Assessing middle and high school students' understanding of evolution with standards-based items

Jean C. Flanagan, Jo Ellen Roseman

AAAS Project 2061

NARST Annual Meeting 2011

Orlando, FL

Abstract

Understanding evolution is essential to science literacy, as it is the foundation for understanding modern biology and contributes to decision making about environmental and health policy issues. While research has shown that most students lack a correct understanding, none of the assessment tools used in this research have been tested on a large scale. A set of standards-based multiple choice items that assess evolution concepts was developed and field tested on 9,419 middle and high school students in 43 states across a broad range of demographics. Common misconceptions were used as the distractors wherever possible. Four key concepts were assessed: similarities and differences between species, a continually changing environment, the mechanism of natural selection, and the common ancestry of life. The overall mean percent correct was 42.3%. High school students (48.1% correct) performed significantly better than middle school students (38.9%). For all students, the idea of common ancestry was significantly harder than the other three ideas. We noted that a subset of the items assessing this concept, which dealt with common ancestry among both similar and different organisms, were especially difficult for students. The misconception that distantly related organisms share no similarities was found to be strong among all students.

Introduction

For about a quarter of U.S. high school graduates, general biology will be the only science course they have completed (NAEP 2005; Berkman & Plutzer 2011). High school biology is therefore a critical 'last chance' in formal schooling for many U.S. citizens to acquire science literacy. For students to make sense of biology, a basic understanding of evolution is fundamental; as Theodosius Dobzhansky famously wrote in 1973: "Nothing in biology makes sense except in the light of evolution." Accordingly, evolution is best viewed as not only a 'step' in understanding biology, but rather a central hub for connecting knowledge about cell biology, development, biochemistry, zoology, botany, ecology, and genetics (Nehm et al. 2009). Understanding evolution is especially important given that in our democratic society, citizens today must grapple with numerous policy issues that require an understanding of evolutionary processes. Environmental, healthcare, and agricultural policy debates address issues such as over-fishing, pesticide resistance in crops, biodiversity loss, and over-use of antibiotics, all of which impact or are driven by evolution. In a further complication, evolution itself – and its place in schools – has been the subject of vigorous politically-charged disputes in the U.S.; however, research has shown that many citizens on both sides of the debate do not actually understand the concept they are debating (Shtulman 2006).

Unfortunately, incorrect understandings of evolution and the process of natural selection are widespread among high school students (Clough & Wood-Robinson 1985; Settlage 1994; Demastes et al. 1995), museum visitors (Spiegel et al. 2006; Evans et al. 2010), college students (Bishop & Anderson 1990; Moore et al. 2002), biology majors (Nehm & Reilly 2007; Abraham et al. 2009), and even teachers (Nehm & Schonfeld 2007) and science graduate students (Gregory & Ellis 2009). An abundance of misconceptions has been documented at all of these education levels (for a full review, see Gregory 2009). Some of the most commonly reported misconceptions are 1) the idea that new traits appear in a population because they are needed for the population to survive (instead of by random mutation), 2) the idea that an individual organism can deliberately change to better fit its environment and pass that trait to its offspring (conflation of acclimation of an individual with adaptation of a population), 3) the idea that all members of a population gradually change to fit the environment (instead of differential survival and reproduction), and 4) the Lamarckian idea that use or disuse of a particular trait causes it to grow or atrophy accordingly and be passed down to offspring in such a state.

Given the importance of understanding evolution and the prevalence of misconceptions, there is a surprising shortage of assessment tools available to researchers and educators (Nehm & Schonfeld 2008). Only two main instruments have been developed: an essay test by Bishop and Anderson (1990) and a multiple-choice test, the Conceptual Inventory of Natural Selection (CINS), by Anderson et al. (2002). Other instruments have been at least partially based on the Bishop & Anderson essay test (e.g. Settlage 1994; Demastes et al. 1995; Nieswandt & Bellomo 2009; Nehm & Schonfeld 2007; Nehm & Schonfeld 2008). Furthermore, both the Bishop & Anderson and the CINS assessments were designed for use with undergraduate college students. To date, no assessment instrument has been developed for middle and high school students, despite the prominent presence of

evolution and natural selection ideas in national standards documents (AAAS 1993; NRC 1996). In addition, existing instruments only probe students' understanding of the mechanism of natural selection, though research has suggested that students may have even more difficulty with understanding macroevolution and common ancestry (Poling & Evans 2004; Catley & Novick 2009; Catley et al. 2010).

Here we report on the findings from a national field test of assessment items linked to ideas about evolution and natural selection. The assessment we report on here differs from prior evolution assessments present in the literature in a number of important ways: 1) it is the only assessment that has been tested on a national scale (participants came from 43 different states) with an extremely robust sample size (n = 9,419), 2) it included students in grades six through twelve, allowing us to see how the sample's ability to answer the items varied from grade to grade, 3) the items were precisely aligned to ideas from the national standards, and 4) it tested students not only on their ideas about natural selection, but also on their ideas about similarities and differences between organisms, Earth's continually changing environment, and common descent. Additionally, the assessment items were developed as part of a larger project to build a bank of valid and reliable assessment items aligned to national standards in a variety of science topics; this project employed a rigorous two-year method for item development per topic (DeBoer et al. 2008).

In this paper we discuss the fit of the items to the Rasch model, differences in student ability by grade, differences in item difficulty for each of the ideas assessed, differences in percent correct for middle and high school students, and the frequency of selection of misconceptions. We also comment on some items revealing interesting patterns of student responses.

Methods

Alignment

The items were carefully designed to align to each of four key ideas related to evolution and natural selection. The key ideas were derived from Chapter 5 of *Science for All Americans* and from Chapter 5 of *Benchmarks for Science Literacy*: Section 5F (Evolution of Life). The key ideas were:

> **Key Idea A**: There are similarities and differences among organisms living today and those that lived in the past.
>
> **Key Idea B**: Environmental conditions have changed in the past and continue to change today.
>
> **Key Idea C**: When inherited traits are favorable to individual organisms, the proportion of individuals in a population that have those traits will tend to increase over successive generations.
>
> **Key Idea D**: Similarities and differences in inherited traits of organisms alive today or in the past can be used to infer the relatedness of any two species, changes in species over time, and lines of evolutionary descent.

A clarification statement was developed for each Key Idea that further unpacked it into specific learning goals (sub-ideas) with well-defined boundaries.

Item development and pilot testing

The clarification statements provided specifications for the development of content-aligned items, where alignment required knowledge of the key idea to be both necessary and sufficient to select the correct choice and reject the distractors (Stern & Ahlgren 2002). Students were asked to select one out of four answer choices; students who chose more than one answer were marked incorrect on the item. Both documented and suspected misconceptions were used as the distractors and whenever possible a misconception was used for each incorrect answer choice. Unless a scientific term was specified in the clarification, it was not used in an item and all other vocabulary was kept simple to level the playing field for English language learners.

The items were pilot tested in the spring of 2009 and revised prior to field testing based on Rasch modeling using WINSTEPS (Linacre 2009), qualitative analysis of students' written comments and an expert review by a panel of scientists and educators. Pilot tests asked students to provide a reason for selecting or rejecting each answer choice, to circle any words they were not familiar with in the stem, and to describe anything that was confusing about the item. If Rasch analysis of pilot test data showed that students of high ability were choosing an incorrect answer, or that students of lower ability were choosing the correct answer more frequently than the higher ability students, we carefully examined students' written comments for evidence of false negatives or false positives, and revised the items based on these findings. Findings from both qualitative and quantitative analysis informed item revision: items that could not be sufficiently improved were eliminated. Occasionally new items were developed to further probe an area of interest.

Field testing

A total of 57 multiple choice items were field tested in the spring of 2010. Each student was given a test form containing 21 items. Test forms were designed to have a high degree of item overlap, so that many of the same items were present on several different forms, and five linking items were present on all forms. In constructing the different forms, we avoided including items that contained information relevant to answering other items on the same form. Six test forms were created, and a reverse-order version of each form (for a total of twelve) provided insurance against students not having time to answer the items at the ends of the tests.

Teacher feedback forms were included in the package that all participating teachers received, asking for information about class designation (e.g. GT, AP, ELL, special needs), the time it took the class to complete the test, the grade in which students at the school receive instruction in each of the Key Ideas on the test, and the textbooks used that cover those ideas.

Sample tested

The field test included a total of 9,419 students from 43 states and the District of Columbia, across a broad range of demographics. Of these, 5,875 students were in grades six through

eight (middle school), and 3,544 students were in grades nine through twelve (high school). Approximately 3,200 students responded to any one test item. Table 1 shows the breakdown of how many students were tested in each grade. About half of the students were male and about half were female. Approximately 8% of the sample indicated that English was not their first language. All students were enrolled in a science class of some kind; this class could have been earth, life, physical, environmental, biology, chemistry, physics or any other science course. Similarly, some students in the sample had previously had, some were currently having, and some had not yet had instruction on biology or evolution and natural selection.

Table 1*: Number of students who took the evolution field test, broken down by grade.*

	6th	**7th**	**8th**	**9th**	**10th**	**11th**	**12th**	MS	HS
N	1613	2225	2035	1281	907	928	430	**5875**	**3544**

Data Analysis

Rasch analysis using WINSTEPS (Linacre 2009) was used to assess the quality of the field test items. Rasch fit statistics showed us the extent to which the items were a good match to the students tested and the extent to which each item correlated with the entire set of items. Percent correct, Rasch item difficulty, and Rasch student ability means were calculated to determine if there were differences from idea to idea or from grade to grade. Either pairwise student's t-tests or a one-way ANOVA with a Bonferroni post-hoc was used to determine statistical significance when comparing means. We calculated the likelihood of selection of each of the misconceptions that appeared in the field test by dividing the number of times a misconception was chosen by the number of times it could have been chosen, averaged over the students answering the questions within this particular idea. In presenting these data, we will sometimes report on an average across all items in which the misconception was included as a distractor, and sometimes on the percent of students selecting a misconception in one particular item.

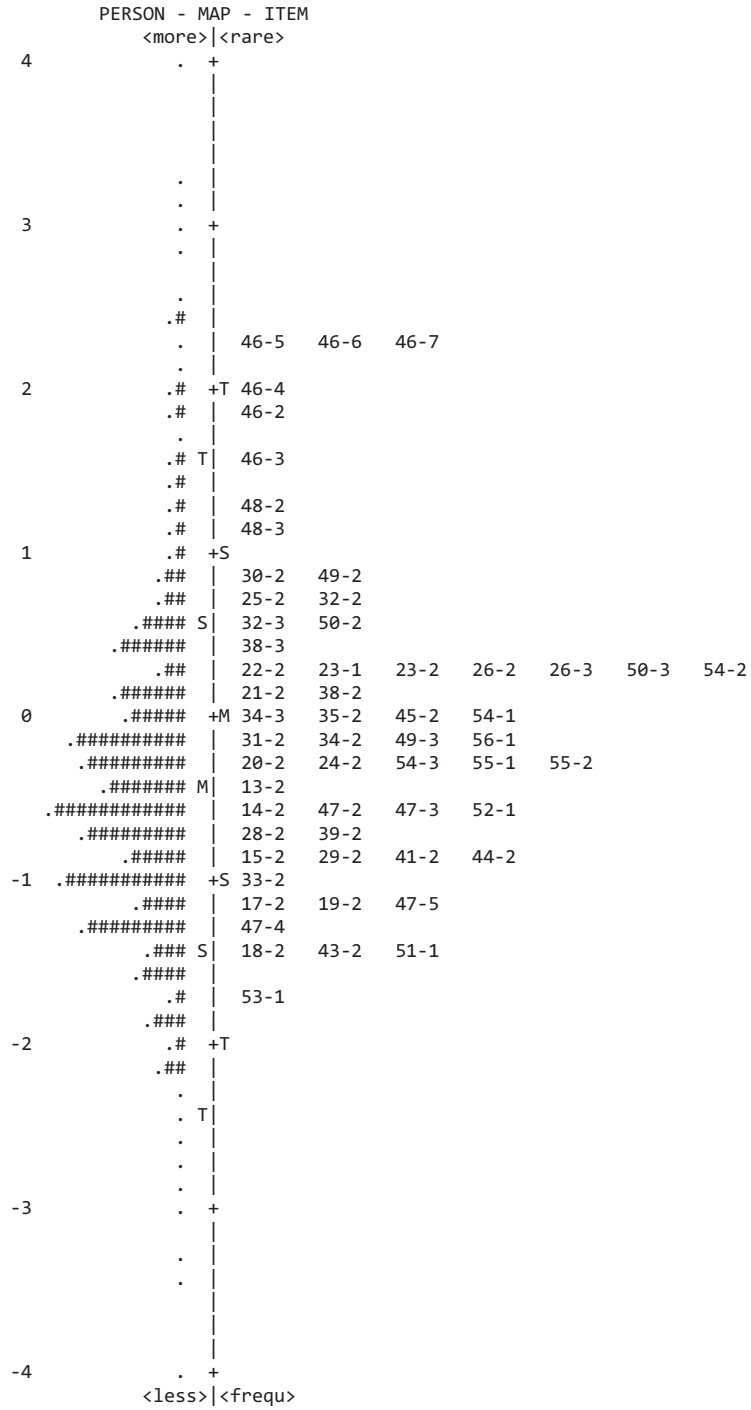
Results

Fit of the items to the Rasch model

Figure 1 shows the person-item map for all 57 items included in the field test. It shows the range of person abilities, based on the set of items, on the left side and the range of item difficulties on the right side. Persons (students) are represented by "#" or "." marks, while items are represented by their numerical identification codes. Ability and difficulty are measured on the same scale, in logits; unlike percent correct statistics, these measurements are mutually independent. Persons and items are arranged on the map such that person ability and item difficulty increase as you go towards the top of the map and decrease as you approach the bottom of the map. The mean of the item difficulties was set at zero. Where item difficulty and person ability match, the person has a 50% chance of answering the item correctly. The person-item map shows that the mean item difficulty is slightly greater than the mean student ability, but still represents an appropriate match to

the sample tested. There is a wide enough range of item difficulties to accurately measure nearly the full range of student abilities, with very few gaps from the low end through the middle to the high end.

Figure 1: *Item-person map showing student abilities on the right and item difficulties on the left. The map shows all 57 items that appeared on the field tests. Each "#" is 65 students, each "." Is 1 to 64 students, and M = mean ability (students)/difficulty (items).*



A summary of the Rasch fit statistics is provided in Table 2. An item separation index of 19.25 with a test reliability of 1.00 indicated an excellent spread of item difficulties and high reliability (Bond & Fox 2007). The somewhat lower values for person separation index (1.57) and person reliability (0.71) are partly explained by the fact that there were many more students than items and relatively few items at the highest and lowest difficulty levels to measure students at the extremes of ability.

The infit mean-square values were in the acceptable range of 0.7 to 1.3 (Bond & Fox 2007) for all items. The outfit mean-square values for all but eight of the items were also in the acceptable range of 0.7 to 1.3. As infit statistics are calculated by giving more weight to responses of persons better matched to the item difficulty, they are generally more meaningful than outfit statistics which are more susceptible to the influence of outliers (Bond & Fox 2007). Therefore, as all the items were in the acceptable range in infit mean-square values, they can be said to have a good fit to the Rasch model.

***Table 2**: Rasch fit statistics.*

	Min	Max	Median
Standard error	0.02	0.10	0.04
Infit mean-square	0.85	1.18	1.01
Outfit mean-square	0.73	2.06	1.02
Point-measure correlation coefficients	0.16	0.50	0.39
Item separation index (reliability)		19.25 (1.00)	
Person separation index (reliability)		1.57 (0.71)	

STUDENT PERFORMANCE

Percent correct

Mean percent correct for all students tested across all items was 42.3%. Mean percent correct for middle school students was 38.9%, and for high school students it was 48.1% (Table 3). Student's t-tests were used to compare the means for middle school students versus high school students on all of the items, as well as each of the key ideas. Overall, high school students performed significantly better than middle school students ($p < .01$). Broken down by key idea, the only statistically significant difference in the performance of high school students and middle school students was in Idea C ($p < .01$).

One-way ANOVA showed statistically significant differences in mean percent correct for all students between key ideas ($F = 6.96$, $p = .001$), and a Bonferroni post hoc test revealed that the mean percent correct for Idea D, which deals with common ancestry, was significantly different from the mean percent correct for Idea A ($p < .05$), Idea B ($p < .01$), and Idea C ($p < .05$).

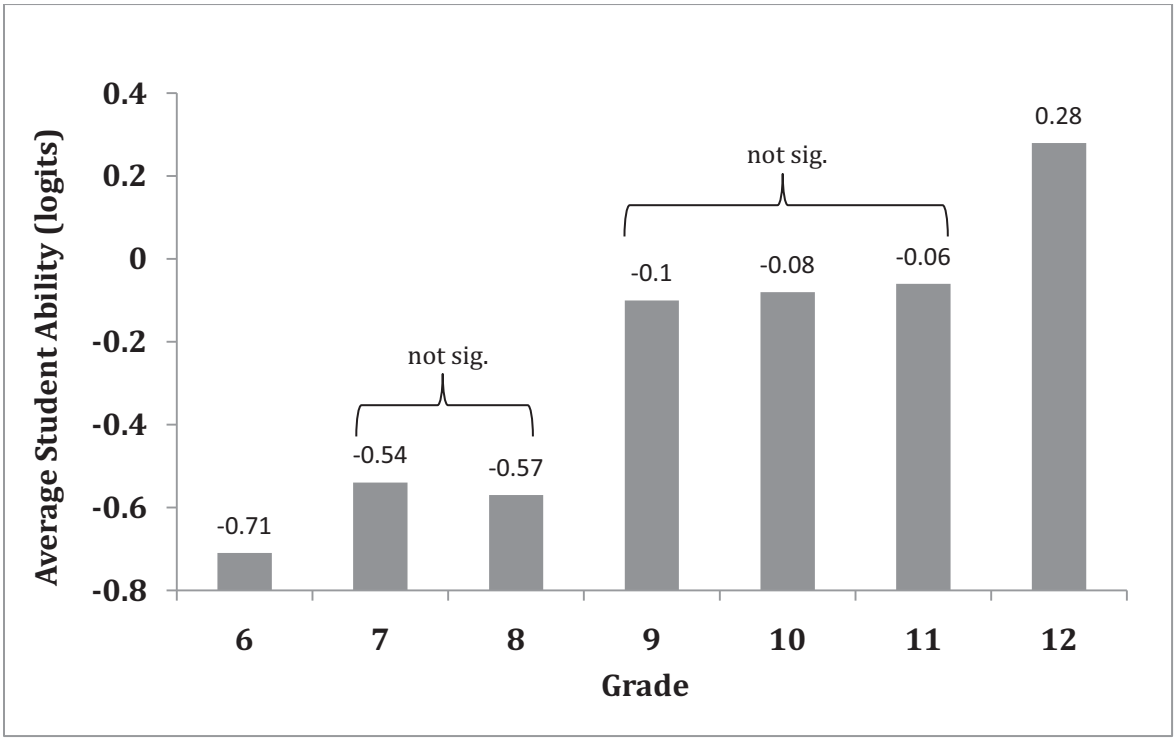
***Table 3**: Mean percent correct by key idea for middle school, high school, and all students.*

	All Items	Idea A	Idea B	Idea C	Idea D
Middle School	38.9%	48.1%	59.0%	39.3%	27.4%
High School	48.1%	55.2%	66.3%	50.0%	35.9%
All Students	42.3%	50.7%	61.7%	43.3%	30.6%

Rasch ability by grade

Significant differences in average student ability by grade were determined by pairwise student's t-tests. For our sample, average student ability by grade as measured by the evolution field test items increases significantly from sixth to seventh grade ($p = 0.000$), levels off between seventh and eighth grade, increases significantly from eighth to ninth grade ($p = 0.000$), levels off from ninth through eleventh grade, and increases significantly again in the twelfth grade ($p = 0.000$) (Figure 2). Possible reasons for this pattern are explored in the discussion section.

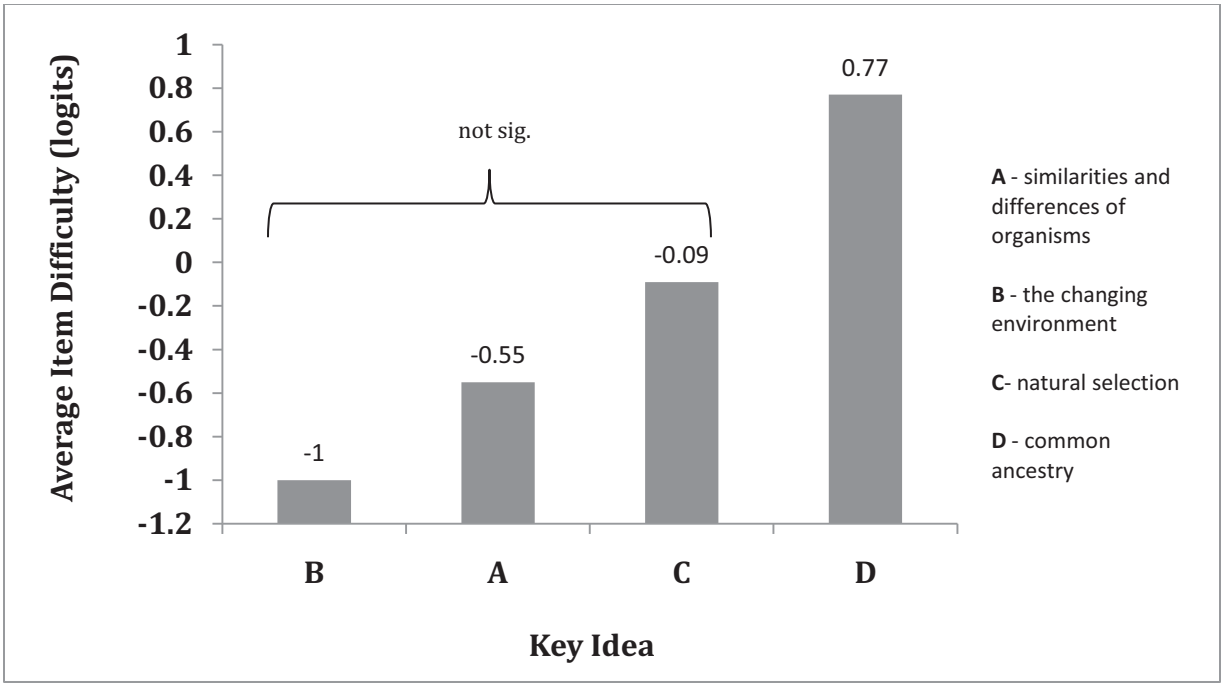
***Figure 2:** Average student ability by grade, measured in logits, for the evolution field test.*



Rasch difficulty of key ideas

Differences in average item difficulty by Key Idea were analyzed with one-way ANOVA, revealing that there were significant differences in difficulty between key ideas ($F = 8.33$, $p = 0.000$). A Bonferroni post-hoc test showed that Idea D was significantly harder for our sample of students than Ideas A and B ($p < 0.01$) and Idea C ($p < 0.05$) (see Figure 3). Although the ideas appear to progress steadily in difficulty from the means (Figure 3), considerable variability in difficulty of items within the means meant that Idea D was the only idea that was consistently more difficult and therefore significant.

Figure 3: *Average item difficulty by key idea.*



Student knowledge and misconceptions

In further analyzing the differences in the difficulties of the key ideas, we examined what students know and what misconceptions they hold within each key idea. In key ideas D and A we designed items with differing contexts that targeted the same knowledge and found a strong effect. For these ideas we will also report on the effects of various item contexts on student response patterns.

Earth's continually changing environment (Idea B): Student knowledge and misconceptions

Idea B, which assessed students' knowledge of Earth's continually changing environment, was perhaps the easiest of the four ideas for our sample, though due to a wide range of difficulties within many of the ideas, it was not significantly different from Ideas A and C in

percent correct or Rasch item difficulty. We probed students' knowledge of past and present changes to the environment and found that 70% of all students understood that the environment changed often in the past and in the present. In an item that probed the timing of these changes throughout Earth's history, 24% of students chose the distractor that 'Many changes to the physical environment happened after life began but before humans appeared; hardly any changes happened before life began or after humans appeared.' While 66% of all students recognized that environmental changes can happen suddenly or gradually, 21% thought that they could only happen gradually.

Natural selection (Idea C): Student knowledge and misconceptions

The idea that individuals of the same species do not vary in meaningful ways has been commonly reported as a misconception (e.g. Greene, 1990; Passmore & Stewart 2004; Anderson et al. 2002; Shtulman, 2006). However, across all the times that this misconception appeared as a distractor in the field test, middle school students chose it an average of 17% of the time, and high school students chose it an average 11% of the time making it one of the less popular misconceptions on the field test. In order to investigate students' conceptions of variation within species, we designed items that asked whether individual members of the same species could vary in certain characteristics that affect survival and reproduction. We found that 59% of students knew that individuals could vary in both their ability to find food and their ability to avoid predators, while only 12% said they could not vary in either of these characteristics. A similar item showed that 55% of students knew that individuals could vary in both their ability to find food and their ability to attract mates, and only 13% said they could not vary in either of these characteristics.

To probe students' ideas about the necessary components of the mechanism of natural selection, we designed three items that asked 'which of the following is needed for natural selection to occur?' In each item, students needed to recognize that 'traits must be inherited from one generation to the next' was the only choice provided that was absolutely necessary for natural selection to occur, even though the distractors could sometimes contribute to evolutionary processes. Across the three items, an average of 38% of students chose the correct answer, while an average of 31% of students chose a distractor that said sudden environmental change was what is needed for natural selection to occur, a misconception that has only recently been reported in the literature (Nehm & Reilly 2007).

In items asking for an explanation for how a population could change over time, one of the most popular misconceptions was the idea that organisms can intentionally develop new inherited traits because they are needed. Across all items that included it as a distractor, middle school students chose this misconception 39% of the time and high school students chose it 33% of the time. While many items included a distractor aligned to this misconception to compete with the correct answer, we also designed an item (Figure 5) that included an answer choice stating that both the misconception and the correct idea were true. Interestingly, this distractor was even more popular (38%) than the misconception with a denial of the correct answer (23%) and the correct answer with a denial of the misconception (27%).

***Figure 5**: An item aligned to Idea C (natural selection). The percentage of students who chose each answer choice is shown in parentheses. The correct answer is in bold text.*

Which of the following is TRUE about how a change in environmental conditions can affect a population of organisms under natural selection? Note: a population is a group of individuals of the same species.

A. It can cause individuals to change their inherited traits to become more suited to their environment, and it can cause a change in the proportion of individuals having certain traits over a few generations. (38%)
B. It can cause individuals to change their inherited traits in order to become more suited to the environment, but it cannot cause a change in the proportion of individuals having certain traits over a few generations. (23%)
C. It cannot cause individuals to change their inherited traits in order to become more suited to their environment, but it can cause a change in the proportion of individuals having certain traits over a few generations. (27%)
D. It can neither cause individuals to change their inherited traits in order to become more suited to their environment, nor can it cause a change in the proportion of individuals having certain traits over a few generations. (13%)

Another item that asked for an explanation for how a population could change over time positioned the ideas that organisms can change themselves because they need to and the Lamarckian idea of use and disuse as competing distractors. The correct explanation of natural selection was chosen by 33% of students, and the misconception that organisms can change themselves because they need to was somewhat more popular (33% chose) than the Lamarckian idea of use and disuse (23% chose). The fourth answer choice said that populations cannot change over time because members of the same species all have the same inherited traits; this distractor was only picked by 9% of students.

Common ancestry (Idea D): Student knowledge, misconceptions and item context effects

The Rasch difficulty and percent correct data both indicate that Key Idea D, which deals with common ancestry, was significantly harder for our sample than the other ideas. There was considerable range in item difficulty across the set of items targeting Idea D, which targeted the idea that all multicellular life shares a common ancestor. Items were designed to probe students' understanding of the idea and also their ability to apply it in a variety of contexts. Students performed best on the items that probed their understanding in the abstract and worst on the items that asked students to use information from the generalization to realize that specific types of organisms shared a common ancestor (Table 4).

Table 4: *Percent correct for middle and high school students on eight items from Idea D targeting the idea that all multicellular life shares a common ancestor. Shaded rows indicate a generalization, while non-shaded rows indicate a specific context.*

Item #	Item Description	Percent Correct	
		Grades 6-8	Grades 9-12
47-5	Living species can share ancestors with other living species, and they can share ancestors with extinct species.	63%	68%
47-3	A species living today and an extinct species could share a common ancestor that lived a very long time ago, even if the two species have few similarities.	50%	63%
49-3	Eagles and owls, which are different types of birds, share a common ancestor.	40%	54%
49-2	All dogs and cats share a common ancestor.	23%	32%
46-3	Cats, dogs, fish, and birds all share an ancient common ancestor.	12%	25%
46-4	Chimpanzees, humans, zebras, and worms all share an ancient common ancestor.	12%	11%
46-7	Chimpanzees, humans, chickens, and oak trees all share an ancient common ancestor.	8%	13%

The most difficult items were those that contrasted pairs of organisms that have many similarities with pairs of organisms that have fewer similarities (for an example see Figure 4). These items had the highest Rasch difficulties (Figure 1) of all the items field tested, and the lowest percent correct. On items of this type, students consistently preferred distractors that aligned with the idea that only organisms with many similarities can share a common ancestor while organisms with fewer similarities cannot.

Figure 4*: Item 46-7 from Table 4, as students would have seen it on the field test. The percentage of students who chose each answer choice is shown in parentheses. The correct answer is in bold text.*

Some organisms, such as a chimpanzee and a human, have many similarities. Others, such as a chicken and an oak tree, have fewer similarities. What is TRUE about the ancestors of these organisms?

A. Chimpanzees and humans share a common ancestor with each other, but chickens and oak trees do not share a common ancestor with each other. (62%)

B. Chimpanzees and humans share a common ancestor with each other, and chickens and oak trees share a common ancestor with each other, but chimpanzees and humans do not share a common ancestor with chickens and oak trees. (14%)

C. Because chimpanzees, humans, chickens, and oak trees are separate species, none of them shares a common ancestor with any other. (13%)

D. Chimpanzees, humans, chickens, and oak trees all share an ancient common ancestor. (10%)

Not only did providing a specific context rather than a general one have an effect on student responses, but the particular specific context used had a noticeable effect on their responses as well. A pair of items asking about the common ancestry of cats and dogs provides a striking example. In one item (Item 46-3, Table 4), the stem is constructed like that in Figure 4, with cats and dogs being the example of similar organisms, and fish and birds being the example of less similar organisms. In response to this item, 41% of students chose the distractor that said that cats and dogs had a common ancestor while fish and birds did not. Another item (Item 49-2, Table 4) had a stem that asked only about the ancestry of cats and dogs. In response to this item, 43% of students chose the distractor that said that all cats shared a common ancestor, and all dogs shared a common ancestor, but cats and dogs do not share a common ancestor with each other. In both items, students chose a distractor that said similar groups of organisms could have a common ancestor while *comparatively* different groups could not; however by following this rule of context-based subjective comparison, the students ended up contradicting themselves by saying in one item that cats and dogs do share a common ancestor, while in another saying that they do not.

A similar item that targeted the same knowledge – but had a small but important difference in the stem – showed a different student response pattern. Item 49-3 (Table 4) asked "Eagles and owls are different types of birds. What do scientists find when they trace the ancestors of eagles and owls?" Students did better on this item than any other item asking about the ancestry of specific organisms: 45% of students chose the correct answer. Only 22% of students chose the distractor that 'all eagles share a common ancestor with each other and all owls share a common ancestor with each other, but eagles and owls do not share a common ancestor with each other.' The first sentence of the stem – "Eagles and owls are different types of birds" – set up a context of similarity, i.e. belonging to the bird group, which allowed students to think they could share an ancestor.

In testing students' ability to recognize that plants and animals share a common ancestor, we investigated whether the inclusion of humans as a type of organism influenced students' response patterns and found an interesting effect. Four items targeted the idea that all plants and animals share a common ancestor (Tables 5 and 6). Two of the items (48-2 and 48-3, Tables 5 and 6) had answer choices that referred to 'all plants and animals including humans' and two (50-2 and 50-3, Tables 5 and 6) had answer choices that simply referred to 'all plants and animals.' For the two items referring to humans, the most popular answer choice was that all animals including humans share a common ancestor, and all plants share a common ancestor, but plants and animals do not share a common ancestor. For the two items referring only to 'all plants and animals,' the most popular answer choice was the correct answer that all plants and animals share a common ancestor with each other. Accordingly, students chose the misconception that plants and animals do not share a common ancestor 60% and 65% of the time in the two items that included humans, and 24% and 22% of the time in the two items that did not mention humans (Table 6). Interestingly, in these cases it was not a belief that humans do not share ancestry with animals (as this idea was included in another less popular distractor) that caused students get the items that mentioned humans wrong, but something about the added consideration of humans as animals that caused them to deny the common ancestry of plants and animals.

Table 5: *Percent correct for middle and high school students on four items from Idea D targeting the idea that all plants and animals share a common ancestor. Shaded rows indicate items testing a generalization, while non-shaded rows indicate items that explicitly included humans with animals.*

Item #	Item Description	Percent Correct	
		Grades 6-8	Grades 9-12
50-3	*Scientists think that* all plants and all animals have a common ancestor with each other.	32%	38%
50-2	All plants and all animals have a common ancestor with each other.	29%	36%
48-3	All plants and animals -- including humans -- came from a common ancestor.	18%	28%
48-2	*Scientists think that* all plants and animals -- including humans -- came from a common ancestor.	18%	24%

Table 6: *Percent of students selecting a distractor linked to the misconception that 'plants and animals cannot share a common ancestor' in response to four items aligned to Idea D. Shaded rows indicate items testing a generalization, while non-shaded rows indicate items that explicitly included humans with animals.*

Item #	Item Description	Percent choosing misconception: 'plants and animals cannot share common ancestor'
48-2	Scientists think that all plants and animals -- including humans -- came from a common ancestor.	68%
48-3	All plants and animals -- including humans -- came from a common ancestor.	65%
50-3	Scientists think that all plants and all animals have a common ancestor with each other.	24%
50-2	All plants and all animals have a common ancestor with each other.	22%

In this same set of items we also probed whether beginning an item with the phrase "scientists think that" influenced students' performance. Pairs of items were designed to be identical except for the presence or absence of this phrase (Table 5 and 6) We did this so as to give an opportunity to students who do not personally 'believe' in evolution the opportunity to demonstrate that they do know what the scientific consensus is. However, we found that the differences in student response patterns were negligible – there was essentially no difference in mean percent correct for either middle or high school students (Table 5).

Similarities and differences of organisms (Idea A): Student knowledge, misconceptions and item context effects

We found that student response patterns in some items in Idea A, which targeted students' knowledge of similarities and differences between organisms, had strong parallels to those in Idea D. The most difficult items in the set were those that probed whether students knew that visibly different organisms belonging to different taxa had both similarities and differences. As with the misconception from Idea D that only organisms with many similarities can share a common ancestor, we found the loosely-defined misconception that 'organisms that have no obvious similarities have no similarities at all' to be similarly strong, yet context-sensitive. Items 54-1, 54-2, and 54-3 (Table 7) asked if pairs of organisms had both similarities and differences, only similarities, or only differences. One item asked students to compare a plant and an animal (54-2), another a vertebrate and an invertebrate (54-1), and another a living vertebrate and an extinct vertebrate (54-3). Overall, the misconception that these pairs of organisms shared no similarities at all was chosen 39% of the time by middle school students and 31% of the time by high school students, making it one of the more popular misconceptions on the field test. As with the common ancestry items, the particular organisms that were mentioned in the item had an impact on student response patterns. In response to both the tree/lizard and the cat/worm item, students chose the misconception that there are no similarities between the organisms an average of 42% of the time. However, in response to the cow/*T. rex* pair, students chose the misconception that there are no similarities only an average of 26% of the time.

Table 7: *Percent correct for middle and high school students on four items from Idea A targeting the idea that there are similarities and differences between all multicellular organisms, living or extinct, across a range of taxa. Shaded rows indicate a generalization, while non-shaded rows indicate a specific context.*

Item #	Item Description	Percent Correct	
		Grades 6-8	Grades 9-12
53-1	There are both similarities and differences between extinct and existing species.	74%	71%
54-3	There are both similarities and differences between a cow and a *Tyrannosaurus rex*.	43%	55%
54-1	There are both similarities and differences between cats and worms.	37%	48%
54-2	There are both similarities and differences between maple trees and lizards.	33%	40%

Discussion

Differences in Rasch ability by grade

For our sample, a significant difference in average student ability was found for students completing the seventh, ninth, and twelfth grades. According to data collected on teacher feedback forms, this pattern roughly reflects the sequence of instruction our sample of students received in biology; students typically completed life science in the seventh grade, biology in the ninth or tenth grade, and an elective science course in the twelfth grade. Excluding forms in which the grade(s) when students are taught evolution ideas were not made clear, we were able to determine when the majority of our sample received instruction in life science or biology. A majority of middle school teachers (about 55%) indicated that life science was taught in the seventh grade in their schools. In high school, teacher data indicated that about 40% of students take a biology course in the ninth grade and about 54% of students take it in the tenth grade. The jump we observed in the ninth grade could be related to the type of students who take biology in ninth grade compared to those who take biology in tenth grade; in at least some school districts, ninth grade biology is taken by honors students whereas tenth grade biology is taken by regular students. As most states require two to three years of science (Education Commission of the States 2006), students in the sixth-eleventh grades were more likely taking a mandatory science class, whereas students in the twelfth grade were more likely taking an elective or advanced science class (though it was not necessarily biology). As the field tests were administered near the end of the school year (April-May), students in each grade were nearing completion of that grade. The curriculum data therefore generally corresponds with the increases in student ability to answer the field test items in the seventh, ninth and twelfth grades. This suggests that although students only show modest gains in understanding of evolution ideas from middle to high school, biology instruction may be having at least a small positive effect.

Rasch difficulty and percent correct by key idea

Average item difficulty was significantly higher for Key Idea D, which dealt with common ancestry, than for the other three ideas. Not surprisingly, the percent correct data corroborates the item difficulty data in showing that Key Idea D was significantly harder than the others for our sample. Key Idea C, which covered natural selection, was the only idea in which high school students significantly outperformed middle school students. This could be a function of increased treatment of natural selection in high school biology compared with middle school life science. The fact that high school students do not show significant improvement over middle school students in Key Idea D suggests that the idea of common ancestry is not being addressed as well in the classroom as natural selection is.

Student knowledge, misconceptions and item context effects

In our assessment of ideas relevant to natural selection, we found that relatively few students hold the commonly reported misconception that all members of the same species are effectively the same. We also found that only a small portion of students (38%) recognize that genetic inheritance is essential for the mechanism of natural selection, while a nearly comparable portion (31%) believe sudden environmental change is necessary for

natural selection. The misconception that individual organisms can change their inherited traits because they need to was one of the most popular on the field test, and many students believed it could co-exist with the correct idea that trait frequencies change over generations in populations. This suggests that many students have 'mixed models' combining correct and incorrect ideas (Nehm & Schonfeld 2010). It also indicates that sufficient connections between the topics of genetics and evolution are not being made; if students understood that genetic mutation is random and not need-based we would expect a much lower frequency of this misconception. Research has indicated that randomness is a particularly hard idea for college students to grasp (Garvin-Doxas & Klymkowsky 2008). While they may be able to identify processes such as diffusion and mutation as random, when asked to apply them to real-world situations students usually cite a driver such as need or intention. This grows out of an underlying belief that random processes are inefficient whereas living things are efficient (Garvin-Doxas & Klymkowsky 2008; Klymkowsky 2011). Preliminary research has demonstrated that reinforcement of the links between genetics and natural selection is very effective in combating need and intention-driven concepts of evolution (Kalinowski et al. 2010).

While students were reasonably good at recognizing general principles of common ancestry, they showed weak understanding when faced with specific pairs of organisms. In the items comparing cats and dogs, students thought cats and dogs shared a common ancestor when fish and birds were included in the item, but thought that cats and dogs did not share a common ancestor when no other organisms were included in the item. These results are in accordance with the findings of Poling & Evans (2004) which showed that in their sample of college-educated adults, the greater the taxonomic distance between pairs of organisms compared, the less likely participants were to agree that species pairs had common ancestors. Furthermore, our items show that the perceived taxonomic distance is completely context-based. Cats and dogs seemed more similar when compared to birds and fish than they did when compared only to each other, and this perceived level of similarity influenced the students' willingness to accept the common ancestry of these organisms. Recent research (Nehm & Ha 2011) has shown evidence for the importance of item features and contexts in natural selection assessment items. Our findings add to this the need to attend to context sensitivity in items about common ancestry. A useful assessment instrument must include a variety of items with different sets of organisms, and differing levels of information provided about the organisms (e.g. "these organisms are both mammals") in order to determine the strength of a student's understanding of the common ancestry of life.

While students may memorize that "all life shares a common ancestor," having them grapple with specific examples shows that they have trouble understanding or accepting common descent on the macro scale. A number of factors could be contributing to students' difficulties with common ancestry. In a study on student responses to different types of evolutionary diagrams, only 37% of students used terms related to ancestry (e.g. ancestor, descendant, common ancestry), while nearly 60% used terms related to anagenesis (e.g. evolved into, became) (Catley et al. 2010). Many diagrams that are not cladograms, such as the popular 'tree of life,' accidentally imply anagenesis (Catley et al. 2010), but even cladograms are ripe for misinterpretation. Many students take them to mean that the closer

two organisms on the 'tips' of the branches are, the more closely related they are, rather than reading back to how recently they shared a common ancestor (Baum et al. 2005). Additionally, more divergent examples may be difficult for students to accept as having common ancestors if they lack an understanding of macroevolutionary time. In a recent study university students were shown to provide wildly varying time ranges for events in the history of earth and the evolution of life, ranging over several orders of magnitude (Catley & Novick 2009). Overall, the data presented here add force to earlier calls for increased attention to macroevolution and common descent in the classroom (e.g. Catley 2006).

Poor student performance on items asking about similarities and differences between specific pairs of organisms sheds additional light on students' difficulties with common ancestry. The apparent inability to summon to mind fundamental similarities (such as being made up of cells) helps explain why many students struggle to see how taxonomically divergent species could possibly share a common ancestor. The fact that students did better on an item comparing two vertebrates (cow and *T. rex*) than they did on items that included invertebrates and plants highlights a continuing need to combat vertebrate-centric ideas about classification (Trowbridge & Mintzes 1988).

Students' fragmented biology knowledge

To further investigate possible explanations for students' problems in realizing that all multicellular organisms share at least some similarities and hence share a common ancestor, we examined field test data we had collected on other biology topics as part of our larger assessment project (AAAS Project 2061 n.d.). For example, items on the heredity topic had probed whether students recognized the underlying similarities in the DNA, cells, and chemical reactions of all multicellular organisms; students lacking this understanding might only rely on external features in judging whether different organisms had similarities and shared a common ancestor. However, we found that more students recognized these cellular and molecular similarities than were able to identify similarities and common ancestry on out evolution field test. From field test data on the topic of heredity, we found that 51% of all students knew that DNA is found in humans, butterflies and trees and 47% of all students knew that DNA is found in humans, dogs, and trees. By contrast, in the evolution field test, only 10% of students knew that chimpanzees, humans, chickens, and oak trees shared an ancient common ancestor. From the field test on the topic of cells, we found that 63% of all students knew that both plant and animals cells make molecules for growth; however, in the evolution field test only 32% of all students recognized that plants and animals share a common ancestor (and only 22% when humans were explicitly included with animals). These pieces of knowledge about cells and heredity seem to exist as fragments that students do not relate to ideas about common descent. This is consistent with the findings of Project 2061's textbook evaluations, which showed that textbooks typically failed to make connections among ideas (Roseman et al. 2008), and with recent reports that evolution is most often taught as a separate unit from the rest of biology, rather than as a central concept with connections to all fields of biology (Nehm et al. 2009).

Conclusions

Here we described the results of a national field test of assessment items aligned to ideas about evolution and natural selection from the national standards (AAAS 1993; NRC 1996). We found that the ability of students to answer the items correctly increased with grade, improving significantly from sixth to seventh grade, from eighth to ninth grade, and from eleventh to twelfth grade. Based on Rasch item difficulty and on percent correct we found that items aligned to the idea of common ancestry were significantly harder than items aligned to natural selection, environmental change, and similarities and differences between organisms. We also found that misconceptions related to common ancestry were among the most prevalent. A closer look showed that many students have difficulty appreciating similarities between taxonomically divergent species and in understanding their common ancestry. The context and wording of the items was also shown to have an effect on student responses.

Forced-choice multiple choice items with misconceptions for distractors gave us the opportunity to gauge the prevalence of misconceptions when they are in direct competition with the correct answer; with open response items, one cannot claim that a student would deny the correct answer, only that the answer the student provided appears to express a misconception. Research has shown that ambiguity of language is particularly problematic in the teaching, learning, and assessment of evolution and natural selection (Moore et al. 2002; Geraedts & Boersma 2006; Nehm et al. 2010), increasing the subjectivity of scoring open-response items. Student comments provided in the pilot test increased our confidence that students' conceptions actually matched the answer choices they picked. Researchers and educators will be able to use these items and take advantage of the efficient and unambiguous scoring inherent to multiple choice knowing that a careful two-year development process has shown them to be valid and reliable. Smaller-scale studies might also make use of our pilot-style format, which asked students to provide a reason for each answer choice and therefore gave us a more complete picture of students' understanding.

Considering the lingering obstacles that evolution education has faced it is perhaps not surprising that our students show a far less than satisfactory understanding of this central concept. According to a recent study, the majority of high school biology teachers in the U.S. (60%) is neither 'for' nor 'against' teaching evolution in schools, and their ambivalence takes many forms; some teach it only in context of microbes, others teach to the state test, emphasizing memorization, and still others 'teach the controversy' (Berkman & Plutzer 2011). However, given its importance in understanding modern science and policy, educators and researchers need to work quickly to diagnose and correct problems in students' understanding of evolution. The assessment items reported on here represent a valuable tool for diagnosing students' thinking and for comparing the effectiveness of instructional approaches or curriculum materials. In addition, the results of the national field test summarized above can help educators, curriculum developers, and researchers zero in on areas that are most problematic for students.

References

AAAS Project 2061 (n.d.) [Pilot and field test data collected between 2006 and 2010]. Unpublished raw data.

Abraham, J.K., Meir, E., Perry, J., Herron, J.C., Maruca, S., Stal, D. (2009). Addressing undergraduate student misconceptions about natural selection with an interactive simulated laboratory. *Evolution Education and Outreach*, 2:393-404.

American Association for the Advancement of Science. (1989). *Science for all Americans*. New York: Oxford University Press.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

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(10):952-978.

Baum, D.A., Smith, S.D., & Donovan, S.S.S. (2005). The tree-thinking challenge. *Science*, 310:979-980.

Berkman, M.B., & Plutzer, E. (2011). Defeating creationism in the courtroom, but not in the classroom. *Science*, 331(6016):404-405.

Bishop, B.A., & Anderson, C.W. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching*, 27(5):415-427.

Bizzo, N.M.V. (1994). From down house landlord to Brazilian high school students: What has happened to evolutionary knowledge along the way? *Journal of Research in Science Teaching*, 31(5):537-556.

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. 2nd Ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Brumby, M. (1979). Problems in learning the concept of natural selection. *Journal of Biological Education*, 13(2):119-122.

Catley, K.M. (2006). Darwin's missing link – a novel paradigm for evolution education. *Science Education*, 90:767-783.

Catley, K.M., & Novick, L.R. (2009). Digging deep: Exploring college students' knowledge of macroevolutionary time. *Journal of Research in Science Teaching*, 46(3):311-332.

Catley, K.M., Novick, L.R., & Shade, C.K. (2010). Interpreting evolutionary diagrams: When topology and process conflict. *Journal of Research in Science Teaching*, 47(7):861-882.

Clough, E.E., & Wood-Robinson, C. (1985). How secondary student interpret instances of biological adaptation. *Journal of Biological Education*, 19: 125-130.

DeBoer, G., Herrmann Abell, C., Gogos, A., Michiels, A., Regan, T., & Wilson, P. (2008). Assessment linked to science learning goals: Probing student thinking through assessment. In R. Douglas, J. Coffey, & C. Stearns (Eds.), *Linking science and assessment*. Arlington, VA: NSTA Press.

Demastes, S.S., Settlage, J., & Good, R. (1995). Students' conceptions of natural selection and its role in evolution: Cases of replication and comparison. *Journal of Research in Science Teaching*, 32(5):535-550.

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

Evans, E.M., Spiegel, A.N., Gram, W., Frazier, B.N., Tare, M., Thompson, S., & Diamond, J. (2010). A conceptual guide to natural history museum visitors' understanding of evolution. *Journal of Research in Science Teaching*, 47(3):326-353.

Garvin-Doxas, K., & Klymkowsky, M.W. (2008). Understanding randomness and its impact on student learning: lessons learning from building the biology concept inventory (BCI). *Cell Biology Education – Life Sciences Education*, 7:227-233.

Greene, E.D. (1990). The logic of university students' misunderstanding of natural selection. *Journal of Research in Science Teaching*, 27(9):875-885.

Gregory, T.R., & Ellis, C.A.J. (2009). Conceptions of evolution among science graduate students. *Bioscience*, 59(9):792-799.

Gregory, T.R. (2009). Understanding natural selection: Essential concepts and common misconceptions. *Evolution Education and Outreach*, 2: 156-175.

Gregory, T.R., & Ellis, C. (2009). Conceptions of evolution among science graduate students. *Bioscience*, 59(9):792-799.

Ha, M., & Cha, H. (2008). Suggestion of a new strategy to teach evolution. Proceedings of the NARST Annual Meeting: Baltimore, MD.

Kalinowski, S.T., Leonard, M.J., & Andrews, T.M. (2010). Nothing in evolution makes sense except in the light of DNA. *Cell Biology Education – Life Sciences Education*, 9:87-97.

Klymkowsky, M. (2011). Why is evolution so hard to understand? *ASBMB Today*, March 2011:14-15.

Linacre, J.M. (2009). WINSTEPS Rasch Measurement Computer Program (Version 3.69.1.8). Chicago: Winsteps.com.

Moore, R., Mitchell, G., Bally, R., Iglis, M., Day, J., & Jacobs, D. (2002). Undergraduates' understanding of evolution: ascriptions of agency as a problem for student learning. *Journal of Biological Education*, 36(2):65-71.

National Center for Education Statistics, U.S. Department of Education, National Assessment of Educational Progress. (2005). High School Transcript Study (HSTS), http://nces.ed.gov/nationsreportcard/hsts/.

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

Nehm, R.H., & Reilly, L. (2007). Biology majors' knowledge and misconceptions of natural selection. *Bioscience*, 57(3):263-272.

Nehm, R.H., & Schonfeld, I.S. (2007). Does increasing biology teacher knowledge of evolution and the nature of science lead to greater preference for the teaching of evolution? *Journal of Science Teacher Education*, 18(5):699-723.

Nehm, R.H., & Schonfeld, I.S. (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching*, 45(10):1131-1160.

Nehm, R.H., Poole, T.M., Lyford, M.E., Hoskins, S.G., Carruth, L., Ewers, B.E., Colberg, P.J.S. (2009). Does the segregation of evolution in biology textbooks and introductory courses reinforce students' faulty mental models of biology and evolution? *Evolution Education and Outreach*, 2:527-532.

Nehm, R.H., & Schonfeld, I.S. (2010). Reply: The future of natural selection knowledge measurement: A reply to Anderson et al. (2010). *Journal of Research in Science Teaching*, 47(3):358-362.

Nehm, R.H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, 48(3):237-256.

Passmore, C., & Stewart, J. (2002). A modeling approach to teaching evolutionary biology in high schools. *Journal of Research in Science Teaching*, 39(3):185-204.

Poling, D.A., & Evans, E.M. (2004). Religious belief, scientific expertise, and folk ecology. *Journal of Cognition and Culture*, 4(3):485-524.

Roseman, J.E., Linn, M.C., & Koppal, M. (2008). Characterizing curriculum coherence. In Y. Kali, M. C. Linn, and J. E. Roseman (Eds.) *Designing Coherent Science Education*. *Implications for Curriculum, Instruction, and Policy.* New York: Teachers College Press.

Settlage, J. (1994). Conceptions of natural selection: A snapshot of the sense-making process. *Journal of Research in Science Teaching*, 31(5):449-457.

Shtulman, A. (2006). Qualitative differences between naïve and scientific theories of evolution. *Cognitive Psychology*, 52:170-194.

Spiegel, A.N., Evan, M., Gram, W., & Diamond, J. (2006). Museum visitors' understanding of evolution. *Museums & Social Issues*, 1(1):69-86.

Stern, L. & Ahlgren, A. (2002). Analysis of students' assessments in middle school curriculum materials: aiming precisely at benchmarks and standards. *Journal of Research in Science Teaching*, 39(9):889-910.

Stern, L., & Hagay, G. (2005). [High school students' conceptions related to speciation and common descent]. Unpublished raw data.

Trowbridge, J.E., & Mintzes, J.J. (1988). Alternative conceptions in animals classification: A cross-age study. Journal of Research in Science Teaching, 25(7):547-571.