BRYCE THOMAS BATTISTI, NIKKI HANEGAN, RICHARD SUDWEEKS
and REX CATES

USING ITEM RESPONSE THEORY TO CONDUCT A DISTRACTER ANALYSIS ON CONCEPTUAL INVENTORY OF NATURAL SELECTION

Received: 4 August 2008; Accepted: 2 November 2009

ABSTRACT. Concept inventories are often used to assess current student understanding although conceptual change models are problematic. Due to controversies with conceptual change models and the realities of student assessment, it is important that concept inventories are evaluated using a variety of theoretical models to improve quality. This study used a modified item response theory model to determine university nonmajor biology students' levels of understanding of natural selection (n=1,192). Using Conceptual Inventory of Natural Selection, we have reported how we applied Bock's modified nominal item response theory model and the distracter test item analysis. We found that the use of this model can define student levels of understanding and identify problematic distracters.

KEY WORDS: biology, concept inventories, item response theory

Introduction

Conceptual change models are theoretically problematic and controversial for science education researchers and practitioners. Conceptual change models expose students' levels of understanding but have shown little to increase achievement (Dillon, 2008; Hewson, 2008; Treagust & Duit, 2008). Constructivist learning theories suggest that as students develop deeper understandings of scientific concepts, they draw conclusions closely aligned with current scientific principles (National Research Council, 2000). While effective practitioners agree with constructivist learning theories, the practitioner is faced with additional problems.

The accountability system that today's society expects of learning institutions appears to be anticonstructivist even though psychometricians work diligently to develop assessments parallel to constructivist teaching practices. This leaves the practitioner in an academic dilemma. Practitioners assess a student's level of understanding at a fixed point in time, but how do they use that same assessment as a diagnostic tool to scaffold learning activities and deepen the student's understanding for improved achievement?

International Journal of Science and Mathematics Education (2010) 8: 845–868

Concept inventories designed to assess students' understandings of scientific concepts are tools practitioners can use to solve their academic dilemma. For example, the Concept Inventory of Natural Selection (CINS) has high validity and reliability rates (Anderson, Fisher & Norman, 2002). The CINS instrument is commonly used with high school, college, and university biology students. Practitioners may use this instrument as a pretest and a posttest for student accountability and to scaffold learning activities.

Concept inventories use common misconceptions as the distracters. With this format, educators can use concept inventories as a guide to develop constructivist learning activities. After individual student misconceptions have been identified through the concept inventory, small student groups can be formed to help students construct deeper understandings through scientific argumentation by having students provide evidence to support their claims (Author, 2008; Mercer, 2008).

The authors examined the CINS and its distracters using Thissen and Steinberg's multiple-choice (MC) model as suggested by Anderson et al. (2002) to (a) describe student levels of understanding and (b) identify problematic distracters as currently written.

Theoretical Framework

Concept Inventory of Natural Selection

Evolution is the foundational theme in biology, and natural selection is the mechanism of evolution that accounts for the unity and diversity of life through natural selection (Cummins & Demastes, 1994; Ferrari & Chi, 1998). Unfortunately, college students' understanding of natural selection is lower than what is expected by most science educators (Bishop & Anderson, 1990; Brumby, 1984; Demastes, Good & Peebles, 1996; Ferrari & Chi, 1998; Greene, 1990; Settlage, 1994). Biology teachers and researchers have documented student responses in an attempt to help students understand natural selection (Aleixandre, 1994; Alters & Nelson, 2002; Jensen & Finley, 1995; Moore, Mitchell, Bally, Inglis, Day & Jacobs, 2002; Passmore & Stewart, 2002; Scharmann & Harty, 1986; Soderburg, 2003). Anderson et al. (2002) constructed a multiple-choice test—the CINS—and used it to assess students' understandings and misconceptions of natural selection among college biology nonmajors. The CINS was designed to assess students' understanding of eight topics in natural selection, plus the origin of species and the origin of variation.

Anderson et al. (2002) field tested the CINS and analyzed the results for the following indices: reliability, validity, difficulty, and discrimination. Each index supported the assertion that the CINS was an effective measure of college students' understanding of natural selection. Anderson et al. (2002) suggested that in the future, an analysis of CINS to determine the rate of distracter choice should be completed using Sadler's (1998) approach where he analyzed the results of a distracter-driven test in astronomy. Sadler used visual representations to determine how prevalent each distracter was for examinees that were at different level(s) of understanding (LOUs) of the subject matter. The study reported identified the relationship between a student's misconception and their LOU.

Measuring Conceptual Constructs

Constructivism underlies much of modern learning theory and posits that conceptual change related to existing concepts is fundamental to the learning process (Driver & Oldham, 1986; Palmer, 1999; Posner, Strike, Hewson & Gertzog, 1982). Scientific knowledge itself can be seen as a set of concepts, or theories, used as explanations for empirical phenomena (Driver & Oldham, 1986; Hatton & Plouffe, 1997; Thissen, Steinberg & Fitzpatrick, 1989). However, some student views are not the same as scientifically accepted concepts. As a result, science teaching becomes a process of elucidating students' current conceptions, facilitating the clarification of incomplete ideas, and structuring constructivist activities to increase student LOU.

Piaget employed interviewing as a technique to identify student views of scientific concepts. The value of interviewing is the interactive format, but it often yields inconsistent results (Johnson & Christensen, 2000; Novak, Mintzes & Wandersee, 2000). Hence, it is rarely cost-effective to question large or diverse groups of students to obtain generalizable results. Essay questions are less time and labor-intensive to administer but, like interviews, are subjectively scored (Novak et al., 2000).

Objectively scored, selected response assessments (in this case multiple-choice tests) provide reliable results and are easy to administer and score with very large groups of students. Tamir (1971) suggested that student interviews be used to delineate common misconceptions, which are then used as distracters for items in multiple-choice tests (Anderson et al., 2002; Sadler, 1998, 2000). Analyses of multiple-choice tests are often performed using the statistical methods classical test theory (CTT) and readily available software. The limitations of CTT include the inability to generalize test results to other groups who take the same test

or to the same group if they take another version of the test. To overcome these limitations, many researchers have turned to item response theory (IRT; Livingston, 2006; McKinley, 1989; Sadler, 1998).

Item Response Theory

IRT is a family of mathematical models used to describe the probability that a person will choose a particular response to a test item as a function of (a) characteristics of the item and (b) the trait level of the examinee (de Ayala, 2009; Embretson & Reise, 2000; McKinley, 1989; Yen & Fitzpatrick, 2006). Dichotomous IRT models produce a single-item response function (sometimes called a *trace line* or *item characteristic curve*) for each item. This monotonically increasing curve graphically shows how the probability of selecting the correct answer increases as a function of the examinee's trait level (e.g., ability, understanding, or whatever trait is being assessed) and one or more statistical properties of the item (e.g., difficulty, discriminating power, or susceptibility to guessing). Dichotomous IRT models are useful for describing the probability that students will select the correct answer to a multiple-choice item. However, these models are not very helpful to users who desire to analyze how the various distracters in each item function.

Instead of generating a single curve for each item, the nominal response model developed by Bock (1972, 1997) produces a separate curve for each distracter plus a curve for the option that is keyed correct. The availability for each option facilitates distracter analysis and provides insight about how the plausibility of each option varies as a function of examinee's LOU. The resulting information provides a basis for making informed decisions about which distracters need to be revised or replaced. This information is especially helpful to researchers attempting to construct distracter-driven tests for the purpose of examining the nature and prevalence of students' misconceptions.

The main limitation of Bock's model is that it makes no provision for the possibility that examinees may have resorted to guessing because they were unsure of the correct answer. Guessing is especially likely for low-ability examinees who lack sufficient understanding to make informed judgments about the relative plausibility of the various alternatives. However, examinees at all levels of understanding may resort to guessing in some form after they have eliminated one or two options that appear to be least plausible.

In an attempt to overcome this limitation, Samejima (1969) proposed a modification to Bock's model. She postulated that each option in a

multiple-choice item will be selected by some proportion (d_k) of the examinees and that this proportion will be combined with the proportion who deliberately chose that option believing that it was correct. The weakness of Samejima's solution is that she developed her model on the assumption that d_k is a constant that would take the same value for all the options associated with a particular item and that this value would equal the reciprocal of the number of options (i.e., $d_k = 1/m_j$). Samejima's approach is tantamount to assuming that examinees who guess always do so by randomly choosing from each of the options with equal probability.

Thissen & Steinberg (1984, 1997) proposed their MC model as an improvement to Samejima's modification of Bock's model. They postulated that "examinees who do not know the correct answer to a multiple-choice item comprise a latent class" and that the persons in this group would generally respond to the item by guessing rather than omitting a response (Thissen & Steinberg, 1997, p. 54). Furthermore, they theorized that each of the distracters would be chosen by some members of this latent class. Based on these assumptions and the idea that some distracters are likely to be more attractive than others to examinees who are misinformed or have partial knowledge, Thissen & Steinberg substituted a variable parameter in place of Samejima's constant (d_k). Since this "don't know" parameter represents a latent category, it is not directly observable. However, it can be estimated using the MC model. The value of the d_k parameter varies from one option to another within a given item. Conceptually, d_k represents the proportion of examinees who do not know the correct answer but select an option by guessing.

In this study, we used MULTILOG (du Toit, 2003; Thissen, 1991) to implement the MC model and estimate parameters for the four options and the d_k category in each of the CINS items. We then used *Excel* to plot separate option response functions (ORFs) for (a) the keyed answer, (b) each of the three distracters, and (c) the "don't know" or guessing category for all 20 items. The purpose for plotting the ORFs was to facilitate distracter analysis and interpretation of results. Readers should remember that the d_k category does not represent an option that was explicitly presented in the test. Rather, it was estimated by the MC model as a way of describing the probability that students at any particular LOU chose one of the options by guessing.

Collectively, the set of ORFs for a particular item illustrates how the probability of selecting each option varies as a function of the students' level of understanding. In addition, the d_k curve for each item shows how the probability of guessing varies as a function of the students' levels of understanding. Generally, the probability of selecting the correct answer

for an individual item is expected to increase as the examinees' understanding increases. Conversely, the probability of choosing a given distracter is expected to decrease as a function of increasing levels of understanding. The graphs for the various items illustrate the degree to which these expectations were satisfied.

Purpose of Study

The purpose of this study was to use Thissen and Steinberg's MC model to analyze college students' responses to each option in the CINS items to (a) identify common misconceptions of evolution that vary as a function of students' understanding, (b) detect problematic distracters, and (c) formulate recommendations to improve the CINS items.

Methods

Student Profile

Responses were obtained from 1,192 students from a class of more than 1,500 students enrolled in Biology 100, an introductory college biology course. Students can take biology 100, a required course for nonmajors, any time during their undergraduate coursework. This results in students at all stages of their degree-seeking instruction enrolling in Biology 100. Of the respondents, 47% were freshmen, 33% sophomores, 12% juniors, and 7% seniors with nearly an equal number of male and female students. Each of the eight topics assessed by the CINS was included in Biology 100 lectures and text reading assignments (Author, 2002). The CINS was administered along with the last class exam of the semester. Students completed the CINS items to receive extra credit points for participation. The possibility of receiving extra credit points for simply completing the test may have motivated some of the students to complete the test without taking it seriously. These students may have randomly selected answers to some items.

Instrument and Procedures

The CINS is a 20-item multiple-choice test designed specifically for use with college biology nonmajors (Anderson et al., 2002):

1. The CINS consists solely of context-dependent items. The 20 items are arranged in four subsets containing eight, five, four, and three items,

respectively. The items within each subset are based on a common scenario or information display that provides a novel context. Examinees are expected to mentally process the information in the display and answer each subset of questions. Haladyna (1992; 1994) and Wesman (1971) refer to items of this kind as "context-dependent item sets". However, other scholars have labeled them as "interpretive exercises" (Ebel, 1951; Linn & Gronlund, 2000).
2. The CINS is a distracter-driven test. The three distracters in each item were deliberately constructed to represent a commonly held misunderstanding.
3. The CINS is a paired item test. Ten pairs consisting of 20 items are designed to test students' understanding of the ten concepts assessed on the CINS.

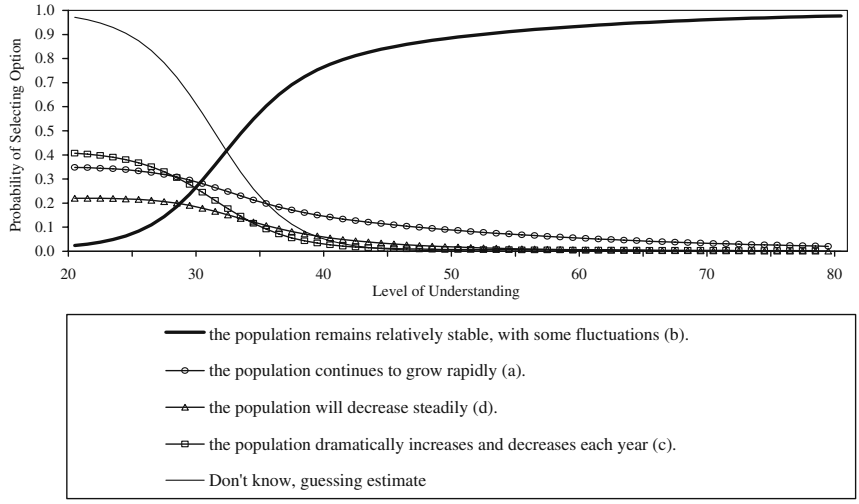
Data Analysis

Student responses to the CINS were imported into the MULTILOG 7.0 program used to estimate all item parameters in this study. These parameters were then exported to Microsoft Excel® to plot the ORFs (Figures 1, 2, 3, 4, 5, 6, and 7).

Item Analysis

ORFs are graphical representations of response patterns for a group of examinees. Figures 1, 2, 3, 4, 5, 6, and 7 show how the probability of selecting a particular option varies as a function of the examinees' LOU as predicted by the Thissen and Steinberg model. Test item stems are displayed above their corresponding ORFs (Figures 1, 2, 3, 4, 5, 6, and 7). The legend below each plot lists the multiple-choice answers presented to the examinees with the alphabetical symbol used to designate the response option shown in parentheses. The top option in each list is the correct answer and is depicted in the graph as a thick solid black line. The other three options in the list are the distracters, and the final option is the "don't know estimate" (d_k) generated by the model. Each figure includes two paired test items; hence, two ORFs are displayed for each natural selection topic covered by the CINS. Although we analyzed all ten item pairs, due to space limitations, we have reported graphic displays for only seven of the ten pairs in this article. The excluded graphs did not provide any new information.

3. Once a population of finches has lived on a particular island with an unvarying environment for many years,



12. Once a population of guppies has been established for a number of years in a real (not ideal) pond with other organisms including predators, what will likely happen to the population?

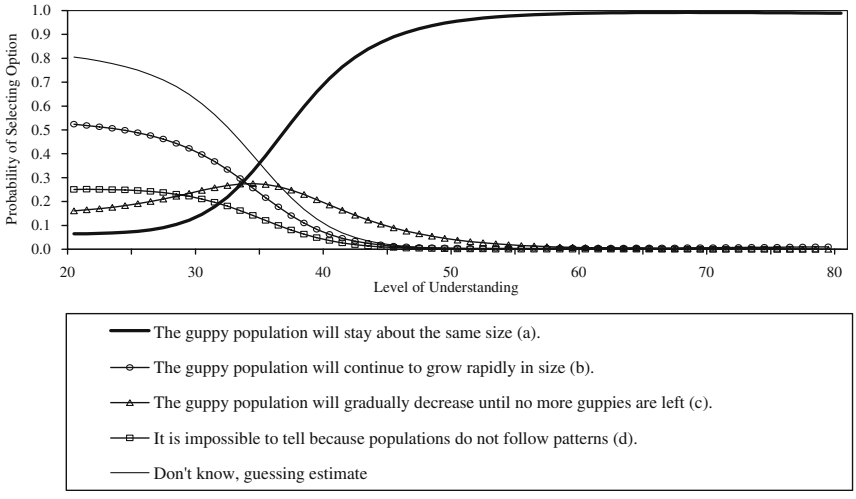
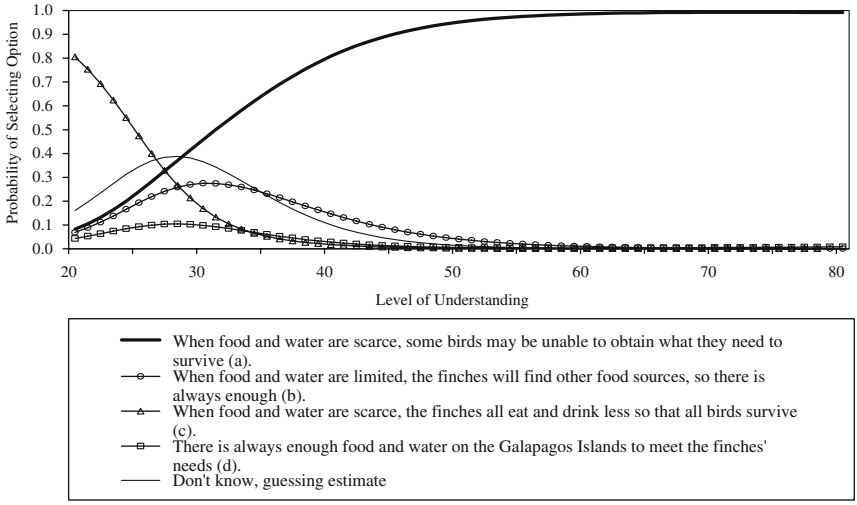


Figure 1. Population stability. Option response functions for items 3 and 12

In the following sections, a description comparing and contrasting the response patterns for the two paired test items with a short explanation of the concept being measured is followed by a discussion of how the various distracters for those items perform at different

2. Finches on the Galapagos Islands require food to eat and water to drink.



14. Lizards eat a variety of insects and plants. Which statement describes the availability of food for lizards on the Canary Islands?

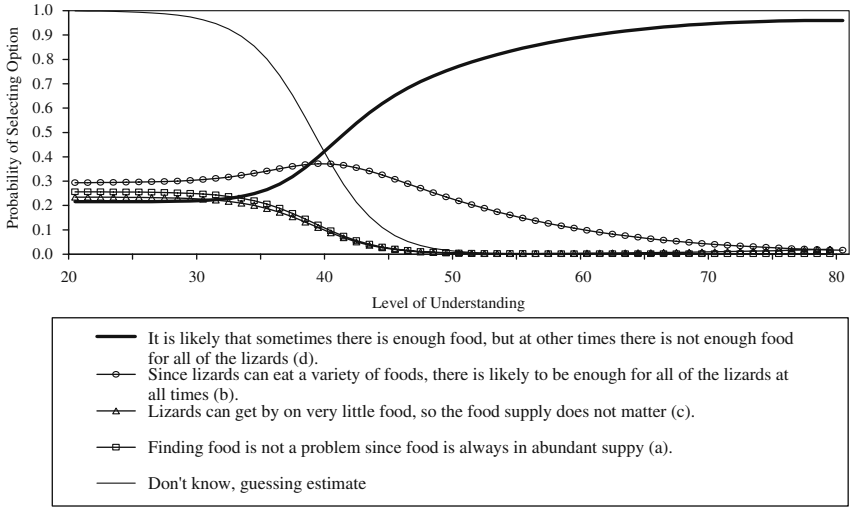
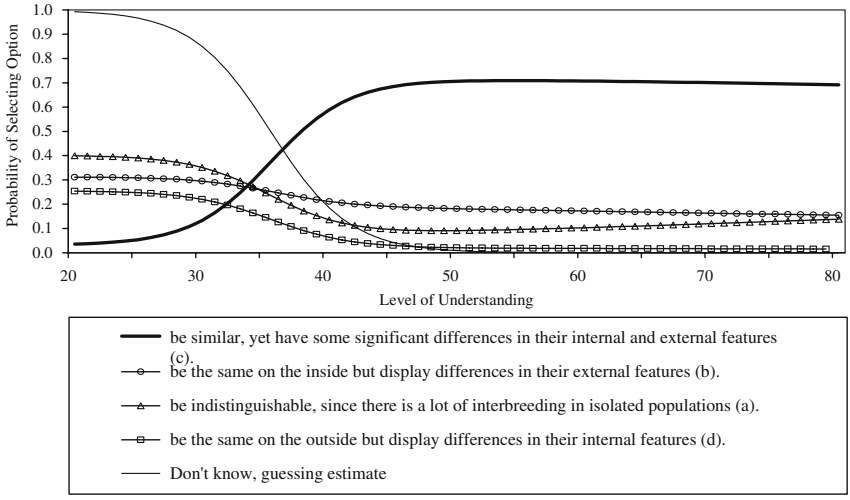


Figure 2. Natural resources. Option response functions for items 2 and 14

LOUs. The difficulty of a multiple-choice item is not only a function of the nature of the problem posed in the item stem but also by the number and quality of the distracters. CINS items include three distracters, but the results of this study show that the various distracters function differently. The trace line for each individual

16. A well-established population of lizards is made up of hundreds of individual lizards. On an island, all lizards in a lizard population are likely to . . .



9. A typical natural population of guppies consists of hundreds of guppies. Which statement best describes the guppies of a single species in an isolated population?

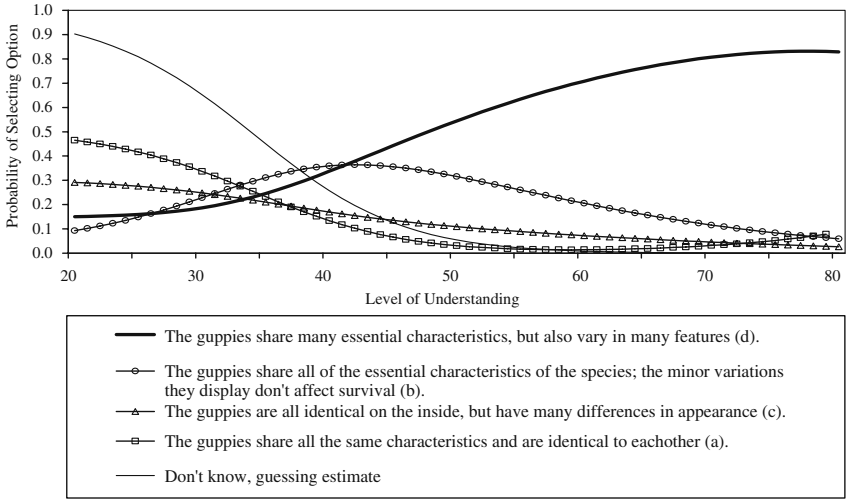
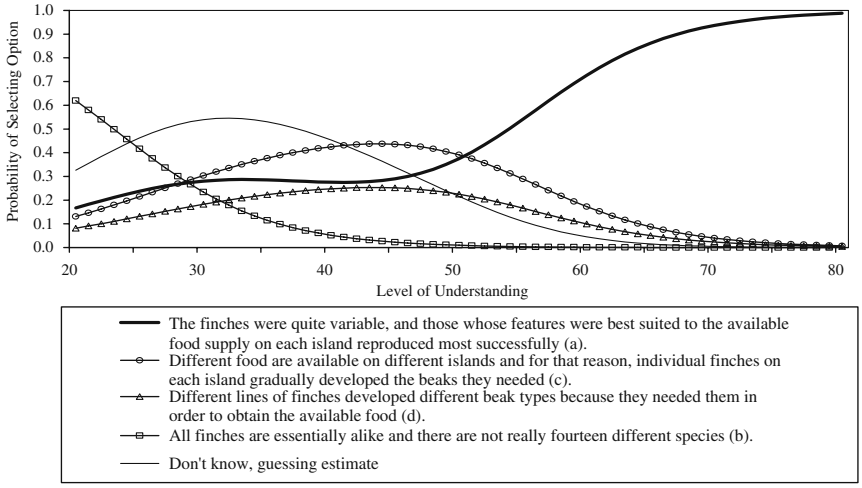


Figure 3. Variation within a population. Option response functions for items 16 and 9

distracter graphically displays how the probability of selection varies as a function of the understanding of the examinees. The items ranged in difficulty from 33 (1.7 standard deviations below the mean) to 57 (0.7 standard deviations above the mean) on the LOU scale.

8. What caused populations of birds having different beak shapes and sizes to become distinct species distributed on the various islands?



20. What could cause one species to change into three species over time?

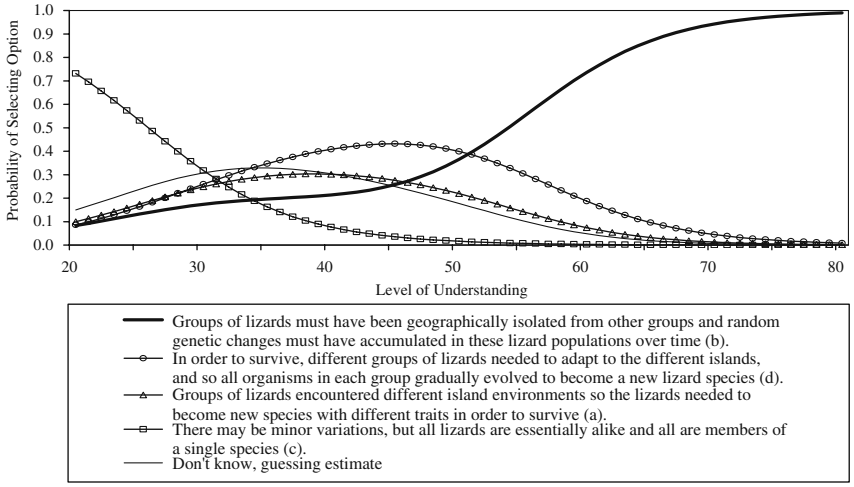
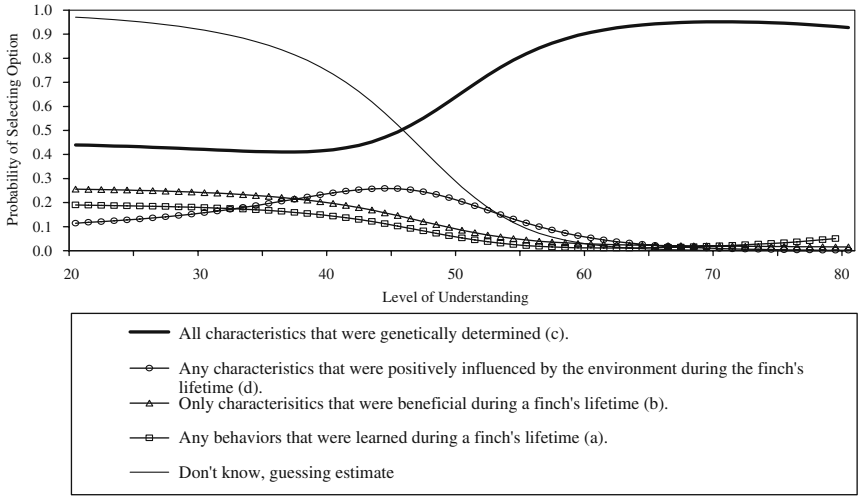


Figure 4. Origin of species. Option response functions for items 8 and 20

Items 3 and 12. Population Stability: "Apart from Seasonal Fluctuations, Most Populations Reach Equilibrium with Their Environment and so Remain Stable in Size." (Figure 1)

The graphs in Figure 1 show that the two items functioned similarly. The curves for the correct option in the two items are similar in shape and

7. What type of variation in finches is passed to the offspring?



17. Which statement best describes how traits in lizards will be inherited by offspring?

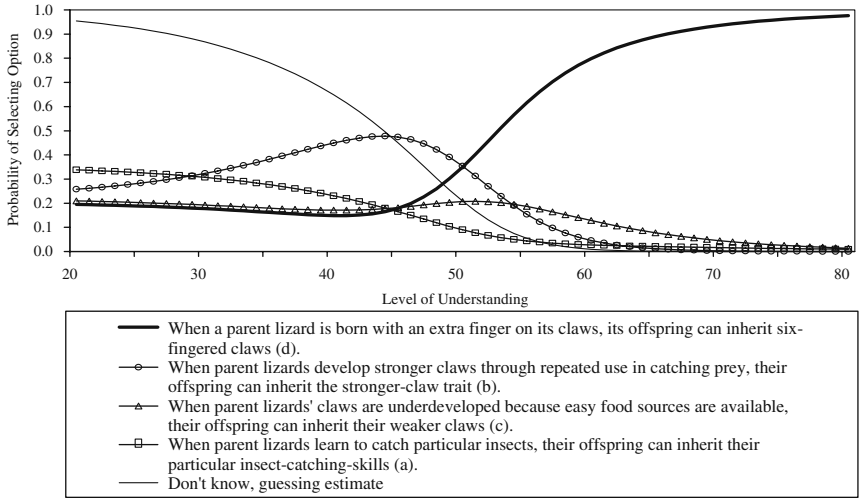
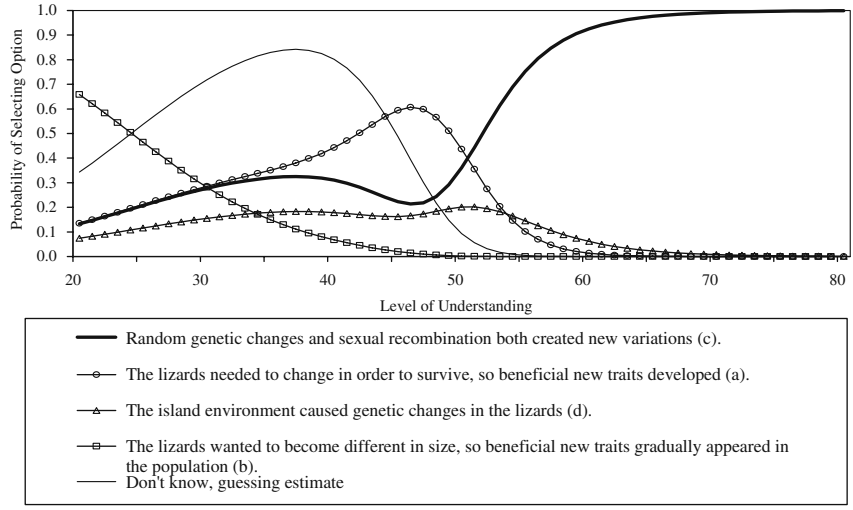


Figure 5. Variation inheritable. Option response functions for items 7 and 17

reasonably close in location, although item 3 is slightly easier as the curve for the correct option is located farther to the left. For both items, students with a LOU greater than 40 had a high probability of answering correctly. The similarity of the two graphs provides evidence that the items measure the same concept, as the test authors intended.

19. According to the theory of natural selection, where did the variations in body size in the three species of lizards most likely come from?



6. How did the different beak types <u>first</u> arise in the Galapagos finches?

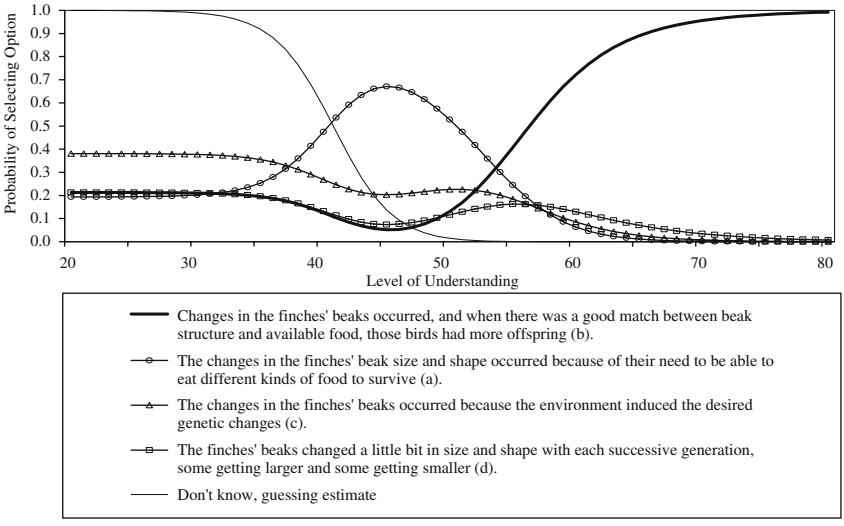
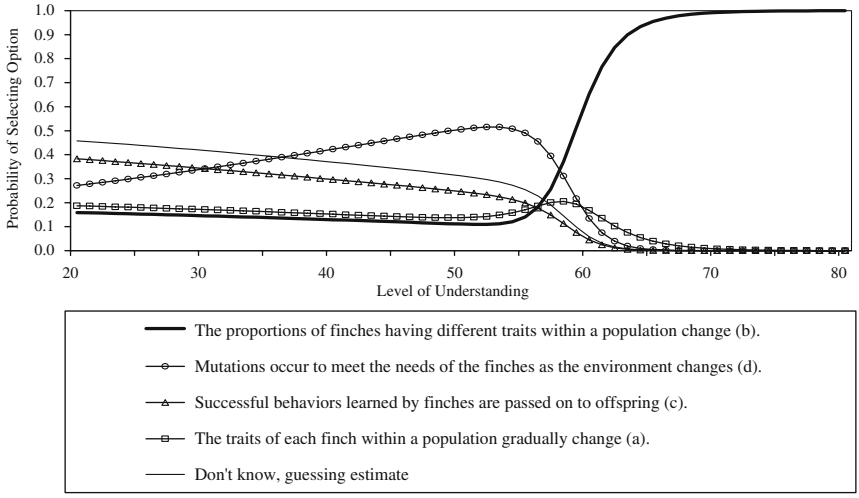


Figure 6. Origin of variation. Option response functions for items 19 and 6

In a well-developed multiple-choice item, the distracters should be attractive to examinees that have an incomplete understanding of the concept being assessed. This is particularly true of the items in distracter-driven tests like the CINS. Although each item in this pair was

4. In the finch population, what are the primary changes that occur gradually over time?



13. In guppy populations, what are the primary changes that occur gradually over time?

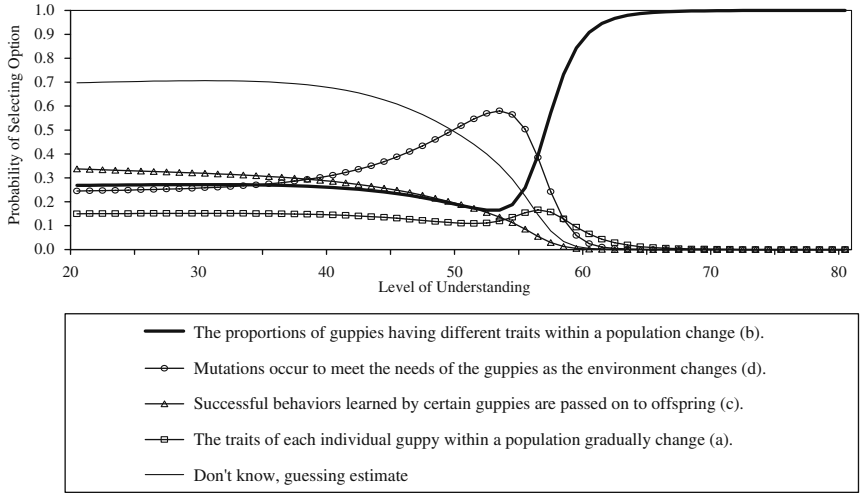


Figure 7. Change in a population. Option response functions for items 4 and 13

relatively easy for the group as a whole, both items functioned quite well in distinguishing between students who understood the target concept and those who did not. The distracters were generally not attractive to students with a LOU greater than 50 while at least two of the distracters were attractive to students with low LOUs. Guessing was prevalent among

students with a LOU less than 30. On item 3, the distracters were more equally attractive than in item 12. For item 12, the distracter that the "population will continue to grow rapidly in size" was most often chosen by students with a LOU less than 30.

Items 2 and 14. Natural Resources: " At any Given Time a Population of Organisms Only has Access to a Limited Supply of Resources such as Nutrients, Water, and Space." (Figure 2)

The ORFs for items 2 and 14 are distinctly different even though the two items were designed to assess the same conceptual topic. The ORF for the keyed answer to item 2 is located farther to the left than the curve for the correct answer to item 14, which indicates that item 2 was easier. The difference in the "don't know" curves for the two items indicates that low LOU students were more prone to select one of the options by guessing when they responded to item 14 compared to their response to item 2. Option c in item 2 represents the incorrect notion that when food is limited, finches will ration and share the available food so that the species will survive. Apparently this distracter was so plausible to low LOU students that many of them chose this option intentionally and did not perceive a need to guess.

Most of the CINS item pairs consisted of items that were rather similar in difficulty. The main exception was the pair that included items 2 and 14. With a difficulty of 33, item 2 was the easiest item on the test, while its companion, item 14, had a difficulty of 43, exactly 1 standard deviation higher. Option d in item 2 is the poorest functioning distracter in the test. The probability that examinees will select this item never exceeds 0.10 and that probability occurs only in a narrow range near the low end of the LOU scale. The poor functioning of this distracter contributes to the fact that item 2 is the easiest item in the test. By way of contrast, option c in item 2 also functions only at the low end of the understanding continuum and has a much higher probability of being selected in the range below 30.

Items 16 and 9. Variation within a Population: "There is Genetic, Physical and Behavioral Variation Among Members of Populations." (Figure 3)

The patterns of the ORFs in the graphs for items 16 and 9 are distinctly dissimilar. The location and shape of the curves for the correct answer to each of the items indicates that item 16 is easier but more discriminating than item 9, at least in the narrow range from 30 to 40 LOU. The most

distinctive feature of item 16 is that the trace line for option c, the correct answer, is practically flat for all LOUs above 35 indicating that it is likely to be selected equally by students with average and high LOU. At least part of the reason for the flatness of the curve for the correct answer is that the curve for distracter b and option c tend to be attractive to students at average and high LOU.

In addition to being less discriminating than item 16, item 9 is also less susceptible to guessing in the low LOU range. Part of the reason for the low discriminating power of item 9 is the attractiveness of distracter b to students in the 35 to 60 LOU segment of the continuum.

Item 16 is noteworthy because it includes two distracters (options a and b) that function across the whole understanding continuum. Neither of these distracters functions at the same level across the continuum, but they have a broader span than the distracters in any of the other items. Another distinctive feature of item 16 is that the trace line for the correct answer never levels out at the low LOU. Generally, the trace line for the correct answer is much higher (much closer to 1.0) at the upper end of the understanding continuum. But distracters a and b together attract a large enough proportion of the students to prevent the correct option from being higher. Apparently, the incorrect and/or incomplete knowledge represented by these two distracters is more pervasive across the continuum than for the other CINS item pairs.

This conclusion is supported by the pattern for the trace lines in item 9, which is intended to assess the same topic as item 16. Distracter b in this item also tends to span the whole understanding continuum, and distracter c spans much of it. Distracter a is more attractive to students at the low end of the continuum, but distracter b spans the whole continuum. While distracter c also becomes progressively less attractive as LOU increases, it also spans much of the middle part of the continuum. Collectively, these functioning distracters combine to keep the trace line for the correct option lower than usual at the upper end of the continuum.

Items 8 and 20. Origin of Species: " Populations of a Single Species can Differentiate Much Over Time such that if Isolated from Each Other They can Eventually Become Distinct Species." (Figure 4)

The stems of these two items are highly similar. Both items assess the same topic, but the issue presented in the stem of item 8 refers to a particular species, while the problem presented in the stem of item 20 is more abstract and generic. The graphs for these two items are strikingly

similar to each other but quite distinctive from the graphs for the other item pairs in the CINS. Both items are relatively difficult compared to the other items in the test.

The graph for each item reveals that different distracters function as the most likely response at the different LOUs. This pattern indicates that understanding/misunderstanding of the targeted concept varies as a function of students' LOU. Another distinctive feature of these two items is that misconceptions are still quite likely even in the 50–65 LOU range.

For both items, distracter d is the most probable for students at the lowest LOU level, but in the LOU range from 30 to 55, two other distracters are more probable. Beginning at a LOU of approximately 50, the correct option becomes the most likely choice. The probability of guessing at each LOU is smaller for both of these items than for most other items in the test.

Items 8 and 20 are two of the more difficult items in the test. Both of these items include complementary distracters that function together to encompass a broader range of the understanding continuum. Each of these items includes one distracter that is attractive to examinees that have low LOU. In addition, each of these items includes two distracters that function across a broad range of the middle part of the continuum. The cumulative probability of selecting either option c or option d in response to item 8 approaches or exceeds 0.70 in the middle of the continuum. Distracters b and d perform similarly in item 20.

Items 7 and 17. Variation Inheritable: " Basic Biological Explanations About the Flow of Biological Information Include the Idea that Much of the Observed Difference Between Members of a Species is Heritable." (Figure 5)

All distracters in each of these items are principles of Lamarckian evolution. The option most often chosen for students of lower understanding levels was acquired beneficial physical traits. In both items, this idea was most often chosen among students with LOUs ranging from 40 to 50.

These two items are about equally difficult, but item 17 has greater power to distinguish between the knowledgeable and less knowledgeable students. In item 7, the correct answer was the most likely to be chosen at all LOUs, even among those who did not know. Both items have a high susceptibility to random guessing in the LOU range of 45 or less. In contrast, the correct answer to item 17 is the option least likely to be selected except by those with an understanding less than 45.

Items 19 and 6. Origin of Variation: "Random Mutations are the Ultimate Source of All Genetic Variation and Sexual Reproduction Acts to Increase the Level of Genetic Variation. Some Mutations are Beneficial, but Most are Neutral or Harmful." (Figure 6)

This is the second most difficult topic in the CINS test. Items 6 and 19 are quite similar in difficulty and discriminating power, but the most salient characteristic of the ORFs for these two items is the distinct wavelike shape of the ORF for distracter a in both items. The probability of selecting this option for examinees in the 40–50 LOU range is so high that it results in a dip in the curves for the correct answer and one or more of the distracters.

Until the 45 LOU on both items, the trace lines for "Needed mutations occur to allow an organism to survive" follow the pattern expected for the correct option and eventually develop into a steep wave. At the same time, the correct options follow the monotonically decreasing pattern expected of distracters and eventually dip.

Item 19 was less susceptible to guessing by low LOU students than item 6. This decreased susceptibility may have been because of the plausibility of option c to the students in this group.

Items 4 and 13. Change in a Population: "Members of Populations Differ from One Another and Hence Respond Differently to Selection Pressures. The Proportions of Alleles in the Population Change Because those More Able to Withstand Selection Pressures are More Likely to Pass Their Genes on to More Offspring." (Figure 7)

Overall, the ORFs for items 4 and 13 display a remarkably similar pattern. The location of the inflection point in ORF for the correct answer to each of these items is located farther to the right on the LOU continuum than for any other pair of items. Hence, items 4 and 13 are the two most difficult items in the CINS test. In addition, the curve for the correct answer increases more steeply and in a narrower range of LOU than for any other items. Hence, these two items are also the most discriminating items in the test. All across the lower half of the LOU continuum (i.e., <55), students are most likely to either guess or to select an incorrect answer. The distracters for each of these two items function quite well for students with a LOU less than 55, but above 65, they hardly function at all. The effective discriminating range for both of these items is in the narrow LOU range between 55 and 65.

Distracter a is particularly interesting in each of these items because it is very attractive to examinees in the middle range of the LOU continuum. These distracters represent the idea that mutations occur to meet the "needs" of the population. This is the same teleological explanation that was frequently chosen in the topics Origin of Species (items 8 and 20) and Origin of Variation (items 6 and 19). In these two items, the correct answer was not chosen except by those with an estimated understanding greater than 60.

Suggested Test Item Improvements

This study revealed several areas that may lead to improvement in the CINS. The following are suggested areas for improvement as categorized by the paired options previously described.

Items 2 and 14: Natural Resources (Figure 2). According to the test makers, their distracters for this pair of items were written to address one misconception: Organisms can always obtain what they need to survive. Since the corresponding options in these two items are conceptually similar to each other, at first glance it is difficult to explain why the ORFs are so different. Three factors may help to explain this difference. First, comparison of the d_k curves for these two items indicates that item 14 is more susceptible to guessing. Second, for students with a LOU less than 30, distracter c in item 2 is much more attractive than its conceptual counterpart in item 14. Third, distracter a in item 14 is more attractive to students in the average LOU range than its counterpart in item 2. This conclusion may seem counterintuitive but is supported by the evidence in the ORFs for these two items. Consequently, one or both of these items need revision.

Items 16 and 9: Variation within a Population (Figure 3). The nonscientific terminology used by the test writers may have reduced the desired specificity in the options included in these two items. For example, it may have been difficult for the students to distinguish between the options "[The organisms] are all identical on the inside, but have many differences in appearance," and "[The organisms] share many essential characteristics, but also vary in many features." Item 16 may be more discriminating simply because its correct option is more specific, "[The organisms are] similar, yet have some significant differences in their internal and external features." Replacement with the terms phenotype and genotype for external and

internal differences, respectively, may increase the difficulty and discrimination of these items.

Items 8 and 20: Origin of Species (Figure 4). Students of lowest understanding had a very high probability of choosing the distracter that designates similar organisms as the same species. This persists up to an understanding level of 30 and is a common creationist explanation for the origin of species—namely that there are few species and many varieties (Dobzhansky, 1973; Miller, 1999). Since no definition for a species is given in the CINS, it is possible that some students understood differences between finch species to be no greater than differences between breeds of dog.

The next most common misconceptions are that (a) individuals change (evolve) because they "need" to become new species and (b) whole populations change (evolve) because they "need" to become new species. These are both teleological explanations for why there is such a diversity of organisms. Only those with high LOUs of natural selection have almost 100% likelihood of choosing the correct option for these items, indicating that the correct option was clear to those who understood how natural selection works. What is striking is how distinctive from the rest of the test and how similar to each other these two items are. This similarly provides evidence that both items measure knowledge of the same topic in the same way.

Items 7 and 17: Variation Inheritable (Figure 5). Perhaps the distracters in item 17, especially the one about the heritability of acquired physical traits, were selected most often because they were more specific than those in item 7. Perhaps item 7 was much easier for those with a low LOU because the correct option for this item uses the phrase "genetically determined" to describe traits that are passed on to offspring. This phrase may have given away the correct option.

Items 19 and 6: Origin of Variation (Figure 6). In item 19, students of lowest LOUs are most likely to select the Lamarckian explanation that the lizards adapt because they "want" to adapt. Had this distracter been included in item 6, it may have performed similarly in this understanding range. Also, in item 19, students of low ability were most likely to choose another teleological distracter; that organisms change because they "want" to change. These distracters exemplify the teleological idea of evolution and natural selection. Those choosing teleological options understand organisms to be evolving toward a goal. Only those with an

understanding greater than 55 favor the correct option that variation's sources are the largely random processes of mutation and sexual reproduction.

Discussion and Recommendations

The traditional way to conduct a distracter analysis is to divide the total number of examinees into three or more groups such as (a) a high scoring group containing about one third of the examinees, (b) the middle scoring third of the examinees, and (c) the lowest scoring third. Then a separate two-way contingency table for each test item of interest is created with the three groups of examinees represented in the rows of the table and a column for each option in the item. Then, one analyzes the percent of examinees in each group who selected each option. Ideally, the percent of examinees in the high-scoring group who select the correct answer should exceed the percent in each of the other two groups. The reverse pattern should be true for each distracter, that is, the percent of examinees in the low: Scoring group should exceed the percent in the middle and high scoring groups (Oosterhof, 1994). This categorical approach to distracter analysis allows the analyst to determine how each option functions for each group of students. If response data for each of the items are available for a large enough sample of students, then the analyst could use four or five groups instead of three to obtain a more fine-grained analysis. But such an expanded analysis still represents a categorical approach.

In contrast, the use of trace lines obtained from Thissen and Steinberg's modification of Bock's nominal response model of IRT permits the analyst to display how the probability of selecting each option, including each distracter as well as the correct answer, varies as a function of the students' overall understanding. If the test has been constructed so that each distracter represents an incorrect or incomplete understanding, then this graphic approach reveals how the probability of selecting each of these incorrect notions varies as a function of the examinees' level of understanding.

Concept inventories are often used to measure students' conceptual constructs, and practitioners will continue to rely on multiple-choice item tests (instruments) practicalities and societal pressures. Instrument designers must act responsibly and continue improvement of concept inventories through rigorous analyses to improve accuracy in student reporting. Practitioners must also be more cognizant of instrument

limitations. We feel it would be pertinent to analyze other concept inventories with Thissen and Steinberg's modification of Bock's nominal item response model and compare those analyses against classical test theory analyses.

REFERENCES

Aleixandre, M. P. J. (1994). Teaching evolution and natural selection: A look at textbooks and teachers. *Journal of Research in Science Teaching*, *31*, 519–535.

Alters, B. J., & Nelson, C. E. (2002). Perspective: Teaching evolution in higher education. *Evolution*, *56*, 1891–1901.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the Conceptual Inventory of Natural Selection. *Journal of Research in Science Teaching*, *39*, 952–978.

Author. (2002). *Biology principles & applications: A syllabus*. San Francisco: McGraw-Hill Primis Custom.

Author. (2008). Disconnections between teacher expectations and student confidence in bioethics. *Science & Education*, *17*(8–9), 921–940.

Bishop, B. A., & Anderson, C. W. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching*, *27*, 415–427.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 23–49). New York: Springer.

Brumby, M. N. (1984). Misconceptions about the concept of natural selection by medical biology students. *Science Education*, *68*, 493–503.

Cummins, C. L., & Demastes, S. S. (1994). Evolution: Biological education's under-researched unifying theme. *Journal of Research in Science Teaching*, *31*, 445–448.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford.

Demastes, S. S., Good, R. G., & Peebles, P. (1996). Patterns of conceptual change in evolution. *Journal of Research in Science Teaching, 33*, 407–431.

Dillon, J. (2008). Discussion, debate and dialog: Changing minds about conceptual change research in science education. *Cultural Studies in Science Education*, *3*(2), 397–416.

Dobzhansky, T. (1973). Nothing in Biology makes sense except in the light of evolution. *The American Biology Teacher*, *35*(3), 125–129.

Driver, R., & Oldham, V. (1986). A constuctivist approach to curriculum development in science. *Studies in Science Education*, *13*, 105–122.

du Toit, M. (Ed.) (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, and TESTFACT*. Lincolnwood, IL: Scientific Software Program.

Ebel, R. L. (1951). Writing the test item. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 185–249). Washington, D.C.: American Council on Education.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum Associates.

Ferrari, M., & Chi, T. H. M. (1998). The nature of naive explanations of natural selection. *International Journal of Science Education*, *20*, 1231–1256.

Greene, E. D., Jr. (1990). The logic of university students' misunderstanding of natural selection. *Journal of Reasearch in Science Teaching*, *27*, 875–885.

Haladyna, T. M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice*, *11*(1), 21–25.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hatton, J., & Plouffe, P. B. (Eds.) (1997). *Science and its ways of knowing*. Upper Saddle River: Prentice-Hall.

Hewson, P. (2008). Conceptions over time: Are language and the her-and-now up to the task? *Cultural Studies in Science Education*, *3*(2), 263–276.

Jensen, M. S., & Finley, F. N. (1995). Teaching evolution using historical arguments in a conceptual change strategy. *Science Education*, *79*, 147–166.

Johnson, B., & Christensen, L. (2000). *Educational research: Quantitative and qualitative approaches.* Boston: Allyn & Bacon.

Linn, R. L. & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Merrill Prentice-Hall.

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 421–441). Mahwah: Erlbaum.

McKinley, R. L. (1989). An introduction to item response theory. *Measurement and Evaluation in Counseling and Development*, *22*, 37–57.

Mercer, N. (2008). Changing our minds: A commentary on 'Conceptual change: A discussion of theoretical, methodological and practical challenges for science education'. *Cultural Studies in Science Education*, *3*(2), 351–362.

Miller, K. R. (1999). *Finding Darwin's God: A scientist's search for common ground between god and evolution.* New York: Harper Collins.

Moore, R., Mitchell, G., Bally, R., Inglis, M., Day, J., & Jacobs, D. (2002). Undergraduates' understanding of evolution: Ascriptions of agency as a problem for student learning. *Journal of Biological Education*, *36*, 65–71.

National Research Council. (2000). *Inquiry and the national science education standards: A guide for teaching and learning (report).* Washington, D.C.: National Academy.

Novak, J. D., Mintzes, J. J., & Wandersee, J. H. (2000). Epilogue: On ways of assessing science understanding. In J. J. Mintzes (Ed.), *Assessing science understanding* (pp. 355–374). New York: Academic.

Oosterhof, A. (1994). *Classroom applications of educational measurement* (2nd ed.). Columbus, OH: Merrill Publishing Co.

Palmer, D. H. (1999). Exploring the link between students' scientific and nonscientific conceptions. *Science Education*, *83*, 639–653.

Passmore, C., & Stewart, J. (2002). A modeling approach to teaching evolutionary biology in high schools. *Journal of Research in Science Teaching*, *39*, 185–204.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*, 211–227.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distracter-driven assessment instruments. *Journal of Research in Science Teaching*, *35*, 265–296.

Sadler, P. M. (2000). The relevance of multiple-choice tests in assessing science understanding. In J. J. Mintzes (Ed.), *Assessing science understanding* (pp. 249–278). New York: Academic.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores [psychometric monograph, 17]*. Iowa City: Psychometric Society

Scharmann, L. C., & Harty, H. (1986). Shaping the non-major general biology course. *The American Biology Teacher*, *48*, 166–169.

Settlage, J. J. (1994). Conceptions of natural selection: A snapshot of the sense-making process. *Journal of Research in Science Teaching*, *31*, 449–457.

Soderburg, P. (2003). An examination of problem-based teaching and learning in population genetics and evolution using EVOLVE, a computer simulation. *International Journal of Science Education*, *25*, 35–55.

Tamir, P. (1971). An alternative approach to the construction of multiple choice test items. *Journal of Biological Education*, *5*, 305–307.

Thissen, D. (1991). MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory (Version 6.0) [Computer program]. Chicago: Scientific Software.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*, 501–519.

Thissen, D., & Steinberg, L. (1997). *A response model for multiple-choice items. Handbook of modern item response theory* (pp. 51–65). New York: Springer.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distracters are also part of the item. *Journal of Educational Measurement*, *26*, 161–176.

Treagust, D. F., & Duit, R. (2008). Conceptual change: A discussion of theoretical, methodological and practical challenges for science education. *Cultural Studies in Science Education*, *3*(2), 297–328.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 81-129). Washington, DC: American Council on Education.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport: American Council on Education.

Bryce Thomas Battisti

Department of Integrated Sciences
Asian University for Women
Chittagong, Bangladesh

Nikki Hanegan and Rex Cates

Department of Biology, College of Life Sciences
Brigham Young University
Provo, UT, 84602, USA
E-mail: nikkihanegan@byu.edu

Richard Sudweeks

Department of Instructional Psychology and Technology, McKay School of Education
Brigham Young University
Provo, UT, 84602, USA