CBE—Life Sciences Education
Vol. 14, 1–11, Spring 2015

Article

The Molecular Biology Capstone Assessment: A Concept Assessment for Upper-Division Molecular Biology Students

Brian A. Couch,* William B. Wood, and Jennifer K. Knight

Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309

Submitted April 23, 2014; Revised November 19, 2014; Accepted November 19, 2014
Monitoring Editor: Alison Gammie

Measuring students' conceptual understandings has become increasingly important to biology faculty members involved in evaluating and improving departmental programs. We developed the Molecular Biology Capstone Assessment (MBCA) to gauge comprehension of fundamental concepts in molecular and cell biology and the ability to apply these concepts in novel scenarios. Targeted at graduating students, the MBCA consists of 18 multiple-true/false (T/F) questions. Each question consists of a narrative stem followed by four T/F statements, which allows a more detailed assessment of student understanding than the traditional multiple-choice format. Questions were iteratively developed with extensive faculty and student feedback, including validation through faculty reviews and response validation through student interviews. The final assessment was taken online by 504 students in upper-division courses at seven institutions. Data from this administration indicate that the MBCA has acceptable levels of internal reliability ($\alpha = 0.80$) and test–retest stability ($r = 0.93$). Students achieved a wide range of scores with a 67% overall average. Performance results suggest that students have an incomplete understanding of many molecular biology concepts and continue to hold incorrect conceptions previously documented among introductory-level students. By pinpointing areas of conceptual difficulty, the MBCA can provide faculty members with guidance for improving undergraduate biology programs.

INTRODUCTION

Recent national reports have recommended that faculty and departments adopt data-driven approaches to transforming the curricula and instructional methods used in their undergraduate programs (Handelsman *et al.*, 2007; American Association for the Advancement of Science [AAAS], 2011). To accomplish this objective, science educators need instruments that measure achievement of the intended outcomes of college instruction. Among the instruments available, concept assessments have gained particular prominence within undergraduate science education communities over the past two decades (Libarkin, 2008). Concept assessments (also called concept inventories) have been developed for several science disciplines and are characterized by an explicit focus on conceptual understanding rather than factual recall. Intended to measure higher-order cognitive processes, these instruments are typically designed for easy administration to large numbers of students by using a multiple-choice (MC) question format that can be machine graded. Concept assessments have been used for many purposes, including informal monitoring of semester-to-semester student performance, formal measurement of the impacts of instructional interventions, and detection of particular incorrect conceptions among students (e.g., Hake, 1998; Marbach-Ad *et al.*, 2010; Smith and Knight, 2012).

A number of concept assessments have been developed for use in biology education (reviewed in D'Avanzo, 2008; Knight, 2010). Some of these instruments cover specific topics, such as diffusion/osmosis, natural selection, plant development, phylogenetic analyses, and host–pathogen interactions (Odom and Barrow, 1995; Anderson *et al.*, 2002;

CBE Life Sci Educ March 2, 2015 14:ar10
DOI:10.1187/cbe.14-04-0071
Address correspondence to: Jennifer K. Knight (jennifer.knight@colorado.edu).

*Present address: School of Biological Sciences, University of Nebraska, Lincoln, NE 68588.

Lin, 2004; Baum *et al.*, 2005; Marbach-Ad *et al.*, 2009). Other instruments align with the content of traditional biology courses, including several for introductory molecular/cell biology and genetics (Garvin-Doxas and Klymkowsky, 2008; Howitt *et al.*, 2008; Smith *et al.*, 2008; Shi *et al.*, 2010; Tsui and Treagust, 2010). Nearly all the existing biology concept assessments were designed for lower-division undergraduate students. As a consequence, departments currently have limited options for assessing the progress of their students beyond the required introductory core series. The lack of resources may, in part, be due to challenges associated with constructing instruments that account for the specialized course work of upper-division students. While commercial tools are available (e.g., the Educational Testing Service [ETS] Biology Major Field Test and the Graduate Record Examination Biology Subject Test), these tests are costly and provide limited feedback to faculty members on specific areas of student difficulty.

Generating questions that adequately reveal the richness and complexity of student thinking is a major challenge in concept assessment development (Smith and Tanner, 2010). Most concept assessments utilize an MC format in which incorrect options (distractors) represent commonly held incorrect student ideas (Treagust, 1988). While this format captures a preference for a certain response option, it provides little additional information on a student's thinking regarding the remaining answer choices. Previous studies have shown that biology students hold a variety of mixed models in which they have both correct and incorrect ideas regarding a particular concept. For example, students may have correct understanding of the process of natural selection, while misunderstanding underlying mechanisms by which new alleles arise within a population (Nehm and Reilly, 2007; Nehm and Schonfeld, 2008; Henson *et al.*, 2012). This situation can lead to an overly optimistic assessment of student abilities when using MC tests, because students who are rewarded for selecting the correct answer may simultaneously hold other incorrect conceptions (Frisbie and Sweeney, 1982).

This problem can be addressed through the use of an alternative question format called multiple-true/false (T/F). In this format, students are presented with a question stem, as in the MC format, but are then asked to answer "true" or "false" to a series of subsequent statements. Multiple-T/F questions are functionally similar to MC questions in which students select all answers that apply. Thus, the multiple-T/F format allows the detection of the mixed models described above, while also better approximating the difficulty of a free-response format (Kubinger and Gottschall, 2007). On a practical level, the multiple-T/F format can be machine graded and provides greater design flexibility, because questions can be written with different combinations of true and false statements.

In this article, we describe the development of the Molecular Biology Capstone Assessment (MBCA), including evidence of its validity from interviews with biology faculty and students and demonstration of its reliability from administration to more than 500 upper-division students at multiple institutions. We report student performance results, which reveal both mixed models and persistent incorrect conceptions, and discuss ways of using the MBCA to gather information about student understanding.

METHODS

Question Development

Students majoring in biology typically take a series of required lower-division courses, followed by a selection of upper-division electives covering advanced topics most relevant to their interests. Because of this typical progression, while all graduating seniors have met certain core course requirements, they likely have diverse levels of understanding of concepts that experts in the field consider important. Our main purpose in constructing the MBCA was to develop an instrument to gauge upper-division student understanding of fundamental biological concepts valued by faculty in the discipline. We envisioned such a tool as being particularly useful for assessing graduating seniors, and so our design process explicitly focused on upper-division students. We followed an iterative question-development process, incorporating multiple cycles of feedback from faculty and students (Table 1). This general approach has been previously described (Adams and Wieman, 2011) and used by others to guide the development of several concept assessments (e.g., Smith *et al.*, 2008; Shi *et al.*, 2010; Kalas *et al.*, 2013; Price *et al.*, 2014).

To establish the scope and general content of the assessment, we began by soliciting ideas from faculty members in our department (Molecular, Cellular, and Developmental Biology). We conducted individual interviews with 21 faculty members in which we asked them to list important concepts that molecular biology majors should have mastered by the time they graduate, regardless of each student's specific upper-division course work. We then held three discussions with groups of faculty members to refine this initial pool and prioritize concepts of central importance. We also informally surveyed the biology misconceptions literature and other concept assessments to identify topic areas that appeared particularly challenging for students. From this process, we compiled a list of 26 concepts and accompanying learning objectives. This list was ultimately pruned to the 18 concepts and 18 specific learning objectives that underlie the final assessment questions (Table 2). Concepts were retained based on whether they were fundamental to a molecular biology core curriculum and eliminated if they were too specific to a particular subdiscipline or taught only in select

Table 1. Overview of the MBCA development process

1. Identify a set of fundamental concepts and learning objectives through individual faculty interviews, faculty roundtable discussions, and literature review.
2. Conduct open-ended interviews to probe student understanding of these concepts.
3. Draft a series of multiple-T/F questions incorporating student ideas as statements.
4. Conduct think-aloud student interviews to ensure question clarity and response validity.[a]
5. Solicit feedback from biology faculty members at multiple institutions for approval of question content.[a]
6. Administer the assessment to upper-division students at multiple institutions.
7. Perform analyses to determine overall student performance, question statistics, and instrument reliability.

[a]Steps 4 and 5 occur concurrently and are accompanied by iterative question revision.

Table 2. Concepts and learning objectives guiding the development of MBCA questions

Question	Concept	Learning objective—Students should be able to:
1	Genetic mutations arise randomly within a population.	Explain how a specific mutation arose in a population that has undergone a change in its environment and exhibits different traits from its ancestors.
2	The differential reproductive success of individual organisms within a genetically heterogeneous population leads to changes in the genetic composition of a population over time.	Draw conclusions from graphical representations to determine how the relative reproductive success of genetically distinct individuals affects the overall genetic composition of a population.
3	Diversity arises from evolutionary processes that cause populations to become reproductively isolated and genetically distinct.	Predict the impact of different factors on the genetic composition of a newly isolated population compared with its parent population.
4	Gene expression is subject to multiple levels of regulatory control.	Distinguish among possible mechanisms for how transcription of a particular gene can vary between cell types.
5	One gene can direct the synthesis of multiple different protein products.	Distinguish among possible mechanisms for how two proteins of different apparent molecular weights can result from expression of a single gene.
6	A cell's history affects its developmental fate and response to its environment.	Distinguish among possible mechanisms that can account for two sister cells responding differently to an identical stimulus.
7	Bacteria, archaea, and eukaryotes exhibit distinct differences in cell structure and function.	Identify structural and functional characteristics of bacteria.
8	The effect of a mutation depends upon the nature of the mutation (base substitution, insertion, deletion, or DNA rearrangement) and its location within a gene.	Predict, using the codon table, how silent and nonsense mutations will affect transcription and translation.
9	A mutation that alters the translated portion of a transcript can affect the resulting protein sequence.	Determine how mutations at different locations within a gene can alter the amino acid sequence of the resulting protein.
10	The output of a signaling pathway depends on the activities of upstream components.	Predict the outcomes of inactivating various components of a known signaling pathway.
11	Chromosome partitioning during meiosis and mitosis affects the genetic identities of the resulting daughter cells.	Explain how a chromosome partitioning error at different stages of meiosis can give rise to a gamete lacking a particular chromosome.
12	Individual molecules can move through a solution in a nondirected manner as a result of thermal motion and random diffusion.	Distinguish between possible mechanisms for explaining how a molecule can travel between cells located multiple cell lengths apart.
13	Closed biochemical systems proceed toward states of lower free energy.	Evaluate the contributions of free energy and entropy changes to the folding of a protein composed of polar and nonpolar amino acids.
14	The rate at which a biochemical reaction approaches equilibrium is governed by the activation energy for that reaction.	Explain in energetic terms why reaction rate changes when the reaction is heated or when a specific enzyme is added.
15	Intermolecular interactions are governed by binding affinity and molecular concentrations.	Interpret results from a binding experiment and predict how variations to the assay would affect the results.
16	Membrane proteins and membrane-enclosed elements maintain fixed topologies as they traffic through different cellular compartments.	Predict which domains of a transmembrane protein will be accessible to the cytosol during trafficking to the cell surface, based on the original orientation of the protein within the endoplasmic reticulum membrane.
17	Genomic markers can be used to identify the molecular bases of phenotypic variation within a population.	Predict the possible locations within the genome of a mutation linked to a particular trait.
18	Genetic traits can by modulated by genetic, epigenetic, and stochastic mechanisms.	Predict possible phenotypes resulting from a cross between a woman homozygous for a recessive X-linked mutation and a man who does not carry the mutation, taking into account the epigenetic phenomenon of X inactivation.

upper-division courses. The concepts identified by faculty members embody broad biological principles; thus, each associated learning objective represents just one of many that might be formulated for a given concept.

Using the concepts and learning objectives as a guide, we developed open-ended questions addressing each learning objective and administered these questions to 7 to 12 upper-division students during individual interviews. We then drafted multiple-T/F questions, each consisting of a narrative stem (sometimes including a graph or diagram) followed by four T/F statements based on results from the student interviews. For example, question 17 asks where a single-nucleotide polymorphism (SNP) linked to a drug response could be located within the human genome. The associated T/F statements reflect student awareness that this SNP could be located within regions that directly influence drug-related signaling and metabolism as well as the incorrect conception that the SNP must directly affect the expression or activity of these genes. In writing questions, we sought to exclude jargon, use only essential biological

Table 3. Summary of MBCA faculty reviews

	Questions with given faculty agreement		
The question is:	≥ 90%	≥ 80%	< 80%
Clear and scientifically accurate	15	3	0
Aligned with the stated concept/learning goal	17	0	1
Appropriate for a graduating molecular biology major	17	1	0

terminology, and follow established item-writing guidelines, such as avoiding answer buzzwords and parenthetical expressions, writing statements of roughly the same length, and using either completely parallel or completely distinct wording structures for each statement (Frey *et al.*, 2005).

Question Revision and Construct Validity

To improve the construct validity of the assessment, we used student interviews, faculty feedback, and pilot administrations to inform the iterative revision of each question. We presented draft questions to upper-division students during semistructured interviews in which individual participants were asked to "think aloud" as they read and answered each question (Anders and Simon, 1980). We collected feedback from a total of 27 students, with each student seeing some or all of the assessment questions. In total, each question was answered by 6 to 19 students during initial interviews, followed by 8 students interviewed on the full set of assessment questions. These interviews helped reveal whether students were correctly interpreting the questions and whether their choices of "true" or "false" were based on appropriate underlying reasons.

We solicited faculty input from 25 different faculty members at different institutions to help revise and validate the questions (Table 3). In total, each question was reviewed by 10–12 faculty members during initial reviews, followed by 10 additional faculty members who reviewed the complete assessment. Faculty members were asked to respond "yes" or "no" to whether each question was clear and scientifically accurate, aligned with the stated learning objective, and was appropriate for a graduating molecular biology major.

Two initial versions of the MBCA were also piloted in an online format to upper-division students during the Fall semester of 2012. The first version was administered to 137 students at three institutions and the second version to 337 students at four institutions. Resulting data were analyzed to determine whether the questions were well targeted and had suitable psychometric properties. Feedback from each of the above processes was incorporated during the revision process, ultimately leading to the final MBCA questions (Supplemental Material 1). Following question construction, we generated descriptions of the underlying knowledge required to correctly answer each question as an aid for interpreting student outcomes (Supplemental Material 2).

The national report *Vision and Change in Undergraduate Biology Education* was published partway through the assessment development process (AAAS, 2011). This report specifies five broad core concepts for biological understanding, which have been further elaborated in subsequent articles (Brownell *et al.*, 2014). To provide users with information on how MBCA questions align with Vision and Change core concepts, we asked 10 faculty members to identify a primary and, if applicable, a secondary core concept addressed by each question. Results from this survey were tabulated by weighting primary assignments as full votes and secondary assignments as half votes (Table 4). For each question, the concept that received the most votes was designated as the primary concept, while any concept receiving three or more votes was designated as a secondary concept.

Table 4. Alignment of MBCA questions to Vision and Change core concepts

	Questions Aligned	
Core concept	Primary	Secondary
Evolution	1, 2, 3	
Structure and function	7, 15, 16	9, 11, 12, 13
Information flow, exchange, and storage	4, 5, 6, 8, 9, 10, 11, 17, 18	
Pathways and transformations of matter and energy	12, 13, 14	2, 15
Systems		6, 10

Final Administration

We administered the final version of the MBCA to students at seven different institutions during the Spring and Fall semesters of 2013 (Table 5). Because most degree programs do not have an existing mechanism to survey graduating seniors, we administered the assessment to students in upper-division courses requiring prerequisite course work in molecular biology, cell biology, and genetics.

Owing to time constraints and content overlap issues, most piloting instructors were unable to allocate class time for the assessment. Therefore, the MBCA was administered to students outside class time using the Qualtrics online survey platform. Instructors were asked to offer the assessment near the end of the semester during a week devoid of other tests or major projects. The assessment was first introduced to students through an in-class announcement made by the course instructor. Students were asked to give their best effort and were told that their participation would help the department improve its educational program (Steedle, 2010). Immediately after class, students received a follow-up email from the course instructor

Table 5. MBCA pilot institution Carnegie classifications[a]

Control	Research activity	Region	*n*
Public	RU/VH	West Coast	227
Public	RU/VH	Mountain West	100
Public	RU/H	Southeast	41
Public	Master's/L	West Coast	41
Public	RU/VH	West Coast	39
Private	RU/H	Northeast	31
Public	Master's/L	Midwest	25

[a]Institutions are ordered by participant number. All institutions offer doctoral degrees. RU = research university; VH = very high research activity; H = high research activity; Master's/L = master's-level, larger programs.

Table 6. Class standings of MBCA student participants

	Freshman	Sophomore	Junior	Senior
Students	0	8	129	365
Percent[a]	0	1.6	25.6	72.4

[a]Note that percentages do not add to 100% because two students did not report their class standing.

with a link to the online assessment, which remained open for 1 wk. Students received a small amount of extra credit (determined by the instructor) for attempting the assessment.

Altogether, the final MBCA was offered to 677 students from nine courses and attempted by 588 students. Five submissions were excluded from analyses, because they lacked answers to at least half of the T/F statements. The MBCA takes approximately 30 min to complete, and in separate in-class, paper-based administrations, we did not observe anyone finish in fewer than 15 min. To account for students who gave only cursory effort to the out-of-class administration, we removed an additional 79 submissions from students who completed the assessment in less than 15 min. The remaining 504 submissions included in the final analysis represent 74% of the students contacted to take the assessment. The last question on the assessment asked students to self-report their class standing, and the results indicated that the analyzed sample consisted almost entirely of upper-division students (Table 6).

Statistical Analyses

Various descriptive statistics were used to characterize student performance on the whole assessment as well as on individual questions and T/F statements. Each T/F statement response was scored as 1 = correct or 0 = incorrect. Students provided responses to 99.5% of all statements; nonresponses were counted as incorrect. For the "fractional" scoring method, overall scores were calculated by summing the number of correct T/F statements for each student and dividing by the total number of statements. For the "all-or-nothing" scoring method, overall scores were calculated by summing the number of questions for which a student answered all four T/F statements correctly and dividing by the total number of questions. Overall score distributions were also generated for individual courses using the fractional scoring method, and the different course means were analyzed using a one-way analysis of variance (ANOVA), followed by Tukey's multiple comparison test between all pairs of courses.

With the multiple-T/F format, item statistics can be determined both for questions and individual T/F statements. At the question level, question difficulty (P_Q) was computed by taking the average percent correct for the four T/F statements comprising the question. Each student was first given a question score based on the number of T/F statements answered correctly. For example, a student answering three of four statements correctly received a question score of 0.75. Question difficulty was then calculated as the average score for each question (note that a higher P_Q value indicates a less difficult question). Question discrimination (D_Q) reflects how well each question distinguishes between high-performing and low-performing students, as defined by overall test scores. Question discrimination was calculated using the following formula: $D_Q = P_{Q(H)} - P_{Q(L)}$, where $P_{Q(H)}$ is the question difficulty for the top third of students and $P_{Q(L)}$ is the question difficulty for the bottom third of students.

Difficulty and discrimination were also calculated at the statement level. Statement difficulty (P_S) was calculated as the fraction of correct responses for each statement. Statement discrimination (D_S) was calculated using the following formula: $D_S = P_{S(H)} - P_{S(L)}$, where $P_{S(H)}$ is the statement difficulty for the top third of students and $P_{S(L)}$ is the statement difficulty for the bottom third of students. Statement difficulties for individual questions were analyzed by one-way ANOVA, followed by post hoc Student's t tests for significance.

Additional statistical analyses were used to estimate instrument reliability, or the consistency with which an assessment measures student performance (Crocker and Algina, 2006). There are several ways of determining reliability, some of which allow calculation from a single test administration, while others require two separate test administrations. Cronbach's alpha (α), an internal reliability coefficient, reflects the internal consistency of student responses by measuring the degree of covariance between all the different items on a test. Values for α can range from 0 to 1, with high covariance between test items producing values closer to 1. Cronbach's α decreases when there is a lack of covariation between test items, such as when low-performing students outscore high-performing students on numerous questions. Cronbach's α was calculated with SPSS software based on statement scores and thus reflects covariation between questions as well as within questions (Dudley, 2006).

A second reliability measure was used, because concept assessments such as the MBCA often test a variety of different conceptual areas, and thus may not produce high internal reliabilities (Smith *et al.*, 2008). Test–retest reliability measures the degree to which a test produces consistent scores over repeat administrations. The resulting stability coefficient (r) can range from 0 to 1, with high stability between consecutive performances producing values closer to 1. The test–retest method is commonly calculated as the correlation between overall scores for a group of students taking the same assessment on two separate occasions, with the assumption that little has been learned or forgotten during the intervening time period. Because these conditions were not possible for our sample, we adopted a modified approach used previously for other concept assessments (Smith *et al.*, 2008). Students enrolled in consecutive semesters of a course are likely to perform similarly on an assessment, and thus they offer an alternative way to gauge instrument reliability. One of the courses in which we administered the MBCA was taught by the same instructor in consecutive semesters, allowing us to gauge the stability of response frequencies for two student groups with similar course backgrounds. Test–retest reliability was determined by calculating the Pearson's correlation for statement difficulties across these two semesters. Overall scores between the two semesters were compared using Student's t tests.

RESULTS

General Performance and Question Statistics

The MBCA contains 18 question stems, and each question has four accompanying T/F statements for a total of 72 statements.

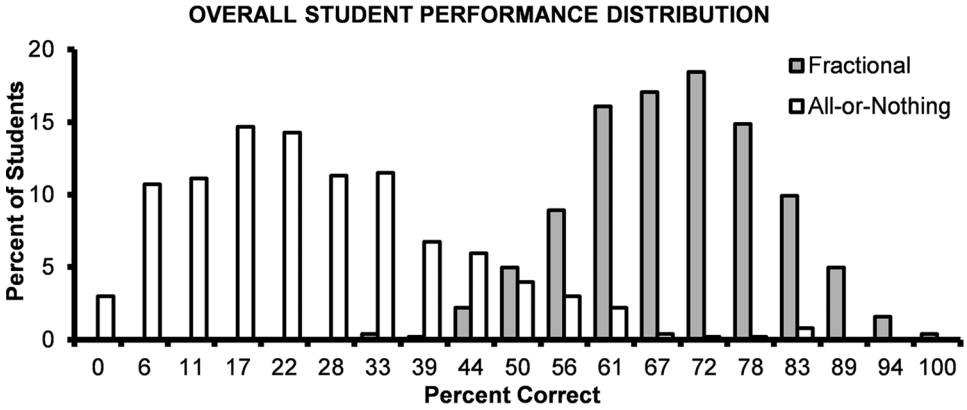

Figure 1. Frequency distribution of overall student scores. Bars represent percent of students having overall test scores within the given percent correct bins. Filled bars indicate student scores using the fractional scoring method, in which students are given credit for each correct T/F statement. Unfilled bars indicate student scores using the all-or-nothing scoring method, in which students are given credit for a question only if they answer all four accompanying T/F statements correctly. Bin labels indicate the upper threshold of each bin. For example, the rightmost bin contains scores greater than 94% and less than or equal to 100%. $n = 504$ students.

The overall performance distribution of 504 students from upper-division courses at seven different institutions is presented in Figure 1. Multiple-T/F questions can be scored in different ways (Gross, 1982; Tsai and Suen, 1993; see *Methods*: *Statistical Analyses* section). Using the fractional scoring method, the overall mean for the entire sample was 67.2 ± 11.5% SD. Using the more restrictive all-or-nothing scoring model, the overall mean was much lower at 25.9 ± 16.1% SD. Individual courses showed a wide range of scores with the fractional scoring method, with overall median scores ranging from 58 to 75% (Figure 2). A one-way ANOVA suggested that the MBCA is capable of detecting differences in the mean performance of students from different courses ($p < 0.001$).

Individual question results provide insights into student achievement levels for different learning objectives (Figure 3). The 18 questions showed a range of difficulties, from 0.77 for questions 5 and 15 to 0.55 for question 12 (large black ovals). We also calculated question discrimination values, which serve as a reflection of how well each question distinguishes high- and low-performing students as defined by overall test scores. Question discrimination values were mostly in the range of 0.2–0.4, with three questions having discrimination values below 0.2. These results indicate that, for the majority of questions, the top third of students outperformed the bottom third of students by 20–40%. The T/F statements comprising each question showed a much broader range of difficulty and discrimination values (small gray dots).

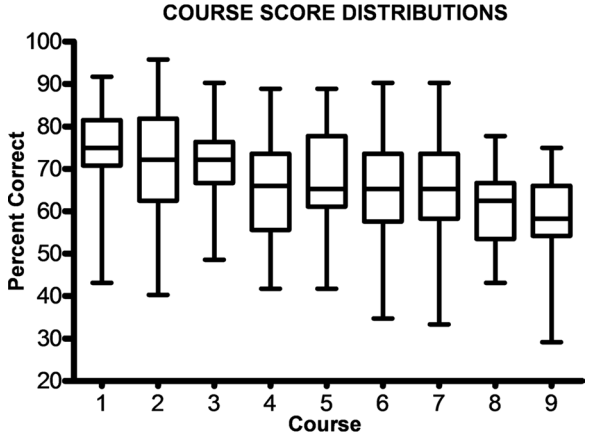

Figure 2. Overall score distributions for individual courses using the fractional scoring method. Central bars represent course median scores, boxes represent inner quartiles, and whiskers represent minimum/maximum scores. Courses are ordered by median scores. ANOVA of individual course means, $F_{(8,495)} = 11.2$, $p < 0.001$. Of the 36 pairwise course comparisons possible, Tukey's multiple comparison test revealed significant differences between 15 different pairs ($p < 0.05$). For example, Course 1 performance was significantly greater than those of Courses 4, 6, 7, 8, and 9.

One of the benefits of the multiple-T/F format is that student performance on individual T/F statements can provide diagnostic information on specific areas of student proficiency or deficiency. A complete listing of all statement difficulties is provided along with the MBCA questions in Supplemental Material 1. A subset of these results exemplifying different response patterns is shown in Figure 3. In some cases, students performed well on one statement, while struggling with the other three statements. Question 3 asks students to evaluate different ways that the genetic composition of a new population can become different from a parent population (Figure 4A). While most students recognized that natural selection can cause changes in allele frequencies within a population (statement 3C), students had difficulty recognizing the potential impacts of other phenomena, such as founder effects, inbreeding, and new mutations (statements 3A, B, and D, respectively). These results suggest that biology students may recognize that natural selection can lead to evolution but fail to grasp additional factors affecting evolution. In other cases, students performed well on three statements but struggled with the fourth. Question 13 asks students how free energy and entropy contribute to protein folding (Figure 4B). For this question, many students correctly predicted that nonpolar side chains will generally be buried within the core of the final protein, that the folding process proceeds toward a state of lower free energy, and that the entropy (disorder) of the polypeptide chain is lower in the folded state (statements 13A, B, and C, respectively). However, roughly half of the students did not indicate that the entropy of surrounding water molecules increases as

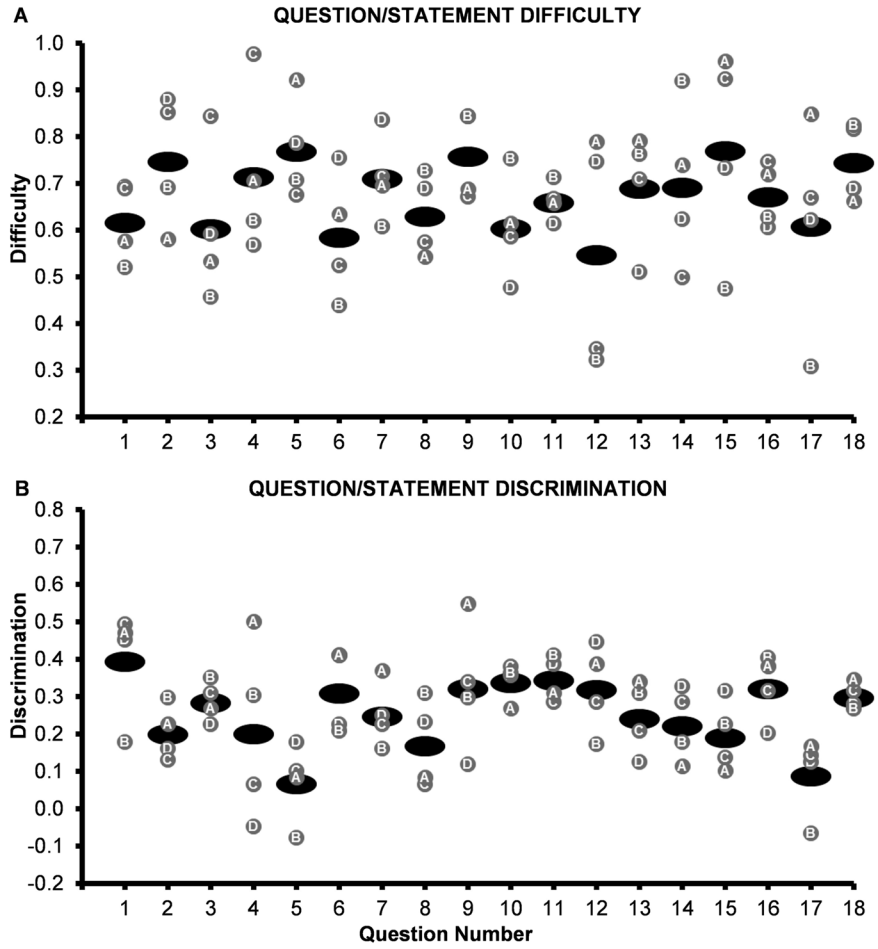

Figure 3. Question/statement difficulty and discrimination. Large black ovals represent (A) difficulty and (B) discrimination values for each question. Small gray dots represent (A) difficulty and (B) discrimination values for the four individual T/F statements comprising each question. Questions are shown in the order they appear on the assessment. Note that a higher difficulty value indicates a higher proportion of correct answers (i.e., an easier question).

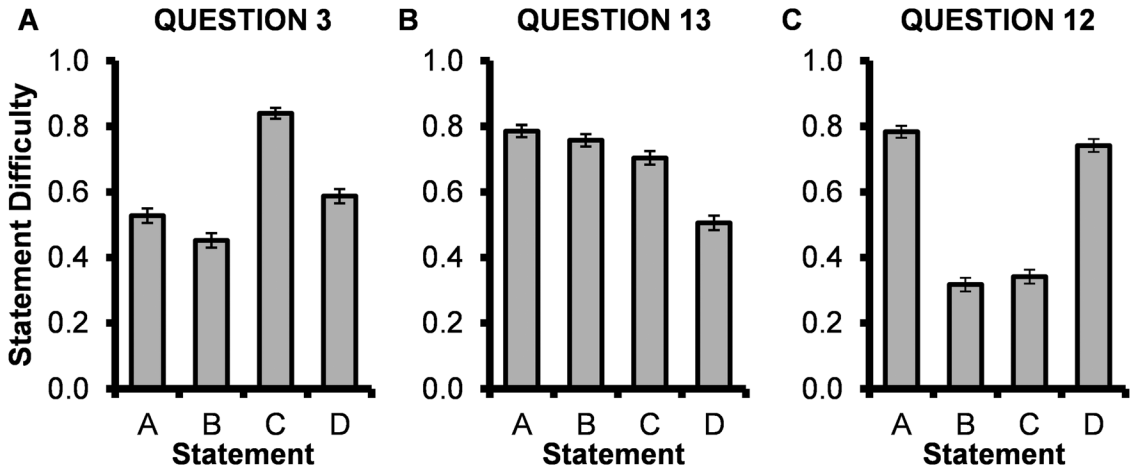

Figure 4. Individual T/F statement difficulty for three selected questions. Bars represent statement difficulties for each different T/F statement ± SEM. (A) For question 3, statement C is significantly higher than all the other statements: ANOVA, $F(3, 2012) = 64.8$, $p < 0.001$; post hoc t tests, $p < 0.001$. (B) For question 13, statement D is significantly lower than all the other statements: ANOVA, $F(3, 2012) = 39.6$, $p < 0.001$; post hoc t tests, $p < 0.001$. (C) For question 12, statements A and D are significantly higher than statements B and C: ANOVA, $F(3, 2012) = 158.1$, $p < 0.001$; post hoc t tests, $p < 0.001$.

hydrophobic side chains aggregate (statement 13D), which suggests that students are missing a critical principle underlying the behavior of hydrophobic substances in water and an important mechanism of protein folding.

Three questions on the MBCA have one true statement and three false statements, and thus they are interconvertible to an MC format. These questions provide an opportunity to demonstrate the increased ability of the multiple-T/F format to capture underlying student reasoning. Question 12 asks students how a secreted ligand travels across a fluid-filled cavity to bind to a target receptor (Figure 4C). The majority of students (74%) correctly selected "true" for statement 12D, indicating they recognized that a molecule can travel via nondirected diffusion. The remaining statements each target a different incorrect conception regarding intercellular diffusion. Interestingly, of the students who correctly answered "true" to statement 12D, 84% incorrectly identified at least one other statement as being true. This same pattern was seen for the other two interconvertible questions, for which 87% (question 8) and 68% (question 15) of students who correctly identified the true statement also incorrectly selected as true at least one of the remaining false statements.

Evidence of Test Reliability

Cronbach's α for the final MBCA administration was 0.80, which falls within a range of acceptable values (Kline, 2000). This value is on par with those of commercially available tests, such as the ETS Proficiency Profile, a general test of critical-thinking, reading, writing, and mathematics for college students (α = 0.78; Liu *et al.*, 2012). Test–retest reliability was 0.93 based on student results from consecutive offerings of the same course (Figure 5). This stability coefficient also falls within a range of acceptable values (Crocker and Algina, 2006) and is similar to those of other published concept assessments, such as the Genetics Concept Assessment (r = 0.93; Smith *et al.*, 2008). Overall student MBCA performance did not significantly differ between the two semesters of this course (semester 1 = 65.3%, semester 2 = 66.4%; $p > 0.05$).

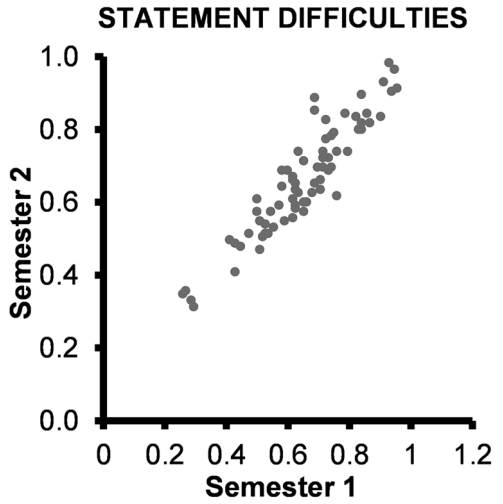

Figure 5. Test-retest reliability. Graph displays statement difficulties from two MBCA administrations in consecutive semesters of the same course. Each gray dot corresponds to the statement difficulty of one of the 72 different statements. Semester 1: n = 112 students; semester 2: n = 115 students.

DISCUSSION

Research Findings and Implications

We developed an assessment to gauge upper-division student understanding of fundamental concepts from molecular biology. The multiple-T/F format used in the MBCA helps capture additional dimensions of student thinking beyond those offered using the traditional MC format. We sought to ensure that our questions were scientifically accurate and that student responses were representative of their underlying thinking by incorporating extensive faculty and student feedback during the development process. Administration of the MBCA to more than 500 students at multiple institutions provides several insights regarding student achievement of the specified learning objectives.

Overall student performance on the MBCA raises some concerns regarding the depth of upper-division student understanding of molecular biology. With the fractional scoring method, students averaged 67%, which only modestly exceeds the 50% average that would result from random guessing (addressed in more detail in *Statistical Criteria* below). Under the more stringent all-or-nothing scoring method, students averaged 26%, indicating that students answer at least one statement incorrectly for most questions. Furthermore, there are several statements for which more than half of the students answered incorrectly (Table 7). Many students displayed incorrect conceptions regarding molecular dynamics and interactions. For example, greater than 50% of students indicated that enzymes catalyze reactions by raising the substrate free-energy level (statement 14C) and that the binding affinity between two molecules changes as a function of their concentrations (for noncooperative binding; statement 15B). Students also had trouble understanding that unequal distribution of cytoplasmic factors such as mRNAs and proteins during cell division can lead to differences in daughter cell behavior (statements 6B and C).

Several concepts that were challenging for students have been highlighted before in other publications and concept assessments. Students generally struggled with interpreting the output of a signaling pathway containing positive and negative regulatory factors (question 10). Consistent with previous reports regarding the difficulty of understanding genetic epistasis, this difficulty was most pronounced for the final statement in which two factors have been rendered nonfunctional (Knight *et al.*, 2013). In some cases, concepts assessed on the MBCA are known to be challenging for introductory students, suggesting that students taking upper-division courses have not yet mastered these concepts. On statement 1A, 43% of students incorrectly indicated that organisms can induce specific mutations to intentionally avoid predators, a naïve conception that has been documented previously (Anderson *et al.*, 2002; Shi *et al.*, 2010). In their work using the Biology Concept Inventory (BCI), Garvin-Doxas and Klymkowsky found that students frequently cite charge attractions and transport mechanisms as ways in which molecules find their cognate receptors and that student performance on this question does not improve markedly across four semesters of biology instruction (Garvin-Doxas and Klymkowsky, 2008). These results are consistent with

Table 7. Common incorrect conceptions among advanced students

Incorrect idea	Statement	% Incorrect
Inbreeding causes new alleles to occur within a population (Anderson *et al.*, 2002).	3B	55
Differences in daughter cell behavior in the early frog embryo are not due to differences in the inheritance of cytoplasmic factors, such as mRNAs[a] (Knight and Wood, 2005).	6B	57
In a linear signaling pathway, the phenotype resulting from making two factors nonfunctional is the same as when only the upstream factor is nonfunctional (Knight *et al.*, 2013).	10D	53
Charged regions on signaling ligands and their receptors can attract each other across long distances (Garvin-Doxas and Klymkowsky, 2008).	12B	68
Motor proteins actively transport signaling ligands across extracellular spaces (Garvin-Doxas and Klymkowsky, 2008).	12C	66
Binding of a substrate to an enzyme raises the free energy of the reactant molecule to the level of the transition state.	14C	51
For noncooperative interactions, binding affinity between two molecules changes as a function of their concentrations.	15B	53
SNPs associated with certain traits must directly cause those traits.	17B	70

[a]Statement 6C, which asks a similar question regarding unequal inheritance of proteins, is answered incorrectly 48% of the time.

our finding that advanced students still struggle with these concepts, as evidenced by the 68 and 66% percent of students who incorrectly answered statements 12B and C, respectively. Conversely, while few students selected random molecular motion as a viable mechanism on the BCI, students taking the MBCA correctly selected this option 74% of the time (statement 12D).

The MBCA's multiple-T/F format allows the collection of fine-grained information regarding the depth of student understanding for a given concept, while still retaining the convenience of machine grading. In many cases, students may correctly recognize one statement as true or false but lack more in-depth understanding of another aspect of this concept. For example, question 4 asks students how a gene can undergo differential expression in different cell types. Students generally indicate that this phenomenon can result from differences in the transcription factors present in each cell type (statement 4C). However, they are less likely to recognize that differences in gene expression can also be due to chemical modifications to DNA (e.g., methylation; statement 4B) or that transcription factor activity can be regulated through posttranslational modifications and interactions with other cellular factors (statement 4D). While the MC format reduces question performance to a single binary outcome, this study and others indicate that student understanding covers a much more diverse spectrum, even for individual learning objectives (Nehm and Reilly, 2007; Nehm and Schonfeld, 2008).

Statistical Criteria

In developing the MBCA, our principal goal was to produce an instrument that could provide information on what students had learned over the course of a molecular biology major. Because most concept assessments, including the MBCA, are designed to sample many different areas of student understanding, psychometric criteria that depend on consistency of student performance across an entire test should be used with caution (Adams and Wieman, 2011). While we intended for our assessment to satisfy established psychometric guidelines, we also weighed these considerations in light of the format and purpose of the assessment. For example, test developers whose sole purpose is to rank-order students based on ability levels typically seek item discrimination values greater than 0.3 for all questions (Doran, 1980). Questions failing to meet this criterion are modified or removed from the assessment. This criterion was developed for MC assessments in which the guess rate is 0.25 and the maximum theoretical discrimination value is 0.75. Because the multiple-T/F format has a guess rate of 0.5, and therefore a maximum theoretical discrimination value of 0.5, we established a working discrimination cutoff of 0.2 for each question. During our early pilot administrations, we took note of any questions or T/F statements with low discrimination values and revised them to the extent possible, while maintaining alignment to the underlying concept and learning objective.

Despite repeated rewordings, several T/F statements continued to have low discrimination values (e.g., statements 4D, 5B, and 17B), which in some cases prevented their respective questions from reaching the 0.2 discrimination threshold. Faculty and student feedback did not reveal any overt issues with the content or interpretation of these statements. Furthermore, interview responses indicated that students indeed struggle with the underlying concepts. For example, statement 17B has a difficulty of 0.30, suggesting that a majority of students do not fully understand the concept of genetic linkage. This confusion could arise from curricula that do not explicitly connect the concept of genetic linkage to genomic applications. These particular statements also share the common feature of requiring careful thought in order to be answered correctly. Thus, it is also possible that low discrimination values are the product of a low-stakes testing environment, in which high-performing students may select an obvious answer without more sophisticated consideration. Because these statements provided valuable information, they were retained in the final version of the assessment. A Cronbach's α of 0.80 suggests the MBCA has an acceptable level of internal reliability, despite the presence of several T/F statements with low discrimination values.

One potential drawback of the multiple-T/F format is the 0.5 guess rate for each T/F statement, which has the potential to limit the range of overall test scores and to influence interpretation of individual statement results. Other researchers have found, however, that this high guess rate is offset by students being able to answer nearly four times as many individual items (T/F statements) compared with MC questions in a similar amount of time. This increase in

the number of separate test items allows for much greater content sampling and increased test reliability (Frisbie and Sweeney, 1982). With regard to the interpretation of MBCA scores, it would have been difficult to achieve high internal reliability if a substantial fraction of students were guessing. In addition, there are several statements, including some on questions later in the assessment, with difficulties above 0.9 or below 0.4, extremes that are statistically unlikely under conditions of significant guessing.

Administration Format

Unlike most concept assessments, which target specific topics or courses, the MBCA integrates concepts from multiple courses and thus ideally should be administered at the departmental level as students are graduating. Unfortunately, most biology departments lack suitable mechanisms for administering an "exit" assessment. For the development process, we elected to administer the assessment through specific upper-division courses, because this provided a feasible way to reach the appropriate student group. Departments interested in assessing students as they leave the major will need to develop additional approaches to recruiting students in an unbiased manner, particularly as calls for student learning outcomes increase in coming years (Middaugh, 2009).

In addition to recruitment, both assessment delivery and student participation incentives must be considered. Because the MBCA's content is not intended to be aligned with a particular course, we administered the assessment in an online format outside class time. Faculty members gave students participation credit as an incentive for completing the assessment. However, there was no external incentive for students to put forward their best effort. We requested that faculty members not award points for correct answers for several reasons: 1) this could be considered unfair, since the specific MBCA content was not covered in that particular course; 2) this would likely promote the utilization of outside resources (e.g., textbooks or websites) and would compromise valid interpretations of student understanding; 3) this would increase the likelihood that assessment answers would be posted to the Internet. While we achieved a 74% participation rate under these low-stakes conditions, several reports suggest that overall student scores may improve under higher-stakes conditions (Wolf and Smith, 1995; Sundre, 1999; Wise and DeMars, 2005; Liu *et al.*, 2012). Depending on their reasons for the administration, departments should consider adopting other incentives, such as rewarding high achievement, that may motivate students to perform at a higher level.

Future Directions

Student performance on the MBCA also raises further questions regarding the extent and trajectory of student learning during a major. At what point do students learn particular concepts, and do they retain these concepts over subsequent years? We are currently working with collaborators at the University of Maine, Arizona State University, and the University of Washington to develop ways to use assessments to monitor learning as students progress through a biology major (NSF DUE-1322364). We anticipate that these efforts will provide key insights into how student understanding changes over time and will help departments structure their programs to better promote student learning and retention of biological concepts.

In addition to the conceptual understanding that is the focus of the MBCA, a successful molecular biology program should aspire to achieve other outcomes, such as improving science process skills and attitudes toward science (Semsar *et al.*, 2011; Gormally *et al.*, 2012; Dirks *et al.*, 2013). Thus, the MBCA represents just one of the many ways in which departments might gauge student outcomes. The ongoing development and use of diverse instruments will allow departments to construct more accurate portraits of student progress and to reflect on how their programs might change to optimize student achievement.

Availability of the MBCA

An administrable version of the MBCA questions (along with an answer key) is available upon request. Potential users should note that the reported test characteristics were collected under the specific conditions reported here and that student performance may change based on assessment location, delivery format, and student incentives. Interested instructors should email their requests to J.K.K., including a link to an institutional Web page where their instructor status can be verified.

ACKNOWLEDGMENTS

This work was supported by the University of Colorado Boulder through the University of Colorado Science Education Initiative (SEI) and a Chancellor's Award for Excellence in STEM Education to J.K.K. This work was also supported by a National Science Foundation Transforming Undergraduate Education in Science, Technology, Engineering, and Mathematics type 2 award (DUE-1322364) to B.A.C., J.K.K., and others. We are grateful to Caleb Trujillo and Sarah Wise, also supported by the SEI, for important research contributions during the early project phases; to Travis Lund, Anjali Krishnan, and Matthew McQueen for advice on statistical analyses; and to Michelle Smith, Sara Brownell, Alison Crowe, and Scott Freeman for critical research discussions. We also thank the faculty members who participated in discussions of learning objectives, reviewed assessment questions, or administered the assessment, and the students who participated in interviews or completed a pilot assessment. This research was classified as exempt from institutional review board review (protocols 0603.081, 0108.9, and 12-0336).

REFERENCES

Adams WK, Wieman CE (2011). Development and validation of instruments to measure learning of expert-like thinking. Int J Sci Educ *33*, 1289–1312.

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC.

Anders K, Simon HA (1980). Verbal reports as data. Psychol Rev *87*, 215–251.

Anderson DL, Fisher KM, Norman GJ (2002). Development and evaluation of the conceptual inventory of natural selection. J Res Sci Teach *39*, 952–978.

Baum DA, Smith SD, Donovan SSS (2005). The tree-thinking challenge. Science *310*, 979–980.

Brownell SE, Freeman S, Wenderoth MP, Crowe AJ (2014). BioCore Guide: a tool for interpreting the core concepts of vision and change for biology majors. CBE Life Sci Educ *13*, 200–211.

Crocker L, Algina J (2006). Introduction to Classical and Modern Test Theory, Mason, OH: Cengage Learning.

D'Avanzo C (2008). Biology concept inventories: overview, status, and next steps. BioScience *58*, 1079–1085.

Dirks C, Leroy C, Wenderoth MP (2013). Science Process and Reasoning Skills Test (SPARST): development and early diagnostic results. Presented at the Society for the Advancement of Biology Education Research (SABER) annual meeting, held 11–14 July 2013, in Minneapolis, MN.

Doran RL (1980). Basic Measurement and Evaluation of Science Instruction, Washington, DC: National Science Teachers Association.

Dudley A (2006). Multiple dichotomous-scored items in second language testing: investigating the multiple true–false item type under norm-referenced conditions. Lang Test *23*, 198–228.

Frey BB, Petersen S, Edwards LM, Pedrotti JT, Peyton V (2005). Item-writing rules: collective wisdom. Teach Teach Educ *21*, 357–364.

Frisbie DA, Sweeney DC (1982). The relative merits of multiple true-false achievement tests. J Educ Meas *19*, 29–35.

Garvin-Doxas K, Klymkowsky MW (2008). Understanding randomness and its impact on student learning: lessons learned from building the Biology Concept Inventory (BCI). CBE Life Sci Educ *7*, 227–233.

Gormally C, Brickman P, Lutz M (2012). Developing a test of scientific literacy skills (TOSLS): measuring undergraduates' evaluation of scientific information and arguments. CBE Life Sci Educ *11*, 364–377.

Gross LJ (1982). Scoring multiple true/false tests: some considerations. Eval Health Prof *5*, 459–468.

Hake RR (1998). Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am J Phys *66*, 64–74.

Handelsman J, Miller S, Pfund C (2007). Scientific Teaching, New York: Freeman.

Henson K, Cooper MM, Klymkowsky MW (2012). Turning randomness into meaning at the molecular level using Muller's morphs. Biol Open *1*, 405–410.

Howitt S, Anderson T, Costa M, Hamilton S, Wright T (2008). A concept inventory for molecular life sciences: how will it help your teaching practice? Aust Biochem *39*, 14–17.

Kalas P, O'Neill A, Pollock C, Birol G (2013). Development of a meiosis concept inventory. CBE Life Sci Educ *12*, 655–664.

Kline P (2000). Handbook of Psychological Testing, New York: Routledge.

Knight JK (2010). Biology concept assessment tools: design and use. Microbiol Aust *31*, 5–8.

Knight JK, Wood WB (2005). Teaching more by lecturing less. Cell Biol Educ *4*, 298–310.

Knight JK, Wood WB, Smith MK (2013). What's downstream? A set of classroom exercises to help students understand recessive epistasis. J Microbiol Biol Educ *14*, 197–205.

Kubinger KD, Gottschall CH (2007). Item difficulty of multiple choice tests dependent on different item response formats—an experiment in fundamental research on psychological assessment. Psychol Sci *49*, 361.

Libarkin J (2008). Concept Inventories in Higher Education Science, Washington, DC: National Research Council.

Lin S-W (2004). Development and application of a two-tier diagnostic test for high school students' understanding of flowering plant growth and development. Int J Sci Math Educ *2*, 175–199.

Liu OL, Bridgeman B, Adler RM (2012). Measuring learning outcomes in higher education: motivation matters. Educ Res *41*, 352–362.

Marbach-Ad G, Briken V, El-Sayed NM, Frauwirth K, Fredericksen B, Hutcheson S, Gao L-Y, Joseph SW, Lee V, McIver KS, *et al.* (2009). Assessing student understanding of host pathogen interactions using a concept inventory. J Microbiol Biol Educ *10*, 43–50.

Marbach-Ad G, McAdams KC, Benson S, Briken V, Cathcart L, Chase M, El-Sayed NM, Frauwirth K, Fredericksen B, Joseph SW, *et al.* (2010). A model for using a concept inventory as a tool for students' assessment and faculty professional development. CBE Life Sci Educ *9*, 408–416.

Middaugh MF (2009). Planning and Assessment in Higher Education: Demonstrating Institutional Effectiveness, Hoboken, NJ: Jossey-Bass.

Nehm RH, Reilly L (2007). Biology majors' knowledge and misconceptions of natural selection. BioScience *57*, 263–272.

Nehm RH, Schonfeld IS (2008). Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. J Res Sci Teach *45*, 1131–1160.

Odom AL, Barrow LH (1995). Development and application of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction. J Res Sci Teach *32*, 45–61.

Price RM, Andrews TC, McElhinny TL, Mead LS, Abraham JK, Thanukos A, Perez KE (2014). The Genetic Drift Inventory: a tool for measuring what advanced undergraduates have mastered about genetic drift. CBE Life Sci Educ *13*, 65–75.

Semsar K, Knight JK, Birol G, Smith MK (2011). The Colorado Learning Attitudes about Science Survey (CLASS) for use in biology. CBE Life Sci Educ *10*, 268–278.

Shi J, Wood WB, Martin JM, Guild NA, Vicens Q, Knight JK (2010). A diagnostic assessment for introductory molecular and cell biology. CBE Life Sci Educ *9*, 453–461.

Smith JI, Tanner K (2010). The problem of revealing how students think: concept inventories and beyond. CBE Life Sci Educ *9*, 1–5.

Smith MK, Knight JK (2012). Using the Genetics Concept Assessment to document persistent conceptual difficulties in undergraduate genetics courses. Genetics *191*, 21–32.

Smith MK, Wood WB, Knight JK (2008). The Genetics Concept Assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci Educ *7*, 422–430.

Steedle JT (2010). Incentives, motivation, and performance on a low-stakes test of college learning. Paper presented at the American Educational Research Association annual meeting, held 30 April–3 May 2010, in Boulder, CO.

Sundre DL (1999). Does examinee motivation moderate the relationship between test consequences and test performance? Paper presented at the American Educational Research Association annual meeting, held 19–23 April 1999, in Montreal, Quebec, Canada.

Treagust DF (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. Int J Sci Educ *10*, 159–169.

Tsai F-J, Suen HK (1993). A brief report on a comparison of six scoring methods for multiple true-false items. Educ Psychol Meas *53*, 399–404.

Tsui C, Treagust D (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. Int J Sci Educ *32*, 1073–1098.

Wise SL, DeMars CE (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. Educ Assess *10*, 1–17.

Wolf LF, Smith JK (1995). The consequence of consequence: motivation, anxiety, and test performance. Appl Meas Educ *8*, 227–242.