Educ Psychol Rev (2014) 26:307–329
DOI 10.1007/s10648-013-9248-9

RESEARCH INTO PRACTICE

Teaching to the Test…or Testing to Teach: Exams Requiring Higher Order Thinking Skills Encourage Greater Conceptual Understanding

Jamie L. Jensen · Mark A. McDaniel · Steven M. Woodard · Tyler A. Kummer

Published online: 16 January 2014

Abstract In order to test the effect of exam-question level on fostering student conceptual understanding, low-level and high-level quizzes and exams were administered in two sections of an introductory biology course. Each section was taught in a high-level inquiry based style but was assigned either low-level questions (memory oriented) on the quizzes and exams, or high-level questions (application, evaluation, and analysis) on the quizzes and exams for the entirety of the semester. A final exam consisting of 20 low-level and 21 high-level questions was given to both sections. We considered several theoretical perspectives based on testing effect, test expectancy, and transfer-appropriate processing literature as well as the theoretical underpinnings of Bloom's taxonomy. Reasoning from these theoretical perspectives, we predicted that high-level exams would encourage not only deeper processing of the information by students in preparation for the exam but also better memory for the core information (learned in the service of preparing for high-level questions). Results confirmed this prediction, with students in the high-level exam condition demonstrating higher performance on both the low-level final-exam items and the high-level final exam items. This pattern suggests that students who are tested throughout the semester with high-level questions acquire deep conceptual understanding of the material and better memory for the course information, and lends support to the proposed hierarchical nature of Bloom's taxonomy.

Keywords Assessment · Bloom's taxonomy · Biology · Testing effect · Test expectancy

Introduction

The American Association for the Advancement of Science (AAAS) and the National Science Foundation (NSF) in their recent (2010) call to action, *Vision and Change*, along with an older

J. L. Jensen (✉) · S. M. Woodard · T. A. Kummer
Department of Biology, Brigham Young University, 401 WIDB, Provo, UT 84602, USA
e-mail: Jamie.Jensen@byu.edu

M. A. McDaniel
Department of Psychology, Washington University, Campus Box 1125, One Brookings Drive,
St. Louis, MO 63130, USA

but still resonating call by the National Research Council, as part of the *National Science Education Standards* (1996), suggests that a more active learning approach can lead to greater gains in true conceptual understanding as well as greater retention in the STEM subjects. Following this advice, many instructors have put forth efforts to transform their instructional approaches to reflect a more active, student-center philosophy. However, many default to previously written exams based on recalling massive amounts of biological facts rather than focusing on the science process skills they are trying to teach (J. L. Momsen et al. 2013). Unfortunately, there are potentially negative consequences from this arrangement. First, the exams do not appear to reflect the goals of these courses, and likely the goals of all instructors: to improve student reasoning and encourage deep conceptual understanding through practices implemented in class as well as through the encouragement of appropriate study habits by students outside of class. The second potential negative consequence is the focus of the present study. The level of exam questions (e.g., a focus on retention of facts) themselves may influence student learning throughout the course. As we develop in more detail below, these influences could be a consequence of learning from testing (see Roediger and Karpicke 2006, for review), from test expectancy effects fostered by the level of quizzes and exams administered in the course (McDaniel et al. 1994; Thiede et al. 2011), or both.

In a recent review of the relative assessment literature, Joughin (2010) discussed the three seminal and most influential works in establishing the influence of assessments on students: Becker, Greer, and Hughes' *Making the grade: The academic side of college life* (1968); Snyder's *The hidden curriculum* (1951); and Miller and Parlett's *Up to the mark: A study of the examination game* (1974). Joughin concluded that there was reasonable support for several tenets surrounding research on assessment's role in student learning: assessments influence a students' distribution of efforts, their approach to learning, and their study behaviors. However, Joughin notes that most of the studies have weaknesses that limit their generalizability. In addition, we note that they do little to inform researchers on the causal mechanisms behind the effects seen. Several additional reviews support Joughin's conclusions showing connections between assessments and student motivation (Harlen and Deakin Crick 2002; Van Etten et al. 2008) and student approaches to learning (Dickie 2003; Struyven et al. 2005).

Based on both the laboratory work in cognitive and educational psychology (e.g., Thiede et al. 2011) and the classroom studies mentioned above, we posit that the nature of the quizzes and unit exams that students receive throughout the semester potentially impact cognitive aspects of their learning, both the effort and strategy they utilize for learning during classroom activities and the type of self-directed learning (e.g., study strategies) in which they participate outside the classroom. Accordingly, we hypothesize that the quiz and unit exam levels uniquely impact student performances on their final exam, even when course content and instruction are held constant.

In this study, assessment levels differed according to Bloom's Taxonomy (Bloom 1984). Bloom's taxonomy of cognitive domains is a well-established framework for categorizing the assessment items into six levels according to the thinking patterns required (hereafter referred to as "Bloom's"). Figure 1 illustrates the revised version of Bloom's (Anderson et al. 2001). It is generally accepted that the first two levels of Bloom's, *remember* and *understand* require only minimal levels of understanding and are considered lower-order cognitive skills (Crowe et al. 2008; Zoller 1993). It has been suggested that the third level of Bloom's, *apply*, is at an intermediate level (Crowe et al. 2008); whereas, the three higher levels of Bloom's (*analyze, evaluate, and create*) require higher-order cognitive skills (Zoller 1993). The taxonomy was originally designed to be a cumulative hierarchy where mastery of the lower levels of the taxonomy were prerequisite to performance on higher levels (Anderson et al. 2001; Krathwohl 2002). For example, a question written at the *analyze* level of the taxonomy would require

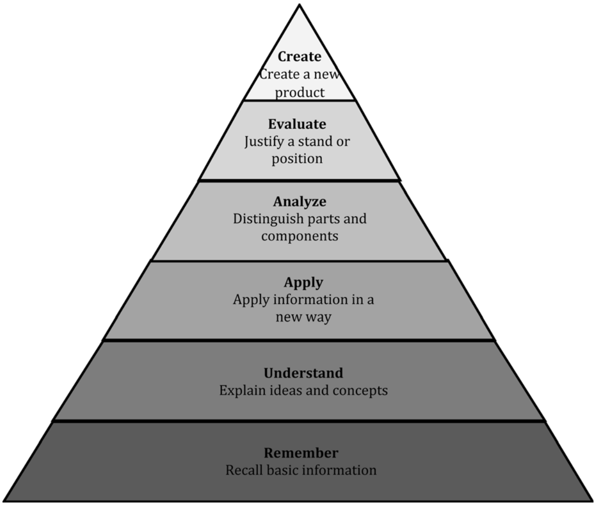

Create
Create a new product
Evaluate
Justify a stand or position
Analyze
Distinguish parts and components
Apply
Apply information in a new way
Understand
Explain ideas and concepts
Remember
Recall basic information

Fig. 1 The revised Bloom's taxonomy

mastery of the basic content (i.e., *remember* and *understand*) in order to perform the analysis required by the question. Many researchers have attempted to test the assumptions of Bloom's model and its hierarchical nature, but the results have been mixed (Hill and Mcgaw 1981; Kropp et al. 1966; Madaus 1973; Seddon 1978).

From an applied perspective, Momsen et al. (2010) established that biology instructors' learning objectives were usually written at higher levels of Bloom's, implying that educators embrace these higher levels as of central importance, whereas basic researchers often emphasize that mastery at lower levels of the taxonomy are critically important as well (e.g., Sternberg, Grigorenko, and Zhang 2008). In the present study, we retain the terminology "high-level" and "low-level" questions to reflect the origins of the distinction we draw, but not to confer implications about the respective putative importance of these levels of assessment and performance.

To test the effect of exam item level (low-level Bloom's or high-level Bloom's questions) on student learning (performance on the final exam), we compared two sections of introductory biology. In one section, all quizzes and unit exams were written at the "Remember" level of Bloom's thereby requiring no more than memorization of material to perform well (for purposes of exposition, we label this the *low-level* condition). In the other section all of the quizzes and unit exams were written at high levels of Bloom's designed to require higher-order cognitive skills (labeled the *high-level* condition). Students in both sections then received an identical cumulative final exam, composed of lower-level questions focused on memory for factual knowledge and higher-level questions focused on application, analysis, and evaluation.

We reasoned that the performance patterns on the final exam could be influenced by the learning gained from the prior quizzes and unit exams themselves (a direct effect of testing) and/or by students adjusting their in-class learning and/or out-of-class study strategies throughout the semester to match the expected assessments (an indirect effect of testing, such as an effect of *test-expectancy*, e.g., see McDaniel et al. 1994). We emphasize at the outset, that similar to many other studies conducted in the classroom examining the influence of quizzing/testing on authentic learning outcomes (e.g., course exams; (McDaniel et al. 2013; McDaniel et al. 2012; Roediger et al. 2011), the current study does not allow determination of whether the source of any test-level (low-level, high-level) effects are a consequence of direct testing effects, text-expectancy effects (an indirect effect), or a host of other indirect effects that might vary across the test-level manipulation (such as effects on metacognitive accuracy, study policies that are more frequent or spaced, exposure to different information). Nevertheless, we can still appeal to the testing effect and test expectancy literatures to anticipate several possible patterns of outcomes from the present test-level manipulation.

The Testing Effect

The quizzing/exam manipulations could directly benefit final exam performance. Termed the 'testing effect', it has been shown that actively retrieving target information is much more effective at enhancing retention of that material than simply re-reading (or restudying) the information. Testing-effect research has generally focused on test questions requiring simple retrieval of facts (S. K. Carpenter and DeLosh 2006; S. K. Carpenter et al. 2008; S. K. Carpenter and Pashler 2007; S. K. Carpenter et al. 2009; Carrier and Pashler 1992; Chan and McDermott 2007; Johnson and Mayer 2009; McDaniel et al. 2007; Rohrer et al. 2010) and aligns most directly with the current low-level quiz/exam condition. The testing effect (and associated processes) for high-level questions has received less attention, but the evidence suggests that testing can increase transfer of knowledge to new questions (Shana K. Carpenter 2012). Related to the present study, McDaniel et al. (2013) found evidence that suggests that quiz questions requiring application of a concept enhance performance on exam questions targeting a different application of the same target concept. Further, these application quiz questions also benefited performance on exam questions targeting memory for the concept terminology (relative to when the terminology was not quizzed). In contrast, quiz questions targeting only memory of concept terminology did not facilitate application of these concepts on a later test.

In the current study, the high-level condition included not only *application* questions but also *analyze* and *evaluate* questions. Cautiously extending from the above preliminary patterns, we reasoned that the high-level quiz/exam questions could directly improve performance on high-level questions on the final exam. We also reasoned that high-level quiz/exam questions could produce performance on low-level final exam items equal to that of the low-level quiz/exam questions. Of course, it remains possible that high-level quiz/exam questions (which included higher levels than application, as used in (McDaniel et al. 2013) would promote substantial elaboration of conceptual information, thereby yielding an advantage on low-level items as well.

By contrast, any direct testing effects in the current study could follow a transfer-appropriate processing (TAP) pattern (e.g., Fisher and Craik 1977; McDaniel et al. 1978) that is that significant benefits are seen when the level of questions on the prior tests align with the level of questions on the final exam when previously tested concepts/constructs are targeted. We acknowledge that the TAP explanation has not fared well in accommodating standard findings from the testing-effect literature in terms of matching/mismatching *question format* (multiple choice vs. short answer questions; S. K. Carpenter and DeLosh 2006; Kang et al. 2007; McDaniel et al. 2007; McDermott et al. in press). Nevertheless, it remains entirely possible that the view would apply to the present circumstances. Here the conditions differ in terms of question *level*, with one level requiring remembering (low-level) and the other level focusing on problem solving (high-level analysis and evaluation). Thus, misaligning the quiz/exam question level with the final exam question level may be associated with profound processing differences, thereby giving students a selective advantage on items on the final exam which align most closely in processes with the quiz/exam items. This leads to the clear possibility of a cross-over interaction such that the low-level condition produces better performance on the low-level final exam items than does the high-level condition, with the reverse true for the high-level final exam items.

Test Expectancy Effect

Recent laboratory findings suggest that test-expectancies stimulate studying that appropriately matches the demands of the anticipated test (Finley and Benjamin 2012; Thiede et al. 2011). Applied to the present context, quizzes and unit exams throughout the semester would provide students with practice on different types of test questions (as in Finley and Benjamin and in Thiede et al.) and could thus stimulate test expectancies that might guide students' study activities. Expectations about how test-expectancy could influence performance hinge on the kinds of preparation (studying) that high-level quiz/exam questions stimulate. One account might suggest that in preparing for high-level quiz/exam questions, a student would focus on problem solving, analysis, and evaluation thereby specifically honing those skills. These honed skills would then lead to an advantage of the high-level condition students on high-level final exam questions; similar advantage would be expected of low-level condition students on low-level final exam questions. This reasoning suggests a transfer-appropriate processing pattern like that described in the above paragraph and reported in laboratory conditions (cf., Thiede et al.).

A more complex account hinges on the idea that in order to perform at higher levels of Bloom's, students must have mastered the lower levels (i.e., remembering and understanding basic terminology specific to the subject) (Bloom 1984). Accordingly, the students' expectancy that high-level Bloom questions will appear on the test would presumably require students to master the basic facts before extending this learning to focus on applying the facts and using them for analysis. Indeed, under this expectancy the basic facts should be elaborated with regard to analysis and evaluation, thereby increasing retention of the facts per se (cf. McDaniel and Donnelly 1996, with regard to astrophysics concepts; also see Mayer 2003, for a review of the positive effects on learning of answering deep-level questions while studying material). Thus, the idea is that students in the high-level condition, having a high-level expectancy, will outperform students in the low-level condition on both low-level and high-level final exam items.

To summarize, a consideration of effects (testing effect and test-expectancy effect) that could plausibly be operative in the current study suggests two possible general patterns. One pattern would reflect selective benefits of each condition on performance on matched final exam items. The other pattern would reflect cascading effects of the high-level quiz/exam questions on both high-level and low-level final exam items. The idea here is that application, analysis, and evaluation encompass processes that benefit retention (memory). Such effects would be evidenced by higher performance of the high-level condition on high-level final exam items, and critically by equivalent or even superior performance on low-level final exam items.

Student attitudes toward assessments are generally negative, regardless of format, given the effect the assessments can have on their course grade. However, students likely enter class with pre-conceptions about assessment format. In an analysis of 77 introductory biology courses (and 9713 assessment items), Momsen et al. (2010) found that 93 % of the items were assessing low-level Bloom's (remember or understand). Given that this is the case, we expect that students, who clearly expect low-level exams from an introductory biology course, will likely be dissatisfied with high-level exams presented in this study, and may express this dissatisfaction in the form of increased negative comments about the exam.

Methods

Ethics Statement

Permission for human subjects use was obtained by the Institutional Review Board of the first author's university and written consent was obtained from all participants.

Course and Participants

The participants were undergraduate students enrolled in two sections (~90 students each) of non-majors general biology at a large private western university. The course, which met three 1-h periods per week, is part of the general education required core and covered the entire biology curriculum, from molecular and cellular biology, to genetics and biotechnology, to evolution and ecology. Both sections were taught in the same inquiry fashion using the learning cycle (Bybee 1993; Lawson 2002): specifically, exploratory activities to introduce each unit followed by term introduction and concept application activities. Homework assignments were identical between sections. Students enrolled in the course are typically non-science majors and range from freshman to seniors.

Experimental Design

A quasi-experimental nonequivalent groups design was utilized. Steps were taken to ensure as much group equivalence as possible among the two treatment groups [i.e., same instructor, identical classrooms, course materials, textbooks (Belk & Maier, *Biology: Science for Life*), resources, curriculum, and expected learning outcomes]. One section was assigned to a low-level (herein referred to as LL; N=84) assessment format and the other section was assigned to a high-level (herein referred to as HL; N=85) assessment format that encompassed all weekly quizzes as well as 3 unit exams throughout the semester. In both conditions, the quizzes consisted of 10 questions of various formats (multiple-choice, fill-in-the-blank, and short answer) administered through the course management system at the end of each unit. The exams were 100 multiple-choice questions and were administered in the University Testing Center, a center where exams are proctored to students outside of class time. Both sections took a common final exam, which consisted of half low-level and half high-level questions. The final exam was comprehensive and tested the same (or similar) concepts but with new scenarios such that no question appearing on a quiz or unit exam was repeated on the final. Low-level items were defined as "Remember" on the revised Bloom scale (Anderson et al. 2001). High-level items were defined as items falling into "Apply," "Analyze," and "Evaluate." [Note that none of the items were "Create" due to the constraints of multiple-choice testing.] These items required students to go beyond a simple understanding of the concepts and use them appropriately, e.g., apply them to a new situation, analyze data and draw appropriate conclusions, or evaluate the validity of information based on these concepts. In order to create a distinct difference between exams, we chose not to include items at the "Understand" level. In summary, low-level items required memory of terms and definitions, whereas, high-level items required remembering a concept's definition and then the use of a higher-order skill using this concept (either application, analysis, or evaluation). Occasionally, a particular low-level item on the low-level unit exam was directly subsumed within a high-level item on the corresponding high-level exam in the alternative treatment (for an example, see Appendix, questions i and xi; both cover haploid number). However, more often, the concepts were related, but the high-level item was not just an extension of the low-level item (for an example, see Appendix, questions vii and xviii; both cover electron shells and atomic bonding, but solicit different aspects within this concept. Selected items are included in the Appendix. Items were classified by three independent researchers trained in Bloom's taxonomy. Items were topic-matched between both exam formats to ensure that the same content was being tested in each condition. In addition, based on the instructor's extensive experience with the material and the course, low-level items were selected to ensure that exams were equally difficult between treatments (confirmed by exam averages, as reported in the Results section).

There were a total of 14 weekly quizzes. Quizzes were taken independently and answers were not discussed in class. Students could choose to visit a teaching assistant outside of class to discuss quizzes and get feedback if they so desired (very few students took advantage of this). However, students were not allowed access to quizzes outside of TA office hours and thus were quizzes were not available to be used as study materials. There were three unit exams spaced evenly throughout the semester. Exams were never discussed in class and students could not take exams home as study materials. We did, however, incentivize students with five points of extra credit (a relatively nominal portion of their 895 total points in the course) to go and discuss their exams with teaching assistants. They were allowed to see the exam but not take it with them. On average, about 50 % of students in each condition chose to take advantage of this opportunity.

Dependent Measures

Initial Reasoning Ability A key factor involved in performance, especially in biology, is scientific reasoning ability. Scientific reasoning ability is correlated with college level biology achievement; it is also very closely related to science process skills (e.g., controlling variables, interpreting data, drawing conclusions) and is highly correlated with a student's ability to perform at higher levels of Bloom's (Lawson et al. 2000b). Thus, students with higher reasoning abilities have an advantage on test items requiring procedural skills (e.g., science process skills). To control for this possibility, we assessed student reasoning ability using Lawson's Classroom Test of Scientific Reasoning Skills (LCTSR, ver. 2000, Lawson 1978) and used it as a covariate in our analysis of achievement scores.

The LCTSR consists of 24 items used to assess initial reasoning ability. Scoring procedures, validity and reliability of the test are discussed in Lawson et al. (2000a). Briefly, scores from 0 to 8 are level 3, or concrete operational thinkers. Scores from 9 to 14 are low level 4, or students transitioning from concrete to formal operations. Scores from 15 to 20 are high level 4, or students transitioning from formal to post-formal operations. Scores from 21 to 24 are level 5, or post-formal operational thinkers. The reasoning test was administered as an in-class assignment at the beginning of the course, and students were given a fixed number of points for its completion.

Achievement Student achievement was assessed by a common course final exam. The exam, informed by several standardized biology exams, consisted of 20 low-level multiple-choice items and 21 high-level multiple-choice items. Items were designed and then categorized into Bloom levels by three independent researchers trained in assessing levels of Bloom's Taxonomy. Items were discussed and modified until all raters came to an agreement on the Bloom's level. Because so many different constructs were being measured by this exam, reliability was difficult to determine and overall internal consistency was not expected to be high. A Cronbach's alpha for the 41 content questions was determined to be 0.66. Students were assigned a low-level achievement score by averaging their performance on the 20 low-level items and a high-level achievement score by averaging their performance on the 21 high-level items.

Attitudes Student comments were taken from end-of-semester student course evaluations. Students were asked to respond to the statement “Evaluations are a good measure of learning” on an eight-point Likert scale from “Very Strongly Disagree” to “Very Strongly Agree.” Students were also given the opportunity to give any additional comments about the course in a free response portion. Comments about the exams were taken from each treatment condition and used to qualitatively judge student satisfaction with the evaluation component of the course.

Results

Reasoning Ability

Students were administered the 24-question LCTSR at the beginning of the semester to assess whether the treatment conditions were equally matched. Student scores indicated that the HL treatment group, on average, had slightly higher reasoning skills than the LL treatment group ($M_{\mathrm{HL}}=19.4$, $M_{\mathrm{LL}}=17.8$, $t_{(167)}=2.61$, $p=.01$, $\eta_p^{2}=.039$). However, both treatment groups had formal reasoning skills (defined by scores above 14; see (Lawson et al. 2000c) and therefore should have been capable of learning the theoretical concepts in the course. Nevertheless, as a consequence of this difference, the LCTSR was used as a covariate in analyses on effects of treatment.

Quiz and Unit Exam Performance

Table 1 displays raw scores on each of the 14 quizzes within each treatment group. Scores on quizzes varied from approximately 5.5 to 9 out of a possible 10 points. Simple comparisons showed that on quizzes 1, 3, 6, 9, and 11, the HL treatment

Table 1 Average quiz scores between treatment conditions

Quiz	Treatment condition		*t*	*p*
	Low-level *M* (*MSE*)	High-level *M* (*MSE*)		
1[a]	8.40 (0.23)	9.04 (0.16)	2.32	0.02
2	7.74 (0.26)	7.90 (0.17)	0.53	0.59
3[a]	7.16 (0.21)	8.03 (0.21)	2.92	<0.01
4	7.38 (0.28)	7.07 (0.18)	0.94	0.35
5[b]	7.96 (0.18)	6.46 (0.25)	4.72	<0.01
6[a]	6.71 (0.17)	7.48 (0.23)	2.69	<0.01
7[b]	7.72 (0.23)	6.96 (0.30)	2.06	0.04
8	6.64 (0.22)	6.42 (0.27)	0.60	0.55
9[a]	7.27 (0.19)	8.15 (0.16)	3.48	<0.01
10	6.48 (0.25)	7.08 (0.27)	1.64	0.10
11[a]	5.58 (0.26)	6.53 (0.28)	2.50	0.01
12[b]	8.12 (0.18)	6.38 (0.23)	5.82	<0.01
13[b]	7.49 (0.24)	6.46 (0.29)	2.72	<0.01
14	6.94 (0.19)	6.66 (0.23)	0.94	0.35

[a] Indicates that the high-level condition performed significantly better than the low-level condition

[b] Indicates that the low-level condition performed significantly better than the high-level condition

group outperformed the LL treatment group; whereas, on quizzes 5, 7, 12, and 13, the LL treatment group outperformed the HL treatment group. The remaining quizzes were statistically equivalent. Thus, no pattern emerged that would indicate that one condition was losing motivation throughout the semester or that one condition was disproportionately benefitting from quizzing.

Table 2 displays mean percentages on each of the 3 unit exams as a function of exam level (low level, high level). Inspection of this table reveals that means were virtually identical across the two types of exams. A 3 (exam number)×2 (treatment condition) mixed model analysis of covariance (ANCOVA), with LCTSR as the covariate, confirmed that mean scores did not differ across the low-level and high-level exams ($F<1$), and that this equivalence held for all three exams [F (2, 332)=1.44, $p>.23$, $\eta_p^2=.009$], for the interaction; also see Table 2 for t-values comparing the two treatments on each of the three exams). Accordingly, any treatment effects on final exam performance are not a function of divergent quiz or unit exam performances across the treatment conditions (LL, HL).

Several effects did emerge from the ANCOVA. Exam percentages varied significantly as a function of exam number, F (2, 332)=4.52, $MSE=.015$, $p=.01$, $\eta_p^2=.027$; Table 2 shows that Exam 2 scores were somewhat higher than Exam 1 and 3 scores. The LCTSR was positively related to Exam scores, F (1, 166)=33.08, $MSE=.06$, $p<.0001$, $\eta_p^2=.17$, and more so for the earlier Exams ($r=.47$ and .39 for Exams 1 and 2 respectively) than for the last Exam ($r=.19$), F (2, 332)=5.19, $MSE=.015$, $p<.006$, $\eta_p^2=.030$ (for the interaction of LCTSR with exam number).

Achievement

Final exam scores were analyzed with a 2 (question level: low level, high level)×2 (treatment) mixed model ANCOVA, with final-exam question level as a within-subjects factor, treatment as a between-subjects factor, and LCTSR as the covariate. The treatment main effect was significant [F (1, 166)=7.15, $p=.008$, $\eta_p^2=.041$[1]] indicating that students who took high-level unit exams (HL treatment, adjusted mean=.54) generally scored higher those students who took low-level unit exams (LL treatment, adjusted mean=.50). Importantly, this treatment main effect did not interact with question level ($F<1$). Figure 2 (unadjusted means) shows that the higher scores for the HL condition were clearly evident for both the low-level final exam questions and the high-level final exam questions. To confirm this observation and because of the theoretical significance of the result, we conducted separate ANCOVAs (with LCTSR as a covariate) on scores on the low-level and high-level question . For the low-level questions, the HL condition (adjusted mean=.63) scored significantly higher than the LL condition (adjusted mean=.59) [F (1, 166)=4.19, $MSE=.02$, $p<.05$, $\eta_p^2=.025$[2]]; and for the high-level questions, the HL condition (adjusted mean=.46) similarly scored higher than the LL condition (adjusted mean=.41) [F (1, 166)=6.32, $MSE=.02$, $p<.02$, $\eta_p^2=.037$[3]].

The overall ANCOVA also indicated that correctly answered more low-level questions than high-level questions, F (1, 166)=34.74, $MSE=.009$, $p<.0001$, $\eta_p^2=.173$. Reasoning level (LCTSR) assessed at the beginning of the semester was positively associated with final exam scores, F (1, 166)=41.50, $MSE=.02$, $p<.0001$, $\eta_p^2=.200$. This association was significantly

[1] Analysis without the covariate (ANOVA) indicates the same effects, F (1, 167)=13.04, $MSE=.01$, $p<.0001$, $h_p^2=.072$

[2] Again, analysis without the covariate (ANOVA) indicates the same effects, F (1, 167)=7.73, $MSE=002$, $p<.01$, $h_p^2=.044$

[3] Again, analysis without the covariate (ANOVA) indicates the same effects, F (1, 167)=12.27, $MSE=.02$, $p<.01$, $h_p^2=.068$

Table 2 Average unit exam scores between treatment conditions

Unit exam	Treatment condition		t	p	η_p^2
	Low-level *M* (*MSE*)	High-level *M* (*MSE*)			
1	69.0 (1.6)	69.9 (2.0)	0.32	>0.75	<0.001
2	76.4 (2.1)	75.1 (1.8)	0.49	>0.62	<0.001
3	68.8 (2.4)	70.7 (2.2)	0.60	>0.55	<0.001

modulated by question level, F (1, 166)=5.65, MSE=.009, p<.02, η_p^2=.033, such that students' reasoning level was more strongly associated with their ability to solve high-level questions (r=.50) than low-level questions (r=.33). This pattern is sensible because performance on high-level items requires both content knowledge and reasoning skills, whereas for low-level items reasoning ability is presumably less necessary.

Attitudes

Response rate on end-of-semester student course evaluations was high (74 % in the LL treatment and 80 % in the HL treatment). Student impressions of the evaluations used in the course were generally lukewarm. In response to the statement, "Evaluations are a good measure of learning," the average response (on an 8-point Likert scale) was a 5.5 in the LL treatment and 5.3 in the HL treatment, between "Somewhat Agree" and "Agree." Open-ended comments were searched for feedback specifically about exams. In the LL treatment, only 8 students chose to comment on the exams, mostly with disappointment in their performance. In the HL treatment, comments were much more plentiful and it appears that students recognized that the exams were testing higher-level thinking skills but resented their poor performance (see Table 3).

Discussion

The assessment level incorporated into the course had a significant impact on students' conceptual understanding and final achievement scores. A striking pattern emerged such that

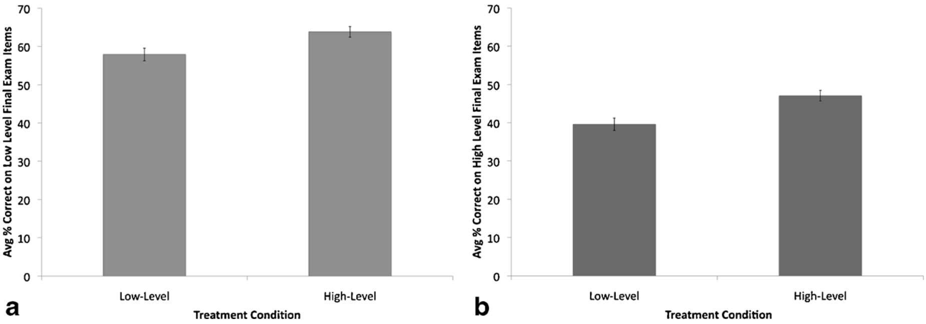


Fig. 2 **a** Low-level and **b** high-level achievement sub-scores for each treatment condition. For both sub-scores, the HL treatment significantly outperformed the LL treatment (p<.05 and p<.02, respectively). Error bars represent 95 % confidence intervals

Table 3 Student impressions of exams

Low-level treatment

I felt that some questions of the test were not covered or not covered enough

She quizzed you on vague things she might have talked about in class for 2 min but on the quiz it was complex.

I did think a lot of the questions on the tests tried to trick me instead of test my actual knowledge.

High-level treatment

I must admit those tests were weighted towards the hardcore application side, but they provided an opportunity to expand my critical thinking skills.

I didn't feel like the tests were a good measure of our learning at all. I felt a lot of it was basic science reasoning, kind of like the science on the ACT exam. The exams were long and time consuming, and I know me and other students thought they tested your test taking ability, rather than the material taught in class

The tests were all very application based so they forced everyone to really try hard with higher thinking skills.

Tests were extremely well written and actually provided opportunities to learn and acquire application skills–a rare feat in a test situation.

The biggest thing for me with this course is that her tests are extremely difficult. I don't understand the point in it when it really doesn't help the student learn or see how much they're learning when they do so poorly.

Comments are representative samples of open-ended response questions on end-of-semester student course evaluations

students who routinely took quizzes and unit exams requiring higher-order thinking not only showed deeper conceptual understanding by higher scores on high-level questions, but also showed greater retention of the facts, as evidenced by higher scores on low-level questions. These effect sizes were in the range of small-to-medium size effects (small effects are considered to be 0.01 and medium are considered to be 0.06). From a practical standpoint, differences in scores between the low-level and high-level conditions were equivalent to approximately a half-grade level (e.g., the difference between a B and B+). Returning to our original theoretical perspectives, these results best fit the expectation outlined in the introduction that high-level quizzing and testing would have cascading effects such that the focus toward high-level deep conceptual understanding (on quizzes/unit exams) would confer benefits on final exam items that targeted deep use of information (application, analyze, evaluate) and exam items that targeted memory of target information. As outlined in the introduction these effects can potentially be explained through both a test expectancy and a testing effect perspective. Below, we also sketch some other indirect consequences of testing that could potentially be involved in the present effects.

Briefly, a test-expectancy perspective would suggest that students adjusted the focus and strategies used in their in-class learning and/or at home studying to best match the demands of the quizzes and exams. Much as students with a surface or achieving study strategy (Biggs 1987) will tailor their study habits to match expectations, if students are presented with assessments throughout the course that require simple memory of facts (low-level exams), they likely focus their cognitive strategies on memorizing terms and definitions while neglecting to practice applying the material. The higher-level exams, on the other hand, would prompt a change in student expectations and as a consequence, a change in the focus of their studying from low- to high-level tasks. In particular, students might have begun to focus their studying on these exercises to integrate, evaluate and apply the material. By focusing on integration and application in their studying, we assume that students (in the high-level quiz/exam condition) would also need to learn the terminology and basic understanding tapped by the low-level final exam questions.

Of course the assessment level manipulation could have fostered a host of other indirect effects. For instance, high-level quizzes and exams could have prompted an increase in the quantity or spacing of study for the final exam than did low-level quizzes and exams (perhaps because the high-level quizzes/exams created the anticipation of a more challenging final exam). Another possibility is that the students in the high-level quiz/exam condition were more likely to seek feedback and review their exams with the teaching assistants than were students in the low-level condition. To inform this possibility, we calculated the percentage of students who consistently met with teaching assistants to review their exams. The results disfavored this possibility as students in the low-level condition met more often with teaching assistants to review exams (M=55 %) than did students in the high-level condition (M=39 %, p=.04). Finally, it could be that the high-level quizzes/exams allowed students to more effectively calibrate their metacomprehension of the material, which in turn might be expected to guide more efficient study policies for the final exam (i.e., more effective deployment of study time; cf. (Thomas and McDaniel 2007).

The present patterns might also be a consequence of testing (on the quizzes and exams) directly affecting learning and consequently performance on the final exam. The idea is that students practiced at performing high-level processing of the material on quizzes/exams better learned conceptual aspects of the material that required applying, analyzing, and evaluating than students practiced at remembering the material (on quizzes/exams). This would be revealed as an advantage on the high-level questions presented on the final exam. Additionally, testing on high-level questions could stimulate processing that enhanced memory for the target information (e.g., McDaniel et al. 2013), thereby possibly playing a role in producing *superior* performance on low-level (memory-based) final exam questions relative to low-level testing. In addition to (or instead of) high-level items stimulating more extensive processing when answering the quizzes and unit exams (than low-level items), it is possible that the high-level items (quiz and unit exam) exposed students to greater amounts of information than did the low-level items, due to the more complex nature of the high-level items (see the Appendix for sample high- and low-level unit exam items). If so, then this additional exposure per se could have contributed to the increased achievement (final exam performance) of students in the high-level condition.[4]

An alternative interpretation to the general idea that the benefits of the high-level condition were a consequence of direct and/or indirect cognitive effects is that students in the low-level condition found the quizzes and unit exams to mismatch the way the course was being taught (an inquiry, problem-based approach); as a consequence their motivation suffered relative to students in the high-level condition. That is, students in the low-level condition may have become less motivated to engage in class and to study than those in the high-level condition, thereby producing lower final exam performance. At least two findings argue against this interpretation. First, if motivation were declining in the low-level condition, we would expect to have seen quizzes and unit exam scores drop throughout the semester. However, quizzes and unit exam scores remained consistent between conditions. Second, the attitudinal data indicated that students in the low-level condition considered the quizzes and unit exams to be typical and at the level expected. Of the 62 students in the low-level condition who submitted end-of-semester student course evaluations, only eight made any reference to exams, and those references were generally regarding the difficulty, not the format. This suggests that factually oriented exams are typical of college science courses and expected by students (as was seen by Momsen et al. 2010). In the high-level treatment, many more references were made to exams and most comments referred to the types of questions on these exams.

[4] We appreciate an anonymous reviewer for raising this possibility.

The student comments lead us to two conclusions. First, students recognized that the exams were testing their ability to think and that this was somewhat novel and unexpected. Second, students were extremely dissatisfied with this type of testing, despite their increased learning. Testing that requires application of material is challenging for students. It requires a real effort to understand the material and not just a cramming session to memorize definitions. Students are rarely exposed to this type of assessment and therefore find it to be uncomfortable and difficult. Many students expressed to the instructor their frustration with their inability to effectively study for these exams. It was our impression that their frustration was not due to a lack of ability but rather to a lack of experience with exams requiring study strategies aimed at deep conceptual learning and critical thinking. As was pointed out in a recent review, students often do not recognize how they learn and thus do not appreciate many beneficial learning tasks (Bjork et al. 2013).

Interestingly, the LCTSR was correlated with performance on all three unit exams as well as the final, especially with high-level items. Teaching through inquiry did not eliminate this correlation. First, the robust association of the LCTSR and high-level item performance indicates that high-level assessment items indeed require scientific reasoning skills in addition to simple content knowledge. Second, it suggests that students with higher reasoning skills have a distinct advantage on high-level assessments. It follows, then, that perhaps more explicit focus on the teaching of scientific reasoning skills is warranted when high-level assessments are used.

Overall, our findings are in line with the hierarchical assumption of Bloom’s taxonomy of knowledge processes. The observed general benefits of the high-level quiz/unit exam condition (relative to the low-level condition) on the final exam suggest that preparing for high-level questions (application, analysis and evaluation) necessitated a mastery of information at lower levels on the taxonomy (memory and comprehension) as well. We also suggest that it is possible that in order to perform on items at higher levels of Bloom’s taxonomy, students must engage in elaborative practices when answering high-level questions that enhance their understanding of the basic terminology. Of course these possibilities await direct evaluation in more controlled experiments. Regardless of the eventual explanation, the findings show cascading effects such that consistent experience with high-level quiz and exam items (as identified by Bloom’s taxonomy) not only facilitates performance on subsequent questions requiring application, analysis, and synthesis, but also facilitates performance on items that require memory and understanding of basic terminology (relative to a steady diet of low-level quiz and exam items).

Conclusions and Educational Implications

The present results reinforce the assumption that assessments inform students of expectations for the course, and further indicate that such expectations can have important consequences for student learning outcomes. Writing exams that require higher-order thinking skills is certainly a challenging task for an instructor, especially utilizing the multiple-choice format that can be easily administered in a large-enrollment course. However, higher-order assessments may be a key factor in stimulating students to effectively acquire a deep understanding of the material, an understanding that supports, not only application, analysis and evaluation, but also better retention of the core facts. By contrast, adopting a more typical (and perhaps easier for the instructor) approach of giving factual recall exams does students a disservice, as these kinds of exams are less likely to foster thinking critically and applying knowledge and further do not appear to even promote acquisition and retention of factual information to the extent stimulated by higher-order exams (given throughout the course).

This study illustrates the importance of higher-level assessment in promoting scientific understanding. Using a backward design (Wiggins et al. 1998), instructors are encouraged to first identify desired results or learning outcomes. In biology, these learning outcomes should certainly include content knowledge, but we suggest they should also include the learning of scientific process skills. The second step in backward design is to determine what evidence would illustrate that learning outcomes were met and to design assessments accordingly. The third step is to design learning activities that align with the assessments. By aligning our learning activities with our assessments and our assessments with our learning outcomes, we provide greater opportunities for students to demonstrate what they have learned. Many science instructors fail to design assessments that effectively test for scientific process skills, often defaulting to test-bank generated exams testing only content knowledge. Assessments should be designed to truly test scientific process skills and as such should be written at higher levels of Bloom's taxonomy. Not only will this assessment format give the appropriate evidence of the attainment of desired learning outcomes, but this study shows that it actually directs student learning, focusing their study efforts on these desired skills and ultimately leading to deep conceptual understanding. Effective assessment can turn students' time inside and outside of class into productive learning time, rather than bouts of rote memorization.

Conflict of Interest The authors declare that they have no conflict of interest.

Appendix

Sample Test Items

Low-Level Unit Exam Items

Term/Concept: Haploid/Meiosis

i. In humans, the haploid number, *n* equals:

   a) 2*n*.
   b) 44.
   c) *23.*
   d) 46.

Term/Concept: Nondisjunction/Meiosis

ii. Trisomy is a consequence of ________________________, the unequal distribution of chromosomes during meiosis.

   a) recombination
   b) cytokinesis
   c) *nondisjunction*
   d) crossing over

Term/Concept: Types of Selection/Evolution

iii. Divergent selection occurs when:

a) low-ranking, “sneaker” males disrupt the mating between a dominant female and a male.
b) individuals with one extreme value of a trait have greater fitness than individuals with the other extreme value of the trait.
c) *individuals with extreme values of a trait have greater fitness than individuals with intermediate values of the trait.*
d) there is selection against the extreme ends of a trait’s distribution.

Term/Concept: Hardy–Weinberg Equilibrium/Microevolution

The pygmies are a group of humans who are thought to have lived continuously in the rainforest of central Africa since the ice age, in geographically separated areas called refuges. There are two distinct groups of pygmies, the eastern and the western, living in two different refuges. The eastern population has an HbS allele frequency of 8 %. The western population has an HbS allele frequency of 18 %.

iv. Are the eastern and western pygmies in Hardy–Weinberg equilibrium?

a) Yes
b) *No*
c) Impossible to tell

Term/Concept: Predicted Allelic Frequencies/Microevolution

v. Given the HbS allele frequency in the western population is 0.18 (q), what is the frequency of the normal allele (HbA) in the western population?

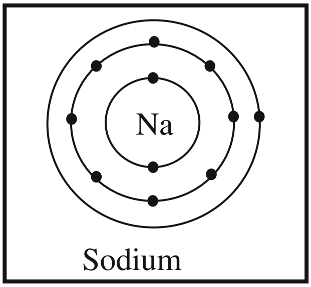
a) 0.2
b) 0.18
c) *0.82*
d) 0.08

Term/Concept: Interpreting the Periodic Table/Atomic theory

17
Cl
35.5

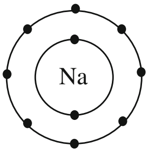
vi. Looking at the figure to the right, how many protons does chlorine have?

a) *17*
b) 18
c) 35.5
d) 71

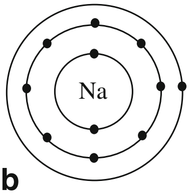


Term/Concept: Electron Shells/Atomic Bonding

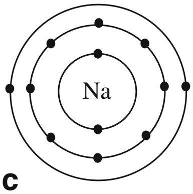
vii. If the normal electron-shell diagram of Sodium (Na) is pictured to the right, what does Na^+ look like?



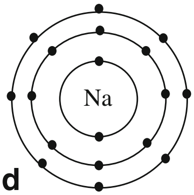
a



b



c



d

Low-level Final Exam Items

Term/Concept: Types of Bonds/Atomic bonding

viii. ____________ bonds deal with the sharing of electrons, while ___________bonds deal with the transferring of electrons completely to another atom.

a) Ionic, Hydrogen
b) Ionic, Covalent
c) Covalent, Hydrogen
d) *Covalent, Ionic*
e) Hydrogen, Covalent
f) Hydrogen, Ionic

Term/Concept: G1/Cell Cycle

ix. Within a Eukaryotic cell cycle, the primary stage during which normal cellular functioning occurs is

a) *G1*
b) S
c) G2
d) Mitosis
e) Meiosis

Term/Concept: Hardy–Weinberg Equilibrium/Microevolution

x. Which of the following violations to Hardy–Weinberg assumptions would be most likely to change the allele frequency of a small population of pigmy monkeys living on an island?

a) Mutation
b) Changing environment
c) *Genetic drift*
d) Gene flow
e) Selection
f) Non-random mating
g) All of the above have an equal chance of moving the population out of Hardy–Weinberg equilibrium.

High-Level Unit Exam Items

Term/Concept: Haploid/Meiosis

You and your spouse are expecting your first child. Because you and your spouse have an extensive history of genetic diseases on both sides of the family, you are concerned about the health of your unborn baby and request that an amniocentesis be performed so that you can prepare yourself for whatever lies ahead. During this procedure, a large needle is inserted through the mother's belly and into the amniotic sack in order to collect a sample of amniotic fluid containing sloughed cells from the embryo. These cells are then grown in a laboratory for about a week in order to analyze the chromosomes.

xi. The DNA from a typical human sperm cell weighs approximately 3.3 picograms (a picogram is a 10^{-12} g). If all chromosomes weighed approximately the same, how much does a typical chromosome weigh?

a) 0.07 picograms
b) *0.14 picograms*
c) 0.28 picograms
d) 75.9 picograms
e) 151.8 picograms

Term/Concept: Nondisjunction/Meiosis

xii. Doctors find that in the majority of embryo's cells, the DNA weighs nearly 6.8 picograms. What is a possible explanation for this?

a) The embryonic cells remained haploid, meaning that the sperm's chromosomes never fused with the egg's and were most likely kicked out with the polar body.
b) The embryonic cells contain all duplicated chromosomes; thus, there must be a problem with Meiosis II and sister chromatids are never separating.
c) *The embryonic cells contain extra DNA indicating a possible trisomy.*
d) The embryo is normal but is definitely a girl, since girls carry an extra X chromosome.

Term/Concept: Types of Selection/Evolution

Zawicki and Witas (2007) studied the C-C chemokine receptor 5 (CCR5), a receptor on the white blood cell that is used by many pathogens to gain access to the cell, including HIV, Bubonic plague, and others. It has been found that a mutation (the deletion of 32 base pairs; designated CCR5-D32) prevents HIV from entering the cell. Individuals who are homozygous for the CCR5-D32 mutation are nearly completely resistant to HIV infection and individuals who are heterozygous have slower disease progression and better survival rates.

xiii. There are no known disadvantages to being homozygous for the CCR5-D32 mutation. Predict what type of selection will most likely occur, given enough time?

a) Stabilizing
b) *Directional*
c) Diversifying

Term/Concept: Hardy–Weinberg Equilibrium/Microevolution

xiv. Scientists have found a high incidence of this allele in European and Mediterranean people. The frequency of the CCR5-D32 allele on the European continent is 10 %. What percentage of the people are likely *carriers* of the CCR5-D32 allele?

a) 90 %
b) 10 %
c) *18 %*
d) 9 %

Term/Concept: Predicted Allelic Frequencies/Microevolution

xv. Given that the CCR5-Δ32 allele confers immunity to the bubonic plague, which of the following phenomena would you predict?

a) *The bubonic plague of the 14th century would closely correspond in geographic distribution as well as timing with the suspected origin and increasing frequency of the CCR5- Δ32 mutation.*

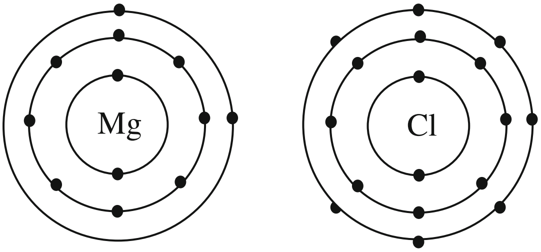
b) The bubonic plague of the 14th century would directly precede (come before) the emergence of the CCR5- Δ32 mutation in the same geographic distribution indicating that the plague caused the mutation of the CCR5 receptor.

c) The bubonic plague of the 14th century would be only in geographic areas where the CCR5- Δ32 mutation was *not* prevalent since those are the people most susceptible to plague infection.

d) The bubonic plague of the 14th century would have emerged shortly after and in the same geographic location as the emergence and increasing frequency of the CCR5- Δ32 mutation indicating that the CCR5- Δ32 mutation increases plague infection rate.

Term/Concept: Electron Shells/Atomic Bonding

A salt is defined as "a compound resulting from the formation of an ionic bond; also called an ionic compound" (Campbell & Reece, 2005). That means that a salt can be formed from any two ions. Below are the electron-shell diagrams of magnesium (Mg) and chlorine (Cl).



xvi. If table salt (the kind you eat) is sodium chloride (NaCl), what is the molecular formula of magnesium chloride?

a) MgCl
b) Mg_2Cl
c) *$MgCl_2$*
d) Mg_2Cl_2

High-Level Final Exam Items

Term/Concept: Predicting Selection/Microevolution

xvii. The incidence of spinal muscular atrophy (an autosomal recessive disease) in the United States is about 1 case in every 17,000. Whereas, in North Dakota, the prevalence is 1 in 6720. Which of the following would support the hypothesis that genetic drift was responsible for the increased allele frequency in North Dakota?

a) There is an abnormally high concentration of mutagenic chemicals in the ground water causing an increased mutation rate in North Dakota.

b) One of the best treatment centers for SMA is located in North Dakota causing migration of SMA carriers into the area at an abnormally high frequency.

c) *The original settlers of North Dakota were a small group of pioneers who happened to have an abnormally high frequency of the SMA allele in their population.*

d) A particular mosquito-born parasite native to North Dakota causes high infant mortality; carriers of the SMA allele are less likely to catch the disease.

Term/Concept: Electron Shells/Atomic Bonding

The hydrogen car runs off the power of hydrogen created from water in the process of electrolysis (i.e., the splitting of water). In this process, water is split into hydrogen and oxygen gases by running an electrical current through the water. This is done by placing two metal rods in the water and attaching them to an electrical source. The cathode donates electrons to the solution; whereas the anode accepts electrons from the solution. The reactions at each node are shown below (keep in mind that these reactions are happening at the same time):

At the Cathode, where hydrogen gas bubbles are produced: $2H^+ + 2e^- \rightarrow H_2$ (gas)

At the Anode, where oxygen gas bubbles are produced: $2H_2O \rightarrow O_2$ (gas)$+4H^+ + 4e^-$

xviii. How many electrons must come from an outside source (like the anode or cathode) to make this reaction run?

a) Two electrons are being donated by the cathode

b) Two electrons are being donated by the anode

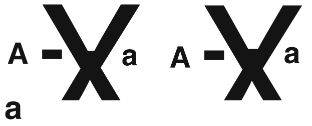
c) Four electrons are being donated by the anode

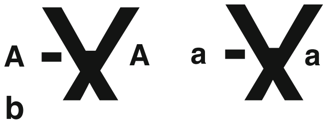
d) Eight electrons are being donated by the anode

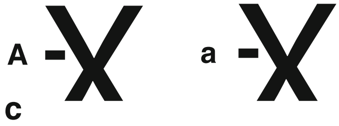
e) *No electrons are being donated, they are all contained within the original water molecule*

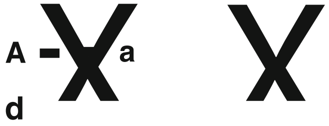
Term/Concept: Homologous Chromosomes/Meiosis

xix. Cystic fibrosis, an autosomal recessive disease, is located on chromosome #7. If you viewed the 7th chromosome pair of an individual who is a carrier for cystic fibrosis during prophase of mitosis, which of the following is the best representation of what you would see?









Term/Concept: Interpreting the Periodic Table/Atomic Theory

xx. Let's say I have a hypothetical element, called "Element J." It tends to make three bonds and the most common isotope has 3 more neutrons than protons. Which of the following is the most accurate representation of Element J?

a

15
J
17.7

b

7
J
16.2

c

5
J
13.0

d

5
J
8.1

References

AAAS. (2010). *Vision and change: a call to action*. Washington: AAAS.

Anderson, G. L., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., et al. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Boston: Allyn & Bacon.

Becker, H. S., Geer, B., & Hughes, E. C. (1968). *Making the grade: The academic side of college life*. New Brunswick: Transaction.

Biggs, J. B. (1987). *Student approaches to studying and learning*. Hawthorn: Australian Council for Educational Research.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444.

Bloom, B. S. (1984). *Taxonomy of educational objectives*. Boston: Allyn and Bacon.

Bybee, R. (1993). *An instructional model for science education: Developing biological literacy*. Colorado Springs: Biological Sciences Curriculum Studies.

Campbell, N. A., & Reece, J. B. (2005). *Biology*, 7th ed. San Francisco: Benjamin Cummings.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279–283. doi:10.1177/0963721412452728.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*(2), 268–276. doi:10.3758/Bf03193405.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*(3), 474–478. doi:10.3758/Bf03194092.

Carpenter, S. K., Pashier, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*(2), 438–448. doi:10.3758/Mc.36.2.438.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology, 23*(6), 760–771. doi:10.1002/Acp.1507.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633–642. doi:10.3758/Bf03202713.

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

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.

Dickie, L. O. (2003). Approach to learning, the cognitive demands of assessment, and achievement in physics. [Reports - Research]. *Canadian Journal of Higher Education, 33*(1), 87–111.

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning Memory and Cognition, 38*(3), 632–652. doi:10.1037/A0026215.

Fisher, R. P., & Craik, F. I. M. (1977). Interaction between encoding and retrieval operations in cued recall. Journal of Experimental Psychology: *Human Learning and Memory. 3*(6), 710–711.

Harlen, W., & Deakin Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review, version 1.1*). Research Evidence in Education Library (Vol. 1). London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Hill, P. W., & Mcgaw, B. (1981). Testing the simplex assumption underlying Blooms Taxonomy. *American Educational Research Journal, 18*(1), 93–101. doi:10.3102/00028312018001093.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*(3), 621–629. doi:10.1037/A0015183.

Joughin, G. (2010). The hidden curriculum revisited: A critical review of research into the influence of summative assessment on learning. *Assessment & Evaluation in Higher Education, 35*(3), 335–345.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modulate the effect of testing on memory retention. *European Journal of Cognitive Psychology, 19*, 528–558.

Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218.

Kropp, R. P., et al., & Florida State Univ. Tallahassee. Inst. of Human Learning. (1966). The construction and validation of tests of the cognitive processes as described in the "Taxonomy of Educational Objectives". (pp. 444).

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching. 15*(1), 11–24.

Lawson, A. E. (2002). *Science teaching and development of thinking*. Belmont: Wadsworth/Thompson Learning.

Lawson, A. E., Alkhoury, S., Benford, R. B. C., & 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*(9), 996–1018.

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

Lawson, A. E., Clark, B., Cramer-Meldrum, E., Falconer, K. A., Sequist, J. M., & Kwon, Y.-J. (2000c). Development of scientific reasoning in college biology: Do two levels of general hypothesis-testing skills exist? [Reports - Research]. *Journal of Research in Science Teaching, 37*(1), 81–101.

Madaus, G. F. (1973). A causal model analysis of Bloom's Taxonomy. *American Educational Research Journal, 10*(4), 253–262.

Mayer, R. E. (2003). *Learning and instruction*. Upper Saddle River: Prentice Hall. 10(4), 253–262.

McDaniel, M. A., & Donnelly, C. M. (1996). Learning with analogy and elaborative interrogation. *Journal of Educational Psychology, 88*(3), 508–519. doi:10.1037//0022-0663.88.3.508.

McDaniel, M. A., Blischak, D. M., & Challis, B. (1994). The effects of test expectancy on processing and memory of prose. *Contemporary Educational Psychology, 19*, 230–248.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4-5), 494–513. doi:10.1080/09541440701326154.

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.

McDaniel, M. A., Friedman, A., & Bourne, L. E., Jr. (1978). Remembering the levels of information in words. *Memory & Cognition, 6*, 156-164.

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

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., III, & McDaniel, M. A. (in press). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. Journal of Experimental Psychology: Applied.

Miller, C. M. L., & Parlett, M. (1974). *Up to the mark: A study of the examination game*. London: Society for Research into Higher Education.

Momsen, J. L., Long, T. M., Wyse, S. A., & Ebert-May, D. (2010). Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. [Reports - Evaluative]. *CBE - Life Sciences Education, 9*(4), 435–440.

Momsen, J. L., Offerdahl, E., Kryjevskaia, M., Montplaisir, L., Anderson, E., & Grosz, N. (2013). Using assessments to investigate and compare the nature of learning in undergraduate science courses. *CBE - Life Sciences Education, 12*, 239–249.

National Research Council. (1996). *National science education standards*. Washington: National Academy.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

Roediger, H. L., III, 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.

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

Seddon, G. M. (1978). The Properties of Bloom's Taxonomy of Educational Objectives for the Cognitive Domain. Review of Educational Research, r.

Snyder, G. (1951). *The hidden curriculum*. New York: Knopf.

Sternberg, R. J., Grigorenko, E. L., & Zhang, L. (2008). Styles of learning and thinking matter in instruction and assessment. *Perspectives on Psychological Science, 3*, 486–506.

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education, 30*(4), 325–341.

Thiede, K. W., Wiley, J., & Griffing, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81*, 264–273.

Thomas, A. K., & McDaniel, M. A. (2007). The negative cascade of incongruent generative study-test processing in memory and metacomprehension. *Memory & Cognition, 35*(4), 668–678. doi:10.3758/Bf03193305.

Van Etten, S., Pressley, M., McInerney, D. M., & Liem, A. D. (2008). College senior's theory of their academic motivation. *Journal of Educational Psychology, 100*(4), 812–828.

Wiggins, G., McTighe, J., & Association for Supervision and Curriculum Development Alexandria VA. (1998). Understanding by Design. (pp. 214): Association for Supervision and Curriculum Development, 1703 North Beauregard Street, Alexandria, VA 22311–1714 (stock number 198199, members:.

Zawicki, P., & Witas, H. W. (2007). HIV-1 protecting *CCR5-D32* allele in medieval Poland. *Infection*, *Genetics and Evolution, 8*(2): 146-151.

Zoller, U. (1993). Are lecture and learning compatible? Maybe for LOCS: unlikely for HOCS (SYM). Journal of Chemical Education 70: 195–197.