TESTING EFFECT AND COMPLEX COMPREHENSION IN A LARGE INTRODUCTORY UNDERGRADUATE BIOLOGY COURSE

by

Christopher L Pagliarulo

_____

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF MOLECULAR AND CELLULAR BIOLOGY

In Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

from the Graduate College of

THE UNIVERSITY OF ARIZONA

2011

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Christopher Lawrence Pagliarulo entitled *Testing Effect and Complex Comprehension in a Large Introductory Undergraduate Biology Course* and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

_____ Date: July 22nd, 2011
Debra J. Tomanek, PhD

_____ Date: July 22nd, 2011
Lisa K. Elfring, PhD

_____ Date: July 22nd, 2011
Frans E. Tax, PhD

_____ Date: July 22nd, 2011
Andrew P. Capaldi, PhD

_____ Date: July 22nd, 2011
Angel C. Pimentel, PhD

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: July 22nd, 2011
Dissertation Director: Debra J. Tomanek, PhD

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the author.

SIGNED: Christopher Lawrence Pagliarulo

ACKNOWLEDGEMENTS

I'd like to thank my advisor, Dr. Debra Tomanek, for the freedom to make plenty of mistakes, for the wisdom and patience to keep picking me back up, and for the hard questions that make this job worth it. There is no better way to learn or to mentor.

I am grateful to Dr. Lisa Elfring for her tireless efforts tackling the data with me and for her tireless cheer and friendship, often when needed most.

Dr. Angel Pimentel has been a trusted advisor, friend, and colleague, whose honest and abundant curiosity about learning and teaching kept the wind in my sails and my eyes on the target.

Dr. Andrew Capaldi and Dr. Frans Tax's thoughtful guidance, mentorship, and encouragement over the last two years were critical to the completion of this manuscript.

I am also grateful for Dr. Hope Jones. For everything, partner. For everything.

For Sidney, always my kindred spirit, my artist, my Sous-chef.

For my parents, who demanded I follow my own course and saw it through with me, never wavering.

And for my new family and my friends, who kept me human and stood by my side.

Finally, I want to thank NASA and NSF for their substantial support and investment in my education. You won't be sorry.

TABLE OF CONTENTS

TABLE OF CONTENTS – *Continued*

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Traditional undergraduate biology courses are content intensive, requiring students to understand and remember large amounts of information in short periods of time. Yet most students maintain little of the material encountered during their education. Poor knowledge retention is a main cause of academic failure and high undergraduate attrition rates. Characterizing strategies that support robust learning is critical for ensuring student success. One such strategy *is testing effect*, the observation that repeated testing can improve the fidelity and durability of retained knowledge more than an equal quantity of restudy. Numerous investigations have described the nature and boundaries of testing effect. Very few, however, have characterized its efficacy in actual classroom practice. The current study investigated whether repeated testing or repeated study affected student retention and understanding of complex biological concepts. The study was conducted in a large (~320 students) introductory biology class. All study conditions and assessments were required components of the course. Student retention of two fundamental molecular biology "big ideas" was targeted; (1) the relationship between genotype and phenotype, and (2) the relationship between gene expression and cell function. Students were randomly assigned to one of three repeated quiz or study conditions. For four weeks, students encountered various combinations of multiple-choice (MC) questions and review material related to big ideas 1 & 2 and/or unrelated lecture topics. Five weeks after the last quiz, all students completed identical MC final exam questions related to both big ideas. To determine the quality of "understanding" assessed by the MC questions, a subset of students also

completed a short answer (SA) test prior to the final exam. Both question formats assessed the same knowledge (2 big ideas) at the same level (comprehension and application). Final exam performance supported the finding that repeated retrieval improves long-term retention of knowledge relative to repeated study. Novel to other previous work conducted at the undergraduate level, the current findings suggest that repeated testing affects student retention and understanding of sophisticated concepts. Careful design and analysis of parallel multiple-choice and short answer questions demonstrated that each can target and elicit similar qualities and types of knowledge.

INTRODUCTION

Undergraduate science curricula require students to retain and make sense of tremendous amounts of information in relatively short periods of time (Wood 2009; Labov, Reid et al. 2010). As students advance, successful learning becomes increasingly dependent upon previous knowledge and understanding. Learners must retain and build upon knowledge and ideas acquired in earlier courses to succeed in more advanced topics, an assumption reflected in the structure of most undergraduate curricula (Bransford, Brown et al. 2000; Brown 2004; van den Broek and Kendeou 2008; Wood 2009; Tibell and Rundgren 2010). Yet studies that have examined undergraduate knowledge retention suggest that students maintain only a small fraction of the material they encounter over the course of their education (Stigler 1963; Kohen and Kipps 1979; Walstad 2001). Significant gaps in learning and/or memory appear soon after instruction and tend to progressively worsen over time (Roediger 2008). For example, undergraduate students asked to recall the lecture content immediately after class were able to remember only 30-40% of the material presented (Kohen and Kipps 1979; Roediger 2008). One week later, retention of the same material dropped to an average 20%. Walsted (2001) tested student retention of basic economics knowledge after completing a two-semester introductory economics course. Course graduates scored only 20% better than control students receiving no instruction, and only 10% better than alumni of the same program (Walstad 2001). Poor knowledge retention affects the academic success of most undergraduate students at some time during their education

and has been identified as a key factor contributing to high student attrition rates (Alexander and Mayer 2010; Mayer 2010; Robertson, Canary et al. 2010; Willcoxson, Cotter et al. 2011). Investigating learning strategies that support retention is a critical step toward maximizing the effectiveness and success of undergraduate education programs.

Testing Effect

One such strategy is "testing effect." First discussed in the literature more than a century ago, testing effect describes the observation that repeated testing improves retention of learned materials significantly more than an equal quantity of restudy (Abbott 1909; Brown 1923; Spitzer 1939; Hogan and Kintsch 1971; Bartlett and Tulving 1974; Roediger and Karpicke 2006; Karpicke and Roediger III 2008; Pyc and Rawson 2010) In other words, students repeatedly quizzed on material are more likely to retain that information with greater fidelity and for longer periods than students who repeatedly re-read or re-study the same information. Numerous carefully controlled studies have demonstrated that the enhanced memory is not simply the product of greater exposure and feedback that accompanies repeated testing, but a distinct artifact of the recalling process itself (Roediger and Karpicke 2006; Kang, McDermott et al. 2007; Rohrer and Pashler 2007; Karpicke and Roediger III 2008).

Testing effect has been observed using a variety of materials, including: lists of words, names, and word-pairs (Spitzer 1939; Hogan and Kintsch 1971; Bartlett and Tulving 1974; Fritz, Morris et al. 2007); pictures and maps (Carpenter and Pashler 2007); facts and definitions (Roediger and Karpicke 2006; Roediger III and Butler 2010; Karpicke and Blunt 2011); and textbook passages and scientific explanations (Butler and Roediger III 2007; Kang, McDermott et al. 2007; McDaniel, Roediger et al. 2007; Larsen, Butler et al. 2008; Karpicke and Roediger 2010; Karpicke and Blunt 2011). The effect is also robust across ages, seen consistently in preschool children, high school and college students, as well as middle-aged adults (Fritz, Morris et al. 2007; Logan and Balota 2008; Karpicke and Roediger 2010; Kornell, Castel et al. 2010).

Not all test formats affect memory equally, however. Questions that encourage more effortful recall, such as short answer or essay formats, produce greater long-term retention than tests emphasizing recognition - such as certain multiple-choice or matching question formats (Larsen, Butler et al. 2009; Karpicke and Zaromb 2010; Pyc 2010; Pyc and Rawson 2010; Roediger III and Butler 2010). For example, Kang *et. al.* (2007) found that students' ability to recall facts and ideas from short articles was significantly enhanced after an intervening short answer test versus either a multiple-choice test or restudy period. In a recent series of elegant studies by Sensenig (2011), students completed various types of intervening multiple-choice exams after studying factual passages. One exam condition required students to first recall and write down

answers to questions prior to seeing answer choices. Another exam condition presented question and answer choices simultaneously. The studies controlled for exposure time and included restudy treatments. Again, students that engaged in the more effortful test (recall, then answer choices) retained significantly more than either the recognition only or restudy conditions (Sensenig 2011). Interestingly, only an "effortful" intervening test appears important for retention. The format of the final test - whether short answer, multiple-choice, parallel, or dissimilar to the intervening tests - has no significant bearing on student performance (Coles 2008; Butler 2009; Karpicke and Zaromb 2010; Pyc and Rawson 2010).

But pursuit of powerful testing effects using multiple-choice tests should not be abandoned just yet. Two important considerations justify continued investigation. One, the use of multiple choice formats for assessment of undergraduate learning is ubiquitous; a product of tradition, ease of implementation, and necessity born from the disproportionate growth of student-to-faculty ratios that expand class sizes and limit resources (Walstad 2001; Mislevy, Steinberg et al. 2003; Buckles and Siegfried 2006; Wood 2009). Multiple-choice tests are well embedded in current instruction culture and their near term popularity is more likely to expand than decline. Two, despite widely held beliefs that multiple-choice assessments measure only basic recognition or comprehension of knowledge, no such innate limitation has been demonstrated (Mislevy, Steinberg et al. 2003; Buckles and Siegfried 2006; Draper 2009; Tsui and

Treagust 2010). In fact, numerous forms of multiple-choice questions (both new and old) have successfully targeted "higher-order" understanding and thinking, such as complex knowledge integration or the ability to apply learning to new situations (Klymkowsky, Gheen et al. 2007; Palmer and Devitt 2007; Zheng, Lawhorn et al. 2008; Foster and Miller 2009; Smith and Tanner 2010; Tsui and Treagust 2010). Alternative examples include two-tiered, case-based, and assertion-reason questions (Mislevy, Steinberg et al. 2003; Wilkinson and Frampton 2004; Buckles and Siegfried 2006; Sampson 2006; Stupans 2006; Williams 2006). Studies comparing long-term memory effects of short answer versus multiple-choice tests have emphasized traditional multiple-choice formats *designed* to elicit limited, discrete knowledge (Roediger 2008; Pyc and Rawson 2009; Karpicke and Zaromb 2010; Roediger III and Butler 2010). Thus, any conclusions about the utility of multiple-choice questions for enhancing retention must be limited to the species of questions examined so far. There is still much to learn about the relationship between question structure, recall, and retention.

Effect in the Classroom

Findings from the last two-decades of testing effect research suggest it holds promise for improving retention in undergraduate education. However, most of this work was carried out in idealized laboratory conditions involving material, learning objectives, and timeframes that have little relevance to the demands of traditional undergraduate courses (Roediger and Karpicke 2006; McDaniel, Anderson et al. 2007; Roediger III and

Butler 2010; Rohrer and Pashler 2010; Rohrer, Taylor et al. 2010; Karpicke and Blunt 2011). However, three recent studies have attempted to extend our understanding of testing effect in the classroom through study designs that incorporated authentic course materials, environments, and/or participation of actual classroom populations in ongoing instruction.

Butler and Roediger (2007) investigated the impact of test format on retention in a simulated classroom environment. Undergraduate participants were shown art history video lectures, immediately followed by either a multiple-choice, short answer, or reading review (restudy) condition. Corrective feedback was supplied for half of the review questions. When retention of the videos was measured one month later, content covered on the short answer tests was recalled significantly more often than material reviewed by multiple-choice test or restudy (Butler and Roediger III 2007). Interestingly, corrective feedback had no effect on retention. McDaniel et al. (2007) integrated a repeated testing strategy into an online undergraduate Brain and Behavior course. Course assessments consisted of fill-in-the-blank questions that emphasized recognition of factual statements describing neuropsychological processes. Students that studied the required course readings using fill-in-the-blank review questions scored significantly better on the mid-term and final post-tests than students who only read review statements. Carpenter, Pashler, and Cepeda (2009) looked at middle school student's retention of US history facts. One week or 16 weeks after initial instruction,

students reviewed the lesson material through testing and feedback, through restudy, or not at all. All students completed a post-assessment nine months later. Student performance scores suggested that the 16 week testing and feedback condition encouraged the greatest retention of facts, almost twice as much as students that received no review at all (Carpenter, Pashler et al. 2009).

In all three studies, students' ability to remember and recall information on a later test was significantly enhanced by testing. And because the contexts better approximated traditional classroom conditions than earlier work, the utility of testing effect strategies in the classroom appears compelling. But two significant caveats limit any conclusions of efficacy that can be drawn.

(1) As in earlier works, these investigations included intervening test conditions to promote recall (see Roediger & Karpicke, 2006 for review). The particular design and implementation of the pre/post-test conditions, however, promoted something more. In each study, student performance was greatest on final assessment questions that paralleled earlier exposure. Meaning, for example, performance on a short answer post-test question was greatest when the parallel intervening test question was also short answer. The earlier test informed the learner *how* information would need to be recalled later and supplied opportunity to practice that exact task. The alternate test, restudy, or

no test conditions omitted these cues. Thus, the experimental conditions that enhanced student retention differed from the alternative treatments in *two* ways; A) the type of re-exposure to target knowledge, and B) the priming of learning and recall to the final assessment format. Because none of the investigations differentiated effect sizes for each of these variables, it is unclear which contributed most to the observed outcomes.

(2) None of the studies' summative post-tests were graded components of a class or course. Butler and Roediger's (2007) assessments were simulations of classroom exams administered to paid student participants. The summative exams in the McDaniel et al (2007) study were optional, offered to students as preparation for graded exams. And the retention of students in the Carpenter et al (2009) study was assessed after the final exam, with no forewarning that it would be. Thus, none of the students investigated were subject to the same pressures (in the form of grades) to learn and retain information as students in authentic classes. Performance had no consequence, which study participants were either aware of (Butler & Roedinger and McDaniel et al) or did not know was a factor (Carpenter et al). Because study conditions did little to elicit motivation and because most students have negative biases about taking tests (Karpicke and Blunt 2011), it is unlikely that participants engaged and benefited

from the repeated test conditions as would students in authentic, high-stakes classrooms.

The issues complicating interpretation of these limited findings leave significant doubt whether testing effect actually occurs in relevant classroom environments, and whether repeated testing strategies offer potential for improving undergraduate instruction and student retention of knowledge.

Goals of Investigation

The current exploratory investigation seeks to address the limitations of previous work in order to better characterize the efficacy of testing effect in actual classroom practice. Through the use of multiple assessment strategies and new methods of coding, this work also explores the relationship between testing effect, test format, and student comprehension of complex ideas. Specifically, the current study investigated whether repeated testing or repeated study is a more effective strategy for retaining students' understanding of biological concepts in a large, undergraduate biology lecture course. Based on previous findings of recall and recognition tasks measured for testing effect, it was predicted that repeated quizzing would significantly improve retention of over repeated study of the same material (measured as loss of performance or percentage retention between pretest and post-test on parallel questions).

In order to help characterize the quality of student comprehension measured by the multiple-choice questions, the study also investigated differences in students' performance of their retained understanding as measured with multiple choice vs. short answer assessments.

METHODS

Participants & Course

Description of participants - Three hundred and thirty undergraduates enrolled in Molecular and Cellular Biology (MCB) 181 participated in the experiment. The class population was 63% female and 47% male. Sophomores made up the majority of students (41%), followed by freshman (38%), juniors (15%), and seniors (6%). Most students listed Pre-Medical / Pre-Veterinary sciences (41%) or Physical & Life Science, Mathematics, or Engineering (39%) as their major. Social Sciences (11%) and non-science or undeclared majors (9%) comprised the remainder. Data from 49 students was excluded due to failure to complete the course and/or all relevant study assessments.

Description of class - MCB 181 is the first required course of most life science related curriculums within the College of Science and Letters at the University of Arizona. The one semester lecture course followed a traditional majors introductory biology syllabus that is aligned with the national consensus of topics (Gregory, Ellis et al. 2011). The class met every Tuesday and Thursday for one hour and fifteen minutes. The average lecture consisted of an hour of PowerPoint style presentations interspersed with group activities and 3-6 “clicker” questions (Mayer, Stull et al. 2009). Students completed weekly online multiple-choice quizzes, three mid-term tests and a comprehensive final exam as part of the course requirements. Quizzes were implemented using the online

course management system - Desire2Learn (Rubin, Fernandes et al. 2010). Unless otherwise stated, all course assessments were multiple-choice format. All experimental conditions and assessments were integrated into the curriculum and the required course evaluations.

Ethical Considerations - According Human Subjects Protection Program (HSPP) regulations, the project was not considered human subjects research because student performance data contained no individual identifier information and was reported in aggregated form. Because no risk of individual privacy breach existed, Institutional Review Board evaluation was not required. The appropriate "Not Human Subjects" HSPP worksheet was completed, authorized, and filed with the Molecular and Cellular Biology Department, University of Arizona.

Procedure

The experiment was a mixed methods study using a 3 (*Type of exposure: Repeated testing, repeated study, and no exposure*) by 2 (*Topic: Big Idea 1, Big Idea 2*) within-subjects design. Figure 1 (below) illustrates the experimental procedure; including how and when study materials were implemented over the course of the semester. Full descriptions of the instruments and subject matter implemented can be found in the *Materials* section below.

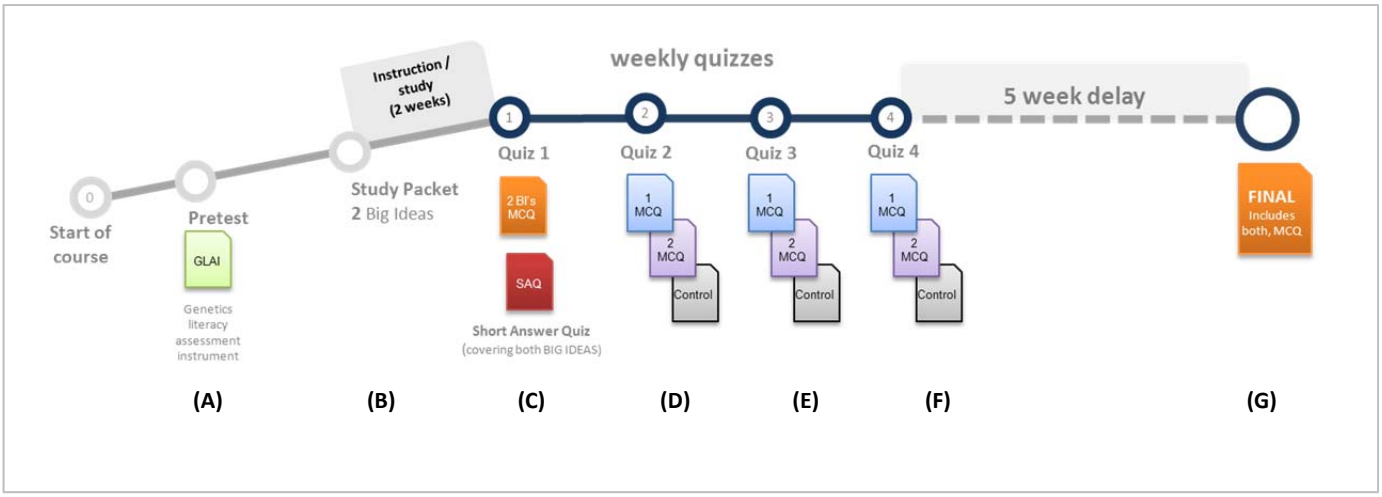


Figure 1: Summary of experiment procedure – **(A)** The first week of class, all students completed a pretest based on the Genetics Literacy Assessment Instrument (Bowling, Acra et al. 2008). Pretest performance data was used to ensure treatment group equivalency. **(B)** In week two, students received a study packet covering both big ideas and instructions about future assessments. **(C)** After a two week study period, all students were assigned an online multiple-choice quiz (MCQ – Quiz 1) that assessed students' comprehension of the big ideas. That same week, 36 students also completed an extra-credit Short Answer Quiz (SAQ) that assessed the same ideas through short answer written responses. SAQ responses were coded and scored based on predetermined *big idea topic learning objectives* and the *Knowledge Integration Scoring Rubric* (DeBoer, Lee et al. 2008; Liu, Lee et al. 2008), both of which are described in the Materials section below. After Quiz 1, students were randomly assigned to one of three repeated testing/repeated study MCQ conditions - *see Table 2: Quiz group conditions*. **(D, E, and F)** In MCQ's 2-4, students in groups 1 and 2 were repeatedly tested or exposed to big ideas 1 & 2. Students in group 3 (control) received questions from unrelated lecture material. Five weeks after Quiz 4, **(G)** all students answered 10 identical multiple-choice questions concerning both big ideas on the comprehensive final exam. The main dependent variable – comprehension retention of each big idea - was calculated as (Final exam % score / Quiz 1 % score). MCQ and final exam retention performance of students also completing the SAQ was calculated separately.

Materials

Pretest - All students completed a sixteen-question pre-test the first week of class to measure previous knowledge of cell biology and genetic concepts. Questions were chosen from the Genetics Literacy Assessment Instrument - Test-retest stability (Pearson): 0.68, Internal reliability (Cronbach's α): 0.995, N=395 (Bowling, Acra et al. 2008). The instrument was designed to assess knowledge relevant to topics covered in MCB 181 and concepts assessed by the current study. Pretest performance data were used to confirm that average previous student knowledge of relevant topics was not significantly skewed across treatment groups.

Two Big Ideas – Student comprehension and retention of two fundamental "big ideas" of molecular biology were targeted in the experiment; (1) the relationship between genotype and phenotype, and (2) the relationship between gene expression and cell function (*see Table 1 below for description of specific learning objectives*). The topics were chosen for their relevance to overall course learning objectives. The difficulty and complexity of the topics allowed for differentiation of student comprehension and retention quality (Lewis, Leach et al. 2000; Lewis, Leach et al. 2000; Lewis and Wood-Robinson 2000; Lewis and Kattmann 2004; Crowe, Dirks et al. 2008; Mazzocchi 2008). Two weeks into the course, students received a six page instructional study packet about the big ideas. The packet emphasized conceptual understanding and the application of knowledge through use of question-answer scenarios (see Appendix B for

examples). Students were informed that questions concerning the big ideas would appear on future weekly multiple-choice (MC) quizzes and on the final exam. Correct quiz responses would receive extra credit points toward overall quiz grade, providing students additional motivation to engage the material.

Table 1: Specific learning objectives for "big idea" topics (emphasized in instructional handout and related assessment questions).

Topic	**Big Idea #1:** The relationship between genotype and phenotype.	**Big Idea #2:** The relationship between gene expression and cell function.
If students fully comprehend this idea, they should be able to:	1. Compare and contrast the definitions of genotype and phenotype. 2. Describe the relationship between a gene and a protein (*e.g.* is there a direct or indirect link? do changes in one affect the other, how?). 3. Describe the relationships between protein function, environment, and phenotype. *4.* Predict whether changes to genes or environment will affect phenotype and *vice-versa.*	1. Describe the relationship between cell structure (*i.e*. parts and layout) and cell function. 2. Draw out a basic gene expression pathway (*e.g.* Central Dogma), including regulation steps. 3. Describe the relationship between gene expression and cell environment. 4. Describe the relationship between gene expression and cell function.
Ideal student responses that demonstrate successful comprehension of learning objectives.	***Question****: Briefly explain the relationship between genotype and phenotype.* ***Student response****: An organism's genotype includes all the genes, coded in DNA, which it inherits from its parents. Phenotype includes all the organism's observable traits such as height, hair color, or how fast or slow its metabolism is. Genes contain specific instructions for how to build proteins. A protein is made of a chain of amino acids bond together in a specific order. Each amino acid has a unique shape and charge. Because the chain is flexible, the amino acids can interact, causing the chain to fold into a particular 3D shape depending on its sequence. The overall structure (shape and charges) of the final protein determines its function. One example is a membrane bound ion channel*	***Question:*** *Briefly explain how one genotype is used to create hundreds of different cell phenotypes.* ***Student response:*** *All 10 trillion cells in our body come from one fertilized egg that copied and divided itself many many times by mitosis. So all the resulting cells have the same exact DNA (same genotype). Human DNA has about 25,000 genes with instructions for how to make proteins. Proteins do things like catalyze reactions, transport molecules, support cell structure, and receive and transfer information.* *The phenotype of a cell depends on the combination of proteins it is*

	protein, which literally looks like and acts like a tunnel with a controllable gate. *Phenotypes like height or rate of metabolism tend to result from the functions of many many proteins working together. Sometimes an organism's phenotype depends on both the genes it has (genotype) AND environmental conditions. For example, being tall requires you have many growth encouraging genes AND proper nutrition when developing. Some phenotypes, like sex, are determined only by genotype (XX or XY chromosomes). Changes to a gene's DNA (like a mutation) can change which amino acids it codes for, resulting in changes to the protein's structure and function. If significant, that change in function might result in an observable difference in phenotype (like disease, a change in behavior, or a heritable adaptation).*	*made of and how they work together. Turning on or off (expressing) specific combinations of genes can create specific cell function. No cell expresses all of its genes at the same time. Instead, cells tend to express only certain sets of genes at certain times in their life cycle.* *Gene expression is a dynamic process that can change based on what is going on outside of the cell or inside the cell This process is regulated by proteins that interact with DNA, either promoting or inhibiting transcription of genes. Changes to regulatory proteins changes their function and what genes are turned on or off. Cells can change gene expression and function in other ways as well, like destroying mRNA transcript after its already made, or sticking active proteins in an organelle where they can't do anything. This active control of gene expression and function allows cells to produce all kinds of different phenotypes from just one genotype.*

Quiz 1 – Two weeks after receiving the instructional study packet, all students completed an identical online multiple-choice quiz (*hereafter labeled Quiz 1*) that assessed comprehension of both big ideas (*see Table 3 for example question*s). Again, all study questions were integrated into the normal weekly quizzes and the 60 item comprehensive final exam to ensure that assessments modeled authentic classroom practice. The format and logistics of the assessments limited question number to 5 per big idea. All quiz questions were chosen or derived from instruments used in previous work to assess comprehension of the same concepts (Lewis, Leach et al. 2000; Lewis, Leach et al. 2000; Lewis, Leach et al. 2000; Lewis and Wood-Robinson 2000; Wood-Robinson, Lewis et al. 2000; Lewis and Kattmann 2004; Duncan and Reiser 2007; Klymkowsky, Gheen et al. 2007; Crowe, Dirks et al. 2008). All online quiz questions appeared to students in random order. Students could attempt Quiz 1 only once. An answer key and corrective feedback was made available after the closing date of the quiz. Quiz 1 performance data served as the initial measure of comprehension that Final Exam performance would be compared to in order to determine final retention performance. Ideally, students would learn study material to mastery prior to subsequent treatments of study or testing. However, due to the conceptual difficulty of material, the restricted timeframe of the course, and the number of students (300+), it was neither logistically possible nor plausible to ensure all students mastered content before moving on.

Quizzes 2-4 (repeated quizzing /restudy conditions) – After Quiz 1, students were assigned to one of three weekly quiz conditions (Table 2: Quiz group conditions). Numerous studies have demonstrated that three repeated testing events is the threshold for significant testing effect to be observed, both in shorter term (hours to days) and longer term (days to months) studies (see Roediger and Karpicke 2006 for review). Therefore, treatment conditions were applied to three weekly quizzes directly following Quiz 1. In each condition, students encountered a specific combination of big idea questions (repeated testing) and/or statements to read and evaluate (repeated study). Quizzes taken by Groups 1 & 2 repeatedly tested only one of the two big ideas (see Table 3, Sections 1 & 2 for examples). Questions assessed similar comprehension as those found on Quiz 1 and the comprehensive final exam. Excessive verbatim repetition of questions across assessments can encourage recognition based responses over more thoughtful comprehension based responses (Haladyna, Downing et al. 2002; Mislevy, Steinberg et al. 2003; Stupans 2006; Momsen, Long et al. 2010), Therefore, an effort was made to vary questions sufficiently to avoid reflexive responding (Klymkowsky, Gheen et al. 2007). This was done, for example, through modification of question and response phrasing (Table 3, Section A) or through novel scenarios that assessed similar comprehension (Table 3, Section B).

Table 2: Quiz group conditions (Description of the questions and restudy conditions found in each group condition).

Group 1	***Group 2***	***Group 3***
3 Big idea 1 MC questions **2** Big idea 2 statements to evaluate **5** lecture MC questions (unrelated)	**2** Big idea 1 statements to evaluate **3** Big idea 2 MC questions **5** lecture MC questions (unrelated)	**10** MC questions assessing unrelated lecture material

As a control for exposure time, students were also asked to evaluate statements about the alternative big idea (see Table 3, Section C for examples). Statements were constructed based on similar controls used in previous investigations (McDaniel, Anderson et al. 2007; Weigold 2008; Carpenter, Pashler et al. 2009; Larsen, Butler et al. 2009; Karpicke and Zaromb 2010; Roediger III and Butler 2010). The cross-counterbalance of retesting and restudy conditions is similar in design to previous instruments used to investigate testing effect in applied contexts (McDaniel, Anderson et al. 2007; Rohrer and Pashler 2007; Agarwal, Karpicke et al. 2008; Kromann, Jensen et al. 2009). .

Table 3: Comparison of big idea questions across study assessments.

<table>
<tr><td colspan="2">Section A: Example of Big idea 1 question variation</td></tr>
<tr><td>Quiz 3, Group 1
What is the relationship between the physical and behavioral traits of an organism and the information contained in the organism's genes?
a) Traits are a product of two or more proteins functioning together.
b) Genes code for the structure/function of proteins.
c) Environmental signals primarily determine how proteins will function.
d) The information in genes code for the specific structure and function of traits.</td><td>Final Exam (all students)
What is the relationship between genes and traits expressed in individuals?
a) Genes determine the structure and function of proteins, which are responsible for individual traits.
b) Genes code for chromosomes, which are responsible for individual traits.
c) Genes code for individual traits, which regulate expression of specific proteins.
d) Environmental signals primarily control expression of genes responsible for individual traits.</td></tr>
<tr><td colspan="2">Section B: Example of Big idea 2 question variation</td></tr>
<tr><td>Quiz 1 (All students)
What is the relationship between gene expression and cell function?
a) Expression of specific combinations of traits encoded by genes determines overall cell function.
b) Gene expression patterns fluctuate significantly until cell function is established.
c) Gene expression patterns control cell protein populations which determine cell function.
d) Cell specialization and function is established through expression of a modified cell genotype.</td><td>Quiz 2, Group 2
A friend gives you some cuttings (clones) from his favorite plant. You pot the cuttings and place one pot in your office and one outside on the patio. After about a week, both plants look healthy and are growing new leaves. You are concerned, however, because the plant growing outside is turning purple while the plant in your office remains green. What is the most likely cause of these different phenotypes?
a) Genotype difference between the cuttings.
b) Expression of different genes influenced by the environment.
c) An accumulation of mutations in the plant grown outside.
d) Differences in adaptability between the cuttings.</td></tr>
<tr><td colspan="2">Section C: Examples of big idea 1 & 2 statements to evaluate (restudy condition as control for exposure)</td></tr>
<tr><td>Quiz 2, Group 2
A phenotype or trait results from the expression</td><td>Quiz 2, Group 1
In multicellular organisms, no cell expresses all of</td></tr>
</table>

of one or more proteins and their associated function. *(There is no wrong answer to this question.)* a) This idea is familiar to me. b) This idea is new or unfamiliar to me.	the genes in its genome. Usually, each cell expresses only a small fraction of its total genes at any one time. *(There is no wrong answer to this question.)* a) This idea is new or unfamiliar to me. b) This idea is familiar to me.

Group 3 quizzes covered current or past lecture material unrelated to the big ideas. Group 3 data served as a control for the influence of lecture on the comprehension and retention of the big ideas. Unlike Quiz 1, Quizzes 2-4 supplied only total performance scores after completion. No answer key or explicit corrective feedback was supplied to any group condition. However, because 5 total quiz attempts were permitted, it was possible for students to infer correct answers over multiple attempts. Attempt number was tracked for each quiz and was considered as a factor in later analysis of retention (*see* Results).

Final Exam - Five weeks after Quiz 4, students took an identical 60 question multiple-choice final exam which included 10 questions parallel to the big ideas topic questions presented on Quiz 1 (5 questions per topic). All students completed the same 10 questions. The main dependent variable – retention of big idea comprehension over the

course of the semester - was calculated as Final exam % score / Quiz 1 % score (per big idea).

Short Answer Quiz (SAQ) - To characterize the quality of comprehension assessed by the MC assessments, a subset of students also completed an extra-credit SAQ . The SAQ was offered the same week of Quiz 1 (*see* Figure 1). All eight questions used for analysis derived from Duncan & Reiser's (2007) short-response instrument designed to elicit student comprehension of topics equivalent to Big Ideas 1 & 2. Students were given one attempt to take the SAQ online. A total of thirty-six students completed the quiz. Students were instructed that total extra credit awarded was based on how well they elaborated and explained (made visible) their current understanding, NOT on the correctness of their answers. Answers that were clearly copied and pasted or derived directly from the textbook would receive no credit. Therefore, there was little motivation for students to look to outside sources to complete the quiz. A sample question and answer was supplied to model the quality of answers receiving "full credit."

Both the MC and SA quizzes implemented in the study were designed to assess the same concepts (big ideas) at the same level (comprehension and application - as measured by Bloom's taxonomy and the Blooming Biology Tool developed by Crowe, Dirks, and Wenderoth) (Buckles and Siegfried 2006; Crowe, Dirks et al. 2008; Zheng,

Lawhorn et al. 2008). Though several studies have demonstrated that MC questions are capable of evaluating sophisticated student comprehension and understanding (Ram, van der Vleuten et al. 1999; Fellenz 2004; Stupans 2006; Williams 2006), MC scores supply limited information about what that comprehension "looks like." Properly designed SA questions can elicit sufficiently elaborate information necessary to characterize the quality and quantity of student knowledge and understanding (Palmer and Devitt 2007). The SA quiz was used to gather this quality of evidence to complement the evidence of student comprehension demonstrated by MC question performance.

Scoring Short Answer Quizzes

The SAQs were scored and coded using the Knowledge Integration (KI) Framework and Rubric. The Knowledge Integration framework is a constructivist based theory of learning used to describe and measure the state of an individual's current knowledge (Ram, van der Vleuten et al. 1999; Zimmerman 2005; DeBoer, Lee et al. 2008; Liu, Lee et al. 2008). The theory assumes that new understanding is achieved through construction and development of cognitive knowledge structures. The structures can be imagined as mental networks composed of discrete facts or ideas connected by a variety of associations and/or relationships. Knowledge structures grow in size and complexity through integration of new information with previously held knowledge (van den Broek and Kendeou 2008; Catts 2009). As learning progresses, the expanding networks are themselves iteratively evaluated, reorganized, and integrated to ensure internal

consistency and fit with new experience (Liu, Lee et al. 2008; Liu, Lee et al. 2010). Comprehension and aptitude result when knowledge structures support meaningful interpretation of new and previously held ideas and experience. (Chen 2006; Cheng 2008; van den Broek and Kendeou 2008; Catts 2009; van den Broek 2010) The larger and more integrated knowledge structures become, the greater the performance and aptitude they are likely to support (Clark and Linn 2003; Catts 2009).

The KI rubric is a simple scale and methodology for measuring the quantity and quality of learner's knowledge structure(s) (Liu, Lee et al. 2008; Liu, Lee et al. 2010) (*see Table 4 below*). It is used in conjunction with assessment instruments that demand recall, explanation, and application of understanding through tasks such as solving a problem or developing an argument (Ram, van der Vleuten et al. 1999; DeBoer, Lee et al. 2008; Lee and Liu 2009). The rubric focuses on evaluation of answer relevancy, explanation power, and connection making between ideas. The output is a numeric score between 1 and 5 that represents all evaluation criteria. The single score allows easy comparison across related items, offering a broad window into the coherence and sophistication of learners' current state of knowledge, as well as their ability to apply that knowledge (Liu, Lee et al. 2010; van den Broek 2010).

Table 4: The 5 scoring levels of the knowledge integration rubric (Liu, Lee et al. 2008)

Score	**Description**
5	Two or more full links represent the "complex" link level (scored as 5).
4	At the "full link" level (scored as 4), students make at least one full link between two relevant and correct ideas.
3	At the "partial link" level (scored as 3), relevant and correct ideas are generated but not elaborated enough to demonstrate how two ideas are connected, meaning that no specific mechanisms or relationships are articulated.
2	At the "no-link" level (scored as 2), students use incorrect ideas and incorrect links.
1	The "no answer" and the "off task" responses are scored as 0 and 1, respectively.

As outlined in Liu *et al.* (Liu, Lee et al. 2008), implementation of the rubric involves two basic steps:

1) Individual facts and ideas within each answer are identified and evaluated for accuracy and relevance to the assessment task.
2) If multiple correct ideas are present, descriptions of connections and/or relationships between ideas are identified and evaluated. Each connection that is explained accurately and fully (*i.e.* includes a specific mechanism or a type of relationship) is scored as a "full link." Connections that are mentioned but not fully explained are scored as "partial links."

Referencing Table 4, a student response that contains correct ideas but no links would be scored as a 2. A correct response that included multiple full links would be scored as a 5. Table 5 illustrates how the Knowledge Integration rubric was utilized to score the SA quiz (see below).

Relationship between MC and SA quiz questions - Comparison of Quiz 1 and SAQ KI scores will help characterize the quality of student comprehension measured by the MC questions used in the study**.** Such a comparison is supported by Liu *et al.'s* (2008) finding that KI scores parallel scores of other assessments - such as multiple-choice questions – if learning objectives and targeted level of understanding are aligned across assessments. Related findings demonstrate that KI scores not only move in parallel with

other assessments, they reliably predict a student's ability to fully explain their reasoning behind MC response choices (Chen 2006; Linn, Lee et al. 2006; DeBoer, Lee et al. 2008; Liu, Lee et al. 2008; Lee and Liu 2009; Liu, Lee et al. 2010). Because both the MC and SA quizzes implemented in this study are designed to assess the same concepts at the same level, a strong correlation between Quiz 1 and SAQ KI scores would suggest that the two assessment instruments are targeting a similar type and quality of knowledge.

Table 5: Scoring SAQ using the Knowledge Integration Rubric.

All 288 SAQ responses (36 quizzes completed / 8 questions per quiz) were transferred from D2L to a MS Excel spreadsheet for analysis. Section **(A)** illustrates scoring of SA quiz question 1, including sample descriptive codes representing facts, ideas, and relationships described in student answers, as well as the overall KI scores applied to each answer. Section **(B)** details the KI scoring of a student response from section A (bottom row).

Section A: **Example of analysis of raw SAQ data downloaded to MS Excel template.**

	Is there a connection between proteins and phenotype? Explain.	Code: genes contain information for building proteins. PF: protein function determines cell function – observable phenotype GeneF: gene information determines function/expression of proteins Exp: Protein expressions affects phenotype Mut: Changes to genes can affect protein structure-function, pheno	
	Since genes are made up of proteins and genes dictate genotype which in turn determines phoenotype, yes there is a connection.	2	Since genes are made up of proteins and genes?
	Proteins and phenotype work together, but not sure as to how, or for what.	2	
	Genotype controls phenotype because genes direct the production of proteins. Proteins in turn, dictate virtually every reaction in the cell and thus are directly responsible for observational characteristics.	5	Genef PF
	Yes; phenotypes are determinant on either one or many different proteins. Proteins have information stored in them to perform different tasks or functions which can be displayed as visible or physical traits. Proteins express the different genes they contain as phenotypes.	4	PF Proteins express the different genes they contain as phenotypes.
	no.	1	
	Yes, there is a connection between the two. Proteins are required for DNA synthesis. Chromosomes are organized structures of DNA and protein. Chromosomes in turn have genes which determine phenotype.	3	Chromosomes in turn have genes which determine phenotype
	Though proteins provide various functions encoded by genetic material, they do not have a direct effect on phenotypic variation among living things. Their functions, though wide and nearly-all-encompassing, cover the microscopic variation among organic material.	3	do not have a direct effect on phenotypic variation
	Proteins influence gene expression which determines what parts of the genotype are expressed as the phenotype. For example, Human growth hormone which is produced by the pituitary gland affects height. People with pituitary dwarfism have a mutation that inhibits the production of human growth hormone which causes them to be shorter than the average person.	4	exp
	Proteins are usually pretty similar to phenotype. Both are based upon your set of genes, called your genotype. Phenotype is the physical manifestation of those genes. At the same time, those genes are what encode the proteins that we have. It is because of those proteins, and how much our genotype says we should have of them, that gives us our physical appearance, and therefore our phenotype.	5	Code exp
	Proteins are responsible for an individual's phenotype in most cases (how a person looks); they are made of a specific sequence of amino acids that were specified by an individual's DNA. In a way, a hierarchy of sorts exists in which we come from; for example DNA is made up of billions of base pairs; every 3 base pairs, when copied by RNA, codes for a specific amino acid (although there are repeats); different combinations of amino acids create different proteins; different proteins fold different ways and therefore have different functions; these different functions now can be expressed and thus account for the wide variation in phenotypes for the different	5	Code Pf exp

Individual student responses to Question 1

Descriptive coding of SAQ responses

SA scoring using Knowledge integration rubric
(see Section B)

Section B: **KI scoring process**

Step 1: Identify and evaluate the facts and ideas within each response.
Step 2: Identify and evaluate any connections and/or relationships made between ideas.

Student answer (from Section A, bottom row): "Proteins are responsible for an individual's phenotype in most cases (how a person looks); **(1)** they are made of a specific sequence of amino acids that were specified by an individual's DNA. In a way, a hierarchy of sorts exists in which we come from; for example DNA is made up of billions of base pairs; every 3 base pairs, when copied by RNA, codes for a specific amino acid (although there are repeats); **(2)** different combinations of amino acids create different proteins; different proteins fold different ways and therefore have different functions; **(3)** these different functions now can be expressed and thus account for the wide variation in phenotypes for the different cells."

Ideas and relationships identified (codes and explanations):

1. **Code**: genes contain information for building proteins.
2. **PF**: protein function determines cell function – observable phenotype
3. **Exp**: Protein expression affects phenotype

In this case, the student used multiple biological ideas correctly and fully described three complex relationships between those ideas. Application of the rubric (*see Table 4 for reference*) to this particular answer would yield:
KI score 5 ***"Two or more full links represent the "complex" link level."***

RESULTS

The purpose of the study is to determine if repeated testing versus repeated study affects retention of complex comprehension. Retention, defined as the ability to recall and apply information, requires that information is first encoded. Due to the complex nature of the big ideas used in the study, the size of the class, and the limited resources of the instructors, it was logistically unworkable to ensure that all students mastered the target material. Therefore, it was necessary to limit comparison groups to those students who scored 60% or greater on Quiz 1 big idea 1 & 2 questions (3 correct out of 5 on each idea, total of at least 6 correct). The approach has been used in earlier memory work to reduce the influence of confounding variables such as motivation and analytic skill (Roediger 2008). Sixty percent was the minimum threshold necessary to distinguish differences in retention across treatment groups and is approximately the average test score on the majority of Intro to Biology exams across sections and semesters. Students completing the extra-credit short answer quiz were also excluded from this population due to the additional exposure and testing of material. The selection criteria reduced treatment group populations to 48 (Group 1), 50 (Group 2), and 27 (Group 3). Using 80% or greater as the selection criteria would have been more optimum, however too few students performed at this level to permit statistical comparison. Unless noted, the following comparisons and analyses utilize the group populations described in Table 6.

Table 6: Population breakdown of students scoring ≥ 60% on Quiz 1 big idea 1 & 2 questions.

Treatment		Description	N
Group	1	Test BI 1 / Study BI 2	48
	2	Test BI 2 / Study BI 1	50
	3	Control	27

Accounting for Previous Knowledge

The Pretest data was collected to ensure that previous knowledge related to the big ideas was equivalent across treatment groups prior to the beginning of the study. Any group bias could significantly affect later performance and prevent meaningful interpretation of retention outcomes. A one-way ANOVA was used to determine if Pretest performance varied across groups, indicating a bias in previous knowledge. No significant variation was detected, $F(2, 124) = .930$, $p = .397$ (see Figure 2 for comparison of Pretest means).

Figure 2: Pretest performance assessing previous knowledge related to big ideas.

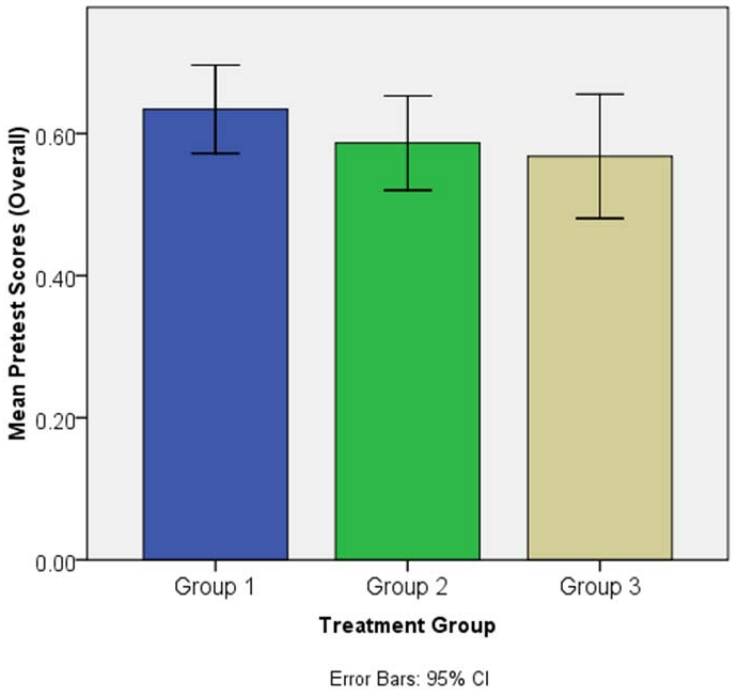
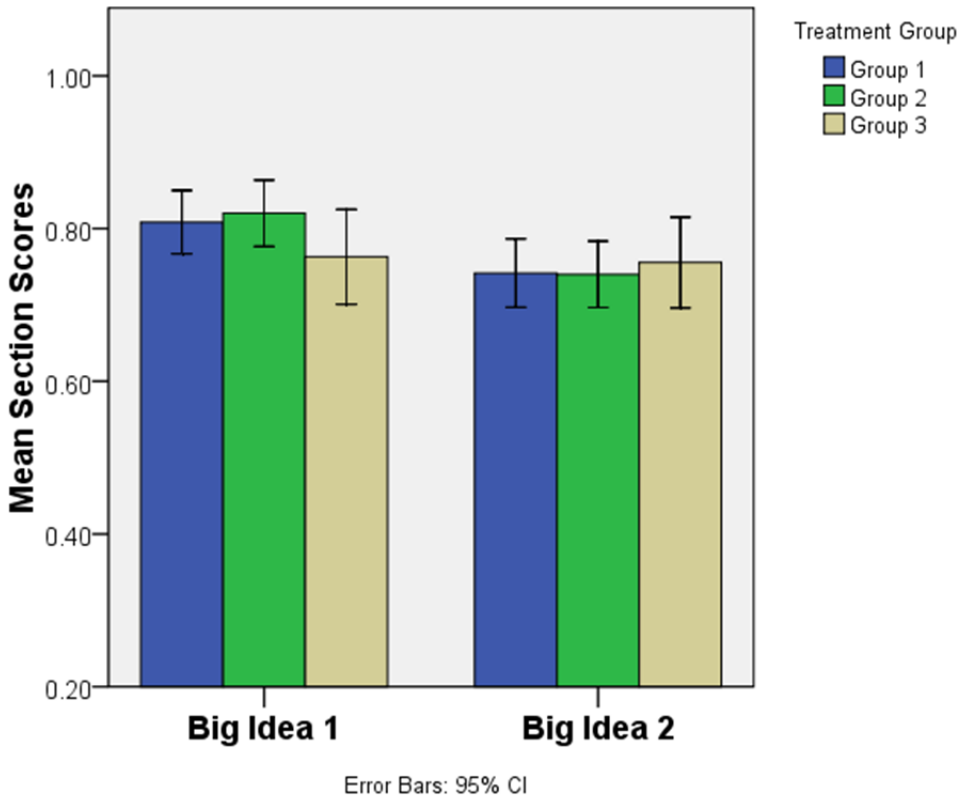


Figure 3: Quiz 1 performance across treatment groups (results clustered by Big Idea, 5 questions each).



Measure of Initial Learning

Student performance on Quiz 1 served as the baseline for comparison to determine if group treatments affected later retention of the big ideas. All students completed an identical quiz containing 5 MC questions targeting each big idea. Again, any differences in group performance on this initial assessment would make meaningful comparison and interpretation of later assessment outcomes difficult. Therefor it was important to determine if significant group performance variation was present at Quiz 1. A one-way multivariate analysis of variance (MANOVA) was conducted on the two dependent variables - performance on each set of big idea questions. MANOVA analysis permitted simultaneous comparison of the two dependent performance variables and offered greater sensitivity to group differences than individual ANOVA analyses, especially in situations where covariation of variables is expected (see Measure of Testing Effect section below). Unless otherwise indicated, all multivariate analyses conducted as part of the study yielded nonsignificant Box's *M*, indicating that the homogeneity of variance-covariance matrix assumption was not violated. In addition, no univariate or multivariate outliers were evident, indicating the necessary assumptions for the MANOVA analysis technique were met.

No significant differences were found among the three groups (Wilks' $\lambda = .977$, $F(4,242) = .714$, $p > .1$), suggesting that initial mastery of the target material was equivalent across groups at the beginning of the study (see Figure 3 for comparison of Quiz 1 means).

Measure of Testing Effect

To determine if re-testing or re-study influenced retention to a greater extent, two approaches were used to compare pre (Quiz 1) and post (Final Exam) performance in order to identify treatment affects. Approach one compared group final exam performance on the big idea questions. Approach two utilized the ratio of Final exam / Quiz 1 performance to further characterize the relationship between treatments and retention.

Comparison of Final Exam Performance – Because no significant differences in Pretest and Quiz 1 performance were found between groups, a direct comparison of Final exam scores between groups was possible (see Figure 4). A MANOVA was conducted in order to determine whether repeated testing versus repeated study afforded a performance advantage on the final exam. As with the multivariate analysis of Quiz 1, performance on each big idea was treated as a dependent variable (5 questions per big idea, 10 total). MANOVA analysis permitted comparison of the dependent variables together. In addition, group 1 & 2 outcomes were likely to covary due to topic and format similarities. MANOVA analysis is capable of detecting group differences under these conditions, while multiple ANOVAs are not. Using Wilks' Lambda, there was a significant difference in scores across the treatment groups, $\lambda = .886$, $F(4,242) = 3.762$, $p < .01$. In order to help characterize group differences, separate ANOVAs were conducted on the

outcome variables. Significant differences in performance were seen on Big Idea 1 across groups, $F(2,122) = 4.784$, $p = .01$, but not Big Idea 2, $F(2,122) = 2.311$, $p > .05$.

To better understand the relationships between the dependent variables, the MANOVA and ANOVAs were followed up with discriminant analysis, which revealed two discriminant functions. The first explained 71% of the variance, canonical $R^2 = .08$, whereas the second explained only 29%, canonical $R^2 = .02$. These discriminant functions significantly differentiated the treatment groups, both in combination (1 & 2), $\Lambda = 0.886$, $\chi^2(4) = 14.659$, $p = .01$, and with the first function removed, $\Lambda = .965$, $\chi^2(1) = 4.332$, $p < .05$. The correlations between outcomes and the discriminant functions suggested big idea 1 performance factored significantly more in function 1 over function 2 (r =.896 for the first function versus r=.443 for the second); performance on big idea 2 was the opposite, factoring significantly more so in function 2 (r=.985) over function 1 (r=-.174). The discriminant function plot (Figure 5) shows that the first function discriminates Group 1 from Groups 2 & 3 (compare group centroids across Function 1 axis - solid line), and the second function differentiates control Group 3 from the two intervention groups (compare group centroids across Function 2 axis - dotted line).

Taken together, these data suggest that repeated testing of big idea 1 was associated with group 1's significantly improved performance over either group 2 or the control. However, the effect size was not strong ($R^2 = .08$). Group 1 & 2 performance on big idea

2 questions did not vary significantly, but was significantly greater than the control - suggesting both treatments contributed to performance, but the mechanism is unclear.

Figure 4: Final Exam performance mean scores across treatment groups (results clustered by Big Idea, 5 questions each).

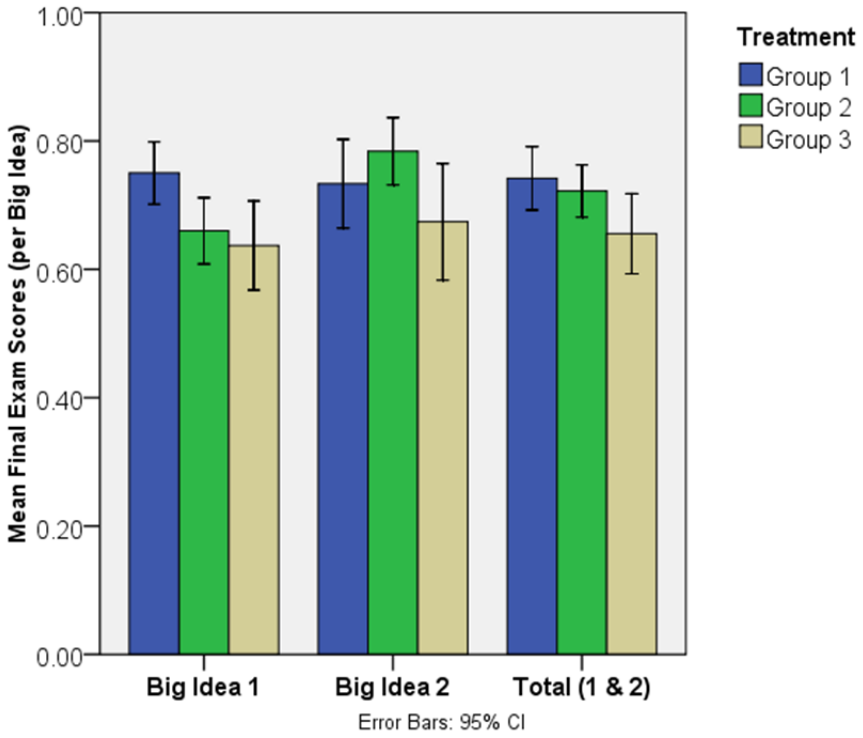
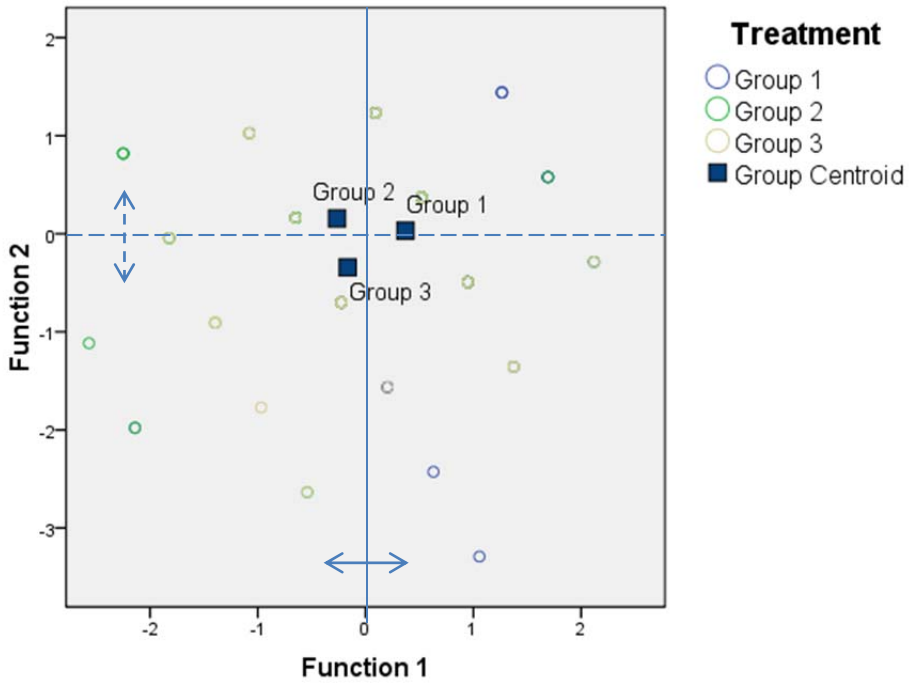


Figure 5: Discriminant function plot - Group final exam performance on big ideas 1 & 2.



Comparison of Conditionalized Retention – To quantify retention across the pre and post assessments, the following ratios of Final Exam performance to Quiz 1 performance were used to conditionalize retention based on initial learning (measured by Quiz 1 performance):

Big Idea (1) Final Exam performance %

Big Idea (1) Quiz 1 performance %

Big Idea (2) Final Exam performance %

Big Idea (2) Quiz 1 performance %

For example, if a student scored 80% on Quiz 1 big idea 1 questions and 60% on equivalent final exam questions, then retention was scored as 75% (.6/.8). (See Figure 6, Percentage Retention means).

As before, a MANOVA was conducted in order to determine whether either treatment yielded a retention advantage, comparing the dependent various together. Using Wilks' Lambda, there was a significant difference in retention across the treatment groups, $\lambda = .918$, $F(4,242) = 2.656$, $p < .05$. However, separate univariate ANOVAs on the dependent variables revealed non-significant differences in retention of Big Idea 1, $F(2,122) = 2.408$, $p > .05$, and Big Idea 2, $F(2,122) = 2.850$, $p > .05$.

The MANOVA and ANOVAs were followed up with discriminant analysis (see Figure 7), which revealed two discriminant functions. The first explained 69% of the variance, canonical R^2 = .06, whereas the second explained only 31%, canonical R^2 = .001. These discriminant functions significantly differentiated the treatment groups in combination (1 & 2), Λ = 0.918, $\chi^2(4) = 10.440$, $p < .05$. However, the second function alone did not significantly differentiate groups, Λ = .973, $\chi^2(1) = 3.316$, $p > .05$.

The correlations between outcomes and the discriminant functions suggested retention of Big Idea 1 factored significantly more in function 2 over function 1 (r = -.647 for the first function versus r=.764 for the second); retention of big idea 2 factors slightly more in function 1 (r=.801) than function 2 (r= .602). The discriminant function plot (Figure 5) shows that the first function discriminates Group 1 from Groups 2 & 3 (compare group centroids across Function 1 axis - solid line), while the second function differentiates control Group 3 from the two intervention groups (compare group centroids across Function 2 axis - dotted line).

Figure 6: Percentage retention of big ideas 1 & 2 from Quiz 1 to Final exam (results clustered by big idea).

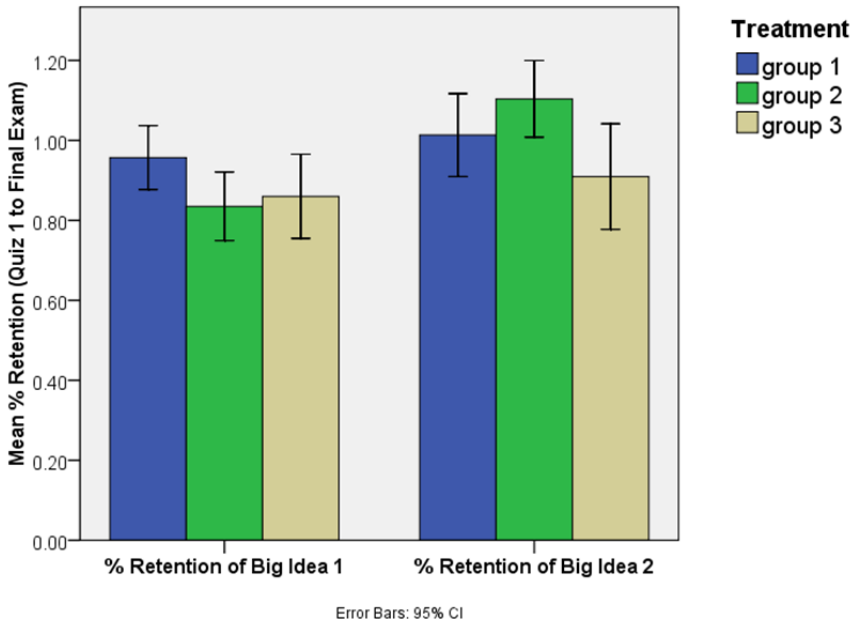
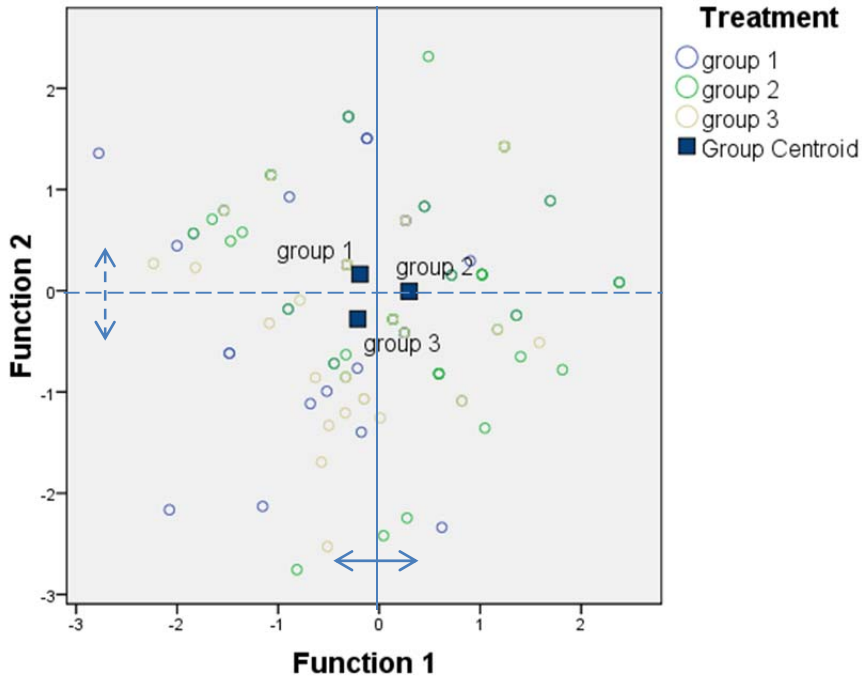


Figure 7: Discriminant function plot – Percentage retention of big ideas 1 & 2 from Quiz 1 to Final exam**.**



Taken together, the retention analyses suggest that repeated testing of big idea 2 significantly affected group 2 retention, but did not distinguish big idea 1 retention across groups 1 & 2. Though the effect size was again not strong (R^2 = .06), it is interesting to note that Group 2's mean % retention was greater than 1.0. An equal increase is not seen in the other groups, suggesting that that repeated testing condition appears to have influenced not just retention, but possibly learning as well. As before, Group 1 & 2's performance was significantly greater than the control. Considering the heavy loading of big idea 2 retention in both discriminant functions (as was similar to big idea 1 final exam performance in the previous analysis), it is likely that the effects of repeated testing of both big ideas were interacting and that both treatments contributed to retention or test performance in general.

Accounting of Quiz Attempts

As described previously, students completed 3 treatment quizzes between Quiz 1 and the Final Exam. Though no direct feedback was supplied by these quizzes, students could attempt each quiz up to 5 times. Because greater exposure to material can influence later retention (Roediger III and Butler 2010), it can be argued that a greater number of attempts could bias performance (through greater exposure and/or retesting of ideas) and confound interpretation of retention outcomes. Therefore, the relationship between total quiz attempts and retention of big ideas 1 & 2 (measured as % of learning maintained pre-post, as described above) was analyzed. Accumulated quiz

attempts for quizzes 2-4 were calculated for each student. Pearson's correlation coefficient, r, was calculated to determine if a relationship existed between total quiz attempts and retention of either big idea. No significant relationship was found between quiz attempts and either retention of big idea 1 ($r = .02$, $p > .05$), or big idea 2 ($r = -.01$, $p > .05$). Therefore, it can be concluded that attempt number did not contribute significantly to the observed performance outcomes.

Relationship Between Multiple-Choice and Knowledge Integration Scores

The multiple-choice questions used in the study were intended to measure students' comprehension of ideas and their ability to explain and apply them. In order to support whether the questions were in fact able to do so, a semi-random subpopulation of students (36) completed both the multiple-choice Quiz 1 and a short answer quiz that required written elaboration of equivalent big ideas topics. The short answers were coded using the Knowledge Integration Rubric (described in Methods), a technique used to identify and quantify the degree of student knowledge integration and the ability to explain and communicate that knowledge. If the multiple-choice and the short answer questions measured equivalent knowledge, one would expect student performance across assessments to be significantly correlated. Knowledge integration and multiple-choice scores were strongly correlated, $r=.53$, p(two-tailed) $<.01$, supporting the idea that both assessments measured similar qualities and quantities of comprehension.

DISCUSSION

The current exploratory study examined whether repeated testing or repeated study is a more effective strategy for improving students' retention of complex biological concepts in a large, undergraduate biology lecture course. Additionally, the study sought to determine if a multiple-choice question format can measure the quality of understanding at a level equivalent to short answer questions. The latter investigation was necessary to bridge the gap between how testing is usually administered in large lecture classes, and how complex understanding is most often measured (Rodriguez 2003; Kuechler and Simkin 2010; Momsen, Long et al. 2010).

Based on previous testing effect findings of recall and recognition tasks, it was predicted that repeated quizzing would significantly improve retention of over repeated study of the same material (Roediger III and Butler 2010).

The analysis measuring impact of testing on retention and final exam performance yielded mixed results. Final exam scores of students scoring 60% or better on the initial quiz were significantly improved for Big idea 1, however univariate and discriminate analysis could not distinguish significant differences between groups 1 & 2 on big idea 2. Both groups performed significantly better than the control, confirming that lecture attendance was not a contributing factor.

Interestingly, when percentage retention between Quiz 1 and Final exam was measured, discriminate analysis revealed the opposite effect. Repeated testing of big idea 2 improved group 2's retention over group 1, however no differences in retention of big idea 1 could be detected across groups 1 & 2. This reversal of significance may be due in part to the fact that repeated testing of either big idea helped performance on big idea 2 final exam questions (see Figure 4). This was not the case for big idea 1 questions. The ideas tested by big idea 2 quiz questions, such as factors that influence gene expression and cell function, are closely related to big idea 1 topics. In fact, it could be argued that big idea 1 ideas surrounding the relationship between genotype and phenotype may be necessary precursors for understanding gene expression and cell function. Thus, not only was big idea 2 question performance was aided by both treatments, group 2 was exposed to more complex ideas and greater expectations in the repeated testing treatment, resulting in greater performance on all questions overall. This is supported by the observation that group 2's retention ratio was above 1.o, suggesting that some amount of learning must have occurred as a result of repeated testing, an effect not seen in the other treatments. Again, both groups scored significantly higher than the control group, suggesting differences were a result of the interventions, not the influence of lecture.

The analysis of group retention differences suggest that repeated testing did improve student performance on the big idea questions over repeated study. The effect was

most distinct with repeated testing of big idea 1. Repeated testing of big idea 2 appeared to influence performance on both types of questions. Because the big idea topics overlapped substantially, it could be anticipated that effecting retention or comprehension of one idea influenced performance on the other. Because independent improvement of retention was seen in at least group 1, it is reasonable to conclude that the observed effects resulted in general from testing and retrieval processes, not greater exposure.

Another key finding was the close relationship between KI scores and multiple-choice performance. The Knowledge Integration Construct and Rubric was specifically designed to measure the quality, integration, and applicability of student knowledge (Liu, Lee et al. 2008; Lee and Liu 2009; Lee, Liu et al. 2011). Multiple studies have demonstrated its value in measuring complex knowledge constructs (DeBoer, Lee et al. 2008; Liu, Lee et al. 2010; van den Broek 2010; Chiu and Linn 2011). However, very few have demonstrated that multiple-choice questions, when designed and implemented appropriately, can reliably measure similar constructs (Stupans 2006; Lee, Liu et al. 2011). In fact, numerous authors have argued that recognition format questions, such as multiple-choice, are incapable of measuring complex knowledge due to inherent limitations of the design (Palmer and Devitt 2007; Nielsen, Buckingham et al. 2008; Zheng, Lawhorn et al. 2008; Foster and Miller 2009; Kuechler and Simkin 2010). Both the SA and multiple-choice formats used in this study were specifically designed to target

and measure equivalent comprehension of knowledge. The finding that short answer quiz KI scores and multiple-choice quiz scores correlate strongly suggests that format itself may pose no inherent limitation. Rather it is the goal and design of the question that determines what can be elicited.

Taken together, these exploratory findings suggest that repeated testing can impact student retention of complex knowledge, and that that the benefits of testing can be realized within an authentic classroom environment using traditional methods and formats of assessment.

Limitations and Next Steps

The intervention and assessments employed in the study were very limited in scale, and the effect size of the observed differences between treatment groups was small and not as consistent as predicated. In addition, overlap of the big idea 1 & 2 topics appears to have resulted to testing effect "bleed over" across treatment groups, somewhat confounding the interpretation of post-test performance. The following experimental design and context factors assist with framing the significance and meaning of the findings. Study improvements and follow up questions are also suggested:

1) *Environment* - The experiment was conducted in a large, complex course environment, not a highly controlled lab as with the majority of related studies

(Roediger and Karpicke 2006; Carpenter, Pashler et al. 2008; Karpicke and Roediger III 2008). All assessments were part of a graded course curriculum and required independent, unmonitorable engagement by students. Numerous factors that can significantly impact student performance, such as discipline, study skills, and motivation, could not be controlled (Bransford, Brown et al. 2000; Duit and Treagust 2003; Brown 2004; Kang, McDermott et al. 2007). These factors could have considerably confounded interpretation, overwhelming detection of any treatment effect. However, findings from this study were significant and for the first time demonstrate that testing effect is relevant to real instructional environments and practice. Future work should attempt to measure other performance factors, such as discipline, determination, and approaches to study to better understand their role and interaction with practices such as repeated testing in effecting and predicting student performance and success (Lizzio, Wilson et al. 2002; Tomanek and Montplaisir 2004; Duckworth, Peterson et al. 2007).

2) *Timescale* – a significant majority of testing effect investigations utilize short duration timeframes (minutes to days) that have limited significance to the demands of a semester long course (Marsh, Roediger et al. 2007; McDaniel, Roediger et al. 2007; Karpicke and Roediger III 2008; Pyc 2010; Sensenig 2011). In the current study, final post-test assessments were carried out 5 weeks after the last treatment quiz, a timeframe similar to semester mid-term schedules or

inter-semester periods. Findings demonstrate that effects of testing can influence retention and recall on timescales parallel to authentic course requirements. Future work should investigate longer-term, more routine practices of repeated testing on retention strength and duration as well as the utility of any resulting memory advantage for improving future learning and understanding.

3) *Feedback* – Feedback strongly influences the quality and efficiency of learning in ways that can mask the influence of testing effect (Hattie and Timperley 2007; Kang, McDermott et al. 2007; Shute 2008). Nearly all testing effect investigations conducted within classroom like environments have included strong feedback components (McDaniel, Anderson et al. 2007; Agarwal, Karpicke et al. 2008; Carpenter, Pashler et al. 2009; Mayer, Stull et al. 2009; Butler 2010; Karpicke and Blunt 2011; Sensenig 2011). Therefore, it is difficult to determine if observed effects of testing are true artifacts of retrieval processes, or are a product of formative, feedback dependent learning. The experimental design in this study purposefully excluded direct feedback in treatment quizzes 2-4. In addition, measures were conducted to ensure that indirect feedback - in the form of multiple quiz attempts - did not influence performance. Thus, the observed difference can be attributed to mechanisms underlying testing effect versus feedback with greater confidence than before. Moving forward, however, some of the most important and exciting questions regarding testing

effect relate to its power to influence not just retention, but learning. Such work should return to the use of feedback, investigating how and why cycles of study, testing and feedback contribute to learning and understanding. Of particular importance today is the question of cost and efficiency, and whether such learning cycles can be mediated by peers, computers, or independently by the learners themselves.

4) *Highly complex subject material* - The knowledge targeted in this study included two of the most complex and difficult concepts covered by traditional Intro to Biology curricula; (1) the relationship between genotype and phenotype, and (2) the relationship between gene expression and cell function (Lewis, Leach et al. 2000; Lewis, Leach et al. 2000; Lewis and Wood-Robinson 2000; Lewis and Kattmann 2004; Duncan and Reiser 2007; Wood 2009). Relative to the focus of previous testing effect studies, these concepts, and the level at which student comprehension was assessed, represent a significant increase in complexity and difficulty (Crowe, Dirks et al. 2008; Wood 2009; Tibell and Rundgren 2010). Students were required to both understand and apply knowledge of numerous individual facts and ideas as well as explain the relationships that connect them into higher level systems and processes. This level of comprehension and ability has not been tested previously, and represents a significant extension in our understanding of the capacity and testing effect and multiple-choice questions to influence and measure complex knowledge. Use of more comprehensive

assessments and more frequent data analysis will permit future studies to map changes in student knowledge and understanding over time. This "microgenetic" type of approach is a powerful means of determining when and how ideas are correctly and incorrectly connected, and what role learning strategies such as repeated testing play in their construction (Schoenfeld, Smith et al. 1993; Siegler 2006).

Relating the Results to Theories of Learning and Cognition

Despite decades of research investigating the nature and boundaries of testing effect, few cognitive or biological mechanisms have been put forward to explain the underlying processes. (e.g., Glover, 1989; Karpicke & Roediger, 2008; Karpicke & Zaromb, 2010; Pyc & Rawson, 2009). The current findings provide evidence that supports current psychological and neurobiological models of memory function (Nyberg 2002; O'Reilly and Frank 2006; Shrager, Kirwan et al. 2008; Kandel 2009; Basak and Verhaeghen 2011; Hintzman 2011). These models predict and help to explain the improvements in retention outcomes observed in the current experiment, and suggest how repeated testing yields improvements in comprehension and complex understanding. The following sections will review these models in light of the current findings.

The Psychology of Testing Effect

Mastery of a skill or topic requires progressive accumulation and integration of understanding and experience (Bransford, Brown et al. 2000). The organization and breadth of the resulting knowledge determines how efficiently and effectively it is retrieved and applied (Brown 1923; Davis 2010; Hintzman 2011; Yasuda, Johnson-Venkatesh et al. 2011). Until recently, the impact of repeated retrieval on these processes was poorly characterized. But in a 2010 study, Zaromb et al. (2010) found significant correlation between testing, retention, and knowledge organization. Student participants memorized lists of words in the presence or absence of repeated testing conditions. Some students were also prompted to use mnemonic or relational information to assist their learning. Comparison of resulting learner retention and knowledge organization revealed a strong relationship between the degree of idea clustering and improved later recall. (Zaromb 2010). The findings suggest that repeated testing may encourage integration of ideas into associated clusters, and that the clustering process improves memory strength and accessibility (Zaromb and Roediger III 2009; Zaromb 2010). The relationship between MC performance, knowledge complexity, and treatment effects observed in the current study also supports the idea that knowledge structure impacts retention and can modified by repeated testing conditions in ways that strengthen comprehension. In fact, although similarities between big idea topics appeared to cause some confounding transfer of

understanding, such spill over in performance would be predicted if clustering were a key mediating process of testing effect.

Landscape model theory offers a compelling and useful description of how the mind builds higher-order knowledge structures that support complex understanding. The Landscape model was originally developed to describe the processes supporting cohesion and comprehension during reading (van den Broek and Kendeou 2008). But the model can be applied successfully to a broad range of learning situations and is a useful framework for understanding the observed enhancements to both memory retention and comprehension of complex concepts observed in the current study (van den Broek 2010).

From a cognition perspective, comprehension is the product of the construction of mental representations in memory that support meaning and sense making (van den Broek and Kremer 2000; van den Broek and Kendeou 2008; Catts 2009). These mental representations are composed of a bounded set of elements or ideas clustered together by the numerous relationships that connect them (Catts 2009). For example, a mental representation of “fishing” might include ideas about fish species, water characteristics, boats, casting, weather, previous experiences and the temporal, episodic, and procedural relationships that hold them all together. Building mental representations

depends on identifying meaningful associations between elements and previous understanding (van den Broek 2010).

This process of connection making occurs almost exclusively in short-term working memory (Basak and Verhaeghen 2011). Our most current understanding of working memory is that it is composed of three functionally unique layers. Layers 1 & 2 represent *attention*. They function as a kind of RAM, the place where information in immediate focus is retained, processed, or purged in real time (Nader and Einarsson 2010; Basak and Verhaeghen 2011). Layer 3 functions as a kind of backburner or overflow buffer where ideas tagged for processing wait passively for reactivation or storage in long-term memory (Basak and Verhaeghen 2011; Basak and Verhaeghen 2011). Layer 1 is our immediate focus, able to maintain only 1 idea at a time. Layer 2 can hold 3-4 additional ideas in an active state. Rapid shifting between layers allows layer 1 to process or connect ideas appropriately (Basak and Verhaeghen 2011). Layer 3 storage is vast but inert. Contents must be drawn into the upper layers to be modified in any way (Basak and Verhaeghen 2011).

Information progresses through layers 1, 2, & 3 as the demands of attention require. Internal standards of coherence, acquired through previous learning and reinforcement, inform working memory whether new information is *making sense (van den Broek and Kremer 2000; Basak and Verhaeghen 2011)*. If information becomes incoherent or

difficult to relate to previous knowledge, the standards will trigger stopping, refreshing, and/or analysis (e.g. stopping to read a difficult sentence or paragraph again). If information is familiar or easy to understand, it moves rapidly in and out of our attention. Landscape model describes this process of creating "big" coherence and comprehension through the small window of working memory (van den Broek 2010).

The confined structural limits of working memory funnels information into a linear and sequential stream (Baddeley 2010; Basak and Verhaeghen 2011). A few ideas in, a few ideas out. No change in external conditions (information organization, instructional quality, or study practice) can overcome this cognitive bottleneck (Mongillo, Barak et al. 2008; Baddeley 2010; Basak and Verhaeghen 2011). For example, even though students in the current experiment were supplied with models and explanations of all the new ideas and relationships they were required to understand (*see* Appendix A *for examples*), their working memory could not *understand* or even perceive the models in their complete state. The representations had to be broken down into tiny pieces, passed through the window of attention, and rebuilt in memory (Hintzman 2011; Rauchs, Feyers et al. 2011). The organization and presentations found in the study packet served to guide the resynthesis process. However, complex topics like gene expression or phenotype will always require significant time and effort to understand because the landscape of ideas is large and the window into the mind is small.

While students engaged with the study material, their working memory maintained approximately 4 active ideas at any one time (Alloway, Banner et al. 2010). Because more than 30 topics were covered, and because the moving window of attention is so small, it could be predicted that ideas presented more distant from each other in the text would be less frequently associated. For example, idea 2 would more likely be associated with idea 5 than idea 17 because ideas 2 & 5 are far more likely to populate working memory at the same time (Basak and Verhaeghen 2011). Even if the relationship between ideas 2 & 17 is obvious or intuitive, working memory may never detect the connection because of the lack of proximity (O'Reilly and Frank 2006; Mongillo, Barak et al. 2008; Alloway, Banner et al. 2010).

This pattern was clearly observed in coding responses to the short answer questions. Overall, student explanations were more likely to mention associations between ideas presented closely together (*e.g.* changes to the amino acid composition of a protein can affect its shape and/or charge) than ideas that were presented or represented more distantly (*e.g.* mutations in a gene can affect the function of the coded protein). In most cases, students that directly connected conceptually distant ideas (*e.g.* gene mutation and protein function, bacteria and chemotaxis, etc.) offered no explanation of intermediate relationships or processes, suggesting they were not able to *make sense* of them (Alloway, Banner et al. 2010). Repeated study or rereading of the same notes or chapter often fails to yield substantial new learning because ideas flow into attention in

the same repeated order, precluding novel combinations of ideas in attention (van den Broek 2010). Test questions can be used, in a sense, to re-order or re-present the learning material to help working memory overcome the physical limitations of the initial presentation. Chapter questions are an example of this mechanism in practice. Students that answer questions as they read through text consistently demonstrate improved comprehension and retention and are better able to identify relationships between topics (Hamaker 1986; Agarwal, Karpicke et al. 2008; Roediger III, Agarwal et al. 2010). For this reason, the repeated quiz questions used in the current study treatments were designed to bring more distant ideas to the attention of students. Because repeated quizzing yielded improved retention, the current study findings appear to support this conception of working memory function and the utility of testing to support student sense making processes.

Standards of Coherency

As discussed above, working memory uses "standards of coherency" to determine when current ideas have been *made sense of* and new ideas can be shifted in (Alloway, Banner et al. 2010; Baddeley 2010; Basak and Verhaeghen 2011). Different standards are applied to different information depending on context, familiarity, motivation, and difficulty (Basak and Verhaeghen 2011). Each set of standards is a learned adaptation to repeated experience, as is the awareness of what standards to apply when. Assessments powerfully shape this process, informing the learner which information is important,

what will need to be re-produced and/or applied, and thus what types of effort and learning approaches will be most rewarded (Bangert-Drowns, Kulik et al. 1991; Ram, van der Vleuten et al. 1999; Roediger and Karpicke 2006; Marsh, Roediger et al. 2007). The current study was designed to control for differences of question quality that may inform coherency standards. Both big idea 1 and 2 repeated questions were designed to elicit higher-order comprehension. Control questions, aligned to the learning goals of the course textbook, focused more on measure of discrete facts and ideas. Groups 1 & 2 each performed better on their respective final exam questions than the other, and both groups performed significantly better than the control group. Though not investigated directly, it is likely that some portion of the performance difference may be due to changes to group 1 & 2 coherency standards (Eley 1992; Gulikers, Bastiaens et al. 2006; Kember, Leung et al. 2008). Repeatedly requiring students in those treatments to use and apply their knowledge may have offered opportunity to hone standards to the demands and expectations of the questions. This may have facilitated greater sense making and comprehension, which in turn supported greater knowledge integration and retention (Zimmerman 2005; Kember, Leung et al. 2008; Basak and Verhaeghen 2011).

Relating the Results to Theories of Memory Biology

For most of the last century, memory was conceived as a type of archive or tape recorder, permanent and objective (Roediger 2008). But recent advances in neurobiology suggest that memories are evolving physical constructs subject to

constant modification and flux (Nader and Hardt 2009; Hardt, Einarsson et al. 2010; Nader and Einarsson 2010; Hoeffer, Cowansage et al. 2011) Memory formation involves changes in neuron protein expressions, activation, and distribution, formation of new synapse connections, strengthening and /or weakening of extant neural networks, and even new cell formation (Lee, Everitt et al. 2004; Hintzman 2011; Hoeffer, Cowansage et al. 2011). If one could watch a student learn the definition of "cell" or "gene" at the cell and tissue level, it would be reflected in new localized synapse network density, clustering, and overall architecture, as well as changes within the neuron cells themselves. The flow of novel information into working memory triggers hippocampal upregulation of protein synthesis and synapse restructuring (J. L. C. Lee, et al., 2004; Nakashiba, Young, McHugh, Buhl, & Tonegawa, 2008; Saxe et al., 2007). Both are necessary steps in new memory formation. Group 1 & 2 quiz questions repeatedly targeted the same knowledge sets, but varied in wording or approach over the course of the treatment. Because of this novelty, each quiz event may have stimulated the processes necessary to form new or strengthened memory traces (Nakashiba, et al., 2008; Nee & Jonides, 2010; Shrager, et al., 2008). Previous studies investigating variation of test questions on resulting testing effect strength suggest that the approach not only improves retention performance, but also influences knowledge flexibility and application (Zaromb, 2010; F. Zaromb & Roediger III, 2009)

Long-term memories are stored primarily in the cerebral cortex by a process called consolidation (Kandel 2009). Working memory, the sight of attention, is primarily localized in the hippocampus. When a memory is accessed (brought to attention in working memory), the underlying neural structure of that memory enters a transient and malleable state, similar to when it was first created (Hardt, Einarsson et al. 2010; Nader and Einarsson 2010) While in this state, the memory is open to modification, editing, strengthening, or weakening (Nader and Einarsson 2010). This process is termed reconsolidation. Connecting new information to preexisting knowledge likely involves macro-level reconsolidation processes affecting both structures . Put simply, the old networks need to change shape in order to be connected to newer networks. Thus, integrating new learning with previous knowledge is a kind of "updating" process that changes the structure of both new and old memories in order to attain the new integrated function. (Boller & Rovee-Collier 1994, Lee 2009, McDaniel & Masson 1985, Sara 2000). Group 1 & 2 treatments brought big ideas into student attention, where the structure of the underlying memories was susceptible to updating and reorganization (Hardt, Einarsson et al. 2010). By repeatedly entering this state, ideas were not only more likely to associate with each other, they were more likely to trigger recall of and associate with previous knowledge as well (Basak and Verhaeghen 2011). This cycle strengthened both the associated memory traces, as well as the overall network in which they resided. Evidence that this type of spreading activation occurs is the phenomena of hypernesia, the sudden ability to remember (accurately) previously

inaccessible memories of a complex event or idea after repeated retrieval of associated memories (Kreher, Holcomb, Goff, & Kuperberg, 2008; Nyberg, 2002).

The ability to retrieve information accurately and quickly is a function of the overall synaptic architecture from which the information is accessed (Kandel 2001; Kandel 2009). As discussed, processes in working memory strive to build coherent knowledge networks. These networks must grow complex enough to provide the new function, but no more so. Associations and activity representing irrelevant information can interfere with memory integrity and overall network function (van den Broek and Kendeou 2008). For example, many age related degenerative diseases that affect memory and cognitive function are associated with greater knowledge disorganization (Artinian, McGauran et al. 2008; Nee and Jonides 2010). To maintain clear and efficient networks, the mind selectively weakens noisy connections and strengthens useful ones. This is another form of macro-scale reconsolidation, a process that occurs and is thought to be a primary function of sleep (Walker and Stickgold 2010; Rauchs, Feyers et al. 2011). The process is guided by comparative analysis of relative memory network strengths – meaning that within localized regions, strong networks get stronger and weak networks get weaker (Rauchs, Feyers et al. 2011). Big idea questions may have informed working memory what information or associations should be prioritized for later reconsolidation and storage processes. For example, studies investigating the influence of post-learning sleep on retention found that information that is consciously prioritized during initial

learning is more likely to be consolidated and later retained. (Durrant, Taylor, Cairney, & Lewis, 2011; Rauchs, et al., 2011). Memory traces and associated networks that are strengthened in this way are more likely to be preserved during later pruning and reconsolidation processes (Rauchs, et al., 2011; Walker & Stickgold, 2010).

Implications for Education

Though the mechanisms of testing effect require significant further study, our current understanding has significant potential to inform current educational practice. For example, the majority of US undergraduate introductory biology curriculums and courses emphasize linear and sequential presentation of material (Momsen, Long et al. 2010). Course assessments tend to be few and far between, stressing memorization and recall of material in the same order and format (Momsen, Long et al. 2010). One factor driving these practices is the pervasive assumption among instructors that associations and higher-order comprehension will form spontaneously once "basic" knowledge reaches a critical mass (Cheesman, French et al. 2007; Wood 2009; Momsen, Long et al. 2010). The landscape model framework helps us see, however, that association is not spontaneous (van den Broek 2010). The process overwhelmingly occurs in the confined space of working memory, directed by effortful and conscious attention. Decades of "transfer" research demonstrates how rarely spontaneous connections form across knowledge domains in the absence of directed learning or support (Barnett and Ceci 2002; Hodkinson 2005; Hager and Hodkinson 2009; Davis 2010). Tests can be used to

encourage repeated retrieval in a variety of combinations, supporting greater knowledge clustering, reinforcement, and retention than studying alone (Lee, Everitt et al. 2004; Baddeley 2010; Hardt, Einarsson et al. 2010; Durrant, Taylor et al. 2011). Such improvements have been observed in a variety of challenging topics, including clinical diagnosis, statistics, biology, and psychology (Bangert-Drowns, Kulik et al. 1991; McDaniel, Anderson et al. 2007; Butler 2009; Roediger III and Butler 2010).

The practice of repeated quizzing in undergraduate courses can have secondary benefits as well. Time between exams in large lecture courses is often significant (Momsen, Long et al. 2010). Most students do not maintain a steady study schedule throughout the course, cramming study time into intensive sessions just prior to exams (Eley 1992; Gulikers, Bastiaens et al. 2006; Gijbels, Segers et al. 2008; Baeten, Kyndt et al. 2010). Cramming strategies tend to result in poor performance and atrocious rates of knowledge retention (Vacha and McBride 1993). The recent neurobiology evidence that informs testing effect gains also suggests why cramming strategies tend to fail. As discussed above, new learning requires creation and restructuring of new and extant synaptic networks (Kandel 2001; Mongillo, Barak et al. 2008). These processes are directed by complex cascades of cell-cell and tissue-tissue signaling and regulation, which require massive influxes and redistribution of enzymes, precursor molecules, nutrients, and ATP to function (Kandel 2009; Lee and Silva 2009). Intensive learning practices, like cramming, often don't work well simply because they exceed the physical

and biochemical limits of these pathways (Lee and Silva 2009). It's as if the lecture keeps going, but the brain's pen ran out of ink. Repeated testing encourages students to adopt more constant and consistent study-behavior (Kember, Leung et al. 2008). This not only increases the overall time students spend with material, but maintains encoding levels that do not overburden the operating capacities of a healthy hippocampus and cerebral cortex (Mongillo, Barak et al. 2008; Lee and Silva 2009; Baddeley 2010). Studies investigating various testing schedules suggest that more frequent testing can, though not always, modify student's approaches to study and learning (Eley 1992; Kember, Leung et al. 2008). Students who do switch to frequent study schedules see significant improvements in general test performance and better able to explain and apply knowledge (Eley 1992; McDaniel, Howard et al. 2009).

Multiple Choice questions

If repeated testing is to become a norm in large undergraduate classrooms, multiple-choice questions will be the most likely tool due to their ease of implementation and efficiency (Mislevy, Steinberg et al. 2003; Fellenz 2004). But a majority of instructors believe MCQs are ill suited to meet demanding learning and assessment goals (Ram, van der Vleuten et al. 1999; Haladyna, Downing et al. 2002; Fellenz 2004; Palmer and Devitt 2007). Most criticism is based on the assumption that MCQs are innately unidimensional, only able to measure the ability to recognize a correct answer from a set of options (Walstad 2001; Williams 2006; Joughin 2010). In fact, the terms multiple-

choice tests and recognition tests are often used interchangeably in education and assessment literature (Joughin 2010). Many instructors assume short answer questions that require free recall and explanation are the only valid means of probing complex understanding (Palmer and Devitt 2007; Kuechler and Simkin 2010).

But MCQs' "recognition" label is a mischaracterization. The design of a question itself determines whether considerable comprehension and careful application and analysis of knowledge is necessary, not whether answer options are offered or not (Mislevy, Steinberg et al. 2003; Suskie 2009). A metaanalysis of 67 studies comparing the relationship between student performance on MCQs and short answer questions found results were variable and inconsistent (Rodriguez 2003). The best predictor of strong correlations was the overlap of question stem design across the two formats (how related the knowledge target was in both formats). In short, MC and short answer questions that were designed to test the same type and quality of knowledge, did. The findings from the current study echo this conclusion. Whether multiple-choice or short answer, the big idea questions used in the current study were specifically designed to measure equivalent, higher-order comprehension of complex biological ideas. Application of the Knowledge Integration Rubric and comparison of MC and short-answer performance suggest the questions performed as designed (Linn, Lee et al. 2006; Liu, Lee et al. 2008).

The findings demonstrate that multiple choice questions can be designed to successfully elicit complex ideas and comprehension. Unfortunately, they are rarely used toward this purpose, as effective tools to enhance learning and retention (Momsen, Long et al. 2010). Good MCQs are difficult and time-consuming to develop and no widely accepted item-writing theories or algorithms exist to guide or hurry the process (Haladyna, Downing et al. 2002). Thus, the current resistance is understandable. Hopefully, as more investigators challenge the assumptions limiting assessment practices, more instructors will reconsider how they utilize the tools they have at hand to promote deeper, more meaningful learning.

CONCLUSION

The current exploratory study investigated the efficacy of repeated testing to enhance retention of knowledge in a large introductory biology classroom. Results support the generalized finding that repeated retrieval improves long-term retention of knowledge relative to repeated study (McDaniel, Roediger et al. 2007; Carpenter, Pashler et al. 2008; Karpicke and Roediger III 2008; Roediger III and Butler 2010). Novel to other work conducted at the undergraduate level, the current findings also suggest that repeated testing, even on a small scale, can affect student retention and understanding of sophisticated higher order understanding, a learning goal often emphasized in course syllabi but rarely assessed by course quizzes and exams (Momsen *et al*, 2010). Careful design and analysis of parallel multiple-choice and short answer questions demonstrate that each can target and elicit similar qualities and types of knowledge.

Student performance on multiple-choice questions was strongly correlated with the degree of knowledge association and integration as measured by the Knowledge Integration Construct (Zimmerman 2005; Liu, Lee et al. 2008). This finding is predicted by current cognitive and neurobiology theories of learning and memory, which suggest that repeated retrieval practices create conditions in working memory that support association formation and strengthening of memory traces in long-term memory (O'Reilly and Frank 2006; Tronson and Taylor 2007; Mongillo, Barak et al. 2008; van den Broek and Kendeou 2008; Silva, Zhou et al. 2009; Baddeley 2010; Basak and Verhaeghen

2011). These processes in turn support retention and later learning (Basak and Verhaeghen 2011). The results of the current study further support the accuracy and usefulness of these models for understanding the mechanisms underlying learning and memory. In addition, they suggest that the use of repeated testing methods may be a valuable tool for more fine grained investigation of the dynamic nature of memory function and structure.

APPENDIX A: EXPLAINING BIG IDEAS - GENE FUNCTION AND REGULATION

Explaining Big Ideas: Gene Function and Regulation

Writing clear explanations to complex questions isn't always easy, especially in biology. To help you think about how to approach it, below are examples of good student answers to the Big Idea questions. Problems like these require piecing together several smaller ideas into a larger whole, or starting with simple concepts and building to more complex ideas. The numerous details can get confusing or get in the way. To stay organized, it's always good to build a simple concept map of what you're trying to understand and explain. I've included the maps used to build the example answers on the second page.

Remember, memorizing these answers doesn't tell you if you understand the ideas. Test questions that ask you to **USE** your understanding to solve problems require that you have some kind of working model of the process in your head. A representation you can use to ask yourself how each part is affected by the others or predict what would happen if something changes. Building, understanding, and tweaking a simplified map is a good way to do that.

Big Idea #1: How genes work (what they are, what they do)

Specific Problem - What is the relationship between genotype and phenotype?
An organism's genotype includes all the genes, coded in DNA, that it inherits from its parents. Phenotype includes all the organism's observable traits such as height, hair color, or how fast or slow its metabolism is.

Genes contain specific instructions for how to build proteins. A protein is made of a chain of amino acids bond together in a specific order. Each amino acid has a unique shape and charge. Because the chain is flexible, the amino acids can interact, causing the chain to fold into a particular 3D shape depending on its sequence. The overall structure (shape and charges) of the final protein determines its function. One example is a membrane bound ion channel protein, which literally looks like and acts like a tunnel with a controllable gate.

Phenotypes like height or rate of metabolism tend to result from the functions of many many proteins working together. Sometimes an organism's phenotype depends on both the genes it has (genotype) AND environmental conditions. For example, being tall requires you have many growth encouraging genes AND proper nutrition when developing. Some phenotypes, like sex, are determined only by genotype (XX or XY chromosomes). Changes to a gene's DNA (like a mutation) can change which amino acids it codes for, resulting in changes to the protein's structure and function. If significant, that change in function might result in an observable difference in phenotype (like disease, a change in behavior, or a heritable adaptation).

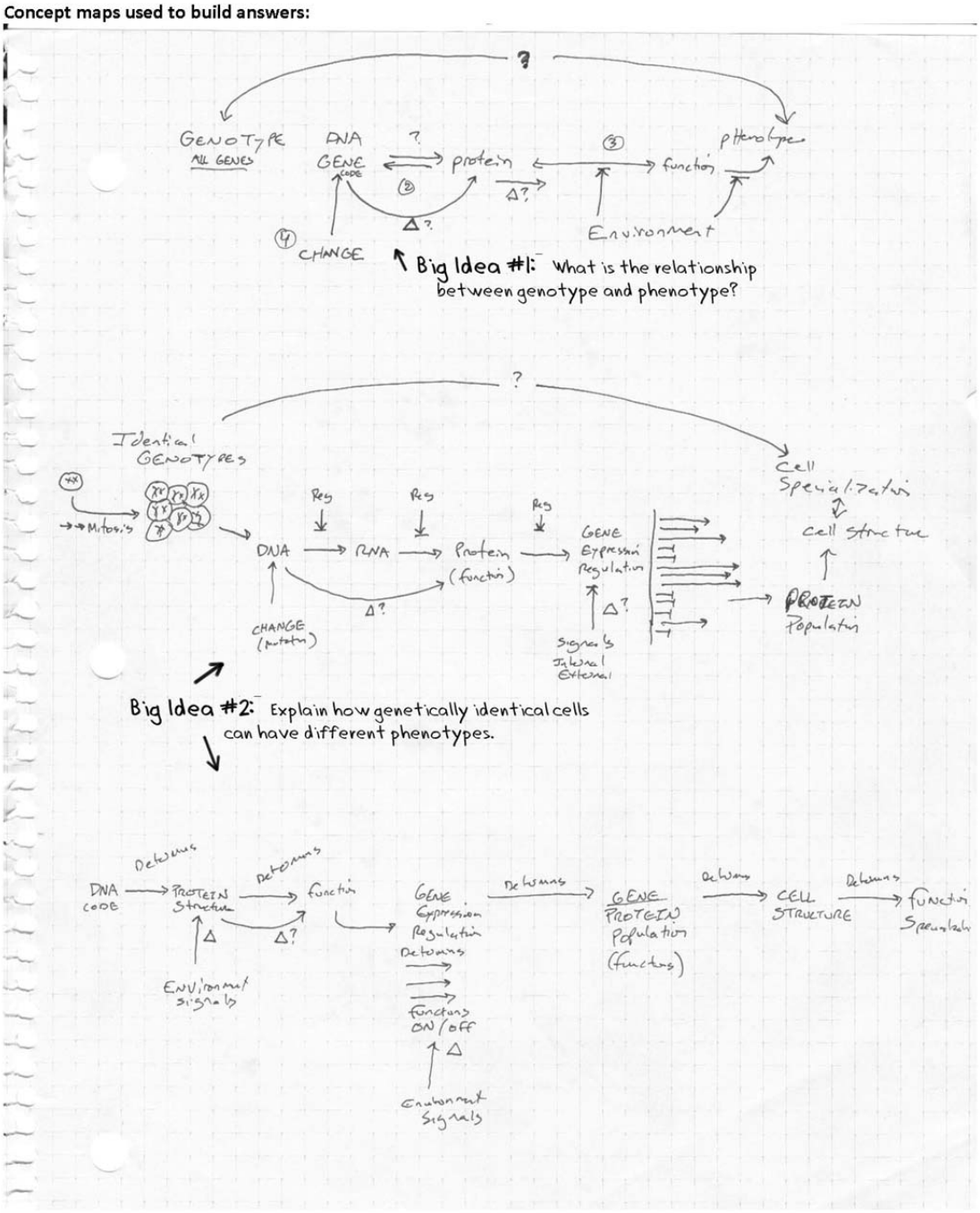
Big Idea #2: Each of your cells share the same genome. Briefly explain how one genotype create hundreds of different cell phenotypes?

All 10 trillion cells in our body come from one fertilized egg that copied and divided itself many many times by mitosis. So all the resulting cells pretty much have the same exact DNA (same genotype). Human DNA has about 25,000 genes with instructions for how to make proteins. Proteins do things like catalyze reactions, transport molecules, support cell structure, and receive and transfer information.

The phenotype of a cell (what it looks like and what it does for a living) depends on the combination of proteins it is made of and how they work together. Turning on or off (expressing) specific combinations of genes from the genome can create specific cell function (like playing particular combinations of notes can create classical music, hip hop, or unbearable nonsense). No cell expresses all of its genes at the same time (it would be like playing all notes of a church organ at once…, disturbing.) Instead, cells tend to express only certain sets of genes at certain times in their life cycle. A stomach cell expresses lots of digestive enzyme genes, while an ear cell expresses many different sets of genes to function.

Gene expression is a dynamic process that can change based on what is going on outside of the cell (chemical signals coming in from neighbor cells, temperature, or the detection of pathogen) or inside the cell (pH levels or how much ATP is available). This process is regulated by proteins that interact with DNA, either promoting or repressing transcription of genes (turning genes "On or Off"). Changes to regulatory proteins (like phosphorylating them, or binding them with an inhibitor protein when a hormone shows up) changes their function and what genes are turned on or off. Cells can change gene expression and function in other ways as well, like destroying mRNA transcript after its already made, or sticking active proteins in an organelle where they can't do anything. This active control of gene expression and function allows cells to produce all kinds of different phenotypes from just one genotype.

Concept maps used to build answers:



APPENDIX B: BIG IDEAS HANDOUT

For more info, check out references in Biological Science (3rd Edition), Freeman

Big Ideas

1: How genes work

Specifically - *What is the relationship between genotype and phenotype?*

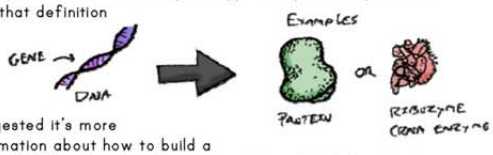
For more info on genotype and phenotype Pg. 267

An organism's **GENOTYPE** refers to its entire collection of genes, coded in DNA. Sexually reproducing organisms like yourself have 2 versions (alleles) of each gene, one from mom and one from dad.

PHENOTYPE describes all the organism's observable traits such as height, hair color, how fast or slow its heart rate is, and even if it is predisposed or resistant to particular diseases.

For more info on genes Pgs. 317 - 322

So what are Genes? We'll it depends on who you're drinking with at the time. If chillin with some fellow biologists at the post National Cell Biology Conference party (yes, biologists can party), at best you could safely say a gene is a bit of DNA that if expressed (used) by the cell, will affect its phenotype and potentially that of the larger organism somehow. Wow, I know. Try using that definition on your next test.

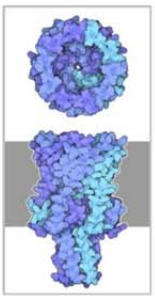


The problem is, the last 10 years of research have suggested it's more complicated than we thought. Most genes contain information about how to build a particular protein. But some don't. Some genes code for RNA's molecules that never get translated, functioning independently in the cell. And of course, genes are always coded in DNA right? Absolutely, except when they aren't. Many viruses, like HIV, use RNA instead of DNA as their information storage molecule. Does this mean those viruses don't have genes? Tell that to the cell they just invaded, forcing it to make its proteins and do its bidding. But for the sake of your sanity (and grade), we'll keep it simple for now: ***A gene is a specific segment of DNA containing information for how to make a protein.***

For more info on proteins Pgs. 44 - 65

Like DNA, proteins are polymers. They're made of chains of amino acid monomers bound together in a specific order. There are hundreds of different amino acid monomers in nature, but Life only uses around 20 to make proteins. Each amino acid has a **unique shape and charge**. Because the overall chain is flexible, the amino acids can interact, causing the chain to fold into a particular **3D shape**, depending on its sequence. The final structure (shape and charges) of the folded protein determines its function, just like how the shape and workings of a machine determine what it does.

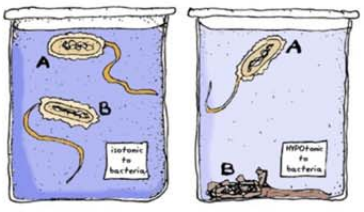
One obvious example of protein function following structure is the bacterial mechanosensitive channel of large conductance (MscL for short). MscL is a membrane protein that regulates internal pressure of a bacteria cell in emergencies. Bacteria can usually control their internal osmotic pressure by pumping ions or sugars in or out (so they don't dry out or explode). But sometimes stuff happens (like rain) that changes the external environment too rapidly to adapt this way. MscL is the emergency pressure valve, allowing small molecules like sugars and amino acids out when pressure inside gets too high. Loss of those small molecules reduces hypertonicity (higher solute concentration inside than outside), preventing water from rushing into the cell and causing it to burst. MscL widens its channel as the surrounding membrane stretches apart from internal pressure. It acts like a pressure valve because it's built like one (can you see the resemblance?).



Bacterial mechanosensitive channel protein (MscI)

Goodsell, David
www.rcsb.org

For more info on hypo & hyper tonic Pg. 106

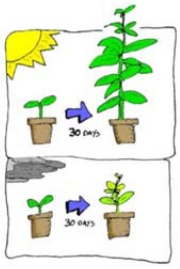


How each individual protein contributes to a cell or organism's phenotype may or may not be obvious. In biology, if you want to figure out what something does, you BREAK it and see what happens (look for a difference in the resulting phenotype). In the case of MscL, bacteria lacking that protein (bacteria B in the beakers) burst and die when dropped in a hypotonic (lower solute) solution. So MscL's contribution to phenotype is pretty clear.

But more **complex phenotypes** like height or heart rate tend to result from the functions of many many proteins working together. Height, for example, has been traced to over a hundred genes in our genome. Because of this complexity, it's hard to quantify how much each gene contributes or what exactly each does. Nonetheless, tallness or shortness runs in families (strongly heritable), so having the right combination of gene alleles is clearly important.

For more info on genes and environment Pgs. 283-285

Sometimes an organism's phenotype depends on both the genes it has (genotype) AND **environmental conditions**. Again, back to height: Being tall requires having many growth encouraging genes AND proper nutrition when developing. If a plant gets plenty of light, water, and nutrients, it will grow as tall and fast as its genes allow. But if sunlight is limited, it won't have sufficient energy to grow. If a kid eats Captain Crunch for breakfast, lunch, and dinner, his body is not getting all the necessary raw material for growth. Even with a champion set of mega-growth genes, the kid's body will lack sufficient building materials to follow through with the instructions.



Genes & Environment: These dudes clearly have a healthy collection of growth encouraging alleles, and ate well as children.

Complex phenotypes that involve hundreds of **genes** and require direct inputs from the **environment** are difficult to attribute to one or the other influence in any accurate quantitative way. However, some phenotypes, like sex, curling your tongue, or sensing certain smells, are determined only by genotype. Usually, such traits can be traced to discrete genes or structures (like Y chromosomes for dudes). If you got the gene, you got the trait. All those pea plant traits Mendel studied were discrete phenotypes, showing up when offspring had or didn't have the appropriate allele. This wasn't an accident, by the way. Mendel studied many traits that behaved in complex ways he couldn't explain using his model of inheritance. He just published the ones that made sense to him.

For more info on mutation Pgs. 347-349

Changes to a gene's DNA (like a mutation) can change which amino acids it codes for. Because each amino acid has a unique shape and charge, a change in just one amino acid can affect how the whole chain folds together, potentially resulting in changes to the protein's structure and function. If significant, that change in function might result in an observable difference in phenotype (like disease, a change in behavior, or a heritable adaptation). Check out the **sickle-cell disease example** in Freeman 347-349.

Changes to phenotype, like working out and getting very muscular, won't change the genes involved in building and regulating muscle because there is no direct information flow from proteins back to RNA or DNA. But can you think of changes to other types of phenotypes that could affect your genotype (usually in bad ways)?

Alright, so do you think you got it? Brains literally can't know they understand something until they're asked to use that new information. No quantity of rereading will help until it has a reason to rethink what's already been patched together. In fact, rereading will convince your brain it actually understands something it doesn't, because it remembers seeing the ideas before. But recognition aint understanding. So, give your brain a bit of a workout and explain the following to a study partner, or piece of paper.

Distinguish genotype and phenotype.
(What does each idea mean and how do they differ?)

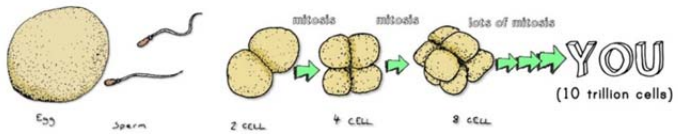
Explain the relationship between a gene and a protein.
(e.g. Is there a direct or indirect link? Do changes in one affect the other, how?)

Explain the relationship between protein function, environment, and phenotype.
(e.g. How does protein activity and function create a phenotype we can see? Would you expect factors like temperature or food availability to affect phenotype? How?)

Predict whether changes to genes or environment will affect phenotype, and vice versa.
(Can changes to genes cause changes to phenotypes? If so, how? Can changes to phenotype caused by a shift in environment cause changes to genotype?)

2: How genes are regulated (and why that matters)

Specific Problem – What explains cell specialization?
All ~10 trillion cells in your body come from one fertilized egg that divided many many times by mitosis. So all the resulting cells pretty much have the same exact DNA (same genotype).



But your body has well over a hundred different cell types carrying out thousands of different processes. If all of your cells are following the same "manual," how can there be so many differences between them?

For more info on cell structure and function Pgs. 130 - 131

If you open up a red blood cell and compare the insides to the guts of a nerve cell, you'll see very different collections of proteins. The **phenotype of a cell** (what it looks like and what it does for a living) depends on the combination of proteins it's composed of and how they work together. When biologists started studying genes to see if and how they directed differences in cells, they noticed something important. **No cell expresses all of its genes at the same time.** In fact, once most cells grow up and pick a career, they tend to express only a small portion of their total genes for the rest of their lives.

So it turns out that **Turning on or off (expressing) specific combinations of genes from the genome can create specific cell function**, similar to how particular combinations and patterns of musical notes can create classical music, hip hop, or unbearable nonsense. You can think of your genome as containing bits of information explaining how to make particular notes (genes) and when to play them (regulatory regions). The end product is a TYPE of music, or in a cell a TYPE of structure and function

REGULTION: So how does regulation work? Sometimes gene regulation functions like sheet music. Genes turn on and off in a specific predictable order over a consistent amount of time. A good example of this is in early development. As pictured above, fertilization of an egg by a sperm sets off a genetic program of repeated cell division that quickly makes one cell into tens, thousands, millions.

But the cells of a multi-cellular organism need to work together nicely, have specialized function, and respond appropriately to external conditions. So gene regulation must also be adaptive. Often, gene regulation functions like **if-then** computer logic statements. They read something like this:

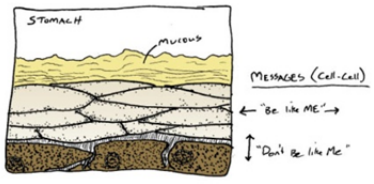
Gene 324E.Chromosome7>::

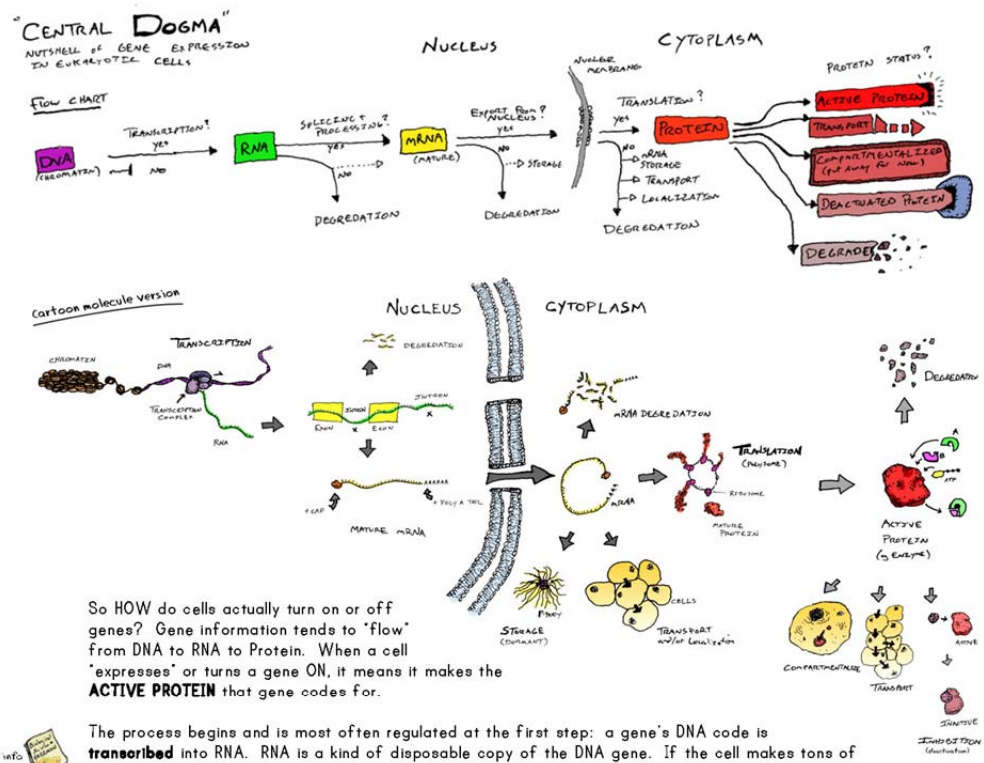
If proteins 1, 3, 5, 7, are active
AND protein 45 and 21a from neighbor cells are at high concentration
AND adrenalin is detected outside the cell
= **TURN ON**
otherwise
OFF::

For more info on signals and expression Pgs. 377 - 378

Put many of these tiny programs together, and you get complex behavior. This criteria based regulation is dynamic, controlling gene expression based on what is going on outside of the cell (e.g. chemical signals coming in from neighbor cells, temperature, or the detection of a pathogen) or inside the cell (e.g. pH levels or how much ATP is available).

Neighbor cells can influence gene expression in each other by secreting messages that direct gene regulation. For example, cells lining your stomach know to express mucous genes because they get signals from cells beside them to "be like me" and cells deeper in the tissue to "don't be like me."





So HOW do cells actually turn on or off genes? Gene information tends to "flow" from DNA to RNA to Protein. When a cell "expresses" or turns a gene ON, it means it makes the **ACTIVE PROTEIN** that gene codes for.

For more info on regulation Pgs. 379 - 384

The process begins and is most often regulated at the first step: a gene's DNA code is **transcribed** into RNA. RNA is a kind of disposable copy of the DNA gene. If the cell makes tons of RNA, many active proteins tend to be produced. This is because the instructions in each RNA can be read and **translated** into a protein hundreds or thousands of times.

But what if conditions change and the cell needs different proteins quickly? As you can see in the models above, the cell has many options. It can destroy or store particular RNA transcript already made so they won't be translated. It can deactivate or compartmentalize proteins already working in the cell so they can't do their job. Or it can destroy proteins that are preventing expression of genes now required.

Cells can take on so many different functions and respond precisely to different situations because they can regulate gene expression in so many ways. This active control of gene expression and function allows cells to produce all kinds of different phenotypes from just one genotype.

If you know it, you should be able to:

Describe the relationship between cell structure (i.e. parts and layout) and cell function.
(e.g. What makes a stomach lining cell look and function different from an earlobe cell?)

Draw out a basic gene expression pathway (e.g. Central Dogma above), including regulation steps.
(How does the cell go from DNA to a protein? If a protein function is no longer needed, what options does the cell have to shut it down?)

Connect gene expression and environment.
(No all genes in a cell are "ON" all the time. What role if any does the environment have in determining which genes are expressed?)

Connect gene expression and cell function.
(Liver cell gene expression looks WAY different from prostate cell gene expression. How does this explain their different function?)

REFERENCES

Abbott, E. E. (1909). "On the analysis of the factors of recall in the learning process." Psychological Monographs.

Agarwal, P. K., J. D. Karpicke, et al. (2008). "Examining the testing effect with open- and closed-book tests." Applied Cognitive Psychology **22**(7): 861-876.

Alexander, P. A. and R. E. Mayer (2010). "INTRODUCTION TO RESEARCH ON INSTRUCTION." Handbook of Research on Learning and Instruction: 245.

Alloway, T. P., G. E. Banner, et al. (2010). "Working memory and cognitive styles in adolescents' attainment." British Journal of Educational Psychology **80**(4): 567-581.

Artinian, J., A. M. T. McGauran, et al. (2008). "Protein degradation, as with protein synthesis, is required during not only long term spatial memory consolidation but also reconsolidation." European Journal of Neuroscience **27**(11): 3009-3019.

Baddeley, A. (2010). "Working memory." Current Biology **20**(4): R136-R140.

Baeten, M., E. Kyndt, et al. (2010). "Using student-centred learning environments to stimulate deep approaches to learning: Factors encouraging or discouraging their effectiveness." Educational Research Review **5**(3): 243-260.

Bangert-Drowns, R. L., J. A. Kulik, et al. (1991). "Effects of frequent classroom testing." The Journal of Educational Research **85**(2): 89-99.

Barnett, S. and S. Ceci (2002). "When and where do we apply what we learn? A taxonomy for far transfer." Psychological bulletin **128**(4): 612-637.

Bartlett, J. C. and E. Tulving (1974). "Effects of temporal and semantic encoding in immediate recall upon subsequent retrieval1." Journal of Verbal Learning and Verbal Behavior **13**(3): 297-309.

Basak, C. and P. Verhaeghen (2011). "Aging and Switching the Focus of Attention in Working Memory: Age Differences in Item Availability But Not in Item Accessibility." The Journals of Gerontology Series B: Psychological Sciences and Social Sciences.

Basak, C. and P. Verhaeghen (2011). "Three layers of working memory: Focus-switch costs and retrieval dynamics as revealed by the N-count task." Journal of Cognitive Psychology **23**(2): 204-219.

Bowling, B. V., E. E. Acra, et al. (2008). "Development and Evaluation of a Genetics Literacy Assessment Instrument for Undergraduates." Genetics **178**(1): 15-22.

Bransford, J., A. Brown, et al. (2000). How people learn, National Academy Press Washington, DC.

Brown, G. (2004). "How students learn." A supplement to the RoutledgeFalmer key guide for effective teaching in higher education series.

Brown, W. (1923). "To what extent is memory measured by a single recall?" Journal of Experimental Psychology **6**(5): 377-382.

Buckles, S. and J. Siegfried (2006). "Using multiple-choice questions to evaluate in-depth learning of economics." The Journal of Economic Education **37**(1): 48-57.

Butler, A. (2009). "Using repeated testing and variable encoding to promote transfer of learning." Thesis.

Butler, A. (2010). "Repeated Testing Produces Superior Transfer of Learning Relative to Repeated Studying." Learning, Memory **36**(5): 1118-1133.

Butler, A. and H. Roediger III (2007). "Testing improves long-term retention in a simulated classroom setting." European Journal of Cognitive Psychology **19**(4): 514-527.

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

Carpenter, S. K., H. Pashler, et al. (2009). "Using tests to enhance 8th grade students' retention of US history facts." Applied Cognitive Psychology **23**(6): 760-771.

Carpenter, S. K., H. Pashler, et al. (2008). "The effects of tests on learning and forgetting." Memory & cognition **36**(2): 438-448.

Catts, H. (2009). "The narrow view of reading promotes a broad view of comprehension." Language, Speech, and Hearing Services in Schools **40**(2): 178.

Cheesman, K., D. French, et al. (2007). "Is There Any Common Curriculum for Undergraduate Biology Majors in the 21st Century?" BioScience **57**(6): 516-522.

Chen, C. (2006). Prompting students' knowledge integration and ill-structured problem solving in a Web-based learning environment, THE UNIVERSITY OF OKLAHOMA.

Cheng, B. (2008). "Generation in the knowledge integration classroom."

Chiu, J. L. and M. Linn (2011). "Knowledge Integration and Wise Engineering." Journal of Pre-College Engineering Education Research (J-PEER) **1**(1): 2.

Clark, D. and M. Linn (2003). "Designing for knowledge integration: The impact of instructional time." Journal of the Learning Sciences **12**(4): 451-493.

Coles, R. (2008). The effect of tests on learning: The role of the opportunity to retrieve, VILLANOVA UNIVERSITY.

Crowe, A., C. Dirks, et al. (2008). "Biology in Bloom: Implementing Bloom's Taxonomy to Enhance Student Learning in Biology." Cell Biology Education **7**(4): 368-381.

Davis, R. A. (2010). "Psychology of learning."

DeBoer, G., H. Lee, et al. (2008). "Assessing integrated understanding of science." Designing coherent science education: Implications for curriculum, instruction, and policy: 153–182.

Draper, S. (2009). "Catalytic assessment: understanding how MCQs and EVS can foster deep learning." British Journal of Educational Technology **40**(2): 285-293.

Duckworth, A. L., C. Peterson, et al. (2007). "Grit: Perseverance and passion for long-term goals." Journal of Personality and Social Psychology **92**(6): 1087.

Duit, R. and D. Treagust (2003). "Conceptual change: a powerful framework for improving science teaching and learning." International Journal of Science Education **25**(6): 671-688.

Duncan, R. and B. Reiser (2007). "Reasoning across ontologically distinct levels: Students' understandings of molecular genetics." Journal of Research in Science Teaching **44**(7): 938-959.

Durrant, S. J., C. Taylor, et al. (2011). "Sleep-Dependent Consolidation of Statistical Learning." Neuropsychologia.

Eley, M. G. (1992). "Differential adoption of study approaches within individual students." Higher Education **23**(3): 231-254.

Fellenz, M. (2004). "Using assessment to support higher level learning: the multiple choice item development assignment." Assessment & Evaluation in Higher Education **29**(6): 703-719.

Foster, D. and H. Miller (2009). "A new format for multiple-choice testing: Discrete-Option Multiple-Choice. Results from early studies." Psychology Science Quarterly **51**(4): 355-369.

Fritz, C. O., P. E. Morris, et al. (2007). "Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning." Applied Cognitive Psychology **21**(4): 499-526.

Gijbels, D., M. Segers, et al. (2008). "Constructivist learning environments and the (im) possibility to change students' perceptions of assessment demands and approaches to learning." Instructional Science **36**(5): 431-443.

Gregory, E., J. P. Ellis, et al. (2011). "A Proposal for a Common Minimal Topic Set in Introductory Biology Courses for Majors." The American Biology Teacher **73**(1): 16-21.

Gulikers, J., T. J. Bastiaens, et al. (2006). "Relations between student perceptions of assessment authenticity, study approaches and learning outcome." Studies in Educational Evaluation **32**(4): 381-400.

Hager, P. and P. Hodkinson (2009). "Moving beyond the metaphor of transfer of learning." British Educational Research Journal **35**(4): 619-638.

Haladyna, T., S. Downing, et al. (2002). "A review of multiple-choice item-writing guidelines for classroom assessment." Applied measurement in education **15**(3): 309-333.

Hamaker, C. (1986). "The effects of adjunct questions on prose learning." Review of educational research **56**(2): 212.

Hardt, O., E. Einarsson, et al. (2010). "A bridge over troubled water: reconsolidation as a link between cognitive and neuroscientific memory research traditions." Annual Review of Psychology **61**: 141-167.

Hattie, J. and H. Timperley (2007). "The power of feedback." Review of educational research **77**(1): 81.

Hintzman, D. L. (2011). "Research Strategy in the Study of Memory: Fads, Fallacies, and the Search for the “Coordinates of Truth”." Perspectives on Psychological Science **6**(3): 253.

Hodkinson, P. (2005). "Reconceptualising the relations between college-based and workplace learning." Journal of Workplace Learning **17**(8): 521-532.

Hoeffer, C. A., K. K. Cowansage, et al. (2011). "Inhibition of the interactions between eukaryotic initiation factors 4E and 4G impairs long-term associative memory consolidation but not reconsolidation." Proceedings of the National Academy of Sciences **108**(8): 3383.

Hogan, R. M. and W. Kintsch (1971). "Differential effects of study and test trials on long-term recognition and recall1." Journal of Verbal Learning and Verbal Behavior **10**(5): 562-567.

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.

Kandel, E. R. (2001). "The molecular biology of memory storage: a dialogue between genes and synapses." Science **294**(5544): 1030.

Kandel, E. R. (2009). "The biology of memory: a forty-year perspective." The Journal of Neuroscience **29**(41): 12748.

Kang, S. H. K., K. B. McDermott, et al. (2007). "Test format and corrective feedback modify the effect of testing on long-term retention." European Journal of Cognitive Psychology **19**(4): 528-558.

Karpicke, J. and J. Blunt (2011). "Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping." Science (New York, NY).

Karpicke, J. and H. Roediger III (2008). "The critical importance of retrieval for learning." Science **319**(5865): 966.

Karpicke, J. and F. Zaromb (2010). "Retrieval mode distinguishes the testing effect from the generation effect." Journal of Memory and Language **62**(3): 227-239.

Karpicke, J. D. and H. L. Roediger (2010). "Is expanding retrieval a superior method for learning text materials?" Memory & cognition **38**(1): 116.

Kember, D., D. Y. P. Leung, et al. (2008). "A workshop activity to demonstrate that approaches to learning are influenced by the teaching and learning environment." Active Learning in Higher Education **9**(1): 43.

Klymkowsky, M., R. Gheen, et al. (2007). Avoiding Reflex Responses: Strategies for Revealing Students' Conceptual Understanding in Biology, American Institute of Physics, 2 Huntington Quadrangle, Suite 1 NO 1, Melville, NY, 11747-4502, USA.

Kohen, A. I. and P. H. Kipps (1979). "Factors determining student retention of economic knowledge after completing the principles-of-microeconomics course." The Journal of Economic Education **10**(2): 38-48.

Kornell, N., A. D. Castel, et al. (2010). "Spacing as the friend of both memory and induction in young and older adults." Psychology and Aging **25**(2): 498-503.

Kromann, C. B., M. L. Jensen, et al. (2009). "The effect of testing on skills learning." Medical Education **43**(1): 21-27.

Kuechler, W. L. and M. G. Simkin (2010). "Why Is Performance on Multiple Choice Tests and Constructed Response Tests Not More Closely Related? Theory and an Empirical Test*." Decision Sciences Journal of Innovative Education **8**(1): 55-73.

Labov, J. B., A. H. Reid, et al. (2010). "Integrated Biology and Undergraduate Science Education: A New Biology Education for the Twenty-First Century?" CBE—Life Sciences Education **9**(1): 10.

Larsen, D., A. Butler, et al. (2009). "Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial." Medical Education **43**(12): 1174-1181.

Larsen, D. P., A. C. Butler, et al. (2008). "Test-enhanced learning in medical education." Medical Education **42**(10): 959-966.

Lee, H. and O. Liu (2009). "Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective." Science Education **9999**(9999).

Lee, H. S., O. L. Liu, et al. (2011). "Validating measurement of knowledge integration in science using multiple-choice and explanation items." Applied measurement in education **24**(2): 115-136.

Lee, J. L. C., B. J. Everitt, et al. (2004). "Independent cellular processes for hippocampal memory consolidation and reconsolidation." Science **304**(5672): 839.

Lee, Y. S. and A. J. Silva (2009). "The molecular and cellular biology of enhanced cognition." Nature Reviews Neuroscience **10**(2): 126-140.

Lewis, J. and U. Kattmann (2004). "Traits, genes, particles and information: re-visiting students' understandings of genetics." International Journal of Science Education **26**(2): 195-206.

Lewis, J., J. Leach, et al. (2000). "All in the genes?-young people's understanding of the nature of genes The young people who took part in this study show only a very limited understanding of the most basic ideas relating to function, structure, and location of genes. Implications for teaching the more complex concepts are considered." Journal of Biological Education **34**(2): 74-79.

Lewis, J., J. Leach, et al. (2000). "Chromosomes: the missing link-young people's understanding of mitosis, meiosis, and fertilisation." Journal of Biological Education **34**(4): 189-199.

Lewis, J., J. Leach, et al. (2000). "What's in a cell?-young people's understanding of the genetic relationship between cells, within an individual." Journal of Biological Education **34**(3): 129-132.

Lewis, J. and C. Wood-Robinson (2000). "Genes, chromosomes, cell division and inheritance-do students see any relationship?" International Journal of Science Education **22**(2): 177-195.

Linn, M., H. Lee, et al. (2006). "Teaching and assessing knowledge integration in science." Science(Washington) **313**(5790): 1049-1050.

Liu, O., H.-S. Lee, et al. (2008). "Assessing Knowledge Integration in Science: Construct, Measures, and Evidence." Educational Assessment **13**(1): 33-55.

Liu, O., H. S. Lee, et al. (2010). "An investigation of teacher impact on student inquiry science performance using a hierarchical linear model." Journal of Research in Science Teaching **47**(7): 807-819.

Lizzio, A., K. Wilson, et al. (2002). "University students' perceptions of the learning environment and academic outcomes: implications for theory and practice." STUDIES IN HIGHER EDUCATION-OXFORD- **27**(1): 27-52.

Logan, J. M. and D. A. Balota (2008). "Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults." Neuropsychology, development, and cognition. Section B, Aging, neuropsychology and cognition **15**(3): 257-280.

Marsh, E., H. Roediger, et al. (2007). "The memorial consequences of multiple-choice testing." Psychonomic Bulletin & Review **14**(2): 194.

Mayer, R. E. (2010). "Applying the science of learning."

Mayer, R. E., A. Stull, et al. (2009). "Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes." Contemporary Educational Psychology **34**(1): 51-57.

Mazzocchi, F. (2008). "Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory." EMBO reports **9**(1): 10.

McDaniel, M., J. L. Anderson, et al. (2007). "Testing the testing effect in the classroom." European Journal of Cognitive Psychology **19**(4): 494-513.

McDaniel, M., H. Roediger, et al. (2007). "Generalizing test-enhanced learning from the laboratory to the classroom." Psychonomic Bulletin & Review **14**(2): 200.

McDaniel, M. A., D. C. Howard, et al. (2009). "The read-recite-review study strategy." Psychological Science **20**(4): 516.

Mislevy, R., L. Steinberg, et al. (2003). "Focus Article: On the Structure of Educational Assessments." Measurement: Interdisciplinary Research & Perspective **1**(1): 3-62.

Momsen, J. L., T. M. Long, et al. (2010). "Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills." Life Sciences Education **9**(4): 435.

Mongillo, G., O. Barak, et al. (2008). "Synaptic theory of working memory." Science **319**(5869): 1543.

Nader, K. and E. Einarsson (2010). "Memory reconsolidation: an update." Annals of the New York Academy of Sciences **1191**(1): 27-41.

Nader, K. and O. Hardt (2009). "A single standard for memory: the case for reconsolidation." Nature Reviews Neuroscience **10**(3): 224-234.

Nee, D. E. and J. Jonides (2010). "Dissociable contributions of prefrontal cortex and the hippocampus to short-term memory: Evidence for a 3-state model of memory." Neuroimage.

Nielsen, R., J. Buckingham, et al. (2008). A Taxonomy of Questions for Question Generation.

Nyberg, L. (2002). "Levels of processing: A view from functional brain imaging." Memory **10**(5): 345-348.

O'Reilly, R. C. and M. J. Frank (2006). "Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia." Neural Computation **18**(2): 283-328.

Palmer, E. J. and P. G. Devitt (2007). "Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions?: research paper." BMC Medical Education **7**(1): 49.

Pyc, M. (2010). Why is retrieval practice beneficial for memory? An evaluation of the mediator shift hypothesis, Kent State University.

Pyc, M. and K. Rawson (2010). "Why Testing Improves Memory: Mediator Effectiveness Hypothesis." Science **330**(6002): 335.

Pyc, M. A. and K. A. Rawson (2009). "Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory?" Journal of Memory and Language **60**(4): 437-447.

Ram, P., C. van der Vleuten, et al. (1999). "Assessment in general practice: the predictive value of written-knowledge tests and a multiple-station examination for actual medical performance in daily practice." Medical Education **33**(3): 197-203.

Rauchs, G., D. Feyers, et al. (2011). "Sleep Contributes to the Strengthening of Some Memories Over Others, Depending on Hippocampal Activity at Learning." The Journal of Neuroscience **31**(7): 2563.

Robertson, S., C. W. Canary, et al. (2010). "Factors Related to Progression and Graduation Rates for RN-to-Bachelor of Science in Nursing Programs: Searching for Realistic Benchmarks." Journal of Professional Nursing **26**(2): 99-107.

Rodriguez, M. C. (2003). "Construct Equivalence of Multiple Choice and Constructed Response Items: A Random Effects Synthesis of Correlations." Journal of Educational Measurement **40**(2): 163-184.

Roediger, H. (2008). "Relativity of Remembering: Why the Laws of Memory Vanished." Annual Review of Psychology **59**(1): 225-254.

Roediger, H. and J. Karpicke (2006). "The power of testing memory: Basic research and implications for educational practice." Perspectives on Psychological Science **1**(3): 181.

Roediger III, H. and A. Butler (2010). "The critical role of retrieval practice in long-term retention." Trends in Cognitive Sciences.

Roediger III, H. L., P. K. Agarwal, et al. (2010). "Benefits of testing memory: Best practices and boundary conditions."

Rohrer, D. and H. Pashler (2007). "Increasing retention without increasing study time." Current Directions in Psychological Science **16**(4): 183.

Rohrer, D. and H. Pashler (2010). "Recent Research on Human Learning Challenges Conventional Instructional Strategies." Educational Researcher **39**(5): 406.

Rohrer, D., K. Taylor, et al. (2010). "Tests enhance the transfer of learning." Learning, Memory **36**(1): 233-239.

Rubin, B., R. Fernandes, et al. (2010). "The effect of learning management systems on student and faculty outcomes." The Internet and Higher Education **13**(1-2): 82-83.

Sampson, V. (2006). "Two-tiered assessment." The Science Scope: 46-49.

Schoenfeld, A. H., J. P. Smith, et al. (1993). "Learning: The microgenetic analysis of one student's evolving understanding of a complex subject matter domain." Advances in instructional psychology **4**: 55-175.

Sensenig, A. (2011). Multiple choice testing and the retrieval hypothesis of the testing effect, COLORADO STATE UNIVERSITY.

Shrager, Y., C. B. Kirwan, et al. (2008). "Activity in both hippocampus and perirhinal cortex predicts the memory strength of subsequently remembered information." Neuron **59**(4): 547-553.

Shute, V. J. (2008). "Focus on formative feedback." Review of educational research **78**(1): 153.

Siegler, R. S. (2006). "Microgenetic analyses of learning." Handbook of child psychology.

Silva, A. J., Y. Zhou, et al. (2009). "Molecular and cellular approaches to memory allocation in neural circuits." Science **326**(5951): 391.

Smith, J. I. and K. Tanner (2010). "The Problem of Revealing How Students Think: Concept Inventories and Beyond." Cell Biology Education **9**(1): 1-5.

Spitzer, H. F. (1939). "Studies in retention." Journal of educational psychology **30**(9): 641.

Stigler, G. J. (1963). "Elementary economic education." The American Economic Review **53**(2): 653-659.

Stupans, I. (2006). "Multiple choice questions: Can they examine application of knowledge?" Pharmacy Education **6**(1): 59-63.

Suskie, L. (2009). Assessing student learning: A common sense guide, Jossey-Bass Inc Pub.

Tibell, L. and C. Rundgren (2010). "Educational Challenges of Molecular Life Science: Characteristics and Implications for Education and Research." Life Sciences Education **9**(1): 25.

Tomanek, D. and L. Montplaisir (2004). "Students' studying and approaches to learning in introductory biology." CBE—Life Sciences Education **3**(4): 253.

Tronson, N. C. and J. R. Taylor (2007). "Molecular mechanisms of memory reconsolidation." Nature Reviews Neuroscience **8**(4): 262-275.

Tsui, C.-Y. and D. Treagust (2010). "Evaluating Secondary Students' Scientific Reasoning in Genetics Using a Two-Tier Diagnostic Instrument." International Journal of Science Education **32**(8): 1073-1098.

Vacha, E. F. and M. J. McBride (1993). "Cramming: A barrier to student success, a way to beat the system or an effective learning strategy?" College Student Journal.

van den Broek, P. (2010). "Using Texts in Science Education: Cognitive Processes and Knowledge Representation." Science **328**(5977): 453-456.

van den Broek, P. and P. Kendeou (2008). "Cognitive processes in comprehension of science texts: The role of co-activation in confronting misconceptions." Applied Cognitive Psychology **22**(3): 335-351.

van den Broek, P. and K. E. Kremer (2000). "The mind in action: What it means to comprehend during reading." Reading for meaning: Fostering comprehension in the middle grades: 1-31.

Walker, M. P. and R. Stickgold (2010). "Overnight alchemy: sleep-dependent memory evolution." Nature Reviews Neuroscience **11**(3): 218-218.

Walstad, W. B. (2001). "Improving assessment in university economics." The Journal of Economic Education **32**(3): 281-294.

Weigold, A. (2008). The relationship between restudying and testing in the short and long term.

Wilkinson, T. and C. Frampton (2004). "Comprehensive undergraduate medical assessments improve prediction of clinical performance." Medical Education **38**(10): 1111-1116.

Willcoxson, L., J. Cotter, et al. (2011). "Beyond the first-year experience: the impact on attrition of student experiences throughout undergraduate degree studies in six diverse universities." Studies in Higher Education **99999**(1): 1-22.

Williams, J. B. (2006). "Assertion-reason multiple-choice testing as a tool for deep learning: a qualitative analysis." Assessment & Evaluation in Higher Education **31**(3): 287-301.

Wood-Robinson, C., J. Lewis, et al. (2000). "Young people's understanding of the nature of genetic information in the cells of an organism." Journal of Biological Education **35**(1): 29-36.

Wood, W. (2009). "Innovations in teaching undergraduate biology and why we need them." Annual Review of Cell and Developmental **25**: 93-112.

Yasuda, M., Erin M. Johnson-Venkatesh, et al. (2011). "Multiple Forms of Activity-Dependent Competition Refine Hippocampal Circuits In Vivo." Neuron **70**(6): 1128-1142.

Zaromb (2010). Organizational processes contribute to the testing effect in free recall.

Zaromb, F. and H. Roediger III (2009). "The Testing Effect in Free Recall is Associated With Enhanced Organizational Processes." Thesis

Zheng, A. Y., J. K. Lawhorn, et al. (2008). "Application of Bloom's Taxonomy Debunks the" MCAT Myth"." Science **319**(5862): 414.

Zimmerman, T. (2005). Promoting knowledge integration of scientific principles and environmental stewardship: Assessing an issue-based approach to teaching evolution and marine conservation, UNIVERSITY OF CALIFORNIA, BERKELEY.