Measuring Knowledge of Natural Selection: A Comparison of the CINS, an Open-Response Instrument, and an Oral Interview

Ross H. Nehm,[1] Irvin Sam Schonfeld[2]

*[1]College of Education and Human Ecology, Department of Evolution,
Ecology, and Organismal Biology, The Ohio State University,
333 Arps Hall, Columbus, Ohio 43210*
*[2]School of Education, Department of Psychology, The City College,
City University of New York, New York, New York*

Received 14 March 2007; Accepted 9 November 2007

Abstract: Growing recognition of the central importance of fostering an in-depth understanding of natural selection has, surprisingly, failed to stimulate work on the development and rigorous evaluation of instruments that measure knowledge of it. We used three different methodological tools, the Conceptual Inventory of Natural Selection (CINS), a modified version of Bishop and Anderson's (Bishop and Anderson [1990] Journal of Research in Science Teaching 27: 415–427) open-response test that we call the Open Response Instrument (ORI), and an oral interview derived from both instruments, to measure biology majors' understanding of and alternative conceptions about natural selection. We explored how these instruments differentially inform science educators about the knowledge and alternative conceptions their students harbor. Overall, both the CINS and ORI provided excellent replacements for the time-consuming process of oral interviews and produced comparable measures of key concept diversity and, to a lesser extent, key concept frequency. In contrast, the ORI and CINS produced significantly different measures of both alternative conception diversity and frequency, with the ORI results completely concordant with oral interview results. Our study indicated that revisions of both the CINS and ORI are necessary because of numerous instrument items characterized by low discriminability, high and/or overlapping difficulty, and mismatches with the sample. While our results revealed that both instruments are valid and generally reliable measures of knowledge and alternative conceptions about natural selection, a test combining particular components of both instruments—a modified version of the CINS to test for key concepts, and a modified version of the ORI to assess student alternative conceptions—should be used until a more appropriate instrument is developed and rigorously evaluated. © 2008 Wiley Periodicals, Inc. J Res Sci Teach 45: 1131–1160, 2008

Keywords: evolution; natural selection; alternative conceptions; assessment

For more than a century, evolutionary theory has without fail demonstrated its explanatory power and practical utility in a diverse array of empirical contexts within a growing number of scientific disciplines (Gould, 2002; Kitcher, 2007). Nevertheless, learners at all levels of the educational hierarchy in the United States are characterized by low levels of understanding and acceptance of evolution, as well as myriad alternative conceptions (e.g., Bishop & Anderson, 1990; Brooks, 2001; Brumby, 1984; Clough & Wood-Robinson, 1985; Dagher & BouJaoude, 1997; Demastes, Settlage, & Good, 1995; Grose & Simpson, 1982; Nehm & Reilly, 2007; Nehm & Schonfeld, 2007; Newport, 2004; Sinatra, Southerland, McConaughy, & Demastes, 2003; Zimmerman, 1987). The increasing importance of evolutionary theory within the biological sciences, paradoxically coupled with growing public resistance to it, has focused considerable attention on the teaching and learning of evolution (Donnelly & Boone, 2007; Nehm, 2006).

The tenacity of evolutionary alternative conceptions in the face of innovative science instruction (Nehm & Reilly, 2007), and the persistence of antievolutionary attitudes among educated adults (Newport, 2004)

Correspondence to: R.H. Nehm; E-mail: nehm.1@osu.edu
DOI 10.1002/tea.20251
Published online 7 May 2008 in Wiley InterScience (www.interscience.wiley.com).

have spurred several research programs to explicate the causes of resistance to evolution, and shed light on solutions to the problem of this resistance. Specifically, scholarly work has begun to explore (1) the precise interrelationships among cognitive, affective, epistemological, and religious variables that contribute to antievolutionary views in individuals of different ages and educational backgrounds (e.g., Colburn & Henriques, 2006; Dagher & BouJaoude, 1997; Ingram & Nelson, 2006; Sinatra et al., 2003; Southerland & Sinatra, 2003); (2) the design, implementation, and evaluation of interventions that promote accurate cognitive models of evolution (Crawford, Zembal-Saul, Munford, & Friedrichsen, 2005; Jackson, Doster, Meadows, & Wood, 1995; Nehm & Reilly, 2007; Scharmann, 1993); and (3) methods for reducing levels of antievolutionary attitudes in students and teachers (Nehm & Schonfeld, 2007; Scharmann & Harris, 1992).

Evolution in general and natural selection in particular continue to function as the conceptual framework upon which the ever-expanding domain of contemporary biological science is built (Gould, 2002; Kitcher, 2007). Likewise, evolution and natural selection serve as unifying concepts that provide curricular coherence for the life science strands of the National Science Education Standards as well as many state standards (Donnelly & Boone, 2007; Lerner, 2000; NRC, 1996; Skoog & Bilica, 2002). Although the impact of the national and state standards pertaining to evolution is debatable (Donnelly & Boone, 2007; Moore, 2001; Skoog & Bilica, 2002), no science or science education organizations have questioned the inclusion, or importance, of evolution and natural selection in these standards.

Growing recognition of the central importance of fostering an in-depth understanding of evolution and natural selection has, surprisingly, failed to stimulate work on the development and rigorous evaluation of instruments that measure knowledge of, and alternative conceptions about, natural selection in learners of different ages and educational backgrounds (Liu & Lesniak, 2005; Nehm, 2006; NRC, 2001). This dearth of attention to assessment makes evaluation of the effectiveness of national and state standards, as well as particular pedagogical strategies used to teach them, difficult if not impossible. Indeed, only two instruments for measuring knowledge of natural selection have been developed: (1) Bishop and Anderson's (1990) essay instrument and (2) the Conceptual Inventory of Natural Selection (CINS; Anderson, Fisher, & Norman, 2002). Both instruments were developed for populations of undergraduate non-majors; no validated instruments are available for use on undergraduate majors or biology teachers.

Despite their extensive use by science educators (Anderson et al., 2002), the CINS and the Bishop and Anderson essay test remain in need of rigorous validation and replication. Indeed, despite criticisms in the literature (Anderson et al., 2002; Nehm, 2006), no empirical work has explored the putative deficiencies of these instruments, and nothing is known about how they differentially capture knowledge of and alternative conceptions about natural selection. Additionally, neither instrument has been employed to measure knowledge of, or alternative conceptions about, natural selection in samples primarily comprised of minority undergraduates. Finally, no work has explored whether these instruments are suitable for the assessment of evolutionary knowledge in biology majors.

This study focuses on the conceptual understanding of natural selection in minority undergraduate biology majors, many of whom will go on to become science teachers, medical professionals, and scientists. Specifically, we attempt to assess biology majors' understanding of, and alternative conceptions about, natural selection using three different methodological tools: (1) the CINS (Anderson et al., 2002), (2) an Open Response Instrument (ORI) that was partly derived from Bishop and Anderson's (1990) essay test (Nehm & Reilly, 2007); and (3) an extended oral interview based on questions from both instruments. We explore how these different tools differentially inform science educators about the knowledge and alternative conceptions that students harbor, and whether these instruments are suitable for use with samples of undergraduate biology majors.

Materials and Methods

Sample Characteristics

This study was conducted using two samples (classes) of biology majors in their second semester at a large minority-serving urban university located in the northeastern United States. Enrollment in the second-semester course was contingent upon successful completion of the first semester of introductory biology, which covered genetics and cell biology. Student data collected in the second-semester course took place after instructional units on evolution and natural selection were completed. The racial and ethnic distribution of students in both samples closely approximated that of science majors at the institution (Hispanic: 32.5%; African-American: 30.12%; Asian: 25.5%; native Americans: 0.09%; non-Hispanic White: 11.75%). Compared to participants in previous evolution-education studies, participants in this study are different: they are mostly minority undergraduate science majors; they are of slightly older age; a greater proportion is female. These differences may restrict the comparability of this study to studies of other student populations and learning contexts.

In Sample G, 100 students voluntarily completed the CINS and ORI (99% participation rate) and 18 students volunteered for associated oral interviews. The mean age of students was 21 years ($SD = 4.37$) and their ages ranged from 17 to 35 years. Females comprised 60.6% of the sample. In Sample N, 82 students completed the ORI only (82% response rate). Instructor-imposed time constraints prevented the implementation of the CINS in Sample N. As in Sample G, the mean age of students in Sample N was 21 years ($SD = 4.24$) and their ages ranged from 17 to 36 years. Sample N was also mostly (60.6%) female. The difference in the proportions of non-Hispanic White students in Samples N and G (10% vs. 15%, respectively) was non-significant; additionally, the percentage of students who self-reported being taught natural selection in high schools (99% vs. 84%) was statistically significant ($p < 0.01$).

Characteristics of the Open Response Instrument (ORI) and the Conceptual Inventory of Natural Selection (CINS)

Two paper-and-pencil instruments were used to measure student knowledge and alternative conceptions of natural selection. The first, referred to as the Open Response Instrument (hereafter: ORI), was developed using questions from earlier studies: three questions from Bishop and Anderson's (1990) essay test and two questions from Nehm and Reilly (2007). The instrument thus comprised five open-ended essay questions: (1) Please define natural selection to the best of your ability; (2) Explain why some bacteria have evolved a resistance to antibiotics (i.e., the antibiotics no longer kill the bacteria); (3) Cheetahs (large African cats) are able to run faster than 60 miles per hour when chasing prey. How would a biologist explain how the ability to run fast evolved in cheetahs, assuming their ancestors could run only 20 miles per hour? (4) Cave salamanders (amphibian animals) are blind (they have eyes that are not functional). How would a biologist explain how blind cave salamanders evolved from ancestors that could see? (5) If biologists wanted to speed up evolutionary change, how would they do it? Students were asked to ''Be as complete as you can'' both on the instrument and in an oral script, and given one-half page of empty space to answer each question. Just prior to the implementation of the instrument, the proctors also announced ''Please answer as completely as possible.'' The ORI was designed to be completed during class in 25 minutes or less, and nearly all students who participated in a prior study were able to provide detailed responses to the questions under these conditions (Nehm & Reilly, 2007).

The ORI was designed to measure undergraduate biology majors' knowledge about natural selection at differing levels of complexity. The five open-response questions on the instrument (see above) were ordered such that they began by requesting familiar concrete knowledge (e.g., ''define natural selection'') and ended with unfamiliar abstract problem-solving questions (e.g., ''how might a biologist try to speed up evolutionary change?''). These questions, all of which focused on natural selection, were designed to parallel Bloom's Taxonomy, which organizes different types of categories of knowledge by levels (Bloom, 1956). For example, Bloom's Level 1 Knowledge category includes cognitive tasks involving the observation and recall of information, such as an awareness of dates, events, places, and major ideas. In contrast, the Level 4 Application category includes tasks such as employing information, methods, or concepts in new situations to solve problems, which is an example of higher-order thinking. Thus, the ORI attempted to gauge students' sophistication at solving different types of evolutionary problems.

The second instrument used in this study was the closed-response (multiple-choice) Conceptual Inventory of Natural Selection (hereafter: CINS) which was developed to measure similar types of knowledge as the ORI (Anderson et al., 2002). The CINS offered one correct response option for each question ($n = 20$) as well as a series of distractors derived from well-documented alternative conceptions (Anderson et al., 2002).

The ORI and CINS differ in several respects: (1) number of questions (5 vs. 20); (2) format of the questions (open- vs. closed-response); and (3) use of alternative conception distractors (absent in the ORI).

Additionally, the ORI was developed in order to measure potential learning gains in biology majors, whereas the CINS was employed and validated on a cohort of non-majors. The two instruments are nevertheless similar in several respects: (1) both attempt to measure knowledge of natural selection; (2) both were designed to be completed in less than 30 minutes; and (3) both were designed for use with samples of college undergraduates.

The content validity of both instruments is reflected in their having been designed by subject area experts to cover key concepts specific to the domain of knowledge we call natural selection (Mayr, 1982; also see the Section called Variable Coding and Tabulation). Moreover, the ORI was developed with the twin content-validational aims of accurately identifying both student knowledge and student alternative conceptions (see Variable Coding and Tabulation Section).

Discriminant Validity Instrument

In order to explore the discriminant validity of the CINS and ORI, we developed a multiple-choice science test on a subject other than evolution. This test was developed using questions from the New York State Regents earth science exams from 2001 to 2006. The Regents exam questions were designed to measure earth science knowledge in high school students who have taken a year-long earth science class. We considered a high-school-level, closed-response test on rocks to be appropriate for our study of discriminant validity for several reasons: (1) the vast majority of undergraduates in our sample did not take the earth science Regents coursework or exams; (2) the test questions are of a level appropriate for first-year college students; and (3) the test was similar in format and structure to the CINS, and therefore might expose students who are good at taking multiple-choice exams despite little knowledge of rocks. The closed-response, ten-question rock test was administered to all oral interview participants ($n = 18$). We assessed discriminant validity by examining the correlation between participant scores on the CINS and ORI and the rock test. Cases of significant correlations between the rock test, on the one hand, and the CINS and ORI, on the other, would call into question the construct validity of the CINS and ORI.

Oral Interviews

Oral interviews were conducted on the voluntary participants from Sample G (see above). The interviews lasted approximately 25 minutes (mean = 24.8 minutes, min. = 15.5, max. = 39.1). These participants exhibited a broad array of scores on both the CINS (30–100%, mean = 66%) and the ORI (58–100%, mean = 80%). To ensure content validity the oral interview was designed such that it comprised four questions, two from Bishop and Anderson's original essay test (1990), one from the CINS, and one from the ORI (1) "A number of mosquito populations no longer die when DDT (a chemical used to kill insects) is sprayed on them, but many years ago DDT killed most mosquitoes. Could you explain why many mosquitoes don't die anymore when DDT is sprayed on them?"; (2) "Seals can remain underwater without breathing for nearly 45 minutes as they hunt for fish. How would a biologist explain how the ability to not breathe for long periods of time has evolved, assuming their ancestors could stay underwater for just a few minutes?"; (3) Question 13 from the CINS (Anderson et al., 2002); and (4) Question 4 from the ORI (see above).

Variable Coding and Tabulation

The first set of variables extracted from the ORI related to student knowledge of the seven key concepts of natural selection (Mayr, 1982; hereafter: key concepts): (1) the causes of phenotypic variation (e.g., mutation, recombination, sexual reproduction); (2) the heritability of phenotypic variation; (3) the reproductive potential of individuals; (4) limited resources and/or carrying capacity; (5) competition or limited survival potential; (6) selective survival based on heritable traits; (7) a change in the distribution of individuals with certain heritable traits.

A coding rubric was developed, piloted, refined, and used to score student responses such that the use of a key concept in an explanation of evolutionary change counted as one point. Thus, an essay response that employed all seven key concepts received seven points. After initial scoring, the essays were blindly recoded using the same rubric in order to test the precision of the coding rubric and the consistency of the raters. Scorer reliability or cross-rater consistency was assessed by calculating the

Pearson correlation coefficients between the number of key concepts identified by each of the two raters. The results indicated statistically significant correlations for both of the ORI questions scored for key concepts ($r = 0.78$, $p < 0.001$; $r = 0.77$, $p < 0.001$). Thus, the scoring rubric appeared to be sufficiently clear, and the raters sufficiently consistent, for reliable coding of key concepts. The coding rubric was used to quantify the presence or absence of the seven key concepts in each of the students' five essay questions (see above). These scores were tallied separately for each question, collectively for each student, and collectively for *all* students. In addition, the number of *different* key concepts used among all five questions (hereafter: key concept diversity) was scored for each student and collectively for all students.

The second set of variables extracted from the ORI related to student alternative conceptions concerning natural selection (hereafter: alternative conceptions). A coding rubric was developed that contained alternative conceptions about natural selection and evolution. These alternative conceptions have been extensively documented in the research literature (e.g., Bishop & Anderson, 1990; Nehm & Reilly, 2007). This rubric was used in order to measure the magnitude and distribution of commonly observed student alternative conceptions (e.g., needs cause evolutionary changes to take place; the use or disuse of traits explains their appearance/disappearance; traits appear only when they are needed; all individuals in a population develop new traits simultaneously, etc., [see Bishop & Anderson, 1990; Nehm & Reilly, 2007]), as well as to capture any novel alternative conceptions elicited by the instrument. Cross-rater consistency in enumerating alternative conceptions was assessed using the Pearson correlation coefficient. Two raters blindly coded the number of alternative conceptions in the salamander essay question. The analysis produced statistically significant correlations between the raters' scores ($r = 0.82$, $p < 0.01$). Thus, the scoring rubric appeared to be sufficiently clear, and the raters sufficiently consistent, for reliable coding of alternative conceptions.

Student responses were scored such that the use of an identifiable alternative conception in an evolutionary explanation counted as one point, with no upper limit to the number of alternative conceptions recognized per essay. Unanticipated (i.e., unspecified on the rubric) alternative conceptions captured in this manner included (1) survival of the fittest means survival of the fittest *species*; (2) "fit" = dominant and "unfit" = recessive, in the allelic sense; (3) "genetic drift" = gene flow between different species; (4) drastic climate change is required for evolution to occur; (5) heritable "compensation" of one trait occurs when another faculty is lost, for example, "super" hearing or smell was attributed to suddenly blind salamanders; and (6) humans are incapable of affecting evolution in any way. The coding rubric was used to quantify the presence or absence of these alternative conceptions, and scores were tallied for each question, collectively for each student, and collectively for *all* students. In addition, the number of *different* alternative conceptions used among all five questions (hereafter: alternative conception diversity) was scored for each student and collectively for all students.

In addition to measuring student performance using separate key concept and alternative conception measures, we used the Natural Selection Performance Quotient (NSPQ) of Nehm and Reilly (2007) to quantify student knowledge and alternative conceptions. The NSPQ takes a ratio of key concept diversity to the sum of key concept diversity plus alternative conception diversity, and multiplies it by the ratio of key concept diversity to total possible key concepts, and thereby produces a single grade-like score (i.e., 0–100). The first term expresses the proportion of the students' answers that were correct, and the second expresses how the correct proportion compared to the most complete possible answer (Nehm & Reilly, 2007). Exponents were chosen to calibrate the NSPQ scale (see Nehm & Reilly, 2007) such that it conformed to our assessment that four key concepts would result in a score above 65 (i.e., passing). In addition to permitting the visualization of student knowledge on a single scale, the NSPQ also distinguishes clearly between students who have problems with their understanding of natural selection, despite displaying significant knowledge, and those students with no alternative conceptions who displayed differing levels of knowledge (Nehm & Reilly, 2007, Table 3, p. 267).

Scoring of the closed-response CINS instrument was more straightforward than scoring the ORI. All correct CINS responses represented key concepts and all incorrect CINS responses represented alternative conceptions. Because the CINS contained the same key concept and alternative conception options in several questions, it was possible to calculate a standardized measure of key concept diversity and alternative conception diversity (in parallel to the ORI diversity variables discussed above). That is, we measured key concept and alternative conception diversity for each student and collectively among all students.

Scoring of the rock test was based on a percentage score, where 100% represented 10 correct items.

The coding of the oral interview was similar to that of the ORI. A coding rubric was developed, piloted, refined, and used to score student responses such that the use of a key concept in an explanation of evolutionary change counted as one point (up to seven total, see above) and the use of an identifiable alternative conception counted as one point, with no upper limit to the number of alternative conceptions recognized per answer. Tallies were made for the number and diversity of key concepts and alternative conceptions per person. Additionally, we calculated an overall interview score: $-1 =$ clear evidence of a faulty mental model of natural selection and numerous alternative conceptions; $0 =$ ambiguous evidence: some correct concepts present, some alternative conceptions present, but unclear evidence whether an accurate mental model of natural selection is being employed; and $1 =$ clear, unambiguous evidence of an accurate working model of natural selection lacking alternative conceptions. Cross-rater consistency of oral interview scores was assessed by calculating Spearman rank-order correlations on overall interview scores (i.e., -1, 0, or 1). Ten randomly selected interviews were recoded by two independent raters and found to be very closely matched ($n = 10$, $r = 0.96$, $p < 0.001$). Finally, we note that the oral interviews were blindly coded prior to comparisons to the ORI, CINS, and rock test scores.

Methodological Framework and Analyses

Both Classical Test Theory (CTT) and Item Response Theory (IRT) were used as analytical and methodological frameworks for our study. Although IRT is currently considered a more advantageous psychometric approach (Bond & Fox, 2001), Anderson et al. (2002) used CTT to evaluate and validate the CINS, and we wanted to compare results from our study to theirs using the same methodological framework. CTT explores item attributes (e.g., item difficulty) within the context of the particular population sampled (Libarkin & Anderson, 2006). IRT, in contrast, assumes that the characteristics of particular instrument items are independent of the abilities of the sample participants (Libarkin & Anderson, 2006). CTT is a traditional psychometric approach that we used as a context for documenting instrument attributes such as reliability and validity. Specific measures included the calculation of item discriminability and difficulty, and factor analyses of response correlation patterns. Variable means and correlations, question discriminability, difficulty, and reliability (Cronbach's α) and principal components analysis (PCA) were calculated for the CINS and ORI data using SPSS 12.0.

Rasch analysis (Boone & Scantlebury, 2006) is an increasingly common IRT approach for investigating the reliability and validity of instruments in science education (Liu & Boone, 2006, chapters therein). Rasch analysis produces conceptually analogous, but empirically different, measures of reliability, validity, and item difficulty than traditional CTT measures (Bond & Fox, 2001). We used WINSTEPS software (Linacre, 2006) to perform the Rasch analysis, therein quantifying item difficulty, redundancy, and person-item relationships (Boone & Scantlebury, 2006). Finally, we explored the concordance and discordance of results derived using traditional and Rasch model analyses.

Results

Open Response Instrument

Quantitative Measures of Student Knowledge Using the ORI. The ORI was used to measure second-semester biology majors' knowledge and alternative conceptions of natural selection. Key concept diversity scores derived from the ORI revealed acceptable averages: 4.33 for Sample N and 3.78 for Sample G. Similarly, in Sample N, 70% of students employed four or more key concepts and in Sample G 58% of students employed more than four key concepts. Thus, on average, biology majors were using about four (out of a possible seven) key concepts of natural selection in their explanations of evolutionary change. We illustrate several examples of participant key concept use below:

> "Natural selection is the unequal survival at reproduction among individuals in a population due to having unique traits that will be favored or unfavored by the environment." (Sample G, #36. key concepts: differential survival, variation).

> “It is likely that cheetahs are competing with other predators while chasing prey. Through mutations, some cheetahs developed the ability to run extremely fast, 60 miles per hour, and it was these cheetahs that were able to find food, survive, have offspring, and pass on their genetic athleticism on to their offspring.” (Sample G, #40. key concepts: causes of variation, heritability, selective survival, resources, competition).

> “The bacteria that had no resistance to a certain antibiotic died quickly, and the bacteria that had some resistance survived and without competition from the bacteria with no resistance, the bacteria with resistance increased in ratio to those without resistance.” (Sample G, #32. key concepts: competition, selective survival, change in the frequency of individuals in a population).

> “Blind cave salamanders must be a product of genetic drift. As their ancestors populated the cave (which was a new territory at first), the genetic variation was reduced compared to that of their home population. Over time this caused a change in the allele frequency. This factor as well as mutation may have led to their having eyes that are not functional. Since the cave is dark the mutation did not hinder their survival” (Sample G, #36. key concepts: genetic variation, inheritance, selective survival based on heritable traits).

Alternative conception scores revealed a different result than key concept results. Mean alternative conception diversity was relatively high in both samples: 1.9 in Sample G and 2.4 in Sample N. Additionally, only 14% of students in Sample G and 30% of students in Sample N employed *no* alternative conceptions on the ORI. NSPQ averages, which provide a single metric encompassing both key concept and alternative conception diversity on a 0–100 scale, were 74 and 79 for Samples G and N, respectively. Overall, while key concept use was relatively high, alternative conceptions formed a large component of biology majors’ explanations of evolutionary change after a year of biology. Four examples of participant alternative conception use are shown below:

> “Natural selection is the process by which organisms survive over a period of time. The species with characteristics either genetic or adapted that benefit their survival are the ones that live while the other less beneficial species dies.” (Sample G, #50. alternative conceptions: species vs. individuals, and acclimation influences survival).

> “Some bacteria have evolved a resistance to antibiotics because they have been exposed to it long enough, and have evolved some way of making their bodies antibiotic resistant” (Sample G, #24 alternative conceptions: environmental exposure causes change).

> “The ability to run fast evolved in cheetahs by their leg structure. The ones who were able to run faster positioned their legs in such a way that over generations of time, it became an environmentally controlled mutation. Those who were able to run fast were able to survive, thus now cheetahs all can run fast because the [illegible] cheetahs are dead with their offspring.” (Sample G, #50 alternative conceptions: inheritance of acquired characteristics).

> “The salamanders didn’t need sight to survive: their eyes became useless and thus blind. Like our pinky, useless and eventually it will fall off.” (Sample G, #54 alternative conceptions: use and disuse).

Key Concept and Alternative Conception Frequencies. Table 1 summarizes the number, percentage, and rank of the seven key concepts of natural selection extracted from the ORI from Samples N and G. The goal of this analysis was to document the most common concepts used by students in their explanations of evolutionary scenarios. Percentages for each key concept category were calculated relative to the *total* number of key concepts mentioned; these percentages therefore do not measure absolute response rates for each key concept in the sample. In both samples, students employed all seven key concepts to varying degrees. Key concept six (“certain phenotypes do better and leave more offspring”) was the most commonly elicited key concept in both samples (91.5% Sample N; 74% Sample G). “Overproduction of offspring” was the least commonly used key concept in Sample N (18%), and “competition” was the least commonly used key concept in Sample G (9%). Considerable variation exists in the relative ranks of key concepts between the two samples, but both samples produced evidence of key concept use for all of the major conceptual components of natural selection (Figure 1).

Table 1
The number, percentage, and rank of the seven key concepts of natural selection extracted from the Open Response Instrument (ORI) from Samples N and G

Key Concept	CINS Concept	Description of Concept with CINS Question Numbers	CINS Responses PER Question	CINS Total Responses	ORI (Sample N)	ORI (Sample G)	Oral Interview	CINS %	Sample N (%)	ORI Sample G (%)	Oral (%)	ORI (N) Rank	ORI (G) Rank	Interview Rank
1	Origin and existence of variation	Random mutations and sexual reproduction produce variations; while many are harmful or of no consequence, a few are beneficial in some environments (6B, 19C); individuals of a population vary extensively in their characteristics (9D, 16C)	29, 64, 54, 89	236	74	61	14	59.0	90.2	61.0	82.4	2	3	2
2	Variation inheritable	Much variation is heritable (7C, 17D)	70, 53	123	44	71	16	61.5	53.7	71.0	94.1	4	2	1
3	Limited survival	Production of more individuals than the environment can support leads to a struggle for existence among individuals of a population, with only a fraction surviving each generation (5D, 15D)	77, 72	149	36	9	5	74.5	43.9	9.0	29.4	5	7	5
4	Biotic potential	All species have such great potential fertility that their population size would increase exponentially if all individuals that are born would again reproduce successfully (1C, 11B)	71, 59	130	28	18	2	65.0	34.1	18.0	11.8	7	6	7
5	Natural resources	Natural resources are limited; nutrients, water, oxygen, etc. necessary for living organisms are limited in supply at any given time (2A, 14D)	83, 58	141	69	42	4	70.5	84.1	42.0	23.5	3	5	6
6	Differential survival	Survival in the struggle for existence is not random, but depends in part on the hereditary constitution of the surviving individuals. Those individuals whose surviving characteristics fit them best to their environment are likely to leave more offspring than less fit individuals (10C, 18B)	81, 83	164	75	74	10	82.0	91.5	74.0	58.8	1	1	4
7	Change in a population	The unequal ability of individuals to survive and reproduce will lead to gradual change in a population, with the proportion of individuals with favorable characteristics accumulating over the generations (4B, 13B)	36, 44	80	29	51	11	40.0	35.4	51.0	64.7	6	4	3
N/A	Population stability	Most populations are normally stable in size except for seasonal fluctuations (3B, 12A)	85, 49	134	0	0	0	67.0	—	—	—	—	—	—
N/A	Origin of species	An isolated population may change so much over time that it becomes a new species (8A, 20B)	49, 51	100	0	0	0	50.0	—	—	—	—	—	—

See text for details on ranking and percentage calculations.

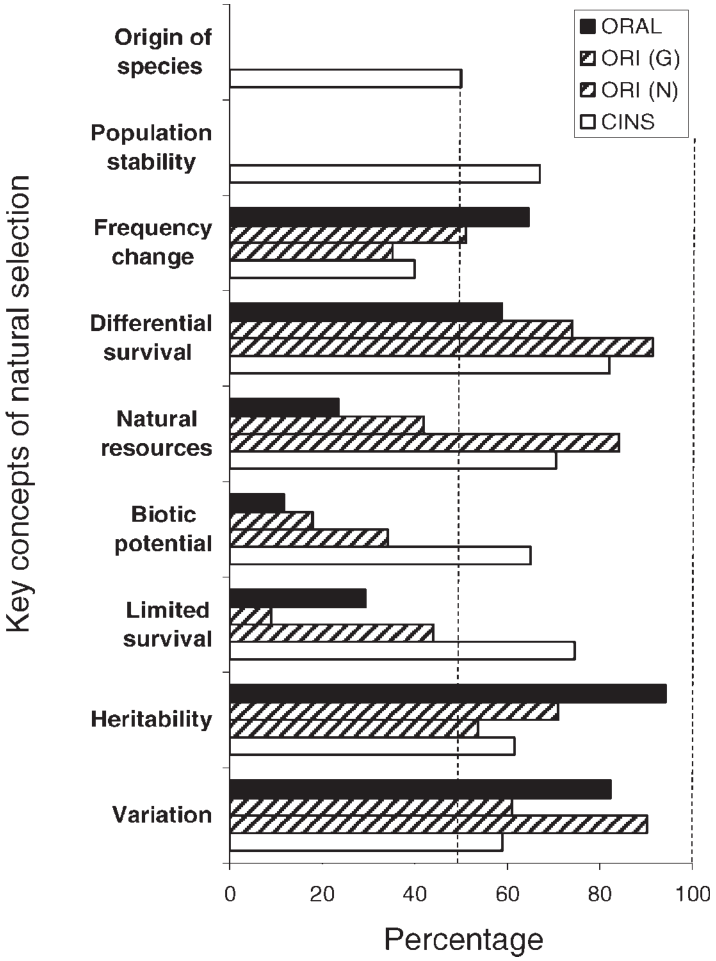


Figure 1. Percentages of "key concepts" of natural selection documented using the Open Response Instrument (ORI) (Samples G and N), the Conceptual Inventory of Natural Selection (CINS), and Oral interview (ORAL). Percentages were standardized among instruments using the total number of key concepts elicited by each instrument.

We also noted the number, percentage, and rank of 31 alternative conceptions (documented from among all of our instruments) from Sample G using the ORI. The conceptual universe of alternative conceptions elicited by the ORI was much more diverse than that of key concepts of natural selection (Table 2). As in Table 1, the percentages for each key concept category were calculated relative to the total number key concepts mentioned; these percentages therefore do not measure absolute response rates for each key concept in the sample. Of the 167 alternative conceptions that we documented using the ORI, the three most common were "use and disuse," "inheritance of acquired traits," and "intention/need," all of which have been well-documented in the literature. The ORI elicited student responses to only 50% of the alternative conception categories that we explored. Additionally, no new categories of alternative conceptions were discovered, although particular questions, such as the "salamander" question, did elicit a large number of consistent but previously undocumented responses relating to the compensatory enhancement and heritability of other senses in response to blindness. Overall, constructed responses from Sample G encompassed a very small universe of alternative conceptions relative to those previously documented in the literature.

Table 2
The number, percentage, and rank of 31 alternative conceptions (documented from among all of our instruments) from Sample G using the Open Response Instrument (ORI)

Misconception Number	Misconception (with CINS Question Number if Applicable)	CINS Responses per Question	CINS Total Responses	ORI Responses	Oral Interview Responses	CINS Frequencies	ORI Frequencies	Interview Frequencies	CINS Rank	ORI Rank	Interview Rank
1	Organisms can intentionally become new species over time (an organism tries, wants, or needs to become a new species) (8C, 8D, 20A, 20D)	28, 22, 14, 28	92	21	4	13	13	10.5	1	3	4
2	Mutations are intentional: an organism tries, needs, or wants to change genetically (6A, 6D, 19A, 19B)	26, 17, 18, 2	63	1	0	8.6	0.6	0	2	13	13
3	Organisms can always obtain what they need to survive (2B, 2C, 2D, 14A, 14B, 14C)	13, 1, 3, 3, 31, 8	59	0	0	8.1	0	0	3	17	17
4	Population level off (1B, 11D, 1D)	9, 28, 20	57	0	0	7.8	0	0	4	18	18
5	Mutations occur to meet the needs of the population (4D, 13D)	29, 25	54	0	0	7.4	0	0	5	19	19
6	Mutations are adaptive responses to specific environmental agents (6C, 15C, 19D)	28, 11, 15	54	4	3	7.4	2.4	7.89	6	8	6
7	Learned behaviors are inherited (4C, 13C)	22, 24	46	0	0	6.3	0	0	7	20	20
8	Changes in a population occur through a gradual change in all members of a population (4A, 13A, 17C)	11, 7, 24	42	6	3	5.8	3.6	7.89	8	7	5
9	Variations only affect outward appearance, do not influence survival (9B, 9C, 16B)	22, 12, 4	38	0	0	5.2	0	0	9	21	21
10	Populations always fluctuate widely/randomly (3C, 12D)	1, 32	33	0	0	4.5	0	0	10	22	22
11	Fitness is equated with strength, speed, intelligence or longevity (10A, 10B, 18A, 18C, 18D)	5, 6, 1, 3, 13	28	1	0	3.8	0.6	0	11	14	14
12	When a trait (organ) is no longer beneficial for survival, the offspring will not inherit the trait (7B, 17B)	10, 14	24	52	8	3.3	31	21.1	12	1	1
13	There is often physical fighting among one species (or among different species) and the strongest ones win (5B, 15B)	9, 15	24	4	0	3.3	2.4	0	13	9	10
14	Populations decrease (3D, 12C)	7, 16	23	0	0	3.2	0	0	14	23	23
15	Traits that are positively influenced by the environment will be inherited by offspring (7D)	18	18	0	0	2.5	0	0	15	24	24
16	Organisms work together (cooperate) and do not compete (5A, 5C, 15A)	2, 12, 2	16	0	0	2.2	0	0	16	25	25
17	All members of a population are nearly identical (9A, 16A)	11, 3	14	0	0	1.9	0	0	17	26	26
18	Traits acquired during an organism's lifetime will be inherited by offspring (7A, 17A)	1, 9	10	27	8	1.4	16	21.1	18	2	2
19	All populations grow in size over time (3A, 12B)	7, 3	10	0	0	1.4	0	0	19	27	27
20	Organisms with many mates are biologically fit (10D)	8	8	0	0	1.1	0	0	20	28	28

21	Not all organisms can achieve exponential population growth (11C)	7	7	0	0	1	0	0	21	29	29
22	Organisms only replace themselves (1A, 11A)	0, 6	6	0	0	0.8	0	0	22	30	30
23	Organisms in a population share no characteristics with others (16D)	4	4	0	0	0.5	0	0	23	31	31
24	The heightening of other senses explains the loss of other sensory structures	n/a	0	17	8	0	10	21.1	24	4	3
25	Change had to happen; death/extinction not recognized as a possibility	n/a	0	3	2	0	1.8	5.26	25	10	8
26	Evolution cannot occur	n/a	0	1	0	0	0.6	0	26	15	15
27	Goal directedness explains change	n/a	0	1	0	0	0.6	0	27	16	16
28	Variants considered to be new species	n/a	0	2	0	0	1.2	0	28	11	11
29	Conflation of species-level and individual-level change	n/a	0	10	0	0	6	0	29	6	9
30	Competition among body parts explains gain or loss of features	n/a	0	2	0	0	1.2	0	30	12	12
31	The environment directly causes change	n/a	0	15	2	0	9	5.26	31	5	7
		Total	730	167	38	100	100	100	—	—	—

See text for details on ranking and percentage calculations.

Table 3
Pearson correlation coefficients of the number of key concepts (shaded boxes) and alternative conceptions (white boxes) elicited among the five open-response questions in the ORI for Sample G (see materials for the complete questions)

Question	Definition	Bacteria	Cheetah	Salamander	Speeding-up
Definition		-0.12 0.25 100	0.03 0.77 100	0.00 1.00 100	0.08 0.42 100
Bacteria	.303(**) 0.00 100		.395(**) 0.00 100	0.10 0.33 100	-0.12 0.24 100
Cheetah	.279(**) 0.01 100	.559(**) 0.00 100		.223(*) 0.03 100	0.06 0.56 100
Salamander	.278(**) 0.01 100	.264(**) 0.01 100	.403(**) 0.00 100		.269(**) 0.01 100
Speeding-up	.200(*) 0.05 100	.324(**) 0.00 100	.373(**) 0.00 100	.269(**) 0.01 100	

▨	Key Concepts
□	Misconceptions

Correlations among ORI Question Responses. Correlation analyses were used to explore whether the five open-response questions in the ORI evoked comparable magnitudes of key concepts and alternative conceptions (Tables 3 and 4). In both Samples G and N, the number of key concepts elicited by the different questions was correlated significantly in many cases (Tables 3 and 4, shaded boxes). In both samples, notable exceptions include the lack of significant correlations between the number of key concepts from the ‘‘speeding-up evolution’’ question and the ‘‘definition’’ question. Additionally, in Sample N, the ‘‘salamander’’ question responses were not significantly correlated with the ‘‘definition’’ question.

In contrast to the results found for key concepts, the five open-response questions in the ORI evoked different magnitudes of alternative conceptions in most cases (Tables 3 and 4, white boxes). In both samples, the number of alternative conceptions elicited by the ‘‘definition’’ question was not correlated with the number of alternative conceptions elicited by any other questions. Likewise, alternative conception responses from the ‘‘bacteria’’ question and the ‘‘speeding-up’’ question were not significantly correlated with one another. The two samples did not provide similar results for several other questions; Sample N, for example, displayed a greater number of significant correlations among questions than did Sample G (Tables 3 vs. 4, white boxes). In most cases, the correlation analyses provided evidence that the five open-response questions elicited comparable magnitudes of key concepts. In contrast, they did not provide evidence that they elicited comparable magnitudes of alternative conceptions.

Difficulty and Discriminability of Open-Response Questions

Difficulty of the ORI questions was calculated using the equation $p = R/T$, where R = the number of students responding correctly to an item and T = the total number of students (Popham, 2006, p. 264). Note that high percentages indicate low difficulty. Because the open-response questions were originally scored using the total number of key concepts and alternative conceptions, it was necessary to convert these scores into binary (i.e., right/wrong) scores. Nehm and Reilly (2007) considered three key concepts of natural selection to be a minimal cut-off for a sufficient answer to an open-ended question about natural selection, and we used this cut-off value. For the separate alternative conception analyses, we viewed any answer that included alternative conceptions to be a wrong answer. Using this approach, we calculated separate difficulty

Table 4
Pearson correlation coefficients of the number of key concepts (shaded boxes) and alternative conceptions (white boxes) elicited among the five open-response questions in the ORI for Sample N (see materials for the complete questions)

Question	Definition	Bacteria	Cheetah	Salamander	Speeding-up
Definition		0.19 0.09 79	0.13 0.24 82	-0.03 0.78 80	0.09 0.43 75
Bacteria	**.420(**)** 0.00 79		**.524(**)** 0.00 79	**.350(**)** 0.00 77	0.20 0.09 74
Cheetah	**.349(**)** 0.00 82	**.583(**)** 0.00 79		**.522(**)** 0.00 80	**.392(**)** 0.00 75
Salamander	0.15 0.19 80	**.383(**)** 0.00 77	**.411(**)** 0.00 80		**.357(**)** 0.00 73
Speeding-up	0.21 0.07 75	**.433(**)** 0.00 74	**.328(**)** 0.00 75	**.355(**)** 0.00 73	

▨ Key Concepts
□ Misconceptions

levels for the five open-response questions from Samples G and N using key concept and alternative conception scores.

As Table 5 illustrates, all five open-response questions from Samples G and N have very high difficulty values using the cut-off of three key concepts. Fewer than 50% of students used the minimum number of key concepts in all questions. Notably, the "speeding-up" question, which we hypothesized would be the most difficult, was among the most difficult in both samples. However, the "definition" question, which we viewed as the least difficult and most concrete recall question, also produced very high difficulty scores. In Samples G and N, using alternative conception scores, the "salamander" question was the only item of high difficulty; more than 50% of biology majors employed alternative conceptions in this question. Overall, the difficulty measures we document indicate that the ORI is challenging for undergraduates who have completed a year of college biology.

ORI question discriminability was calculated using the equation of Popham (2006, p. 266): $D = P(h) - P(l)$, where $P(h)$ = question difficulty of the high scoring group, and $P(l)$ = question difficulty of the low scoring group. For our analyses, we divided the sample along the median (i.e., we used 50% groups; Popham, 2006, p. 266). We calculated discriminability values using both key concepts for each of the five ORI questions. We considered questions with a discriminability value > 0.30 to be acceptable (Popham, 2006, p. 266). We also applied Popham's equation to the occurrence of alternative conceptions for each of the five ORI questions. Although his equation is intended for use with correct responses, we explored whether alternative conceptions could be used to discriminate between groups.

As Table 5 illustrates, most questions have marginal discriminability values, using key concept and alternative conception scores. Notable are the extremely low scores of the "speeding-up" question (Table 5, bottom rows). While many questions were found to be of moderate discriminability, the speeding-up question stood out as being extremely challenging for the biology majors in our samples.

CINS

Quantitative Measures of Student Knowledge Using the CINS. The CINS was used to measure second semester biology majors' knowledge and alternative conceptions of natural selection on a 0–100 scale. Sample G ($n = 100$) had an average CINS score of 62.9% (min. = 20, max. = 100, SD = 19.9). Most students

Table 5

Discriminability (DI) and difficulty values for the five open-response questions from Samples G and N using key concept and alternative conception scores

Question	High KC Sample G	High KC Sample N	Low KC Sample G	Low KC Sample N	KC DI Sample G	KC DI Sample N	KC Difficulty Sample G	KC Difficulty Sample N
Definition	0.18	0.29	0.02	0.10	0.16	0.20	0.10*	0.20*
Bacteria	0.50	0.51	0.14	0.27	**0.36**	0.24	0.32*	0.39*
Cheetah	0.54	0.44	0.00	0.10	**0.54**	**0.34**	0.27*	0.27*
Salamander	0.12	0.22	0.00	0.02	0.12	0.20	0.06*	0.12*
Speeding-up	0.04	0.17	0.02	0.10	0.02	0.07	0.03*	0.13*
Question	High MIS Sample G	High MIS Sample N	Low MIS Sample G	Low MIS Sample N	MIS DI Sample G	MIS DI Sample N	MIS Difficulty Sample G	MIS Difficulty Sample N
Definition	0.74	0.90	0.50	0.80	0.24	0.10	0.62	0.85
Bacteria	0.98	0.90	0.54	0.39	**0.44**	**0.51**	0.76	0.65
Cheetah	0.82	0.85	0.34	0.17	**0.48**	**0.68**	0.58	0.51
Salamander	0.44	0.66	0.24	0.17	0.20	**0.49**	0.34*	0.41*
Speeding-up	0.86	0.85	0.82	0.71	0.04	0.15	0.84	0.78

High = high scoring average of the sample for each question; low = low score averages for each question. KC = key concepts, MIS = alternative conceptions. Note that we also applied Popham's equation to the occurrence of alternative conceptions for each of the five ORI questions. The numbers in columns 2–5 bearing on alternative conceptions represent the proportions of students who employed at least one alternative conception in answering the question. Although his equation is intended for use with correct responses, we explored whether alternative conceptions could be used to discriminate between groups.

Bold values = high discriminability.

*High difficulty.

found the alternative conception distractors so compelling that most received unsatisfactory scores. Overall, based on these scores, the CINS appears to be challenging for the second semester biology majors in our sample.

Key Concept and Alternative Conception Frequencies. Table 1 summarizes the number, percentage, and rank of the seven key concepts of natural selection extracted from the CINS from Sample G. The goal of this analysis was to identify the most common correct concepts used by students. In Sample G, students employed the seven key concepts of natural selection to varying degrees (Table 1). In addition, students responded correctly to the items bearing on ‘‘population stability’’ (CINS items 3B and 12A) and ‘‘origin of species’’ (items 8A and 20B) 67% and 50% of the time, respectively. Recall that we did not consider these latter two ideas to be key concepts of natural selection, and they were never mentioned in the ORI or oral interviews (see below). Key concept six (‘‘certain phenotypes do better and leave more offspring’’) was the most commonly elicited key concept on the CINS (82%). The second and third most abundant key concepts were ‘‘limited survival’’ (74.5%: items 5D and 15D) and ‘‘natural resources’’ (70.5%: items 2A and 14D).

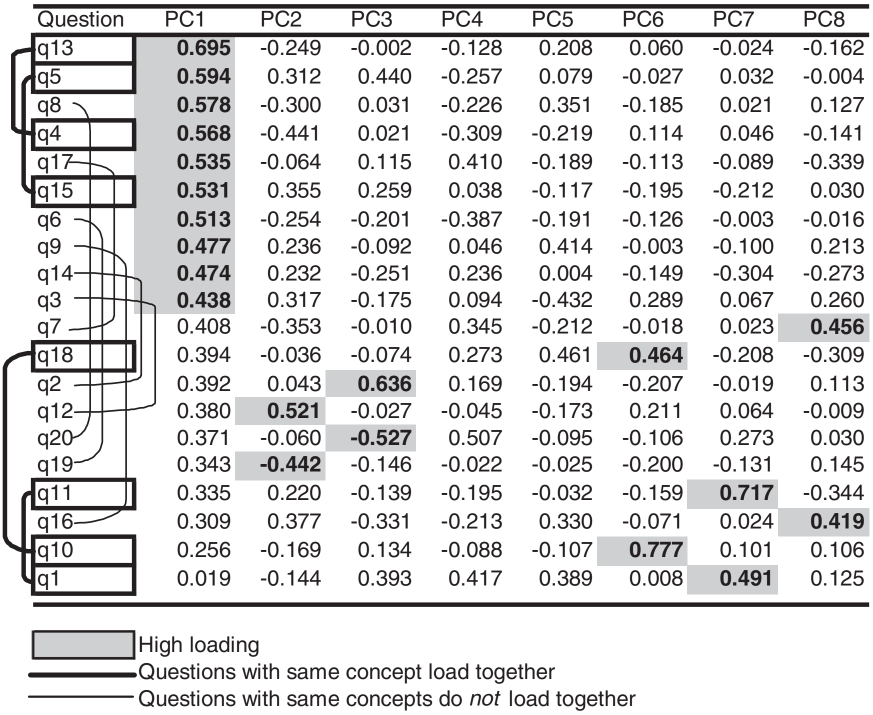
The conceptual universe of alternative conceptions elicited by the CINS was much more diverse than that documented in the ORI (Table 2). We noted the number, percentage, and rank of 31 alternative conceptions. Of the 730 alternative conceptions that we documented in the CINS, the three most common were alternative conception #1 (‘‘intention/need relating to speciation,’’ CINS distractors 8C, 8D, 20A, and 20D), alternative conception #2 (‘‘intention/need related to genetic change,’’ items 6A, 6D, 19A, 19B), and alternative conception # 3 (‘‘resources,’’ items 2B, 2C, 2D, 14A, 14B, and 14C). The CINS elicited alternative conception responses to all but 25% of the categories we documented. Interestingly, four of the alternative conception categories *not* documented by the CINS were among those that formed the ten *most* abundant using the ORI.

Correlation Structure of CINS Questions. PCA was used to explore correlation patterns among the CINS items. A PCA of varimax rotated scores for Sample G produced eight components with eigenvalues > 1.0. These eight components collectively accounted for 65.9% of the variance in the data set. The rotated component matrix values are shown in Table 6. All 20 items had loadings > 0.4 on at least one component, with PC1 explaining the most variation in the dataset (12.3%). PC1 had the highest loadings for items 3, 4, 5, 6, 8, 9, 13, 14, 15, and 17. Items 12 and 19 loaded most highly on PC2, items 2 and 20 on PC3, items 10 and 18 on PC6, items 1 and 11 on PC7, and items 7 and 16 on PC8. Although each of the ten key concepts of natural selection used in the CINS was represented twice (producing a total of 20 possible correct key concept answers), many questions about the same key concept did not load highly on the same components (see Table 6). Notable exceptions of questions that *did* load together on the same component included ‘‘biotic potential’’ (items 1 and 11), ‘‘differential survival’’ (items 10 and 18), ‘‘limited survival’’ (items 5 and 15), and ‘‘change in population’’ (items 4 and 13). Thus, unlike Anderson et al.’s (2002) sample of non-majors, we did *not* find strong support for the different components representing distinct evolutionary concepts in biology majors. Rather, we found one factor that included a highly correlated suite of key concepts.

Difficulty and Discriminability of CINS Questions. As Table 7 illustrates, all 20 multiple-choice questions from Sample G have relatively low difficulty values (i.e., a high percentage of students correctly answered the questions), with the exception of questions 4 and 6. Overall, these results suggest that the CINS is moderately challenging for the undergraduates in our sample. Discriminability (DI) was calculated using the equation of Popham (2006, p. 266; see above). DI values > 0.30 were considered acceptable (Popham, 2006, p. 266). As Table 7 illustrates, about 40% (8/20) of questions have marginal discriminability. Notable are the extremely low scores of questions 1, 10, and 16 (Table 7, bold). In Anderson et al.’s (Anderson et al., 2002) study of CINS discriminability in college non-majors, they found questions 4 and 9 to have marginal values, but we did not. The internal consistency reliability of the CINS was satisfactory, with Cronbach’s $\alpha = 0.78$.

Rasch Analyses of CINS Responses. We used Rasch analysis (Boone & Scantlebury, 2006) to explore item difficulty, redundancy, and person-item patterns for the CINS dataset from Sample G. The CINS dataset

Table 6
A PCA of CINS question scores produced eight components with eigenvalues > 1.0



Question	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
q13	**0.695**	-0.249	-0.002	-0.128	0.208	0.060	-0.024	-0.162
q5	**0.594**	0.312	0.440	-0.257	0.079	-0.027	0.032	-0.004
q8	**0.578**	-0.300	0.031	-0.226	0.351	-0.185	0.021	0.127
q4	**0.568**	-0.441	0.021	-0.309	-0.219	0.114	0.046	-0.141
q17	**0.535**	-0.064	0.115	0.410	-0.189	-0.113	-0.089	-0.339
q15	**0.531**	0.355	0.259	0.038	-0.117	-0.195	-0.212	0.030
q6	**0.513**	-0.254	-0.201	-0.387	-0.191	-0.126	-0.003	-0.016
q9	**0.477**	0.236	-0.092	0.046	0.414	-0.003	-0.100	0.213
q14	**0.474**	0.232	-0.251	0.236	0.004	-0.149	-0.304	-0.273
q3	**0.438**	0.317	-0.175	0.094	-0.432	0.289	0.067	0.260
q7	0.408	-0.353	-0.010	0.345	-0.212	-0.018	0.023	**0.456**
q18	0.394	-0.036	-0.074	0.273	0.461	**0.464**	-0.208	-0.309
q2	0.392	0.043	**0.636**	0.169	-0.194	-0.207	-0.019	0.113
q12	0.380	**0.521**	-0.027	-0.045	-0.173	0.211	0.064	-0.009
q20	0.371	-0.060	**-0.527**	0.507	-0.095	-0.106	0.273	0.030
q19	0.343	**-0.442**	-0.146	-0.022	-0.025	-0.200	-0.131	0.145
q11	0.335	0.220	-0.139	-0.195	-0.032	-0.159	**0.717**	-0.344
q16	0.309	0.377	-0.331	-0.213	0.330	-0.071	0.024	**0.419**
q10	0.256	-0.169	0.134	-0.088	-0.107	**0.777**	0.101	0.106
q1	0.019	-0.144	0.393	0.417	0.389	0.008	**0.491**	0.125

▨ High loading
━ Questions with same concept load together
— Questions with same concepts do *not* load together

These eight components collectively accounted for 65.9% of the variance in the data set. The component matrix values are shown below. Bold boxes indicate highest loadings on each component (as per Anderson et al., 2002). Lines connect questions containing the same key concept of natural selection. Note that many questions testing the same concept (e.g., q8 and q20, on the origin of variation) do not load on the same component.

matches the Rasch analysis requirement of having sample sizes ≥100 and item number ≥20 (Bond & Fox, 2001). Prior to interpreting the item and person logit scores from the analysis, we explored whether our dataset fits the Rasch model. In a dataset with good fit, person and item mean squares are expected to be 1.0. The mean infit and outfit for "persons" from our dataset are 1.00 and 1.06 respectively, and for "items" are 0.99 and 1.06, respectively. The mean standardized infit and outfit for persons (0.0 and 0.1) and items (0.0 and 0.2) are close to the expected value of 0.0. These values indicate good fit and suggest moderate levels of item redundancy. The standard deviation of the standardized infit for "persons" and "items" is 0.9 and 1.1, respectively; both are below the 2.0 cut-off suggested by Bode and Wright (1999). Separation values for "persons" (1.57) and "items" (3.74) are greater than the suggested minimum cut-off value of 1. Overall, then, our CINS data from Sample G appear to fit the Rasch model.

The fit of individual CINS items to the Rasch model is shown in Table 8. Several authors (e.g., Bond & Fox, 2001) consider items with mean-square fit (MNSQ) values between 0.8 and 1.3, and standardized *z* values (ZSTD) >2 or <-2, to be indicative of poor item fit. As shown in Table 8, CINS items 1, 5, and 13 were discordant with model predictions based on both MNSQ and ZSTD values. Item 1 tested knowledge about biotic potential, item 5 tested limited survival, and item 13 tested changes in population frequency. Notably, the parallel questions about the same concepts (items 11, 15, and 4) were not characterized by poor fit with the model. Interestingly, traditional analysis also uncovered problems with item 1, which was characterized by a low discriminability value (Table 8). CINS items 5 and 13, however, had acceptable discriminablity values and difficulty values using traditional analyses.

Table 7
Discriminability (DI) and difficulty values for the 20 CINS questions

Item	High Group	Low Group	DI	Difficulty (%)
q1	0.78	0.64	**0.14**	71
q2	0.94	0.72	0.22	83
q3	0.98	0.72	0.26	85
q4	0.52	0.20	0.32*	36
q5	0.96	0.58	0.38	77
q6	0.42	0.16	0.26	29
q7	0.88	0.52	0.36	70
q8	0.76	0.22	0.54	49
q9	0.80	0.28	0.52*	54
q10	0.88	0.74	**0.14**	81
q11	0.72	0.46	0.26	59
q12	0.66	0.32	0.34	49
q13	0.72	0.16	0.56	44
q14	0.76	0.40	0.36	58
q15	0.90	0.54	0.36	72
q16	0.98	0.80	**0.18**	89
q17	0.76	0.30	0.46	53
q18	0.96	0.70	0.26	83
q19	0.80	0.48	0.32	64
q20	0.70	0.32	0.38	51
			Average:	63

High group = high scoring average of the sample for each question; low = low score averages for each question. KC = key concepts, MIS = alternative conceptions.

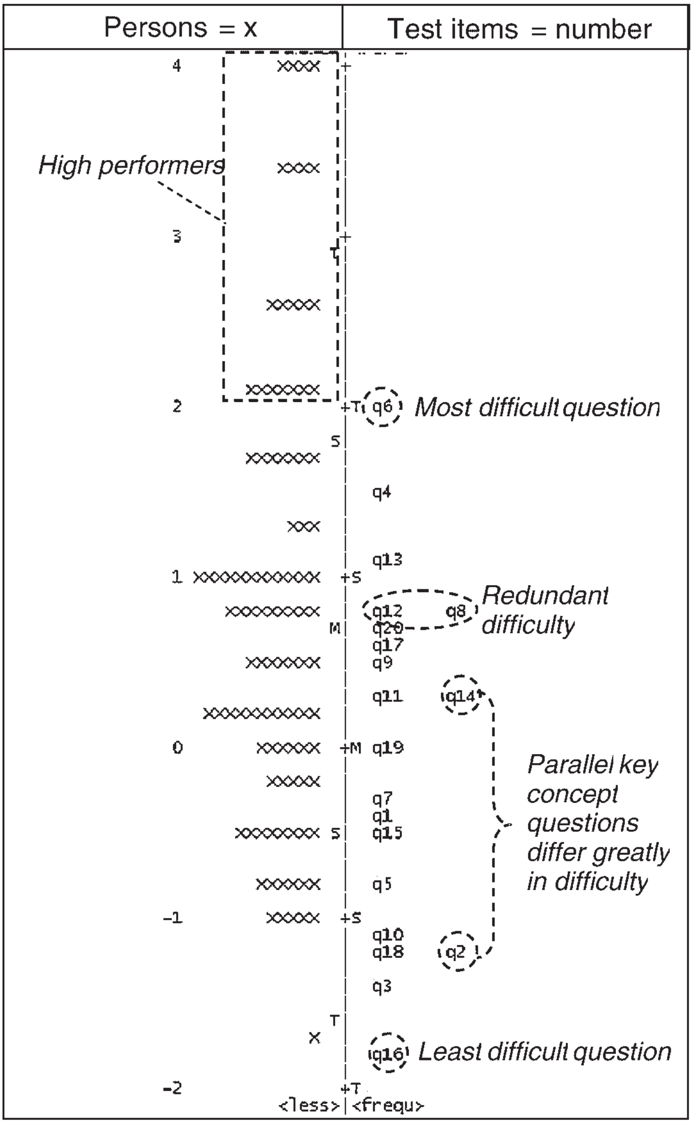
*Questions identified by Anderson et al. (2002) as poor based on low DI values.

Table 8
CINS item fit statistics derived from the Rasch analysis

CINS Item	Raw Score	Measure	Model SE	Infit (MNSQ)	Infit (ZSTD)	Outfit (MNSQ)	Outfit (ZSTD)	PTMEA (Corr.)
1	67	46.08	2.44					0.14
2	79	37.90	2.85	0.91	−0.50	1.75	1.90	0.33
3	81	36.20	2.98	0.90	−0.50	0.68	−0.80	0.38
4	32	65.22	2.44	0.95	−0.40	0.89	−0.60	0.55
5	73	42.30	2.59		−1.50		−1.40	0.49
6	25	69.65	2.61	0.94	−0.40	1.01	0.10	0.55
7	66	46.67	2.42	1.00	0.10	1.03	0.20	0.39
8	45	58.00	2.30	0.90	−1.00	0.88	−0.90	0.55
9	50	55.37	2.29	0.99	0.00	1.01	0.10	0.47
10	77	39.46	2.75	1.05	0.40	1.40	1.20	0.27
11	55	52.74	2.30	1.12	1.40	1.10	0.70	0.38
12	45	58.00	2.30	1.13	1.40	1.21	1.50	0.39
13	40	60.68	2.34					0.64
14	54	53.26	2.30	1.00	0.00	0.93	−0.40	0.46
15	68	45.48	2.46	0.90	−1.00	1.17	0.80	0.44
16	85	32.22	3.35	0.99	0.00	0.81	−0.30	0.27
17	49	55.89	2.29	0.95	−0.60	0.92	−0.50	0.51
18	79	37.90	2.85	0.97	−0.10	0.72	−0.80	0.36
19	60	50.05	2.34	1.10	1.10	1.02	0.20	0.38
20	47	56.94	2.30	1.08	0.90	1.08	0.70	0.43
Mean	58.9	50.00	2.53	0.99	0.00	1.06	0.20	—
SD	16.9	10.02	0.29	0.12	1.10	0.34	1.30	—

Table 9 displays a person-item map, which visually summarizes several aspects of the Rasch analysis of the CINS dataset. The distribution of persons (on the left) and CINS items (on the right) are illustrated on the same logit scale. Persons at the *same* position along the scale as a particular item have a 50% chance of answering the item correctly. Questions of equivalent difficulty lie at the same point on the logit scale (e.g., questions 8 and 12, 11 and 14, and 2 and 18). Individuals located *above* an item, however, have an even greater chance of answering the item correctly (i.e., the item is likely to be easier for these individuals). Those persons

Table 9
A person-item map derived from a Rasch analysis of CINS question responses from Sample G



The distribution of persons (on the left) and CINS items (on the right) are illustrated on the same log interval (−2 to 4) scale.

located *below* an item have a less likely chance of answering it correctly (i.e., the item is more difficult for them).

Overall, Table 9 demonstrates that the distributions of CINS questions and persons are generally well matched, except at the high end of the logit scale. Here, questions of sufficient difficulty are lacking. As Table 9 illustrates, CINS item 6 (origin of species) is the most difficult on the logit scale, whereas nearly 15% of the sample lies above the level of this question. The overall pattern in Table 9 indicates that the CINS sufficiently differentiates the persons in the sample at lower performance levels, but does not differentiate students at the highest performance levels. Additionally, it reveals that equivalent key concept questions in the CINS are not of comparable difficulty. For example, CINS items 19 and 6 (about ‘‘the origin of species’’), and items 2 and 14 (about ‘‘population stability’’) have greatly different difficulty levels.

In summary, the CINS data demonstrate good fit with the assumptions of the Rasch model and indicate that the instrument is appropriately matched in difficulty level to the sample of biology majors studied here. Several instrument items of equivalent difficulty are present (e.g., items 2 and 18), and there was a lack of items of sufficient difficulty to distinguish high performers. Additionally, CINS items 1, 5, and 13 fit poorly to model predictions. Finally, the ten paired key concept questions were not in most cases of equivalent difficulty, although some exceptions were noted (e.g., items 10 and 18).

Oral Interview

Oral interviews with 18 participants from Sample G lasted on average 24.6 minutes (min. = 15.5, max. = 39.1). These participants displayed a broad distribution of knowledge and alternative conception scores on the ORI and the CINS. The average CINS percentage score among interviewees was 66.4 (min. = 30, max. = 100), whereas the average NSPQ score was 80 (min. = 58, max. = 100). Qualitative interview scores averaged 0.2 (min. = −1, max. = 1). We provide examples from two participants (subjects Q and R) to illustrate differences in student understanding documented in the oral interviews.

Participant Q, who received a composite score of +1, demonstrates knowledge of several key concepts of natural selection in question 1: genetic variation produced by cross breeding and rapid generation time, competition for resources among individuals, and differential survival based on genetic variability. He does not demonstrate any alternative conceptions.

Interviewer: A number of mosquito populations no longer die when DDT, which is a chemical used to kill insects, is sprayed on them, but many years ago DDT killed most mosquitoes. Could you explain why many mosquitoes don’t die anymore when DDT is sprayed on them?

Participant Q: OK . . . ah, well, the DDT used to kill a lot of the mosquitoes because the mosquitoes didn’t have a lot of resistance to the DDT. But after a certain amount of time all of the mosquitoes that were more vulnerable to the DDT died and the mosquitoes that weren’t vulnerable to the DDT survived . . . they had access to more of the stuff in the environment and less competition from the mosquitoes that didn’t have resistance to DDT, they survived and were able to take the niche of the mosquitoes that died to DDT.

Interviewer: . . . What is that resistance?

Participant Q: . . . In a population such as mosquitoes they . . . the generations reproduce pretty quickly, like you know, thousands of larvae, and it allows for a wide genetic pool, if anything, so since there is a wide genetic pool there is a greater chance of a mosquito, or any mosquito at all, developing resistance.

Interviewer: When you say wide genetic pool, what do you mean by that?

Participant Q: There is a lot of genetic variability

Interviewer: Do you know why there is a wide genetic variability?

Participant Q: Cross breeding.

In contrast to participant Q, participant R demonstrates a limited understanding of natural selection and several alternative conceptions. His overall interview score was −1. Specifically, we interpret his responses to indicate that traits acquired during an organism's lifetime can be passed on, and in some cases amplified, and that this in part explains biotic change in mosquitoes. Notably, he incorporates numerous experiences from watching television and his daily life into his conceptual explanations for DDT resistance in mosquitoes.

Interviewer: (mosquito question, same as above).

Participant R: . . . Well, if at first the DDT killed most mosquitoes and now it's not killing them [any] more, then a possible explanation would be that when they first started exposing the mosquitoes to the DDT they didn't have any . . . their immune system was not that strong to fight the DDT. As time went on they developed some kind of resistance to the DDT . . . they passed this kind of, um, newly evolved resistance on to the next generation so . . . passing on this trait from generation to generation . . . it will start becoming stronger and if the DDT is used on them it wouldn't kill them . . .

Interviewer: Can you tell me a little bit more about how that [resistance] would happen, in general terms?

Participant R: . . . I was watching the discovery channel and there was a man who said he could develop resistance to the venom of a snake . . . so he started to gradually use little bits of this venom and started injecting venom into his system and from time to time he would increase the amount of venom he took into his system . . . he got bitten by the snake and to the surprise of the doctors this man actually had some kind of resistance to that venom, in comparison to a normal person who would just die . . . my guess would be that at first the mosquitoes . . . from time to time they kept exposing them to this kind of chemical . . . those will develop some kind of resistance to this kind of chemical for them to survive.

Interviewer: This guy who was injecting the snake venom . . . would one of his children or a certain percentage of his children also be resistant to the venom/immunity thing?

Participant R: They might be, but not all the way . . . as he got a resistance.

Interviewer: Why?

Participant R: . . . If we consider the fact that they are children and their immune systems are not that strong, but if we compared them to the same age children . . . if you compare both sides they will have a little bit of resistance.

Interviewer: His family's kids might have not as much as he has, but they will have more than a group of kids that got none?

Participant R: Yeah.

Interviewer: Back to the DDT example . . . Do you think before DDT was invented, whatever was different about the mosquitoes who didn't die when they were exposed to DDT, were they different back then, or did they just become different when they were exposed to DDT . . . Did the DDT make some of them change?

Participant R: When they introduced the chemical obviously it wouldn't kill all of them . . .

Interviewer: Why not?

Participant R: Well, in my house recently we had roaches . . . [we introduced bait] and to my surprise, I mean, it reduced the percentage of the roach more than before I introduced the bait, so I was kind of thinking, why didn't it kill the entire race of roach? The partial conclusion that I could draw is that maybe the places that I put the bait did not fully expose the roach . . . maybe the first batch that got introduced the bait got more of the food . . . so as time went on the concentration went on . . . if the next batch of roaches get exposed to it they might have a 50–50 chance of survival . . . just like this man didn't start off injecting a whole lot of venom into his system . . . that would be my conclusion.

Collectively, all seven key concepts of natural selection documented in the ORI and CINS were also mentioned by interviewees (Table 1, Figure 1). The CINS topics of "population stability" and "origin of species," which are not considered key concepts of natural selection here, were not mentioned by any interviewees. We also scored the number, percentage, and rank of 31 alternative conceptions documented from the ORI and CINS in the interview sample. The percentages for each alternative conception category were calculated relative to the *total* number of responses. Of the 38 alternative conceptions documented in the interviews (Table 2) the three most common were alternative conception #12 ("use and disuse"), alternative conception #18 ("inheritance of acquired traits"), and alternative conception #24 ("sensory compensation"). The interviews elicited responses to only 50% of the alternative conception categories that we explored. Additionally, no new categories of alternative conceptions were discovered. Notably, the ten most commonly ranked alternative conceptions in the oral interviews were identical to the ten most commonly ranked alternative conceptions in the ORI. Overall, the oral interview results were most similar to the ORI, but they also shared a large number of commonalities to the CINS.

Correlations among Instrument Variables

In order to determine whether the two paper-and-pencil instruments provided related measures of natural selection knowledge and alternative conceptions, we calculated correlations among several different measures derived from the ORI and the CINS. The NSPQ, which quantifies ORI scores on a single 0–100 scale using *both* key concepts and alternative conceptions, was significantly correlated with the overall CINS scores (i.e., the percentage of correct responses on a 0–100 scale; $n = 100$, $r = 0.58$, $p < 0.001$). Several other measures also produced significant correlations between the ORI and CINS. CINS percentages were significantly correlated with the ORI key concept diversity measure (see above; $n = 100$, $r = 0.61$, $p < 0.001$). In addition, ORI alternative conception diversity (see above) was significantly correlated with the number of incorrect CINS responses (i.e., the number of alternative conception distractors chosen; $n = 100$, $r = 0.42$, $p < 0.001$). In summary, the ORI and CINS appear to be measuring related information using several different measures of both knowledge and alternative conceptions of natural selection.

We considered the oral interview to be the most meaningful, detailed, and thorough analysis of student knowledge. Correlations between the oral interview scores and the paper-and-pencil instruments were performed to validate instrument measures. In the correlational analyses involving the oral interview, we used one-tailed statistical tests in view of the lack of power given the 18 undergraduates who were interviewed. Like the pattern found in the larger sample, the correlation between the NSPQ and CINS percentage score was significant using the smaller sample of interview subjects alone ($n = 18$, $r = 0.45$, $p < 0.05$). Similarly, the oral interview score was significantly correlated to both the NSPQ score ($n = 18$, $r = 0.74$, $p < 0.01$) and the CINS percentage score ($r = 0.68$, $n = 18$, $p < 0.01$). The rock test (see above), which was administered to all interview participants and used for discriminant validity purposes, was not significantly correlated with any knowledge or alternative conception measure ($-0.30 < r < 0.19$).

Discussion

Despite a growing body of research in evolution education, comparatively little attention has been directed towards the rigorous development and evaluation of instruments that measure knowledge of and alternative conceptions about evolution and natural selection in learners of different ages and educational backgrounds (Liu & Lesniak, 2005; Nehm, 2006; NRC, 2001). We used three different methods, the CINS, the ORI essay test, and an oral interview, to assess biology majors' understanding of and alternative conceptions about natural selection, as well as the validity and reliability of each approach (Table 10). Overall, both the ORI and CINS could serve as replacements for the labor-intensive process of oral interviews. Both produced comparable measures of key concept diversity and, to a lesser extent, key concept frequency. By contrast, the ORI and CINS provided clearly different assessments of both alternative conception diversity and frequency, with the ORI producing a richer description of alternative conception diversity. Regarding key concepts, the ORI results were completely concordant with oral interview results. Both the CINS and the ORI included items that could

Table 10
Summary table comparing aspects of the three methods for measuring student knowledge and alternative conceptions of natural selection

Topic	Oral interview	ORI	CINS
Reliability	Strong support. Scorer reliability established for key concepts and misconceptions	Strong support. Scorer reliability established for key concepts and misconceptions	Moderate support. Acceptable alpha coefficient. Rasch analysis measures also indicate acceptable reliability
Validity	Strong support. Content validity. Convergent validity evidence includes significant correlations with ORI and CINS. Discriminant validity evident by non-significant correlation with rock test scores	Strong support. Content validity. Convergent validity evidence includes ORI NSPQ scores being significantly correlated with both CINS scores and oral interview scores. Discriminant validity evident by non-significant correlation with rock test scores	Strong support. Content validity. Convergent validity evidence includes CINS scores significantly correlated with both ORI NSPQ scores and oral interview scores. Discriminant validity evident by non-significant correlation with rock test scores
Item characteristics	Not applicable	Low discriminability and high difficulty characterize several questions; "speeding up" question too difficult for sample	Low discriminability and high difficulty characterize many questions; no questions distinguish among high performers; parallel concept questions differ greatly in difficulty; too many items of equivalent difficulty
Measurement of key concepts of natural selection	All seven key concepts documented; frequencies most similar to ORI	All seven key concepts documented; frequencies most similar to oral interview results	All seven key concepts documented; frequencies least similar to other measures
Measurement of misconceptions of natural selection	Universe of misconceptions significantly smaller than found using the CINS, but excellent concordance with ORI results	Universe of misconceptions significantly smaller than found using CINS, but excellent concordance with oral interview results	Greatest diversity of misconceptions documented, but poor concordance with both ORI and oral interview patterns
Instrument strengths	All concepts and misconceptions were captured by the ORI and CINS	Appears to document extant misconceptions most accurately. Provides results most similar to an oral interview	Documents key concepts most efficiently
Instrument weaknesses	None	Omission errors problematic; interpretation of responses can be difficult but this can be mitigated by scorer practice with a rubric	May not accurately represent the extent of misconceptions. Permits guessing
Implementation strengths	None	None	Rapid results, scoring, and interpretation
Implementation weaknesses	Highly impractical for use in large samples	Impractical for use in large samples	None

be characterized by (a) low discriminability, (b) high overlapping difficulty, and (c) mismatches with the sample.

Criticisms of the Bishop and Anderson Test

Anderson et al. (2002) provided a series of criticisms of Bishop and Anderson's (1990) open-response instrument and used these criticisms as justification for the development of the CINS. Overall, Anderson et al. criticized the Bishop and Anderson instrument items as being: (1) simple; (2) hypothetical; (3) abstract; (4) and unable to probe student understanding of ecological and genetic principles central to natural selection. If these criticisms were justified, one would expect that compared to the CINS the Open Response Instrument (ORI), which is largely derived from the Bishop and Anderson test, would provide comparatively less reliable and valid information about undergraduate students' knowledge and alternative conceptions of natural selection. We explored this question empirically and were not able to support Anderson et al.'s argument: both the CINS and ORI produced comparable and complementary measures of students' knowledge of natural selection. No data in our study provided clear evidence for the superiority of the CINS.

Our study did provide, however, evidence that calls into question some of Anderson et al.'s criticisms of the Bishop and Anderson instrument. First, based on high difficulty scores for individual questions and the marginal composite instrument scores from our two samples (i.e., the NSPQ, key concept diversity, and alternative conception diversity), it appears that the ORI is not simple, but rather is sufficiently challenging for college biology majors. Second, we found ample evidence that the ecological and genetic principles that underlie the theory of natural selection were indeed elicited by the ORI. All seven "key concepts" of natural selection were documented using the ORI; these included the causes of phenotypic variation (such as mutation and recombination) and ecological principles (such as competition and differential survival based on resources; Table 1, Figure 1). Thus, the criticisms of both simplicity and the inability to elicit fundamental principles were not supported by our study.

We regard the remaining two criticisms of the Bishop and Anderson test—being hypothetical and abstract—as potentially accurate but likewise characteristic of Anderson et al.'s CINS. For example, when we consider an individual concept on the CINS, such as the differential survival of guppies in tropical streams, we wonder whether having a student read a paragraph about guppies makes the subsequent questions about them less hypothetical or abstract relative to, for example, the differential survival of cheetahs on the African savannah (discussed in the ORI). Although neither we nor Anderson et al. (2002) investigated this question empirically, we consider the contemplation and explanation of all evolutionary scenarios to be necessarily abstract because they involve the simultaneous consideration of multiple variables in contexts in which students typically have had no direct experience. Additionally, both the ORI and CINS prompt students to ponder evolutionary patterns and processes that occur during timescales that are often inaccessible to direct observation and, more importantly, beyond student contemplation (i.e., hundreds, thousands, or millions of years). The CINS questions do, however, deal with evolutionary scenarios likely taking place over shorter timescales than those discussed in the ORI. Nevertheless, we argue that the Bishop and Anderson (1990) essay test is not unique in prompting students to answer hypothetical and abstract questions, regardless of whether the questions are based on hypothetical or abstract exemplars. It may be argued, however, that using hypothetical questions that parallel actual situations, but vary from them in significant ways, have the potential to be misleading and could actually reinforce or propagate alternative conceptions among students. No studies to our knowledge have investigated this issue. In summary, we find, at best, some support for Anderson et al.'s criticisms of the Bishop and Anderson instrument but argue that some of their criticisms may also apply in to the CINS.

Validation of the CINS

Anderson et al. (2002) used correlations between scores obtained from seven oral interviews and overall CINS scores as their primary method for validating their instrument. While this small sample size is troublesome, primarily because it is the only evidence supportive of the validity of their instrument, exclusive of response patterns themselves. We found Anderson et al.'s interpretation of their oral interview data problematic. On page 966, for example, Anderson et al. (2002) report that:

> “We were encouraged by our finding that the seven undergraduate biology majors whom we interviewed all demonstrated an accurate understanding of natural selection. This suggests to us that it is possible for students to learn about natural selection and that we should be much more successful than we currently are with nonmajors.”

Table 2 in Anderson et al. (2002, p. 960) lists the seven participants’ interview scores and CINS scores. Interview scores were calculated by coding post-interview transcripts for correct and incorrect “utterances.” According to Anderson et al.’s Table 2, 43% (3/7) of participants received what can only be interpreted as failing scores on the interview and 57% (4/7) received failing scores on the CINS. Thus, it is unclear why Anderson et al. (2002) considered these students to have an “accurate understanding of natural selection” or why they were “encouraged” by this result. Additionally, in the abstract of their paper, Anderson et al. (2002, p. 952) indicate that these seven students were non-majors, whereas the above quote suggests that these seven students were biology majors. If other sources of instrument validation were provided, these confusing interpretations would be less troubling. Anderson et al.’s use of only seven student interviews to validate the CINS, and the confusing interpretation of these data, spurred us to investigate the validity of the CINS more rigorously.

A final issue of concern with the validation of the CINS was the absence of evidence of discriminant validity. Instrument validation typically requires demonstrating that measures not hypothesized to be related to the measured construct are in fact not significantly correlated to scores derived from instrument responses. In other words, discriminant evidence supports the interpretation that the wrong construct is not being measured (Ary, Jacobs, & Razavieh, 2002, p. 558). Because Anderson et al. (2002) did not provide evidence regarding discriminant validity in their validation of the CINS, our study used a closed-response test about rocks to establish discriminant evidence for both the CINS and the ORI (Table 10).

Comparing and Contrasting the CINS, ORI, and Interview

Figure 2 visually illustrates the concordance of key concept and alternative conception elicitation from among the CINS, ORI, and oral interview data. Figure 2 demonstrates the excellent concordance of key concept elicitation using all three methods. As the figure shows, all three methods elicited the same seven key concepts of natural selection. This result validates the inference that the three methodologies captured information about the same construct, and indicated that, in general, the method of elicitation was not correlated to the content extracted. However, as Figure 1 illustrates, the magnitudes of participant responses did differ markedly in some cases (e.g., “limited survival” was mentioned by the undergraduates in nearly 80% of CINS responses but in only 30% of interview responses). Thus, while the CINS, ORI, and oral interview elicit the same types of knowledge among respondents, they do not always reflect different strands of knowledge to the same degree. Taking into consideration the effort required to execute, score, and interpret the ORI and oral interview data, our results indicate that the CINS would be the most efficient method for measuring knowledge of the key concepts of natural selection in the sample of biology majors that we studied. These results also indicate that the CINS may, however, overestimate students’ working knowledge of the key concepts of natural selection (Figure 1).

Unlike the results for key concepts, alternative conception elicitation was significantly related to methodology (Figure 2). Only five of the same alternative conceptions (of 31) were uncovered by all three methods (alternative conceptions 1, 6, 8, 12, and 18; see Table 2). Notably, the oral interviews did not elicit any unique alternative conceptions (Figure 2). Overall, however, the oral interview alternative conception magnitudes were most similar to the ORI responses (Table 2). In terms of alternative conceptions uncovered, the oral interview provided equivalent construct concordance with the ORI and CINS; that is, three alternative conceptions, although not the same three, were elicited by both the instrument and the ORI and CINS (Figure 2).

Used together, the paper-and-pencil ORI and CINS appear to provide an excellent replacement for the time-consuming process of oral interviews. While the two paper-and-pencil instruments provided generally comparable measures of the key concepts of natural selection, they did *not* provide equivalent measures of alternative conception diversity or magnitude. Until a new instrument is developed, we recommend that both the CINS *and* ORI be used to measure the distribution and magnitude of alternative conceptions.

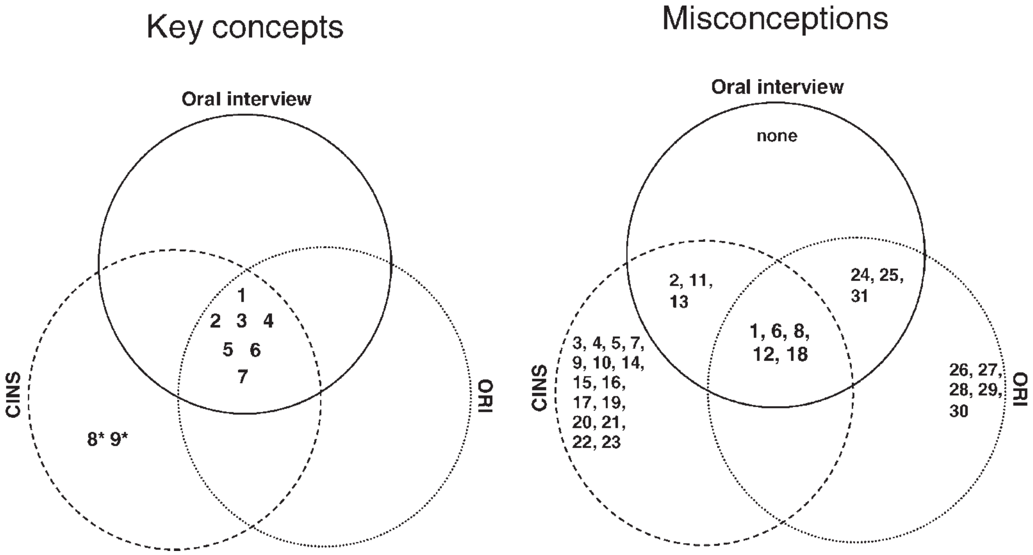


Figure 2. A Venn diagram illustrating the distribution of key concepts and alternative conceptions elicited among the CINS, ORI, and oral interview. See Table 1 for key concept number descriptions and Table 2 for alternative conception number descriptions. * = not considered a key concept of natural selection in this study, but considered as such by Anderson et al. (2002).

Criterion-Referenced Nature of the CINS

Anderson et al. (2002) noted (p. 959) that the CINS is a criterion-referenced test. Although Anderson et al. demonstrated a degree of content validity for the CINS, they did not adduce evidence that the CINS possessed *other* key properties of criterion-referenced tests. For example, a key property of criterion-referenced tests is their capacity to document growth in student knowledge over the course of a period that starts before a learning experience begins (e.g., prior to a learning module devoted to natural selection) and ends once the learning experience concludes (e.g., at the conclusion of the natural selection module). Validity research on a criterion-referenced test could be enhanced with a control group that is not exposed to the target learning experience (King, 1975). Anderson et al. (2002) provide no validity evidence bearing on the sensitivity of the CINS to growth in knowledge of natural selection. By contrast, Nehm and Schonfeld (2007) demonstrated that an essay test adapted from Bishop and Anderson (1990) is sensitive to knowledge growth.

Criterion-referenced testing downplays psychometric reliability because psychometric reliability depends on between- and within-person variance components, as those components bear on the test's ability to consistently discriminate among individuals (Carver, 1974). Since criterion-referenced testing anticipates little pretest variance (most students displaying below-standard performance) and little post-test variance (most students displaying above-standard performance), reliability in this context is reasonably evaluated with alternate forms assessing pre-to-post-test gains. Anderson et al. (2002) provide no evidence bearing on this particular type of reliability. By contrast, Nehm and Schonfeld (2007) demonstrated consistency in pre-to-post-course gains using alternate measures of knowledge of natural selection (and concomitant reductions in alternative conceptions). Finally, creators of criterion-referenced tests typically create a cut-off or standard score that reflects appropriate mastery of the knowledge taught. Anderson et al. (2002) did not provide such a score. By contrast, Nehm and Reilly (2007), using an adaptation of Bishop and Anderson's (1990) test, arrived at a cut-off score that reflected a minimum level of understanding of natural selection.

Carver (1974) pointed out that a criterion-referenced test "may be referenced to a normative group, and a norm-referenced test, to a criterion" (p. 512). Millman and Popham (1974) indicated that criterion-referenced tests have conventional psychometric uses such as predictive validity. Moreover, estimates of variance for reliability purposes are suitable to criterion-referenced tests such as the CINS (Cronbach, 1975). Haladyna (1975) underlined the point that educators who use criterion-referenced tests should be concerned with psychometric reliability because errors of measurement can cloud the educators' ability to distinguish true passes, false positives, true fails, and misclassified fails, particularly among students who score near the cut-off. Although the current paper was not directly concerned with the CINS as a criterion-referenced test, the paper's focus on the test's psychometric properties provides a foundation for future research on the CINS "edumetric" properties.

Instrument Format and Knowledge Measures

Considering that the same student population was used to measure evolutionary knowledge, the differences in knowledge and alternative conception scores that we document may be a result of different question formats (i.e., open- vs. closed-response). Bridgeman (1992, p. 253) outlined the three major advantages to open-response test items: (1) reduction of measurement error associated with random guessing; (2) elimination of unintended corrective feedback, that is, if an expected correct answer is not present among the items, the student knows that a change in strategy is required to solve the problem; and (3) problems cannot be solved by working backwards from the answers.

Research has also demonstrated that open-response items may measure different cognitive characteristics than closed-response items. Traub and MacRury (1990), for example, in their literature review pertaining to proficiency measurement using open- and closed-response testing, concluded that the reasons for differences between the two approaches were not clear, and recommended that both item types be used in knowledge assessment. Traub and MacRury went on to argue that one should *not* assume that both methods assess the same cognitive abilities. Likewise, Kuechler and Simkin (2003) found that open- and closed-response items on computer programming exams did not correlate highly, and likely tested different cognitive processes. Collectively, this work argues for the inclusion of open-response items in knowledge assessments.

In contrast, other work calls into question the benefits of open-response test items. Lukhele, Thissen, and Wainer (1994), for example, found that the College Board's AP chemistry and history tests' open-response questions added little information beyond that provided by the multiple-choice sections. Lukhele et al. (1994) went on to question the basic premise that open-response items are more useful and meaningful than multiple-choice items. In a quantitative meta-analysis of construct equivalence, Rodriguez (2003) found that when stem-equivalent items are employed, multiple-choice and constructed response measures tend to correlate highly. Likewise, Bridgeman (1992, p. 253) found that despite differences in format, open- and closed-response items produced "remarkably similar correlational patterns." Overall, there is evidence that open-response items themselves may not provide more meaningful measures of knowledge; aspects of the questions themselves may account for these differences.

Educators who favor essay tests look to advantages that include the test's usefulness in assessing the student's ability to relate facts and principles to each other, organize knowledge, and write in clear, accurate prose. These advantages do not come without costs. First, time is a premium, and frequently students have insufficient time to demonstrate the above abilities. Second, compared to multiple-choice tests, essay tests tend to be less economical and less efficiently scored (Bennett, 1993). Third, Anastasi (1976) pointed out that because the student writes an essay for a teacher who presumably knows much more about the details of the subject matter than the student, there is a tendency for the student to develop an approach to writing in which obscure ideas are written in a telescoped style that is accessible to the teacher; that style unfortunately carries over to the student's writing for the general reader. Fourth, although this can be mitigated by the application of rubrics and other strategies, there is greater subjectivity in scoring essays than in scoring multiple-choice tests. Fifth, students prefer multiple-choice tests to essay tests (Zeidner, 1993). Sixth, constructed-response tests, including essay tests, compare less favorably to multiple-choice tests in terms of predictive validity (Bennett, 1993).

In summary, while there is no consensus on the equivalency of open- and closed-response items, it is clear that equivalent measures may in some cases be derived using different item formats, and that the greater similarity between items (i.e., stem equivalency) increases the degree of correlation between measures derived from them. The CINS and ORI did not share stem-equivalent questions or formats, and this difference alone may account for the discrepancies between the results that we document.

Practical Considerations

Many of the advantages and disadvantages of open- and closed-response instruments outlined by Kuechler and Simkin (2003, p. 396) apply to the CINS and ORI. In our study using the CINS, we found that a large sample (~100) may be scored and analyzed in about an hour, whereas scoring a comparable sample of open-response instruments requires more than 20 hours. Additionally, scoring the ORI requires considerable expertise and training, even with a well-designed grading rubric. To ensure the reliability of essay scoring, the ORI also requires two graders, which adds additional time, effort, and training. Finally, in order to calculate knowledge and alternative conception measures (e.g., the NSPQ and diversity scores), the ORI requires data entry and statistical analyses.

The ORI carries additional and more serious disadvantages than time. Students' aversion to writing may in some cases lead to limited responses or errors of omission, and poor writing skills may hamper clear communication, preventing the instructor from recognizing the extent of the student's knowledge; both situations will produce scores that inaccurately reflect student knowledge. Additionally, because of the time required to construct responses to the ORI, it is not possible to test a broad array of content knowledge. Collectively, these factors are likely to deter researchers from implementing the ORI in large samples.

In addition to the general limitations of open-response instruments discussed above, several specific limitations were found to characterize the ORI. Several questions had marginal discriminability and high difficulty (Table 5), and the "speeding-up evolution" question appeared to be particularly challenging for the first-year biology majors in our sample (Table 5). Thus, in the student population studied here, the ORI appears to have intrinsic limitations. It is important to point out that relaxing the benchmark of three key concepts, which Nehm and Reilly (2007) considered to be indicative of understanding natural selection, and relaxing the benchmark of one alternative conception as being indicative of a wrong answer, would significantly alter both discriminability and difficulty values and thereby modify this interpretation. Nevertheless, relaxing the scoring benchmarks would not transform the ORI into a significantly less difficult test for the biology majors that we studied.

CINS Difficulty and Appropriate Populations for Testing

The CINS was originally designed to be of use in measuring knowledge of and alternative conceptions about natural selection in undergraduate non-majors, but we found it to be well-suited to our sample of first-year biology majors. This interpretation is also in line with data from the original CINS study, where very-low average CINS scores were reported for undergraduate non-majors (Anderson et al., 2002, p. 963). Indeed, the average scores for non-majors in the two samples Anderson et al. studied were failing (41/100, $n = 110$, and 52/100, $n = 96$). Although Anderson et al. (2002) concluded that their test was well-suited for non-majors, the average performance scores for their samples, in combination with the marginal CINS scores reported in the present study of majors, suggest that their instrument may be better suited for first-year biology majors. The marginal discriminability and moderate difficulty of many CINS questions documented in our study (and several items in Anderson et al.'s original study), suggest that this test is very difficult for undergraduate non-majors.

The Rasch analyses performed here also provide useful information on the fit between the instrument items and the sample who took it. As Table 9 illustrates, the biology majors in our samples are well-matched to the distribution of items. The Rasch analyses also provide useful information on the difficulty of particular CINS items and how the instrument could be improved. Specifically, many questions of redundant difficulty (e.g., items 8 and 12) could be altered to provide a more precise measure of student knowledge using this instrument. Likewise, if the test is used to measure knowledge in first-year biology majors similar to those studied here, questions that differentiate high performing students will need to be added.

Finally, some researchers may be inclined to use pairs of CINS items that measure the same key concept of natural selection (e.g., items 2 and 14 concerning "population stability") to measure pre-post-knowledge gains associated with particular instructional interventions. This inclination should not be acted upon, however, as the Rasch analysis revealed that parallel items about the same key concepts of natural selection (e.g., items 2 and 14) have greatly different difficulty levels. Thus, dividing these parallel pairs of questions into pre- and post-tests could significantly bias learning gain measures. Likewise, the ‘‘salamander’’ and ‘‘cheetah’’ questions from the ORI are also not of comparable difficulty, and are therefore not well-suited for pre-post-testing. In contrast, the bacteria and cheetah questions are most comparable. It would be worthwhile, however, to develop parallel pairs of CINS items of comparable difficulty about the same key concepts of natural selection for use in pre-post-testing. Likewise, developing a new question of comparable difficulty to the ORI ‘‘salamander’’ question would be beneficial.

Knowledge, Alternative Conceptions, and Sociocultural Contexts

Longstanding work in education and cognition has indicated that science understanding emerges from complex interactions in localized social, cultural, linguistic, and naturalistic contexts (e.g., Atran, 1990; Vygotsky, 1978). It has also been shown, however, that some alternative conceptions in science are widespread and transcend particular racial, cultural, and naturalistic contexts (e.g., Azizoglu, Alkan, & Geban, 2006; Chiu, 2005; Wandersee et al., 1994). Because ours is the first to study knowledge and alternative conceptions of natural selection in a sample comprised primarily of underrepresented minorities, it provides an opportunity to explore whether knowledge and alternative conceptions of natural selection that have been extensively documented in primarily white, middle-class samples are in fact more widespread (Scheurich & Young, 1998).

Overall, our results suggest that while the magnitudes of alternative conceptions of natural selection may differ among students from different cultural, ethnic and/or class backgrounds, in most instances the alternative conceptions themselves do not. For example, Nehm and Schonfeld’s (2007) study of 44 biology teachers, most of whom were not from underrepresented groups, displayed all of the alternative conceptions documented in the present study. Bishop and Anderson’s (1990) study sample, Anderson et al.’s (2002) study sample, and many other samples (e.g., Brumby, 1984; Clough & Wood-Robinson, 1985; Ingram & Nelson, 2006) also comprised mostly white non-Hispanic students, and they likewise documented many of the alternative conceptions uncovered here. Finally, while the results of this study suggest that alternative conceptions of natural selection transcend racial, ethnic and/or class boundaries, they do not imply that the same pedagogical strategies and curricular frameworks will be equally effective for ameliorating alternative conceptions of natural selection in these different student populations.

References

Anastasi, A. (1976). Psychological testing. New York: Macmillan.

Anderson, D.L., Fisher, K.M., & Norman, G.J. (2002). Development and evaluation of the conceptual inventory of natural science. Journal of Research in Science Teaching, 39, 952–978.

Ary, D., Jacobs, L.C., & Razavieh, A. (2002). Introduction to research in education (6th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Atran, S. (1990). Cognitive foundations of natural history. Cambridge: Cambridge University Press.

Azizoglu, N., Alkan, M., & Geban, Ö. (2006). Undergraduate pre-service teachers’ understandings and misconceptions of phase equilibrium. Journal of Chemical Education, 83, 947–953.

Bennett, R.E. (1993). On the meaning of constructed response. In R.N. Bennett & W.C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 1–28). Hillsdale, NJ: Erlbaum.

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

Bloom, B. (1956). Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain. New York: David McKay.

Bode, R.K., & Wright, B.D. (1999). Rasch measurement in higher education. In J.C. Smart & W.G. Tierney (Eds.), Handbook of Theory and Research, Volume XIV. New York: Agathon Press.

Boone, W.J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. Science Education, 90, 253–269.

Bond, T.G., & Fox, C.M. (2001). Applying the Rasch Model: Fundamental measurement in the human sciences. Mahwah NJ: Lawrence Erlbaum Associates.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple choice formats. Journal of Educational Measurement, 29, 253–271.

Brooks, D.J. (2001). Substantial numbers of Americans continue to doubt evolution as explanation for origin of humans. (March 5, 2001) Available online at: http://www.gallup.com/poll/releases/pr010305.asp.

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

Carver, R.P. (1974). Two dimensions of tests: Psychometric and edumetric. American Psychologist, 29, 512–518.

Chiu, M. (2005). A national survey of students' conceptions in chemistry in Taiwan. Chemical Education International, 6, 1–8.

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

Colburn, A., & Henriques, L. (2006). Clergy views on evolution, creationism, science, and religion. Journal of Research in Science Teaching, 43, 419–442.

Crawford, B.A., Zembal-Saul, C., Munford, D., & Friedrichsen, P. (2005). Confronting prospective teachers' ideas of evolution and scientific inquiry using technology and inquiry-based tasks. Journal of Research in Science Teaching, 42, 613–637.

Cronbach, L.J. (1975). Dissent from Carver. American Psychologist, 30, 602.

Dagher, Z.R., & BouJaoude, S. (1997). Scientific views and religious beliefs of college students: The case of biological evolution. Journal of Research in Science Teaching, 34, 429–445.

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

Donnelly, L.A., & Boone, W.J. (2007). Biology teachers' attitudes toward and use of Indiana's evolution standards. Journal of Research in Science Teaching, 44, 236–257.

Gould, S.J. (2002). The Structure of Evolutionary Theory. Cambridge: Belknap of Harvard University Press.

Grose, E.C., & Simpson, R.D. (1982). Attitude of introductory college biology students toward evolution. Journal of Research in Science Teaching, 19, 15–24.

Haladyna, T.M. (1975). On the psychometric-edumetric dimensions of tests. American Psychologist, 30, 603–604.

Ingram, E., & Nelson, C. (2006). Relationship between achievement and students' acceptance of evolution or creation in an upper-level evolution course. Journal of Research in Science Teaching, 43, 7–24.

Jackson, D.F., Doster, E.C., Meadows, L., & Wood, T. (1995). Hearts and minds in the science classroom: The education of a confirmed evolutionist. Journal of Research in Science Teaching, 32, 585–611.

King, D.J. (1975). Control groups. American Psychologist, 30, 602.

Kitcher, P. (2007). Living with Darwin. Oxford: Oxford University Press.

Kuechler, W.L., & Simkin, M.G. (2003). How Well Do Multiple Choice Tests Evaluate Student Understanding in Computer Programming Classes? Journal of Information Systems Education, 14, 389–399.

Lerner, L.S. (2000). Good science, bad science: Teaching of evolution in the states. Washington, DC: Thomas B. Fordham Foundation.

Linacre, J.M. (2006). WINSTEPS: Rasch measurement computer program. Chicago: Winsteps.com.

Libarkin, J.C., & Anderson, S.W. (2006). The Geoscience Concept Inventory: Application of Rasch analysis to concept inventory in higher education. In X. Liu & W.J. Boone (Eds.), Applications of Rasch measurement in science education (pp. 45–73). Minnesota: JAM Press.

Liu, X., & Boone, W.J. (2006). Applications of Rasch measurement in science education. Minnesota: JAM Press.

Liu, X., & Lesniak, K. (2005). Students' progression of understanding the matter concept from elementary to high school. Journal of Research in Science Teaching, 89, 433–450.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement tests. Journal of Educational Measurement, 30, 234–250.

Mayr, E. (1982). The growth of biological thought. Cambridge, MA: Harvard University Press.

Millman, J., & Popham, W.J. (1974). The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 11, 137–138.

Moore, R. (2001). Teaching evolution: Do state standards matter? Reports of the National Center for Science Education, 21, 19–21.

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

National Research Council (NRC). (2001). Knowing What Students Know. Washington, DC: National Academies Press.

Nehm, R.H. (2006). Faith-based evolution education? Bioscience, 56, 638–639.

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

Nehm, R.H., & Schonfeld, I.S. (2007). Does increasing biology teacher knowledge about evolution and the nature of science lead to greater advocacy for teaching evolution in schools? Journal of Science Teacher Education, 18, 693–794.

Newport, F. (2004). Third of Americans Say Evidence Has Supported Darwin's Evolution Theory; Almost half of Americans believe God created humans 10,000 years ago. Gallup Organization, November 19.

Popham, W.J. (2006). Assessment for educational leaders. New York: Pearson.

Rodriguez, M. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. Journal of Educational Measurement, 40, 163–184.

Scharmann, L.C., & Harris, W.M. (1992). Teaching evolution: Understanding and applying the nature of science. Journal of Research in Science Teaching, 29, 375–388.

Scharmann, L.C. (1993). Teaching evolution: Designing successful instruction. The American Biology Teacher, 55, 481–486.

Scheurich, J.J., & Young, M.D. (1998). Coloring epistemologies: Are our research epistemologies racially colored? Educational Researcher, 26, 4–16.

Skoog, G., & Bilica, K. (2002). The emphasis given to evolution in state standards: A lever for change in evolution education? Science Education, 86, 445–462.

Sinatra, G.M., Southerland, S.A., McConaughy, F., & Demastes, J.W. (2003). Intentions and beliefs in students' understanding and acceptance of biological evolution. Journal of Research in Science Teaching, 40, 510–528.

Southerland, S., & Sinatra, G. (2003). Learning about biological evolution: A special case of intentional conceptual change. Chapter 11. In: G. Sinatra & P.R. Pintrich (Eds.), Intentional conceptual change Mahwah, NJ: Lawrence Erlbaum Associates.

Traub, R.E., & MacRury, K. (1990). Multiple-choice vs. free-response in the testing of scholastic achievement. Unpublished manuscript, The Ontario Institute for Studies in Education. Published in German. In: K. Ingenkamp & R.S. Jager (Eds.), Tests und trends 8: Jahrbuch der Paaizgogischen Diagnosrik (pp. 128–159). Weinheim & Basel: Beltz Verlag.

Vygotsky, L.S. (1978). Mind and society: The development of higher mental processes. Cambridge, MA: Harvard University Press.

Wandersee, J.H., Mintzes, J.J., & Novak, J.D. (1994). Research on alternative conceptions in science. In: D. Gabel (Ed.), Handbook of Research on Science Teaching and Learning (pp. 177–210). Simon & Schuster Macmillan: New York.

Zeidner, M. (1993). Essay versus multiple choice type classroom exams: The student's perspective. In B. Nevo & R.S. Jager (Eds.), Educational and psychological testing: The test-takers outlook (pp. 67–82). Toronto: Hogrefe & Huber.

Zimmerman, M. (1987). The evolution-creation controversy: Opinions of Ohio high school biology teachers. Ohio Journal of Science, 87(4), 115–125.