Development and Validation of the Conceptual Assessment of Natural Selection (CANS)

Steven T. Kalinowski,[†]* Mary J. Leonard,[‡] and Mark L. Taper[†]
[†]Department of Ecology and [‡]Department of Education, Montana State University, Bozeman, MT 59717

ABSTRACT

We developed and validated the Conceptual Assessment of Natural Selection (CANS), a multiple-choice test designed to assess how well college students understand the central principles of natural selection. The expert panel that reviewed the CANS concluded its questions were relevant to natural selection and generally did a good job sampling the specific concepts they were intended to assess. Student interviews confirmed questions on the CANS provided accurate reflections of how students think about natural selection. And, finally, statistical analysis of student responses using item response theory showed that the CANS did a very good job of estimating how well students understood natural selection. The empirical reliability of the CANS was substantially higher than the Force Concept Inventory, a highly regarded test in physics that has a similar purpose.

INTRODUCTION

One of the most striking features of life on Earth is how well-suited organisms are for the environments they inhabit. Polar bears have white fur that provides camouflage on ice; lions have claws and teeth that allow them to catch prey; and anteaters have long tongues that help them feed on ants living underground. Such adaptations are familiar to everyone but historically were difficult for scholars to explain. Until the 19th century, the prevailing explanation for these traits was that they were given to organisms by a designer (e.g., Paley, 1802). In the early 19th century, Lamarck (1809) provided an alternative explanation that proved incorrect, and then in the mid-19th century, Darwin (1859) developed the explanation that is accepted by biologists today: species are adapted to their environments because individuals possessing traits most suited to those environments are more likely to survive and pass those traits to their offspring. This is the core concept of natural selection and one of the most important ideas in biology (Dobzhansky, 1973).

One of the remarkable features of natural selection is how simple it is. Coyne (2009, p. xvi) described it as "staggeringly" simple. Chown (2013) called it "breathtakingly" simple. And when Huxley first heard it, he famously remarked, "How extremely stupid [of me] not to have thought of that!" (Huxley, 1887). This simplicity suggests that natural selection should be easy to teach. Some faculty seem to believe this. For example, at our university, an award-winning biology professor recently remarked that natural selection "takes only thirty seconds to teach"!

Decades of research show this opinion could not be more wrong: natural selection is one of the most difficult topics in biology to teach (e.g., Bishop and Anderson, 1990; Nehm and Reilly, 2007; Andrews *et al.*, 2011). It may be tempting to blame religion, but this does not seem to be the case (Bishop and Anderson, 1990; Demastes *et al.*, 1995; Ingram and Nelson, 2006). Usually, the main challenge instructors face while teaching natural selection is student misconceptions (for a review, see Gregory, 2009). By "misconception," we mean an understanding or explanation for a scientific concept that differs from what is known to be scientifically correct (National Research Council,

Ross Nehm, *Monitoring Editor*
Submitted June 24, 2015; Revised June 1, 2016; Accepted June 3, 2016
CBE Life Sci Educ December 1, 2016 15:ar64
DOI:10.1187/cbe.15-06-0134
*Address correspondence to: Steven T. Kalinowski (skalinowski@montana.edu).

2012, p. 58). Students have diverse misconceptions about natural selection, but most are variations of the belief that evolution occurs because individual organisms adapt during their lifetimes (Brumby, 1984; Bishop and Anderson, 1990; Bardapurkar, 2008; Nehm and Schonfeld, 2008). Students believe organisms change because they need to change, because organisms use or do not use body parts, or because organisms are affected by their environment—and that these changes cause evolution. Such misconceptions are remarkably resistant to instruction. Even when instructors use lessons carefully designed to help students restructure their understanding of natural selection, many students retain at least some of their misconceptions (Nehm and Reilly, 2007; Andrews *et al.*, 2011; see also Table 1).

Because misconceptions regarding natural selection are common and resistant to instruction, instructors need to carefully monitor how students think about natural selection. This is not easy for instructors to do: students think about natural selection in complex ways. For example, a student may understand how cheetahs evolved to run fast but not how whales lost their hind limbs (Andrews *et al.*, 2011; Nehm and Ha, 2011). Or a student might understand selection in plants but not animals. Such misunderstandings will not be evident unless a student is asked the right question. Instructors and researchers, therefore, need carefully designed instruments to assess how students understand natural selection.

We believe that an instrument seeking to accurately assess how well students understand the basic process of natural selection should have at least three characteristics. First, the instrument should be able to detect as many misconceptions as possible (Nehm and Schonfeld, 2008). We will not review all the misconceptions regarding natural selection that have been documented (see Gregory, 2009), but they are many—and some have received little study. Second, a well-designed instrument should include questions relating to as many evolutionary contexts as possible. By "context," we mean a biological feature of a question that influences the way students think about it in a way that is different from how an expert would think about it. For example, students often explain the evolutionary origin of new traits differently than the loss of traits (Nehm and Ha, 2011).

TABLE 1. Questions on the CANS organized by concept and form

	Proportion correct		Factor loading	
Question number. Species: Description of question	**Pre**	**Post**	**Within**	**Across**
Evolution				
1. Anteater: Trait gain (long tongue)	0.68	0.94	0.76	0.66
15. Saguaro: Trait gain (long roots)	0.59	0.87	0.77	0.79
21. Mosquito: Trait gain (pesticide resistance)	0.46	0.76	0.89	0.98
7. Anteater: Trait loss (teeth)	0.21	0.62	0.75	0.75
24. Mosquito: Trait loss (pesticide resistance)	0.24	0.51	0.77	0.77
10. Bowhead: Describe evolutionary changes (thick skull)[a]	0.76	0.96	0.78	0.73
9. Bowhead: Role of environment in evolution (blubber)	0.68	0.96	0.87	0.84
17. Saguaro: Role of individual change in evolution (waxy skin)	0.17	0.51	0.95	0.88
Mutation				
5. Anteater: Origin of beneficial trait (claws)	0.48	0.74	0.93	0.93
19. Saguaro: Origin of beneficial trait (spines)	0.42	0.68	0.98	0.95
11. Bowhead: Properties of mutations	0.27	0.60	0.90	0.89
23. Mosquitoes: Properties of mutations	0.22	0.57	0.93	0.93
Inheritance				
3. Anteater: What traits are inherited?[a]	0.47	0.53	0.34[b]	0.80
8. Anteater: Effect of use on next generation (tongue)	0.56	0.90	0.99	0.85
14. Bowhead: Comparing parents to offspring (skull)	0.38	0.72	0.60	0.65
22. Mosquitoes: Effect of environment on next generation	0.20	0.57	0.85	0.86
Selection				
2. Anteater: Environmental stress (food shortage)	0.66	0.78	0.97	0.89
16. Saguaro: Environmental stress (drought)	0.81	0.93	0.98	0.73
6. Anteater: Competition in an ideal environment	0.27	0.40	0.24[b]	0.19[b]
12. Bowhead: Exponential growth in empty habitat (graph)	0.28	0.48	0.84	0.27[b]
20. Saguaro: Role of chance in evolution (seed dispersal)	0.65	0.73	0.99	0.09[b]
Variation				
4. Anteater: What traits vary?	0.53	0.69	0.50	0.24[b]
13. Bowhead: Cause of variation	0.72	0.81	0.56	0.05[b]
18. Saguaro: Describe variation[a]	0.64	0.82	0.49[b]	0.24[b]

Columns "Pre" and "Post" show the proportion of students pre- and postinstruction who answered the question correctly. The last two columns specify factor loadings in IRT analyses. "Within" refers to factor loadings restricted to each set of questions (Figure 4). "Across" refers to factor loadings for IRT analyses when 24 all questions were included in a one-dimensional IRT analysis (Figure 5).

[a]Adapted from the CINS.

[b]Factor loading < 0.5.

There has not been a lot of research on what evolutionary contexts students view as important, but we assume there are many. Third, a well-designed instrument should assess important concepts using questions that have a variety of forms (see Spiro *et al.*, 1988). By "form," we mean a type, structure, or format of a question for assessing a concept (examples are given within the text). This should help ensure that results are not strongly influenced by the way questions are constructed.

A few instruments are currently available to assess how students think about natural selection (Bishop and Anderson, 1990; Anderson *et al.*, 2002; Nehm *et al.*, 2012). The two most influential are the CINS (Conceptual Inventory of Natural Selection; Anderson *et al.*, 2002) and the ACORNS (Assessing Contextual Reasoning about Natural Selection; Nehm *et al.*, 2012). The CINS is a 20-question, multiple-choice test designed to assess 10 concepts related to natural selection. The distractors for these questions were crafted to appeal to students having a variety of misconceptions. ACORNS questions are open response and all have the same form. Here is an example: "How would biologists explain how a living bed bug species with resistance to a pesticide evolved from an ancestral bed bug species that lacked resistance to the same pesticide?" Student responses are graded according to how students incorporate six concepts in their response and whether students show signs of three common misconceptions. Instructors can select from many taxa and traits to make questions.

The CINS and ACORNS have been valuable for studying how students think about evolution and are useful for the purposes for which they were designed. However, they do not satisfy the three criteria we described above for thoroughly assessing how students think about the core principles of natural selection. The CINS does not do this, because it was designed to be broad rather than deep. It assesses 10 concepts, which means it has only two questions per concept. Furthermore, the CINS was not designed to assess evolutionary reasoning in multiple contexts. In addition, there is very little variation in the form among the questions. The ACORNS shares this last limitation. All of its questions have the same form.

We have developed a new instrument, the CANS (Conceptual Assessment of Natural Selection; see the Supplemental Material), to assess how well students in introductory college biology courses understand the basic process of natural selection. The CANS is a 24-question, multiple-choice instrument that assesses five concepts related to natural selection: variation, selection, inheritance, mutation, and how these concepts work together to cause evolution (Table 1). The CANS was designed to assess core concepts that are most closely related to understanding natural selection. Related topics such as extinction or speciation were not included. Nor does the CANS include questions on advanced topics such as sexual selection, evolutionary constraints to adaptation, or the molecular basis of evolution. An important feature of the CANS is that it asks students to explain the evolutionary origins of traits that are obvious adaptations—but that students might attribute to non-Darwinian causes of evolution associated with common misconceptions. For example, the CANS asks students to explain how anteaters evolved long tongues because students might plausibly believe anteaters evolved long tongues by constantly stretching their tongues into anthills.

The goal of the work presented here was to evaluate the validity of the CANS. That is to say, we sought to assess how well the CANS measures the extent to which introductory biology students understand the basic process of natural selection. We followed commonly accepted standards to collect validity evidence (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, cited hereafter as AERA *et al.*, 2014). This included review by an expert panel, student interviews, and statistical analysis of student responses. As we present it below, this body of evidence supports the validity of the CANS as an assessment of how individual introductory biology students think about core concepts relating to natural selection.

AN OVERVIEW OF THE CANS

The CANS (see the Supplemental Material) assesses how students think about five concepts related to natural selection: variation, selection, inheritance, mutation, and evolution. Defining these concepts in a manner that creates a meaningful taxonomy of natural selection is not easy. Each of these concepts has multiple components or is related to other concepts in ways that are difficult to represent (Figure 1). Therefore, it is probably most productive to think of these concepts as constellations of ideas that are related in complex ways. We dealt with this complexity by writing "core" definitions (discussed in the following paragraphs) for each concept, and then identifying other concepts students should know in order to reason effectively about natural selection (Figure 1). Once we did this, we wrote questions for the CANS that attempted to assess the network of concepts in a reasonable way.

The CANS has a relatively simple structure (Table 1). It is divided into four parts; each part asks questions relating to a different species: anteaters, bowhead whales, saguaro cacti, and mosquitoes (Table 1). These four species were chosen because they offered opportunities to ask students questions relating to a variety of evolutionary contexts that might influence student thinking. Anteaters were chosen because they provide an opportunity to expose misconceptions regarding use and disuse. Bowhead whales were chosen to expose misconceptions regarding the effect of the environment. Saguaro cacti provide an opportunity to explore student thinking about plants. And, finally, mosquitoes provide an opportunity to explore how students think about resistance to disease or pesticides.

Three questions on the CANS assess variation. We defined variation as the concept that the phenotypes of individuals for most traits vary in populations. Here is a question that assesses this in a direct manner (correct answer underlined):

> 4. A biologist captures ten healthy, adult male anteaters and compares them to each other. Which of the following traits are likely to be different among the anteaters?
>
> a. The length of the femur ("thigh") bone.
> b. The rate at which their livers break down toxins naturally found in ants.
> c. The stickiness of the saliva on the tongues of anteaters.
> d. Two of the above.
> e. <u>a, b, and c</u>

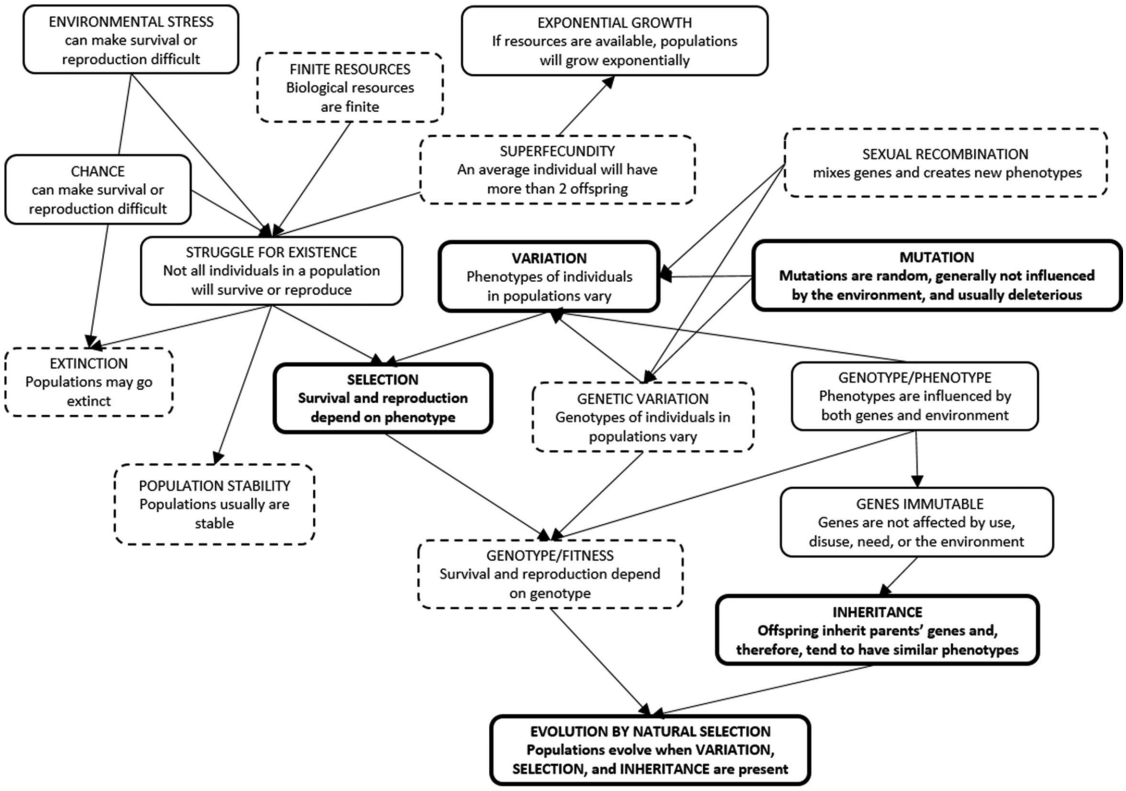

FIGURE 1. The concept map of natural selection used to guide development of the CANS. Concepts in boxes with a bold border are core concepts. Concepts in boxes with a dashed line are not assessed on the CANS.

The other two variation questions on the CANS assess this concept but also incorporate additional concepts. Question 13 assesses whether students understand that the phenotypes of individuals vary and that this variation is influenced by both genes and the environment. Question 18 assesses whether students recognize that phenotypic variation among individuals affects the ability of individuals to survive and reproduce.

Five questions on the CANS assess how students think about selection. We defined selection as the concept that the survival and reproduction of individuals is often affected by their phenotype. Question 16 assesses this concept in the most direct manner:

> 16. During the time period saguaro cacti were evolving to their current form, there were years with very little rain. What likely happened to the saguaro cacti during the driest years?
>
> a. The saguaro cacti managed to obtain the water they needed.
> b. Saguaro cacti with the shortest roots died.
> c. The saguaro cacti survived with less water than normal.
> d. The saguaro cacti grew longer roots.

The CANS also includes questions relating to competition (question 6), the role of chance in evolution (question 20), and exponential growth (question 12). These concepts do not appear to be as closely connected to the core concepts of natural selection as other concepts on the CANS (see the *Results* section), so there is only one question on the CANS for each of these concepts.

Four questions on the CANS assess inheritance. We defined inheritance as the concept that offspring inherit genes from their parents and therefore tend to be phenotypically similar. The inheritance questions on the CANS include three forms. The simplest of the inheritance questions assesses whether students recognize offspring tend to have phenotypes similar to their parents. Here is the question:

> 14. Consider a baby whale born during the time the ancestors of modern bowhead whales were evolving thicker skulls. When this whale grows up, how will its skull compare to the skulls of its parents?
>
> a. When the baby whale grows up, its skull will probably be slightly thicker than the skulls of its parents.
> b. When the baby whale grows up, its skull will probably be similar to its parents.
> c. When the baby whale grows up, its skull will probably be slightly thinner than the skulls of its parents.
> d. Skull development is affected by many factors, so we cannot predict how this whale's skull will grow.

Two of the inheritance questions assess whether students understand that genes are not changed by use/disuse or by the environment. Here is one of those questions:

> 22. Consider a female mosquito that was exposed to DDT during the years a population was evolving resistance to DDT. She survives and later lays a cluster of eggs. How will her exposure to DDT likely affect her offspring?

a. Her exposure to DDT will give her offspring increased resistance to DDT.
b. Her exposure to DDT will have no effect on her offspring.
c. The effect on her offspring of her exposure to DDT cannot be predicted.

Finally, the inheritance questions on the CANS includes a question (3) that assesses whether students realize that reproducing organisms pass on genetic traits regardless of whether the trait is beneficial or harmful.

Four questions on the CANS relate to mutation. We defined mutation as the concept that genetic changes to individuals are caused by random genetic mistakes, are generally are not influenced by the environment, and are usually deleterious. Two mutation questions ask how new traits originated. Here is one of these questions:

19. The ancestors of modern saguaro cacti did not have long and sharp spines. Consider the first ancestor of saguaro cacti to grow spines that were as long and sharp as the spines on saguaro cacti living in Arizona today. Why did this cactus grow such sharp spines?

a. It was fortunate a genetic mistake gave it extra sharp spines.
b. The cactus needed sharper spines to stop animals from eating it.
c. Animals chewing on the cactus caused it to grow sharper spines.
d. Mutations changed the DNA of this cacti because it was injured by an animal.
e. The hot climate caused this change.

The other two mutation questions explicitly asked students what was true about mutations and included distractors that incorporated a variety of misconceptions. Here is one of those questions:

23. What was most likely true regarding the genetic mutations that occurred during the years mosquitoes were evolving resistance to DDT?

a. The number and effect of mutations that occurred was not influenced by DDT.
b. Most of the mutations that occurred helped the mosquitoes survive.
c. The number of mutations occurring in the population increased when DDT was first applied, and then decreased when the mosquitoes finished adapting.
d. The mutations occurred because mosquitoes needed to survive.

We defined evolution as the concept that evolution by natural selection is caused by interaction between variation, selection, inheritance, and mutation. The CANS includes eight questions that assess whether students understand evolution. Five of these questions ask students to describe what caused species to evolve adaptations. Three of these five questions relate to trait gain and two relate to trait loss. Here is one of the trait gain questions:

1. Which of the following is the best description of how anteaters evolved long tongues?

a. Anteaters grew long tongues because they needed to reach inside ant hills.
b. Anteaters grew long tongues because they constantly stretched their tongues.
c. Random mutations occurred because anteaters needed to change.
d. Each year, anteaters with the longest tongues were most likely to live.
e. Changes like this depend on many factors, so it is impossible to answer.

The CANS also includes three other evolution questions. One of these (9), asks students about the role the environment plays in evolution, and one (17) asks students what role individual responses to the environment (like suntanning) play in evolution. Finally, question 10 asks students to describe evolutionary change.

METHODS

We used standard methods for developing and validating the CANS (Adams and Weiman, 2011; AERA *et al.*, 2014). For the purpose of this paper, we will describe the process as having two distinct stages. In the first stage, we used a variety of iterative approaches to develop questions and put together a complete version of the CANS. In the second stage, we conducted a more formal validation of the CANS. All aspects of this work were reviewed and approved by the Institutional Review Board for human subjects research at Montana State University (MSU).

We will not present data from the initial development phase of CANS but will briefly mention six types of data we found helpful. We began the development of many questions by asking students open-ended questions about natural selection concepts. This was necessary for concepts that have not received much study (e.g., variation, mutation). Second, we drafted multiple-choice questions and asked students to select an answer and explain in writing why they selected the answer they did. This provided us with valuable insights on how students interpreted questions and helped us refine and eliminate questions that were confusing or not accurately capturing student thinking. Third, we wrote alternative versions of some questions and administered them to large numbers of students to test whether students answered each version differently. Fourth, we administered early versions of the CANS to students in advanced biology courses (who, presumably, had a strong understanding of natural selection). This helped identify questions that were confusing. Fifth, we performed pre- and postinstruction testing to identify questions that were not answered more successfully after instruction—and, therefore, might be assessing something different from we intended. Finally, we used item response theory (IRT) to evaluate how well sets of questions on the CANS assessed the same concept.

After we developed a draft of the CANS that appeared to work well, we examined its validity using four sources of evidence (Adams and Weiman, 2011; AERA *et al.*, 2014). First, we recruited five experts in evolution education to review the CANS and to evaluate whether the questions were relevant to the concepts being tested and provided good sampling coverage of those concepts. Second, we interviewed 20 students to verify that their responses to questions on the CANS provided an accurate reflection of how they thought about the concept being tested. Third, we compared test scores on the CANS before and after instruction. Fourth, and last, we used IRT to examine the structure and reliability of the CANS.

Expert Review

Our expert review panel, hereafter called the "expert panel," consisted of five prominent education researchers with research experience directly relevant to constructing the CANS. We asked each panel member to evaluate two aspects of the CANS: 1) how well questions on the CANS sampled (or covered) the concepts of variation, selection, inheritance, evolution, and mutation; and 2) how relevant each question was to the concept it assessed. In addition, we asked reviewers for comments on the clarity of the questions and any additional comments they could provide. We used responses from the panel to make modest revisions to the CANS before conducting the next step of student interviews.

Demographics

We used student interviews and in-classroom testing for several aspects of our investigation. Students participating in this research were enrolled in an introductory biology course (BIOB 170) at MSU. The course is designed for biology majors and covers three main topics: evolution, organismal diversity, and ecology. The evolution section of the course was taught by S.T.K. The course meets twice a week for 75-minute lectures and has a weekly laboratory. There are no prerequisites for BIOB 170, and 90% of the students in the course had not taken the companion introductory biology course at MSU, which covers biochemistry, genetics, and cell biology. Enrollment in the course Fall semester 2015 was 262. Of these students, 54% were female and 93% were Caucasian. The largest minority population in the course was Native American (2%). Eighty-four percent of the students were planning on majoring in science, mathematics, or engineering.

Student Interviews

The second step in our formal validation of the CANS was to interview 20 students. Each interview lasted ~30 min and students were paid $20 for their participation. The goal of each interview was to determine whether each question on the CANS provided an accurate depiction of how students thought about natural selection. We used a semistructured interview protocol to examine how students were interpreting questions (Crowl, 1996). During the interview, students were asked to read CANS questions silently to themselves, select an answer, and then explain to the interviewer (S.T.K.) why he or she selected that answer. The interviewer then asked at least one follow-up question to verify whether the student's answer to the question was an accurate reflection of his or her thinking. If the student answered the question correctly, the interviewer selected another answer and asked why he or she did not select one of the incorrect answers. If the student did not answer the question correctly, the interviewer asked why he or she did not select the correct item. As needed, the interviewer asked additional follow-up questions to explore how well the student's thinking was reflected by the answer he or she chose. For example, if the student used a scientific term such as "mutation" or "fitness" the interviewer might ask what that term meant to the student.

Learning Gains

The third step in our validation process was to compare scores on the CANS before and after instruction relating to natural selection. The CANS was designed to measure how well students understand natural selection, so scores should go up after instruction relating to natural selection. We administered the CANS to students enrolled in BIOB 170 Fall semester 2015 in the lecture before natural selection was introduced. Students then received approximately 6 hours of instruction relating natural selection. This instruction incorporated several active-learning activities designed to help students correct misconceptions (see Kalinowski *et al.*, 2013). After instruction, students answered all 24 CANS questions on an exam. We used the normalized gain statistic of Hake (1998) to measure learning gain for the class and a Wilcoxon signed-rank test to assess the statistical significance of the gain observed.

IRT

The fourth step in our validation process was to use IRT to examine the psychometric properties of the CANS. This analysis had two general goals. First, we examined the internal structure of the CANS, and second, we estimated how well the CANS measures students' understanding of natural selection.

A brief summary of IRT may be helpful for interpreting the results we present. IRT, also known as modern test theory, is a set of statistical models that relate the probability of a student answering a question correctly with how well the student understands the concept being tested (for an introduction, see Ayala, 2008). There are many variations of IRT models; we used the three-parameter (3PL) model in most of our analysis. This model, which assumes the probability that the *i*th student answers the *j*th question correctly, is

$$P_{ij} = g_j + \left(1 - g_j\right)\frac{\exp\left[\alpha_j\left(\theta_i - \delta_j\right)\right]}{1 + \exp\left[\alpha_j\left(\theta_i - \delta_j\right)\right]} \quad (1)$$

The most important parameter in Eq. 1 is θ_i. θ_i is a measure of how well the *i*th student understands the concept being tested. θ is usually referred to as "ability." Ability is a latent trait and cannot be directly observed: it must be estimated from how students answer questions. As Eq. 1 shows, students with high ability are expected to have a higher probability of answering a question correctly than students with low ability. θ is measured in standard deviations from the mean. Therefore, if student ability in a classroom is normally distributed, 95% of students will have an ability greater than –2 and less than 2. The parameter δ_j measures the "difficulty" of the *j*th question. Students are expected to have a lower probability of answering a difficult question than an easy question. Negative values of δ_j indicate a question is less difficult than average; positive values of δ_j indicate a question is more difficult than average. The parameter α_j is called the "discrimination" of a question. This parameter indicates the slope of the logistic-like curve specified by Eq. 1. This is an important parameter, because the slope of the curve is proportional to how much information a question has for estimating ability. Questions with steep slopes are more useful for estimating ability than questions with low slopes (see Figure 2 for examples). If a question has a very steep slope, virtually all students with an ability lower than the threshold will answer the question wrong, and all students with an ability higher than the threshold will answer the question correctly. Alternatively, if a question has a slope of zero (no discrimination), it provides no information: all students are expected to have the same probability of answering the question correctly. This would be

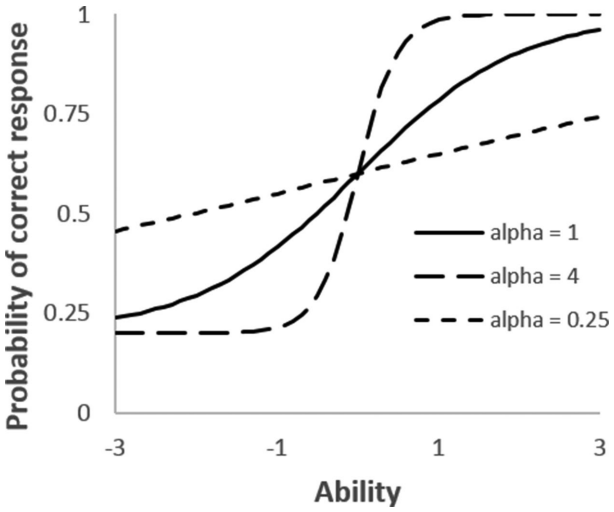

FIGURE 2. Example of IRT traces illustrating different values of α (discrimination).

expected if the question is unrelated to the ability. Values of α can be converted into loadings for each question in a one-parameter factor analysis. Such loadings quantify how much variation in student scores is explained by the latent ability of students. The last parameter in Eq. 1, g_j, is the "guessing" rate (or "pseudoguessing" rate) of a question. This is the probability that a low-ability student will answer a question correctly. If a student with low ability randomly selects an answer to a multiple-choice question from among k options, g will equal $1/k$. The guessing rate can be higher than this if low-ability students exclude some potential answers, and can be lower than this if students are strongly attracted to one or more wrong answer. If the guess rate is estimated for each question, Eq. 1 is a 3PL model. Alternatively, g can be set to $1/k$, in which case the model has two parameters.

Reliability is a useful measure of how precisely a test measures the concept it is assessing, and is defined and calculated differently in IRT than in classical test theory (e.g., Kline, 2005). Reliability is usually defined as the proportion of variance in test scores in a class that can be attributed to real variance in ability (as opposed to sampling error). Cronbach's alpha is a common statistic for estimating reliability. In IRT, the primary method for describing the reliability of a test is an information curve. Information in this context is the reciprocal of the sampling variance of estimates of ability. High values of information indicate that a test is able to estimate ability well. Unlike Cronbach's alpha, which is a single number, information curves depict how well a test estimates ability across a range of abilities. This is useful because most tests are informative only for a certain range of ability. The CANS, for example, was designed for estimating the ability of college students in introductory biology courses and should identify which students in these courses have a basic understanding of natural selection and which do not. The CANS would not be useful for differentiating between students having more advanced understandings of natural selection. Roughly speaking, Cronbach's alpha summarizes a test's information curve with a single value. This representation of a test has less information than an information curve but can be convenient for comparing tests. The IRT equivalent to alpha is empirical reliability (also called "person separation reliability"; see Eq. 6 in Adams, 2005).

Our discussion of IRT models has thus far assumed student responses to a question are a function of a single (one-dimensional) ability. If ability has multiple dimensions, multidimensional IRT models are available. Fitting multidimensional IRT models to test score data requires more data (questions and students) than we have available. We are not going to do much with multidimensional IRT models in this investigation but would like to point out what frequently happens when a unidimensional model is fitted to data that have multiple dimensions. If the multidimensionality is not extreme or abilities are correlated, a unidimensional model may fit the data well. On the other hand, if student responses to some questions are a function of one ability and other questions tap other abilities, and IRT analysis is unidimensional, some questions may load strongly on the ability estimated by IRT and the others may not (this will be seen in our data below.

We used the R statistical package mirt, version 1.8 (Chalmers, 2012), for all of our IRT analysis. Parameters were estimated using maximum likelihood and the expectation-maximization (EM) algorithm. With one exception, we used the 3PL model. Unless otherwise indicated, all analyses refer to student responses on the CANS before instruction.

The first goal of our IRT analysis was to examine the internal structure of the CANS. The goal of this type of analysis is typically to confirm that questions on a test are related to each other in the manner predicted by theory or the principles used to design the test (AERA *et al.*, 2014; Rios and Wells, 2014). This, unfortunately, is difficult for us to do with the CANS because there is little theory available to predict how the 24 questions on the CANS should be related. The CANS contains questions assessing five concepts: evolution, mutation, inheritance, selection, and variation. It is not known whether these five concepts are all components of a single construct (natural selection) or whether these are five independent concepts. This uncertainty makes it impossible to check whether the CANS, as a complete instrument, has the internal structure it should have. Despite this ambiguity, we can make some a priori statements about the structure of the CANS. All the questions relating to each topic on the CANS are intended to assess a single concept (i.e., all the mutation questions are intended to assess whether students understand mutation). Questions within each topic, therefore, should estimate the same ability. We checked this by fitting a 3PL IRT model to each set of questions and estimating its loadings in a factor analysis. There are only three variation questions on the CANS. This is not enough questions to estimate three parameters. Therefore, we used a two-parameter IRT model for the variation questions. In this analysis, we assumed the guessing rate was equal to $1/k$.

The next question we addressed with IRT analysis was whether all the questions on the CANS could reasonably be modeled as assessing a single construct. We did this by fitting a one-dimensional 3PL IRT model to our data and examining model parameters and item fit. Model fit was assessed with Pearson's chi-squared statistic (Orlando and Thissen, 2000). We used IRT to calculate item and test information curves and to estimate the empirical reliability of the test. Finally, we calculated Cronbach's alpha for the CANS to facilitate comparisons with other instruments for which this statistic

has been presented. We used the psych statistical package (version 1.6.4) for R to do this.

RESULTS

The expert panel conducted a thorough review of the CANS. In addition to rating the questions according to the criteria we provided (sampling coverage and relevance), all panelists provided extensive comments on the CANS.

The first task assigned to the expert panel was to rate how well the CANS sampled the concepts of variation, selection, inheritance, evolution, and mutation on a scale of 1 (not at all adequately) to 5 (very adequately). The panel rated the evolution questions highest (average = 4.3), followed by mutation (4.0), variation (3.7), inheritance (3.4), and selection (2.6). The selection component of the CANS was faulted for not adequately assessing some advanced topics: trade-offs between traits, and pleiotrophy. There was a wide range of opinion among the panelists regarding how well questions on the CANS sampled concepts. Two panelists gave the CANS an overall average sampling score of 4.6 (out of 5), while two others gave the CANS an average score of 2.0.

The second task assigned to the expert panel was to assess how relevant each question on the CANS was to the concepts it was supposed to assess (i.e., variation, selection, inheritance, evolution, and mutation). Panelists generally gave high relevancy scores to questions on the CANS: the average score for relevance was 4.2 (out of 5). However, again, there was notable variation among the panelists. Five of the 24 questions on the CANS received a relevancy rating of 5 from one member of the panel and 1 from another member.

Student interviews consistently showed that CANS questions provided accurate reflections of how students thought about evolution. Aside from one exception (noted below), students seemed to interpret questions on the CANS in the manner intended. This is not to say the questions were flawless. The interviews did reveal multiple minor editorial ways to improve the clarity of the questions. These changes generally involved replacing a word or inserting a qualifier into a sentence. There was one exception to this. We had to rewrite question 20 a few times until we were confident students were interpreting the question as we intended.

Two hundred and twenty-three students in BIOB 170 completed the CANS before instruction and 266 students completed the CANS on the exam after instruction. Two hundred and eighteen students completed both tests. All the testing data presented here is from these 218 students.

Scores on the CANS were higher in BIOB 170 after instruction (Table 1 and Figure 3). Before instruction, the average score on the CANS was 47%. After instruction, the class average was 71%. This corresponds to a normalized gain of 0.45. The median score on the CANS increased from 11 to 18 ($p < 0.0001$).

IRT analysis showed that almost all of the questions relating to each of the five concepts covered by the CANS seemed to measure the same concept (Table 1, "Within" column). For example, all eight of the evolution questions loaded heavily on a single ability (presumably evolution by natural selection). This can be seen by the relatively steep slopes of the curves in the item characteristic curves for these questions (Figure 4A). A couple of questions did not load well. Question 6 had the lowest loading. This was a selection question. This low loading

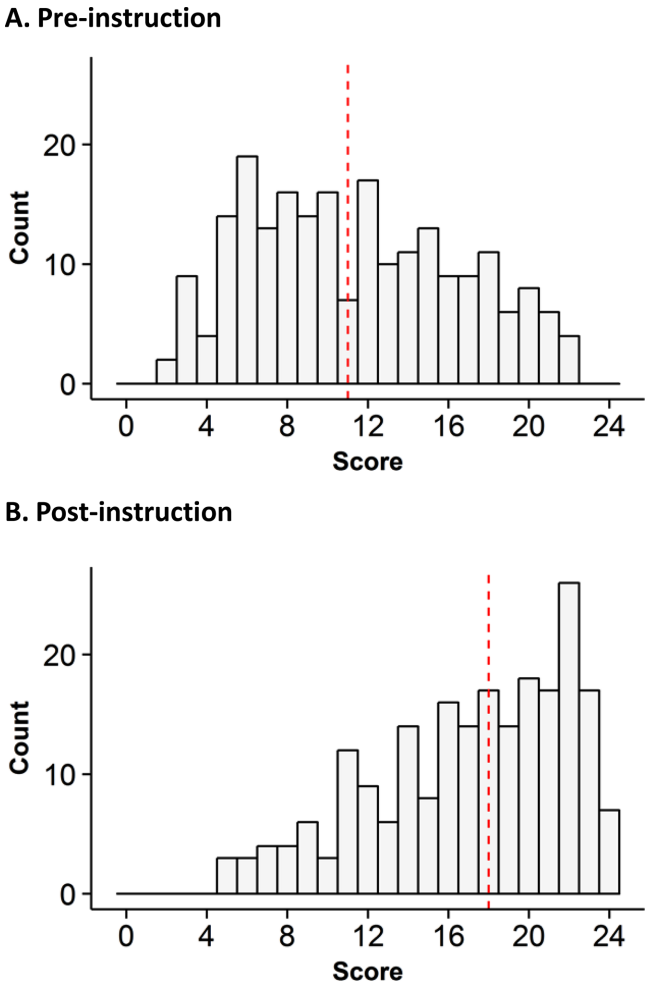

FIGURE 3. Pre- and postinstruction raw scores on the CANS in an introductory biology course.

may have been sampling error, because this question performed well in a previous analysis that included additional selection questions. Question 3 also had a relatively low loading. This question asked students what traits in anteaters are inherited. Interviews showed students struggled to differentiate among the distractors. However, the question loaded well in a unidimensional IRT analysis of all questions (see below), so it does seems to measure something related to natural selection.

IRT analysis of all 24 questions on the CANS together showed that 18 of the 24 questions loaded well on a one-dimensional ability (Table 1 and Figure 5). These 18 questions included all eight evolution questions, all four mutation questions, all four inheritance questions, and two of the five selection questions (questions relating to stressful environments). Three of the five selection questions did not load well, nor did any of the variation questions. There were no obvious goodness-of-fit problems. Only one question (question 20) had a p value < 0.05.

The information curve (Figure 6) for the CANS showed the test had high levels of information (>) from approximately –2 to +2 standard deviations from the mean and does a better job of estimating ability for high-ability students than low-ability students. The empirical reliability of the CANS was quite

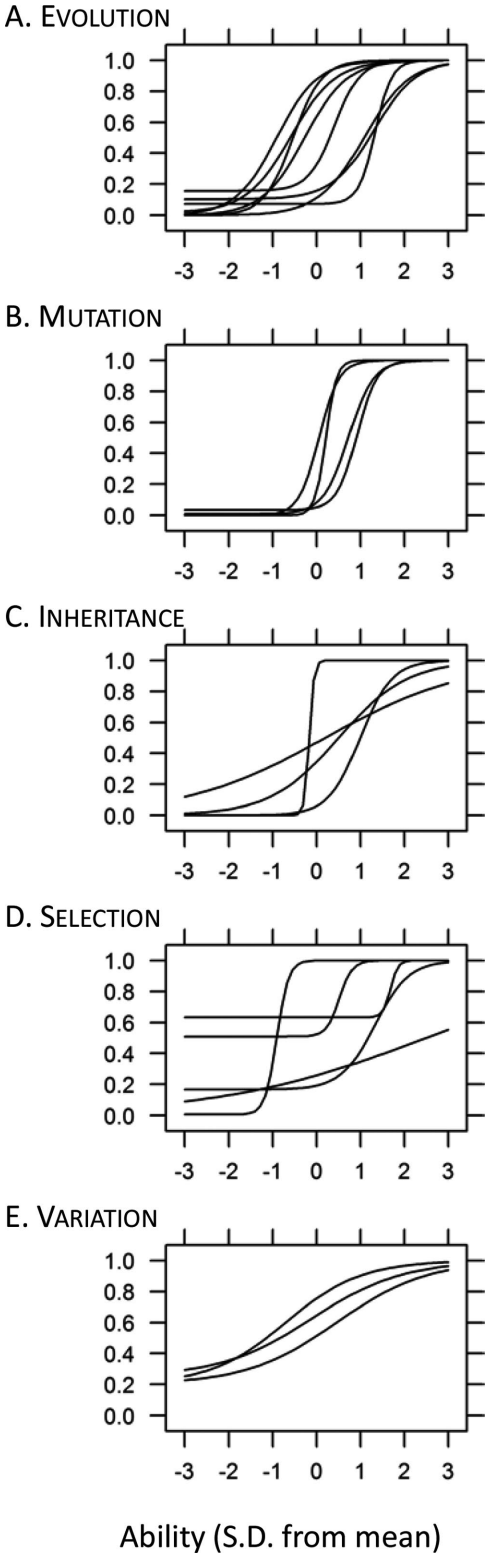

FIGURE 4. IRT traces for questions relating to each concept on the CANS.

high. It was 0.88 before instruction and 0.87 after instruction. Cronbach's alpha was 0.85 before instruction and 0.86 after instruction.

DISCUSSION

We developed and validated the CANS, a set of 24 multiple-choice questions created to assess how well college students in introductory biology courses understand the basic process of natural selection. Unlike previous instruments, the CANS assesses how students think about natural selection in multiple evolutionary contexts using multiple forms of questions. The expert panel that reviewed the CANS concluded that its questions were relevant to natural selection and generally did a good job sampling the specific concepts they were intended to assess. Student interviews confirmed questions on the CANS provided accurate reflections of how students think about natural selection. And, finally, IRT analysis showed that the CANS did a very good job of estimating student ability.

We estimated the empirical reliability of the CANS in our classroom to be 0.88. This means that 88% of the variance of test scores in our classroom can be attributed to differences in understanding among our students, and 12% can be attributed to measurement error. This appears to be very good—at least in comparison with other tests. The most highly regarded concept inventory in college science education research is probably the Force Concept Inventory (FCI) of Hestenes *et al.* (1992). The FCI is a 30-question multiple-choice test that measures how well students understand Newtonian motion. We compared the reliability of the CANS and FCI by administering the FCI to 197 students in a trigonometry-based introductory mechanics course taken by biology majors at MSU. These students took the FCI during the second week of the semester as part of their first weekly laboratory session. The empirical reliability of the FCI in this course was 0.83. This means the FCI has ~42% more measurement error than the CANS (17 vs. 12%). We also compared the empirical reliability of the CANS with published estimates of the reliability of the Scholastic Aptitude Test (SAT). The mathematics portion of the SAT has a reliability of 0.93, and the critical reading portion of the test has a reliability of 0.92 (Ewing *et al.*, 2005).

Our results have implications for a long-running debate regarding what students need to know in order to learn natural selection. Three concepts have been advocated as particularly important for learning natural selection: variation in populations, exponential growth, and genetics. We will discuss each of these possibilities in turn.

Mayr (e.g., 1982, chap. 11) has argued that typological thinking or essentialism—which we will provisionally define as the tendency to dismiss the importance of variation among individuals in a population—was a historical obstacle to the discovery of evolution. It is reasonable to wonder whether students face the same challenge: if students do not see variation among individuals, natural selection would not make sense. There are some correlative data to this hypothesis (Shtulman and Schulz, 2008). Our results, however, do not seem to support it: student responses to the variation and evolution questions were largely independent. More work is clearly needed to sort this out.

The relevance of exponential growth to natural selection was first noted by Darwin (1859) and discussed extensively by Mayr (1982, chap. 11): all species have the potential for exponential population growth; if such growth is not present, it is likely that many, if not most, of the individuals in a population do not survive and reproduce. This provides potent opportunity for natural

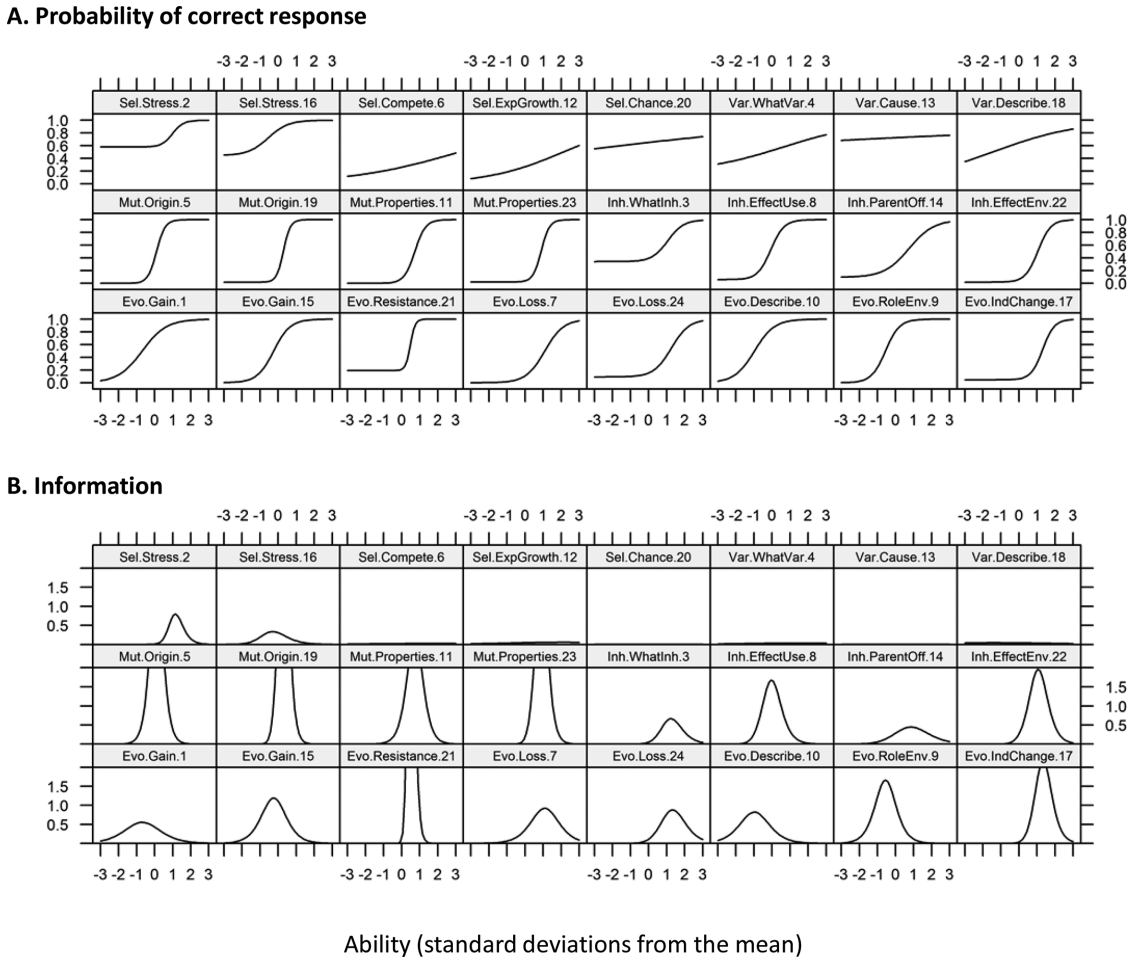

FIGURE 5. IRT traces (A) and information curves (B) for all 24 questions on the CANS analyzed using a one-dimensional 3PL model. The labels in each panel indicate the concept being tested, the form of the question, and the question number. For example, "Evo.Gain.1" indicates the question is question number 1 and is an evolution question relating to trait gain (see Table 1 for more descriptive labels).

selection, and recognizing this should help students understand natural selection. There is very little evidence for this in our data. The exponential growth questions loaded very weakly with the evolution questions in our IRT analysis. These concepts appear to be independent in the minds of our students. Our study was not designed to thoroughly explore the relationship between these concepts; again, further work is needed.

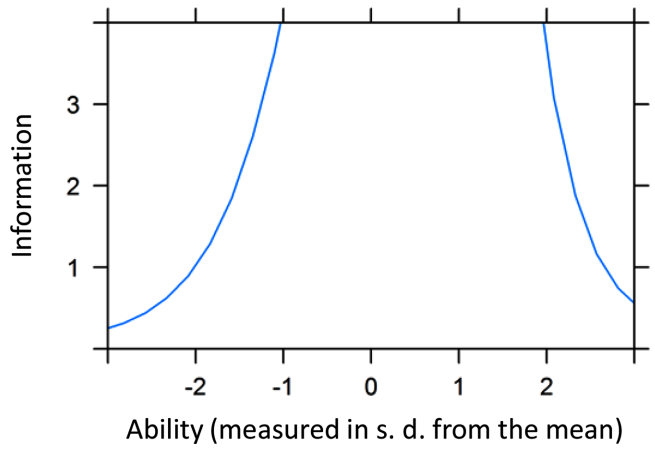

FIGURE 6. The information curve for the CANS estimated before instruction in an introductory biology course.

Many student misconceptions relating to natural selection appear to be based on misunderstandings of inheritance (Kalinowski *et al.*, 2010). Understanding inheritance and mutation, therefore, might help students understand natural selection. Our IRT analysis supports this hypothesis. All the questions relating to mutation and inheritance loaded strongly with the evolution questions.

The previous discussion illustrates the need to better understand how many dimensions there are to student understanding of natural selection. Our results suggest 18 questions on the CANS can reasonably be thought of as assessing a one-dimensional concept (natural selection), but it was also clear that this was not true for six of the questions on the CANS. We fitted multidimensional IRT models to our data to identify other dimensions but did obtain results that seemed meaningful, possibly because the questions we used did not assess other dimensions well, we did not include enough questions to estimate different dimensions well, or our sample size of 218 was too small.

Instructors and researchers should find the CANS useful in a few ways. Most simply, instructors could use questions from the CANS as topics for discussion in a classroom or as a source for exam questions. The CANS should also be useful to instructors for formative or summative assessment. For example, data from our classroom (Table 1) suggest that our students do not have a strong understanding of trait loss, mutation, exponential growth, or competition in stable populations. Researchers might use the CANS to compare scores and learning gains in different classrooms or to quantify how well individual students understand natural selection.

We will conclude this paper with a few comments about the importance of assessment in biology education. While we were validating the CANS, one member of our expert panel described the CANS as "yet another concept inventory." Such fatigue seems common and would be justified if existing instruments provide instructors and researchers with the tools they need to assess student thinking. However, when we talk to biology education researchers, we sense dissatisfaction with existing concept inventories. This is a bad combination of beliefs. Assessment of student thinking is the foundation of biology education research. If existing instruments are not sufficient, the biology education research community needs more research on concept inventories, not less. This will probably require intensive effort to understand how students think about important biology concepts and how students interpret specific questions (e.g., Rebello and Zollman, 2004; Weston *et al.*, 2015). It is hoped that the CANS will help stimulate such work and motivate researchers to create even better instruments for assessing student thinking. If the entire biology education community worked together on this, the work load need not be burdensome.

ACKNOWLEDGMENTS

This work was funded by the National Science Foundation (award 1432577). We thank three anonymous reviewers for comments that improved this article.

REFERENCES

Adams RJ (2005). Reliability as measurement design effect. Stud Educ Eval 31, 162–172.

Adams WK, Weiman CE (2011). Development and validation of instruments to measure learning of expert-like thinking. Int J Sci Educ 9, 1289–1312.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). Standards for Educational and Psychological Testing, Washington, DC: American Educational Research Association.

Anderson DL, Fisher KM, Norman GJ (2002). Development and evaluation of the Conceptual Inventory of Natural Selection. J Res Sci Teach 39, 952–978.

Andrews TM, Leonard MJ, Colgrove CA, Kalinowski ST (2011). Active learning not associated with student learning in a random sample of college biology courses. CBE Life Sci Educ 10, 394–405.

Ayala RJ (2008). Theory and Practice of Item Response Theory (Methodology in the Social Sciences), New York: Guilford.

Bardapurkar A (2008). Do students see the "selection" in organic evolution? A critical review of the causal structure of student explanations. Evol Educ Outreach 1, 299–305.

Bishop B, Anderson C (1990). Student conceptions of natural selection and its role in evolution. J Res Sci Teach 27, 415–427.

Brumby MN (1984). Misconceptions about the concept of natural selection by medical biology students. Sci Educ 68, 493–503.

Chalmers RP (2012). mirt: a multidimensional item response theory package for the r environment. J Stat Softw 48, 1–29.

Chown M (2013). What a Wonderful World, London: Faber and Faber.

Coyne J (2009). Why Evolution Is True, New York: Viking.

Crowl TK (1996). Fundamentals of Educational Research, 2nd ed., Madison, WI: Brown & Benchmark.

Darwin C (1859). On the Origin of Species, London: John Murray.

Demastes SS, Settlage J, Good R (1995). Students' conceptions of natural selection and its role in evolution: cases of replication and comparison. J Res Sci Teach 32, 535–550.

Dobzhansky T (1973). Nothing in biology makes sense except in the light of evolution. Am Biol Teach 35, 125–129.

Ewing M, Huff K, Andrews M, King K (2005). Assessing the Reliability of Skills Measured by the SAT (Research Note 24), New York: College Board.

Gregory T (2009). Understanding natural selection: essential concepts and common misconceptions. Evol Educ Outreach 2, 156–175.

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.

Hestenes D, Wells M, Swackhamer G (1992). Force Concept Inventory. Phys Teacher 30, 141–166.

Huxley TH (1887). On the reception of the "Origin of Species.". In: 1887. The Life and Letters of Charles Darwin, Including an Autobiographical Chapter, vol. 2, ed., ed. F Darwin, London: John Murray, 179–204.

Ingram EL, Nelson CE (2006). Relationship between achievement and students' acceptance of evolution or creation in an upper-level evolution course. J Res Sci Teach 43, 7–24.

Kalinowski ST, Leonard MJ, Andrews TM (2010). Nothing in evolution makes sense except in the light of DNA. CBE Life Sci Educ 9, 87–97.

Kalinowski ST, Leonard MJ, Andrews TA, Litt AR (2013). Six classroom exercises to teach natural selection to undergraduate biology students. CBE Life Sci Educ 12, 483–493.

Kline TJB (2005). Psychological Testing: A Practical Approach to Design and Evaluation, Thousand Oaks, CA: Sage.

Lamarck J (1809). Zoological Philosophy, trans. H Elliot, Chicago: University of Chicago Press.

Mayr E (1982). The Growth of Biological Thought, Cambridge, MA: Harvard University Press.

National Research Council (2012). Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering, Washington, DC: National Academies Press.

Nehm RH, Beggrow EP, Opfer JE, Ha M (2012). Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. Am Biol Teach 74, 92–98.

Nehm RH, Ha M (2011). Item feature effects in evolution assessment. J Res Sci Teach 48, 237–256.

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

Nehm RH, Schonfeld I (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.

Orlando M, Thissen D (2000). Likelihood-based item fit indices for dichotomous item response theory models. Appl Psychol Meas 24, 50–64.

Paley W (1802). Natural Theology, Boston: Gould, Kendall, & Lincoln.

Rebello NS, Zollman DA (2004). The effect of distracters on student performance on the force concept inventory. Am J Phys 72, 116–125.

Rios J, Wells C (2014). Validity evidence based on internal structure. Psicothema 26, 106–116.

Shtulman A, Schulz L (2008). The relation between essentialist beliefs and evolutionary reasoning. Cogn Sci 32, 1049–1062.

Spiro RJ, Coulson RL, Feltovich PJ, Anderson DK (1988). Cognitive flexibility theory: advanced knowledge acquisition in ill-structured domains. In: Tenth Annual Conference of the Cognitive Science Society, ed. V Patel, Hillsdale, NJ: Erlbaum, 375–383.

Weston M, Haudek KC, Prevost L, Urban-Lurain M, Merrill J (2015). Examining the impact of question surface features on students' answers to constructed-response questions on photosynthesis. CBE Life Sci Educ 14, ar19.