treatment 2 was to evaluate the testing effect in conditions in which the requirement to retrieve content knowledge from memory is removed for the final exam. To do this, we administered the same final exam to a different section of an identical course to that in treatment 1, but allowed students to take it open book and open note, thus affording students open access to any content required by the exam. In several studies, researchers administered open-book exams to provide content to students in order to free their minds up for more complex problem solving and found higher achievement in an open-book treatment (e.g., Schumacher *et al.*, 1978; Moore and Jensen, 2007; Williams and Wong, 2009; Agarwal and Roediger, 2011; Stowell, 2015). Researchers presumed that the increased achievement in an open-book treatment is due to the availability of content (Teodorczuk *et al.*, 2017). If the testing effect seen on high-level questions were a factor of being better able to recall content, then we would expect the testing effect on high-level items to be mitigated in an open-book treatment. If, however, the testing effect were due to a better ability to use the cognitive skills necessary to apply the content at higher levels of Bloom's taxonomy, then the testing effect would remain, regardless of the availability of content.

TABLE 1. Unit exam scores for each treatment

	Closed book		Open book	
	Mean	SD	Mean	SD
Exam 1	82.26	8.97	81.11	7.80
Exam 2	78.73	11.18	78.22	13.52
Exam 3	76.49	13.72	73.14	14.27

METHODS

Subjects

We recruited 149 students enrolled in three daytime sections of an introductory biology course for nonmajors at a private university in the western United States to participate in the study. It should be noted that our population consists of high-achieving students (average entering ACT score is 27) who are relatively homogeneous in ethnicity (primarily white, non-Hispanic) and culture. Thus, appropriate considerations should be made when drawing conclusions from this study. The three sections were held back-to-back at 11 am, 12 pm, and 1 pm. The course is part of the general education requirements for the university; thus all students not majoring in the life sciences are required to take the course. The course enrollment is a generalized, representative sample of the university student body. Participants ranged from first years to seniors and came from a variety of disciplines outside the life sciences. The Institutional Review Board at the primary author's institution reviewed and approved this study. Students granted their consent for participation. Treatment 1 consisted of 117 students (58 in one section and 59 in the other), 10 of whom declined to sign a permission form. Treatment 2 consisted of 45 students in a third section, three of whom declined to sign a permission form. All sections were taught by the same instructor in the same classroom using the same Course Design, as described below. Data from two sections were combined in the analyses for treatment 1. Data from the third section were used for treatment 2. For comparison, average unit exam performances for each treatment (all of which were closed book) are listed in Table 1.

To measure group equivalence between treatments, we included three metrics administered to both the closed-book and open-book students. 1) To measure students' initial scientific reasoning ability, we administered Lawson's Classroom Test of Scientific Reasoning (LCTSR v. 2000; Lawson, 1978) at the beginning of the semester. The LCTSR consists of 24 items assessing various aspects of scientific reasoning in a content-independent manner. Lawson *et al.* (2000b) discuss scoring procedures, validity, and reliability of the test. 2) To measure the biology content knowledge with which students entered the classroom, we gave students a short test of biology content called the Biology Knowledge Assessment (BKA). The BKA consists of 26 multiple-choice questions targeting basic biological content consistent with an introductory course and was designed by the authors. Reliability of the instrument was low (Spearman-Brown coefficient = 0.51), thus it was only used to establish a baseline level to assess group equivalence. It was not used in a pretest/posttest design to determine student learning. 3) We compared student performance on the first unit exam of the course. Students scored statistically equivalently on all three measures (see Table 2).

Course Design

We patterned the course after Bybee's 5-E learning cycle (Bybee, 1993). This means that the course structure included two phases: a phase during which basic concepts are constructed through an exploratory, inquiry-based framework (encompassing the *engage*, *explore*, and *explain* portion of the 5-E learning cycle), followed by a concept application phase, usually given as homework, wherein what they have learned is applied to novel contexts to strengthen their conceptual understanding (encompassing the *elaborate* phase of the 5-E learning cycle). Students completed follow-up clicker or online quizzes at the completion of each lesson to accomplish the *evaluate* portion of the learning cycle. This means that each class had an accompanying homework assignment and approximately every other class had a quiz. Students completed a minimum of 3 hours outside class for the three 50-minute class periods spent in class per week. This time estimate does not include study time spent by students independently. In addition, we included suggested textbook readings with each lesson; however, we did not build in accountability for reading. We provided students with a list of all learning outcomes for each class period.

The course comprised three units: ecology and mechanisms of evolution; genetics and the cell cycle; and cells, chemistry, and metabolism. Students took three unit exams throughout the semester and could choose one of two formats, mini-exams or full exams. Exams consisted of multiple-choice questions written entirely at application level or above of Bloom's taxonomy (Bloom, 1984). Full exams included 75 questions taken all at once at the end of a 4-week unit in the university's testing center facility. If students chose to take mini-exams, they took four equal portions of the 75 questions at the end of each week of the unit (i.e., they took the same items, just split into four smaller pieces). Students were given this option to accommodate different testing styles and test anxiety. This decision was based on several studies that suggested that some students

TABLE 2. Descriptive statistics for group equivalency measures[a]

Measure of group equivalence	Treatment	Participants (*n*)	Mean (%)	SD (%)	*t*(*df*), *p*
Exam 1	Closed book	107	82.3	9.0	0.73(147), 0.47
	Open book	42	81.1	7.8	
LCTSR	Closed book	89	79.0	17.1	0 < 0.01(117), 1.00
	Open book	27	79.0	16.4	
BKA	Closed book	91	43.9	11.9	0.11(114), 0.92
	Open book	28	44.2	14.5	

[a]Not all students completed all pretests. Numbers of participants from each treatment are indicated. Independent samples *t* tests reveal no differences between groups, as indicated.

benefit from taking more frequent tests on less material (see Leeming, 2002; De Paola and Scoppa, 2011; Sedki, 2011; Phelps, 2012). Students were given the opportunity to choose which method they would use with each unit exam. Approximately a third chose full exams consistently, a third chose mini-exams consistently, and a third changed their method at least once during the semester. We compared unit exam performance between students who chose mini-exams and those who chose full exams and found no statistically significant differences on any of the unit exams. In addition, all students took a comprehensive final exam at the end of the semester that is Outcome Measure described below. Again, we found no differences on any component of the final exam between those who chose mini-exams and those who chose full exams. Five components made up a student's overall grade in the course: 25% homework assignments, 25% class participation, 25% unit exams, 6% quizzes, and 19% final exam.

Outcome Measure

To evaluate testing effects, we designed a comprehensive final exam that included 124 multiple-choice items. Twenty-four of the items consisted of the LCTSR used as an initial measure of group equivalence and administered on the final to test for changes in scientific reasoning. Twenty of the items were extension questions to test for transfer of reasoning used for a different experiment and were deleted from analysis in the current experiment. We designed the remaining 80 items for the present experiment, and experts evaluated and grouped them into four categories. To categorize items, three independent researchers evaluated the items and grouped them into Bloom's categories. Researchers met and discussed the items until preliminary agreement was reached. We then verified our categorization with one additional rater who independently assessed the items. Agreement between original raters and the new independent rater was 86.2%. We gave items where disagreement remained to two additional raters for independent rating. Agreement between raters rose to 93.8%. We removed one item due to a printing error, three items for ambiguity of Bloom's level due to disagreement among raters, and two items due to lack of discriminating power (i.e., they were too easy).

The remaining 74 items showed 100% agreement between raters. These items were as follows: 22 low-level questions covering content used previously in a high-level unit exam (LL tested), 24 low-level questions covering content never seen in a previous exam (LL untested), 16 high-level questions covering content previously tested on a high-level unit exam (HL tested), and 12 high-level questions covering content never seen in a previous exam (HL untested). See Figure 1 for an illustration; see Appendix A in the Supplemental Material for a list of content covered in tested and untested items and Appendix B for a complete listing of high-level tested and untested questions for comparison. To statistically determine whether tested and untested items were of equal difficulty, we would have had to administer the items to a new set of students as unit exam items throughout the semester before our study, such that no testing was done before them, a task that proved prohibitive. However, we attempted, by expert opinion, to keep items of equivalent difficulty and spread of Bloom's levels. We have included these items in Appendix B in the Supplemental Material so that the reader can compare them. We did not construct the unit and final test questions such that there were equal numbers of apply, analyze, and evaluate questions within each unit or final exam nor were there equal proportions of these questions across the unit exams and the final exam. Rather, the test questions were constructed to align with the course goals to generally require higher-level thinking. Sometimes a particular type of question (application, analyze, evaluate) might have more straightforwardly assessed higher-level thinking for particular content than another type. If so, we used that type of higher-level question, rather than construct a different type of higher-level question that seemed strained for particular content. This practice aligns with recommendations for multiple-choice question design to test high-order thinking (Zimmaro, 2016).

We ran a Cronbach's alpha for the LL items combined (0.71) and the HL items combined (0.68) and determined that they were adequate, given that these items tested a wide range of biology content. None of the tested items were exact repeats of unit exam items. They simply used content that was previously tested using a different item. It is important to note that tested content on unit exams may have appeared in more than one item. On the final, however, each tested content item was only tested once. As a reminder, the differentiation between low- and high-level items was based on Bloom's taxonomy (Bloom, 1984). The bottom two tiers of the taxonomy, knowledge and comprehension, define low-level items. These items require only content knowledge or LOCS (Zoller, 1993). The next three tiers, application (high-level only), analysis, and evaluation, define high-level items. We did not include "create" items, as exams were multiple choice. These items require both content knowledge and HOCS (Zoller, 1993), defined as critical thinking and problem solving. In treatment 1, this exam was taken in the university testing center with no time limit (except the time limitations of the testing center). In treatment 2, students were allowed open access to their textbook as well as any notes taken during the course. They were not given access to the Internet. We gave students in the online treatment a 6-day window with unlimited time to allow for ample time to look up any required content.

During each unit, we provided students with a list of learning outcomes in preparation for unit exams. However, the high-level unit exams covered only a portion of these learning outcomes. Thus, on the final exam, "tested" items were those that covered content previously included on a unit exam. Generally, low-level tested items asked students to recall content that was required by items on prior high-level unit exams, whereas high-level tested items closely resembled items seen on the unit exams, in content and skill, but with the scenario changed. "Untested" items on the final exam were those that used content not previously tested on a unit exam but required in the course learning outcomes and covered in class activities. High-level untested items covered similar critical-thinking skills as those targeted by the high-level tested items but were applied to content not covered on a unit exam (but again, covered by expected learning outcomes and class activities). Low-level untested items on the final exam covered content not previously included in a high-level unit exam item. For example, we asked students to be able to apply all mechanisms of evolution (i.e., natural selection, mutation, genetic drift, gene flow, and nonrandom mating) in the course learning outcomes, and we taught all of those topics during class. However, the unit exam

may have only tested students on genetic drift and natural selection. Thus, a tested final exam item would also test students on genetic drift or natural selection; whereas an untested final exam item would test them on mutation, gene flow, or nonrandom mating. Tables 3 and 4 illustrate sample high- and low-level tested and untested items. While it was impossible to counterbalance items given the classroom setup (i.e., making tested items untested for some students, and vice versa) and thus eliminate the possibility of an item effect, the tested and untested nature of the items was an arbitrary choice by the instructor. Thus, while the possibility of an item effect cannot be definitively ruled out, we would expect that the random selection of items as tested versus untested would approximately equalize item difficulty.

Data Analyses

To determine whether a testing effect was demonstrated on high-level items requiring problem solving (i.e., application, analysis, and evaluation), we compared student scores in treatment 1 on high-level tested final exam items with high-level untested final exam items. Similarly, to determine whether testing using high-level questions (on unit exams) could extend an increased content retention to low-level items, we compared student scores on low-level tested final exam items with low-level untested final exam items. We performed post hoc power analyses using *G**power to determine the power to detect an effect of testing for high-level items and low-level items, with our *p* value set to 0.025 for these simple effects tests to control for alpha inflation. (Note that, for each of these within-subject comparisons, power is determined in part by the overall correlation between the two conditions; we provide those correlations here so that the power analyses are completely transparent.) For high-level items (r (106) = 0.501), the power to detect a medium-sized effect (i.e., effect size $f = 0.25$) was 0.997, and for low-level items (r (106) = 0.407), the power was 0.993. We also performed a two-way repeated-measures analysis of variance (ANOVA) with level (HL and LL) and tested or untested as within-subject variables using the SPSS statistical package v. 21 (2012). Effect sizes are reported as partial eta-squared values (η_p^2). These values can be interpreted as the proportion of variance in the dependent variable that is related to this particular factor, partialing out the variance of the other factors. By convention, the cutoffs for small, medium, and large effect sizes are 0.01, 0.06, and 0.14, respectively (Cohen, 1988).

To determine whether a testing effect was present in treatment 2 (the open-book treatment), we first conducted a two-way repeated-measures ANOVA and separate one-way repeated-measures ANOVAs for each item level paralleling that from treatment 1. Similar to treatment 1, the power (determined by *G**power) to detect a medium-sized testing effect with *p* value set to 0.025 for LL items (r (41) = 0.719) was very high (0.974) and for HL items (r (41) = 0.805) was sufficiently high (0.805). We then combined the data from both treatments to determine the effect of requiring retrieval of content on the final exam versus not requiring retrieval on low- and high-level item performance for open- versus closed-book final exams. We used a mixed-model ANOVA with level (LL and HL) and tested nature (tested or untested) as within-subject variables and whether the exam was open or closed book as a between-subjects variable. We used simple effects tests to follow up significant interactions, applying a Bonferroni correction to account for alpha inflation.

TABLE 3. Sample high-level tested and untested items

	Unit exam item	Final exam item
Tested item (content: photosynthesis)	You are trying a risky experiment with your newly purchased beta fish. You obtain a large glass jar with a lid and fill it with water, soil, an *Elodea* plant (an aquarium plant), some plankton (little invertebrates who eat algae), some algal spores (photosynthetic, single-celled eukaryotes), and your fish. You then seal the lid and hope for the best. What ingredient are you missing in order for this ecosystem to thrive? a) Oxygen b) A food source for your fish c) *Sunlight* d) Carbon dioxide	If green algae cells in a buffer solution containing only inorganic salts are placed in a sealed container at room temperature with excess carbon dioxide gas and exposed to light, the cells will a) *Live for many hours and multiply.* b) Live for several hours, but fail to multiply because there is no source of carbon in the buffer solution. c) Live for several hours, but fail to multiply because no oxygen is present. d) Die rapidly, because no oxygen is present.
Untested item (content: cell cycle regulation)	(No unit exam item)	The passage of a cell through the checkpoints of the cell cycle is tightly controlled by the manufacture of a protein called cyclin. As cyclin concentrations build up, they bind to an ever-present enzyme, cyclin-dependent kinase (cdk) that activates the cell cycle. To turn the cell cycle off, cyclin is destroyed. Cancer would most likely be caused by a) An inactivation of the cyclin geneAn inactivation of the *cdk* gene a) An overactivation of the cyclin gene b) *An overactivation of the cdk gene* c) Both *a* and *b* would cause cancer d) Both *c* and *d* would cause cancer

TABLE 4. Sample low-level tested and untested items[a]

	Unit exam item	Final exam item
Tested item (content: mutations in gene expression)	The mRNA written below is the sequence of the hypothetical *Billy* gene. 5′Cap-AUGGCCAAUCCGCUCCGAAGUGGCGCGUGU-CUGUGAAAAAAAAAAAAAAAA-3′ What kind of mutation would I cause if I changed the highlighted letter to an A? a) Silent point mutation b) *Missense point mutation* c) Nonsense point mutation d) Frameshift	A genetic mutation which changes the resulting amino acid in the coded protein from an Arginine (Arg) to a Serine (Ser) is a a) Silent mutation b) *Missense mutation* c) Nonsense mutation d) Frameshift mutation e) Spontaneous mutation
Untested item (content: photosynthesis alternative pathways)	(No unit exam item)	Due to the extreme heat, desert plants have modified processes of carbon fixation (the gathering of CO_2 for photosynthesis). Which process do they use? a) C3 photosynthesis b) C4 photosynthesis c) *CAM photosynthesis* d) Photorespiration

[a]As a reminder, all unit exams were high-level. Thus, unit exam questions are example high-level questions that covered the low-level content.

Finally, to directly determine whether an open-book exam mitigated the HL testing effect observed with a closed-book exam, we conducted a two-way mixed-factor ANOVA on HL items only. A power analysis to detect an effect was again performed using *G*power with a *p* value set to 0.05 (with the correlation between HL tested and untested items, r (148) = 0.504). The power to detect a medium-sized interaction (effect size $f = 0.25$) was extremely high (0.999), and the power to detect even a small-sized interaction (effect size $f = 0.10$) was not unreasonable (0.683; note these power values are the same for detecting a main effect of testing). For completeness, we conducted a parallel ANOVA on LL items only (with the correlation between LL tested and untested items, r (148) = 0.556). The power to detect a medium-sized interaction was again outstanding (0.999), and the power to detect a small-sized interaction was relatively good (0.730).

RESULTS

To evaluate the simple main effects of testing in treatment 1, we conducted a one-way repeated-measures ANOVA for each item level, which indicated that a testing effect was present on high-level items ($M_{HLtested}$ = 61.45, SD = 14.17; $M_{HLuntested}$ = 45.17, SD = 17.47; $F(1,106) = 109.90$, $p < 0.001$, $\eta_p^2 = 0.51$) but not on low-level items ($M_{LLtested}$ = 60.62, SD = 12.7; $M_{LLuntested}$ = 60.55, SD = 15.31; $F(1,106) = 0.002$, $p = 0.96$, $\eta_p^2 < 0.001$). (We used a Bonferroni-corrected p value, setting the level of significance to $p = 0.025$, to account for alpha inflation.) This apparent interaction between item level and the presence of a testing effect was confirmed by the two-way repeated-measures ANOVA: $F(1,106) = 61.97$, $p < 0.001$, $\eta_p^2 = 0.37$, for the interaction between the level of the final exam question and the tested or untested nature of the content (see Figure 2).

Results exhibited the same pattern for the open-book final exam as for the closed-book final exam, with increased success due to previous testing on high-level items but no increased content retention on low-level items. A two-way repeated-measures ANOVA on the open-book treatment showed a main effect of the level of question, indicating that students performed better on low-level items than on high-level items ($F(1,41) = 14.78$, $p < 0.001$, $\eta_p^2 = 0.265$, see Figure 3). Importantly, it also revealed a significant interaction between the item level and tested nature of the item ($F(1,41) = 19.26$, $p < 0.001$, $\eta_p^2 = 0.32$). We found a testing effect on high-level items ($M_{HLtested}$ = 64.88, SD = 16.21; $M_{HLuntested}$ = 51.19, SD = 19.44; $F(1,41) = 23.56$, $p < 0.001$, $\eta_p^2 = 0.37$, see Figure 3), but no testing effect on low-level items ($M_{LLtested}$ = 71.10, SD = 14.33; $M_{LLuntested}$ = 70.93, SD = 16.55; $F(1,41) = 0.009$, $p = 0.925$, $\eta_p^2 = 0.01$).

The 2 × 2 × 2 mixed-factor ANOVA on the combined data revealed no interaction between tested and untested items and the exam treatment (open vs. closed book; $F(1,147) = 0.72$, $p = 0.40$, $\eta_p^2 < 0.01$), suggesting that the testing effect was equally present in both conditions. It also showed that students performed better on low-level items than on high-level items, but that an interaction was present between this effect and whether they took the exam open or closed book, $F(1,147) = 9.38$, $p = 0.003$, $\eta_p^2 = 0.06$. Figure 3 shows that access to the book benefited students more on low-level items ($M_{Open\text{-}book\ LL}$ = 71.0 vs. $M_{Closed\text{-}book\ LL}$ = 60.6) than on high-level items ($M_{Open\text{-}book\ HL}$ = 59.0 vs. $M_{Closed\text{-}book\ HL}$ = 54.5). Post hoc analyses reveal that having the book and notes improved student scores only on low-level items (LL tested, $p < 0.001$; LL untested, $p < 0.001$; HL tested, $p = 0.07$; HL untested, $p = 0.20$). Additionally, consistent with the ANOVAs for treatment 1 and 2 separately, we found a significant interaction between the level of the item and prior testing (tested, untested), $F(1,147) = 60.82$, $p < 0.001$, partial $\eta^2 = 0.29$, and this interaction did not differ in magnitude regardless of whether the exam was taken open or closed book (a level by tested by open/closed interaction, $F(1,147) = 0.50$, $p = 0.48$).

Finally, separate 2 (test–untested) × 2 (open–closed book) ANOVAs comparing LL items only and HL items only found no interactions (LL: $F(1,147) < 0.01$, $p = 0.97$; HL: $F(1,147) = 0.72$, $p = 0.40$). The testing effect for HL items was again robust, $F(1,147) = 97.01$, $p < 0.001$, $\eta^2 = 0.40$, and importantly, there was no testing effect for LL items, even with reasonable power to detect a small-sized effect (0.730).

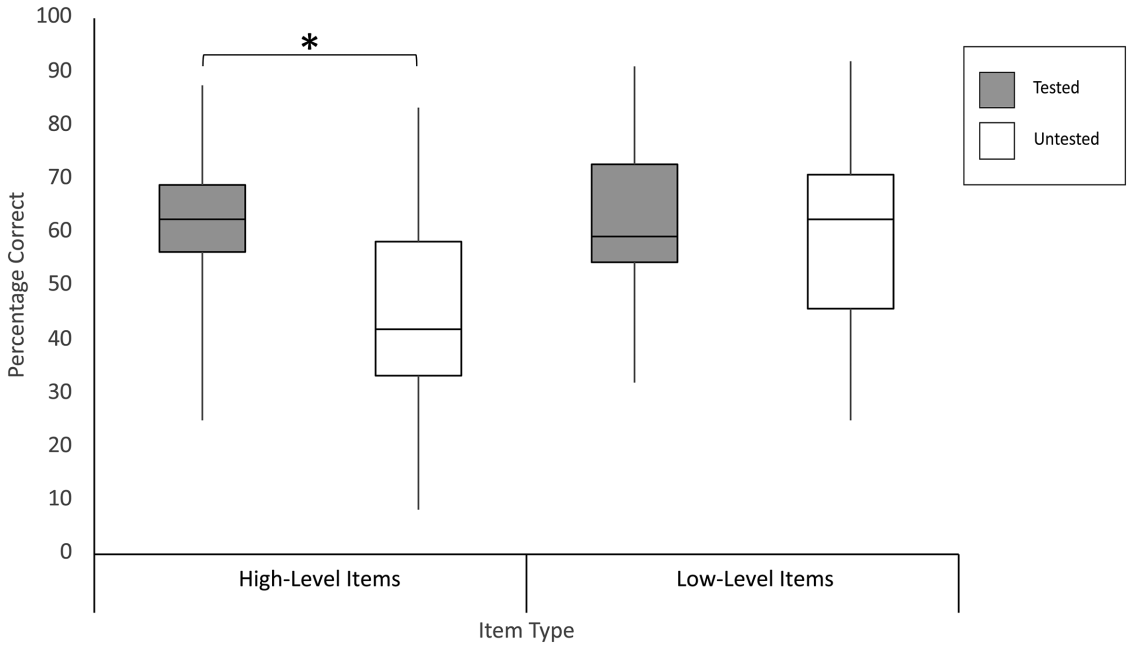

FIGURE 2. Scores on high- and low-level items are shown. "Tested" indicates that the item covered content previously tested on a unit exam. "Untested" indicates that the item covered content not previously tested on a unit exam. Error bars represent mean standard errors. *$p < 0.001$.

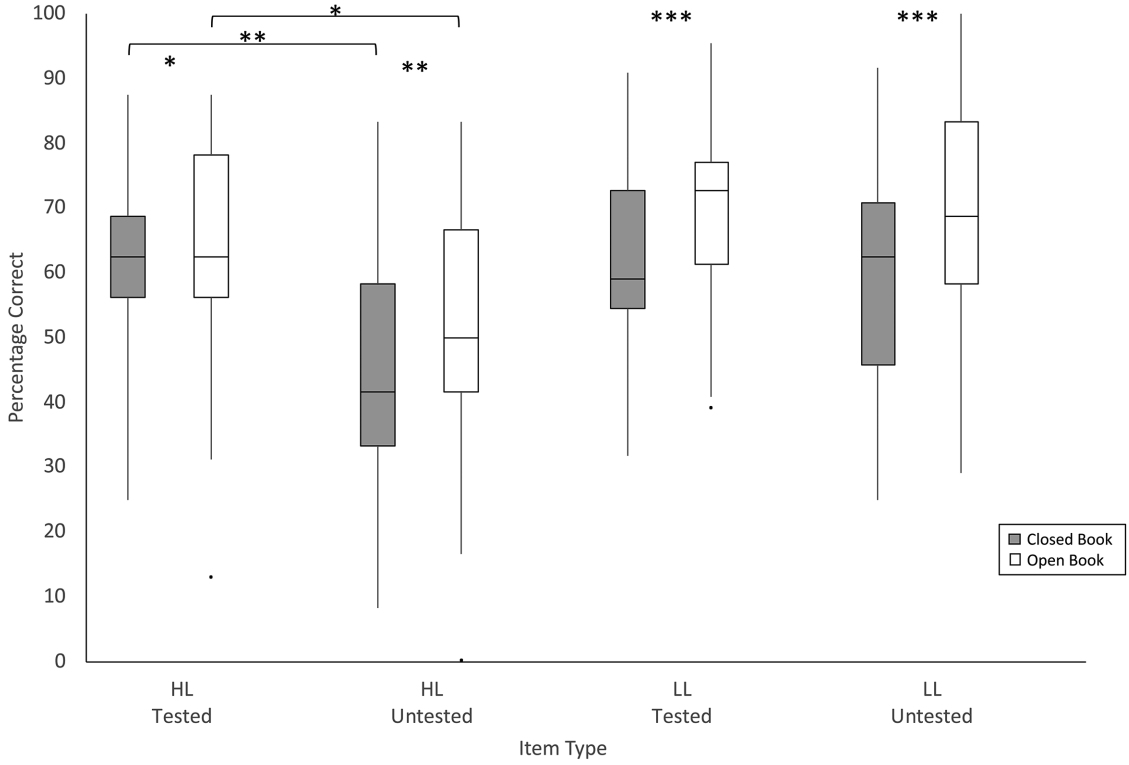

FIGURE 3. Scores on high- (HL) and low-level (LL) items are compared between the closed- and open-book treatments. "Tested" indicates that the item covered content previously tested on a unit exam. "Untested" indicates that the item covered content not previously tested on a unit exam. Error bars represent mean standard errors. *$p < 0.10$; **$p < 0.05$; ***$p \leq 0.001$.

DISCUSSION

Our results indicate that a testing effect was present on high-level items. Students performed better on final exam items that covered content previously tested by a high-level unit exam item over content not covered on a previous exam. This adds further support to previous studies that have shown some effects of testing using questions requiring the application of content (cf. Carpenter, 2012; McDaniel *et al.*, 2013; Thomas *et al.*, 2018) and extends it to higher levels of Bloom (e.g., analysis and evaluation). This led us to question the extent to which this high-level testing effect was a consequence of increasing access of relevant content, increasing facility with requisite reasoning skills in using the content, or a combination of both.

From the open-book treatment, we see that performance on low-level items increased with available access to content. As was expected, low-level items were almost completely dependent upon content knowledge; what little students missed was likely due to a lack of effort in finding the content or perhaps due to a misunderstanding of a question. Time on task was not recorded and could be another potential source of differences to be considered. However, performance on high-level items did not benefit significantly by having content readily available, suggesting that content knowledge is not sufficient to perform at higher levels of Bloom's taxonomy. This suggests that high-level items have additional components involved, which several researchers would suggest are HOCS (Zoller, 1993; Crowe *et al.*, 2008). Even more intriguing is the persistence of the testing effect on high-level items despite the availability of content. This strongly suggests that learners are using cognitive processing skills across domains.

The testing effect did not enable greater retention of low-level content knowledge. That is to say, there was no measurable difference in performance between tested (previously tested with high-level questions only) and untested low-level items in either treatment. Whether this is because both tested and untested content retention benefited equally from high-level testing or because neither benefited is impossible to distinguish using the current procedure. However, in light of a recent review outlining the limited or absent transferability of the testing effect (Pan and Rickard, 2018), we would suggest it is likely the latter explanation, that neither benefited. We recommend further experimentation.

To recapitulate, our study reveals several important findings: 1) The testing effect extends to items encompassing the higher levels of Bloom's taxonomy, and 2) a testing effect on high-level Bloom's items does not translate to better low-level content retention. After considering how these findings extend the present testing-effect literature, we amplify on a new view of performance on high-level questions, which incorporates content knowledge, cognitive skills, *and* the ability to connect skills to the content that we suggest our findings warrant. By designing items on both tested and untested content at both low and high levels of Bloom's on a final criterion test, we detected a testing effect on high-level items. These results are similar to those found in McDaniel *et al.* (2013) with middle school students, in which application of a concept in a testing situation facilitated application of this same concept to a new situation on a follow-up test. This suggests that testing may facilitate further application of content acquisition. Because our high-level questions included analysis and evaluation items (in addition to application items), our study extends these previous findings to even higher levels of Bloom's taxonomy. Further, it is important to note that, unlike many testing effect studies, this study was done in an authentic classroom in which initial testing of the material took the form of somewhat high-stakes unit exams (~27% of the course grade), as opposed to lower-stakes formative testing that is many times used in testing effect studies (e.g., McDaniel *et al.*, 2012, 2013; Trumbo *et al.*, 2016; Thomas *et al.*, 2018). Studies in a laboratory setting show that stakes may influence the magnitude of the testing effect, the higher the stakes, the less the testing effect is present (Hinze and Rapp, 2014).

In this study, the effect of testing that we saw on high-level items did not extend to better retention of low-level content knowledge. As a reminder, throughout the semester, students took high-level exams covering a variety of content. We identified the content involved in these questions and used it to create low-level, recall-type items identified as tested items on the final exam. In addition, we used content encompassed in lesson learning outcomes that did not appear in any high-level unit exam questions to design low-level, recall-type items that represented untested items on the final exam. If it were the case that using pertinent content to answer high-level questions (on previous unit exams) allowed students to better recall that content later (a direct effect, similar to that seen in Carpenter and DeLosh, 2006) or to better prepare for the final exam in regard to that content (an indirect effect), we would have expected to see higher scores on tested low-level items over untested low-level items on the final exam. However, this possible testing effect for low-level items did not emerge in the current experiments.

There are several potential explanations for this finding. First, it is important to remember that, in this study, we deviate from a common design, in that low-level items were not explicitly tested as such (i.e., as low-level questions) on prior tests. Instead, in our experiments, the prior tests were always high-level questions. Thus, the result is not surprising in terms of prior research on the testing effect. However, the absence of a testing effect for low-level items might be surprising from the perspective that Bloom's taxonomy could reflect a hierarchy such that answering a high-level question requires consideration of that content at the lower level (remembering, understanding). There was a nominally higher score on low-level tested items (relative to untested), but these differences were slight and did not reach statistical significance (collapsed across experiments, $LL_{tested} = 65.9$ vs. $LL_{untested} = 65.7$). One possibility is that the effect is too small to detect given our sample size. It is noteworthy, however, in the analysis that collapsed across experiments, that the power to detect even a small testing effect for low-level items was reasonable (0.730); still, the testing advantage was not significant. It may also be that, given the high-level nature of the course and unit exams, with little to no emphasis on learning low-level items, the effect was diluted. In many basic laboratory studies, participants are tested directly on the "remembering" level for the target content (i.e., low-level Bloom taxonomy items); whereas, in this authentic study, we are making the assumption that, by performing on high-level test items, students necessarily had to remember the low-level content (see e.g., Jensen *et al.*, 2014). However, this does not seem to strengthen the memory of this content on a later

test where it is explicit. Perhaps explicit testing is necessary to achieve a detectable effect. A review of the current literature suggests that the testing effect may have limited transferability (see Pan and Rickard, 2018). It may also suggest that Bloom's taxonomy may not be as hierarchical as originally proposed (Anderson *et al.*, 2001; Krathwohl, 2002). In a recent study (Pan *et al.*, 2019), a lack of a connection was shown: researchers found that, even explicit testing on low-level terminology that resulted in a clear testing effect did not lead to an increase in conceptual learning on higher-level items. Alternatively, it is also possible that Bloom's is hierarchical but without explicit testing, the testing effect fails to come through.

High-Level Testing and Cognitive Skill

In a review of the literature, Carpenter (2012) found that testing on content facilitates the transfer of the content knowledge to new situations across concepts after varying intervals of time between the initial test and the recall, across testing formats (e.g., free response vs. multiple choice), and across knowledge domains of both target concepts and rules for problem solving. In the case of the present study, we extend this transfer of "rules" to include high-level cognitive skills. As Karpicke and Aue (2015) point out, practicing retrieval enhances learning of complex materials in educational settings, yet the boundaries of the testing effect as they apply to interaction of learned processes can be further illuminated through classroom-appropriate studies such as this one.

Higher-order items are certainly not content independent; they also require knowledge of content specific to the question. According to the revised taxonomy (Anderson *et al.*, 2001), the two main components required to perform on a high-level item are knowledge and process. To determine which of these (or both) is playing a role in a high-level testing effect, we attempted to control for content knowledge by making it readily available through an open-book final exam. Our results indicate that content is not sufficient to improve high-level performance. Students in the open-book treatment did not experience an increase in performance on high-level questions when content was available. However, the testing effect remained. This suggests that enhanced content knowledge is not exclusively responsible for a testing effect on high-level questions and that cognitive skill (i.e., cognitive processes) is more likely the key component. Interestingly, the application of these skills to specific content in a testing situation did not translate to better performance when that skill was applied to different content, suggesting that testing strengthens only the direct application of the skill to the specific content.

Limitations

Many limitations arise when conducting such experiments in an authentic classroom setting. Student behaviors in between class sessions, students' extenuating circumstances, and the many other factors that play a role in overall student performance cannot be tightly controlled. Thus, conclusions drawn are preliminary and potentially bounded by a multitude of factors that remain unmeasured. That being said, authentic application of the testing effect provides practical evidence for real-world pedagogical practices.

Interestingly, because control of student behavior between class periods is not possible in this authentic situation, it makes it difficult to distinguish between a direct and indirect testing effect (for additional discussion, see McDaniel *et al.*, 2013). Thus, while it appears from our data that initial testing on content provides direct benefit on the final exam, it is possible that students placed more study emphasis on materials that had been tested over materials that had not. However, had this been the case, we would have expected to see better performance on the associated low-level content, which we did not see. That being said, it is debatable as to whether students possess the sophistication and metacognition necessary to appropriately study for high-level questions such that the benefit would be evident. Certainly, further study is warranted. Again, however, this is a practical application of testing and would suggest that, if an instructor wants students to spend the time studying information, that information ought to appear on a test. Prior research supports the idea that testing directly encourages study behaviors (Leeming, 2002).

Another potential confound is that we taught material in between sessions. We did not explicitly control which examples we used for which topics and whether these examples more closely resembled tested or untested items from the exams. Thus, it is possible that some classroom exposures may have unduly influenced certain items on the test over others, making tested items potentially easier. However, it is unlikely that all exposures in class were on tested items, causing an increase in that score over untested items. As noted previously, it was also impossible to statistically determine equivalence of difficulty between tested and untested high-level items. We tried to design the items to be equivalent, but perhaps untested items proved more challenging. However, it is may be that they proved more challenging due to the testing effect, as we would suggest.

One puzzling finding is that, when we gave students open access to their notes and the textbook, students still did not achieve 100% or even close (they scored around 70%) on low-level items. There are many possible reasons that can be offered for this finding. First, students did not have access to the Internet, so the information was not necessarily very easy to access, although it was all available in course materials and text. Another plausible reason is that many students may have had incomplete class notes (i.e., the material was presented in class, but they failed to write it down). Finally, it is possible that students were overconfident in their understanding and simply did not bother to check themselves. Prior studies suggest that students who expect an open-book exam may put less effort into studying or preparing for the test (Agarwal *et al.*, 2008; Ioannidou, 1997). Knowing they have the ability to rely on outside resources may lead to overconfidence and less effective preparation (Anaya *et al.*, 2010; Durning *et al.*, 2016). But then, when it came time to actually find the answers, students were unprepared. Or perhaps they lacked the desire to check themselves, given that this was a general education requirement, and an "A" is not necessarily required. However, this latter explanation is less likely, given the high-achieving, highly competitive student body at the present institution. And given our high-achieving, motivated students, it is possible that these findings may differ with more heterogeneous student populations. Specifically, it is possible that less motivated or underprepared students may not experience the same effect of testing on high-level items as our current student body, especially if more

guessing was involved in the test-taking process. In addition, less motivated, or less prepared students may not experience the metacognition that often occurs when a student gets questions wrong and works to better understand the materials (i.e., the indirect testing effect). Therefore, it is recommended that readers take our population into consideration when interpreting our results and that similar studies be run on other student populations and in other educational contexts to confirm these findings.

In addition, it should be noted that our sample sizes were rather small (although a sample of more than 100 used to confirm the high-level testing effect is fairly standard). We performed post hoc power analyses to try to justify our results given our sample sizes, but certainly, additional studies are warranted. Furthermore, this study was conducted in a single semester with a single, rather homogeneous group of students, as stated earlier. Again, studies to broaden the applicability of our results are certainly warranted.

CONCLUSION

Whereas, the revised Bloom's taxonomy only takes into account the content and skills necessary to perform on a given test item, evidence from this study suggests that a third component may be at play: the ability to connect cognitive skills to specific content. We suggest that it is this third component that the testing effect influences, at least in part. The testing effect is a result of a strengthening of the student's ability to make the connection between content and skill and not necessarily an increase in content alone or skill alone. This relationship would suggest that performance on high-level items requires three distinct, yet interacting, components: content knowledge, cognitive skill, *and* the ability to connect skills to the specific content (see Figure 4). Without this third component, students may often miss the mark, even though they have both the content and the skills available to them. This ability to connect the skill to the content is what focuses student thinking toward the target application. In essence, this allows the student to extend his or her content knowledge into the upper levels of Bloom's taxonomy. Without that practiced connection, content knowledge remains only accessible at lower Bloom's levels.

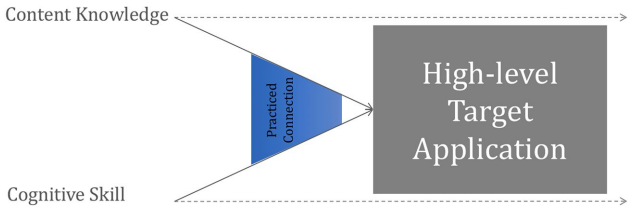

FIGURE 4. A proposed model of performance on high-level exam questions that incorporates content knowledge, cognitive skills, *and* the ability to connect skills to the content. The testing effect may be a result of a strengthening of the student's ability to make the connection between content and skill and not necessarily an increase in content or skill individually. This relationship would suggest that performance on high-level items requires three distinct, yet interacting, components: content knowledge, cognitive skill, *and* the ability to connect skills to the content.

One may question whether this ability to connect is dependent upon a tested situation. Interestingly, in the current study, students performed application activities in the form of homework after every class period. These homework assignments were designed to cover all of the learning outcomes for a given unit, not just the ones that appeared on the test. Thus, it appears that practicing the application of skills to content in an untested, low-stakes manner is not enough to invoke the ability to make the connection on a final criterion test. Instead, it may be the opportunity to make the connection in a high-stakes testing condition, with the opportunity for students to review and receive feedback on their performances, that strengthens that connection over the connections made during class and homework. Thus, to provide the most optimal learning experience for students, to ensure that the concepts taught are transferable into higher levels of Bloom's, we propose that students not only be given opportunities to use the content at higher levels of Bloom's in class and on homework, but also that they be tested in a high-level manner that allows them to practice connecting skills to content in a way that will endure.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Jerry Johnson for extensive feedback on the article and undergraduate researchers for help in design 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., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861–876.

Agarwal, P. K., & Roediger, H. L. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory*, *19*, 836–852.

Anaya, L., Evangelopoulos, N., & Lawani, U. (2010). Open-book vs. closed-book testing: An experimental comparison. Paper presented at: 2010 Annual Conference & Exposition (Louisville, KY). Retrieved April 17, 2020, from https://peer.asee.org/16901

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Boston: Allyn & Bacon.

Association of American Universities. (2013). *Framework for systemic change in undergraduate STEM teaching and learning*. Washington, DC: AAU Undergraduate STEM Education Initiative.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*, 4–16.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology*, *26*(5), 1118–1133.

Bybee, R. W. (1993). *Reforming science education*. New York: Teachers College, Columbia University.

Carpenter, S., & Delosh, E. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*, 268–276.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*(3), 474–478.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1563–1569.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*, 279–283.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*, 438–448.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642.

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 431–437.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*, 553–571.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in bloom: Implementing Bloom's taxonomy to enhance student learning in biology. *CBE—Life Sciences Education*, *7*, 368–381.

Daniel, D. B. (2012). Promising principles: Translating the science of learning to educational practice. *Journal of Applied Research in Memory and Cognition*, *1*, 251–253.

De Paola, M., & Scoppa, V. (2011). Frequency of examinations and student achievement in a randomized experiment. *Economics of Education Review*, *30*, 1416–1429.

Dobson, J. L., & Linderholm, T. (2015). Self-testing promotes superior retention of anatomy and physiology information. *Advances in Health Sciences Education*, *20*, 149–161. https://doi.org/10.1007/s10459-014-9514-8

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, *24*, 58–61.

Durning, S. J., Dong, T., Ratcliffe, T., Schuwirth, L., Artino, A. R., Boulet, J. R., & Eva, K. (2016). Comparing open-book and closed-book examinations: A systematic review. *Academic Medicine*, *91*, 583e599.

Hanham, J., Leahy, W., & Sweller, J. (2017). Cognitive load theory, element interactivity, and the testing and reverse testing effects. *Applied Cognitive Psychology*, *31*, 265–280.

Hill, P. W., & McGaw, B. (1981). Testing the simplex assumption underlying Bloom's taxonomy. *American Educational Research Journal*, *18*, 93–101.

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

Ioannidou, M. K. (1997). Testing and life-long learning: Open-book and closed-book examination in a university course. *Studies in Educational Evaluation*, *23*, 131–139.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1441–1451.

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, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*(3), 621–629.

Kang, S. H. K., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, *18*, 998–1005.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*, 317–326.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*, 212–218.

Kropp, R. P., & Kropp, R. P. (1966). *The construction and validation of tests of the cognitive processes as described in the taxonomy of educational objectives*. Tallahassee: Institute of Human Learning and Department of Educational Research and Testing, Florida State University.

Larsen, D. P., Butler, A. C., & Roediger, H. L. I. II. (2009). Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial. *Medical Education*, *43*, 1174–1181.

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, *15*, 11–24.

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

Lawson, A. E., Clark, B., Cramer-Meldrum, E., Falconer, K. A., Kwon, Y. J., & Sequist, J. M. (2000b). The development of reasoning skills in college biology: Do two levels of general hypothesis-testing skills exist? *Journal of Research in Science Teaching*, *37*, 81–101.

Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review*, *27*, 291–304.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*, 210–212.

Madaus, G. F., Woods, E. M., & Nuttall, R. L. (1973). A causal model analysis of Bloom's taxonomy. *American Educational Research Journal*, *10*, 253–262.

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

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

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

Moore, R., & Jensen, P. A. (2007). Do open-book exams impede long-term learning in introductory biology courses? *Journal of College Science Teaching*, *36*, 46–49.

Nguyen, K., & McDaniel, M. A. (2014). Using quizzing to assist student learning in the classroom: The good, the bad, and the ugly. *Teaching of Psychology*, *42*, 87–92.

Orr, R., & Foster, S. (2013). Increasing student success using online quizzing in introductory (majors) biology. *CBE—Life Sciences Education*, *12*, 509.

Pan, S. C., Cooke, J., Little, J., McDaniel, M. A., Foster, E. R., Connor, L. T., & Rickard, T. C. (2019). Online and clicker quizzing on jargon terms enhances definition-focused but not conceptually focused biology exam performance. *CBE—Life Sciences Education*, *18*(4), ar54. doi: 10.1187/cbe.18-12-0248

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*, 710–756.

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing*, *12*, 21–43.

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*, 382–395.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233–239.

Rowland, C. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432–1463.

Schumacher, C. F., Butzin, D. W., Finberg, L., & Burg, F. D. (1978). The effect of open- vs. closed-book testing on performance on multiple-choice examinations in pediatrics. *Pediatrics*, *61*, 256–261.

Seddon, G. M. (1978). The properties of Bloom's taxonomy of educational objectives for the cognitive domain. *Review of Educational Research*, *48*, 303–323.

Sedki, S. S. (2011). Student performance on exam frequency: A comparative study of St. Mary's University and the American University of Sharjah. *Journal of International Education Research*, *7*, 1–4.

Stowell, J. R. (2015). Online open-book testing in face-to-face classes. *Scholarship of Teaching and Learning in Psychology*, *1*, 7–13.

Talanquer, V. (2014). DBER and STEM education reform: Are we up to the challenge? *Journal of Research in Science Teaching*, *51*, 809–819. doi: 10.1002/tea.21162

Teodorczuk, A., Fraser, J., & Rogers, G. D. (2017). Open book exams: A potential solution to the "full curriculum"? *Medical Teacher*, *40*, 529–530.

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

Trumbo, M. C., Leiting, K. A., McDaniel, M. A., & Hodge, G. K. (2016). Effects of reinforcement on test-enhanced learning in a large, diverse introductory psychology course. *Journal of Experimental Psychology: Applied*, *22*, 148–160.

Williams, J. B., & Wong, A. (2009). The efficacy of final examinations: A comparative study of closed-book, invigilated exams and open-book, open-Web exams. *British Journal of Educational Technology*, *40*, 227–236.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition*, *3*, 214–221.

Zimmaro, D. M. (2016). *Writing good multiple-choice exams*. Austin: University of Texas at Austin Faculty Innovation Center. Retrieved March 1, 2020, from https://facultyinnovate.utexas.edu/sites/default/files/writing-good-multiple-choice-exams-fic-120116.pdf

Zoller, U. (1993). Are lecture and learning compatible? *Journal of Chemical Education*, *70*, 195.