JOURNAL OF RESEARCH IN SCIENCE TEACHING

Item Feature Effects in Evolution Assessment

Ross H. Nehm, Minsu Ha

College of Education and Human Ecology, Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, 1945 N. High St., Columbus, Ohio 43210

Received 21 January 2010; Accepted 28 October 2010

Abstract: Despite concerted efforts by science educators to understand patterns of evolutionary reasoning in science students and teachers, the vast majority of evolution education studies have failed to carefully consider or control for item feature effects in knowledge measurement. Our study explores whether robust contextualization patterns emerge within particular evolutionary reasoning contexts, and the implications of these patterns for instruction, assessment, and models of cognition. We test four hypotheses regarding item feature effects on undergraduate biology majors' evolutionary reasoning using a sample of 1,200 open response explanations of evolutionary change across items differing in context and scale but standardized by taxon and trait. Evolutionary explanations were atomized into a series of scientific and naïve biological elements and tallied among prompts and their features. We documented clear, significant, and predictable item feature effects on evolutionary explanations. Tasks involving evolutionary trait loss elicited a significantly greater number of naïve biological elements than evolutionary trait gain tasks in all contexts, including: within species comparisons, between species comparisons, animal prompts, and plant prompts. Tasks involving between species evolutionary comparisons, regardless of gain or loss, animal or plant, always produced significantly more naïve biological explanatory elements than within species comparisons. For items prompting explanation of trait gain, the use of the core concepts of natural selection were not influenced by the hierarchical level of the task (within or between species). Explanations of trait gain were also the least sensitive to scale and context. Core concepts of natural selection were always deployed less frequently in cases of evolutionary trait loss (within and between species, in animals and plants). We discuss a series of implications of these findings for curriculum, instruction, and assessment. J Res Sci Teach

Keywords: assessment; measurement; evolution; item features; context; misconceptions; natural selection; undergraduates; coherence

A large body of work in cognitive psychology and cognitive science has explored how assessment task formats, features, contexts, and semantic structures constrain and facilitate knowledge retrieval (Chi, Feltovich, & Glaser, 1981; diSessa, Gillespie, & Esterly, 2004; Gentner & Toupin, 1986; Greeno, 2009; Sabella & Redish, 2007; Silver, 1979). Domain-specific studies in science education have likewise explored how assessment tasks differentially reveal the composition and coherence of student knowledge and alternative conceptions (Clough & Driver, 1986; Jones, Carter, & Rua, 2000; NRC, 1996; Ozdemir & Clark, 2009; Palmer, 1999; Taasoobshirazi & Glynn, 2009). Despite the well-established findings from these research communities that assessment task features significantly control knowledge elicitation in many science domains—particularly physics and chemistry—remarkably little empirical work has investigated these issues in the biological sciences (Clough & Driver, 1986).

After more than 30 years of concerted efforts by science educators to assess patterns of evolutionary reasoning in science students, college undergraduates, teachers, and the general public (e.g., Clough & Driver, 1986; Crawford, Zembal-Saul, Munford, & Friedrichsen, 2005; Deadman & Kelly, 1978; Demastes, Good, & Peebles, 1996; Evans et al., 2010; Goldston & Kyzer, 2009), the vast majority of such studies

Contract grant sponsor: National Science Foundation REESE; Contract grant number: 0909999.
Correspondence to: R.H. Nehm; E-mail: nehm.1@osu.edu
DOI 10.1002/tea.20400
Published online in Wiley Online Library (wileyonlinelibrary.com).

continue to haphazardly and inconsistently employ a diverse array of context features in items designed to measure evolutionary knowledge and alternative conceptions. This motivates three simple questions: (1) Can particular prompt features/contexts be identified that constrain or facilitate particular evolutionary reasoning patterns? (2) If clear context effects can be established, how should assessments be designed to profitably make use of such findings? And (3) If significant context effects can be established, what implications do such findings have for decades of work attempting to assess student and teacher knowledge of evolution? We begin by briefly reviewing major findings on the effects of prompt features in science education broadly, and subsequently focus on biology and evolution in detail.

Context Effects in Science Assessment Tasks

Research in many science domains has explored how assessment item features associate with knowledge elicitation patterns. In a study of thermal equilibrium, Clark (2006, pp. 521–522) found that ''context apparently played a significant role in fostering and cueing. . . multiple contradictory ideas.'' Also in chemistry, Wason and Shapiro (1971) found that student ideas about combustion were related to the material sources (e.g., wood, metal) used in assessment tasks. Likewise, in a study of chemical knowledge, Jones et al. (2000) found that ''. . .the variation and idiosyncratic application of prior experiences was individually contextual and therefore incredibly diverse.'' In a study of conceptions of volume, Potari and Spiliotopoulou (1996) found that student ideas were highly dependent upon the shape, mass, and material used in assessment items.

Within the domains of physics—and to a lesser extent earth science—a large and broadly disseminated body of work has investigated numerous theoretical and methodological issues relating to item context effects on knowledge elicitation (Chu, Treagust, & Chandrasegaran, 2009; Ozdemir & Clark, 2009). Overall, this work has shown that different physical features (such as pulleys or ramps) testing the same scientific concept typically cued significantly different student conceptions (Bryce & MacMillan 2009; Chi et al., 1981; Sabella & Redish, 2007).

In the domain of biology, Clough and Driver (1986) appear to be the first to explicitly acknowledge and explore putative context effects in evolutionary explanations, although earlier studies may have done so implicitly (e.g., Brumby, 1979; Deadman & Kelly, 1978). Across evolutionary prompts, Clough and Driver (1986, p. 490) found ''. . .considerable consistency. . . in the use of the accepted scientific framework, but little consistency in use of identifiable alternative frameworks.'' In another detailed exploration of evolutionary explanations, Settlage and Jensen (1996) looked for, but failed to find, a consistent contextual factor—a so-called Disney effect (empathy for cute creatures)—in natural selection item responses. They did, however, find that parallel items elicited substantially different response patterns.

Samarapungavan and Wiers (1997) explored the consistency of children's ideas about the origin of species using different prompts. Unlike nearly all of the other studies we reviewed, these authors concluded that context inconsistency was generally absent. In a study of Korean students, Ha, Lee, and Cha (2006) used forced-choice items to explore putative differences in naive conception choice across evolution prompts that differed in context. They found somewhat predictable patterns of co-variation: use/disuse and intentionality ideas were significantly greater in items using human and animal examples, whereas teleological explanations were significantly greater in items using plants.

More recently, Nehm and Reilly (2007) documented how biology majors' responses to different open-response evolutionary scenarios produced significantly different use patterns for so-called key concepts of natural selection and misconceptions. Similarly, Nehm and Schonfeld (2008)—using Item Response Theory—also documented significant performance differences between ''parallel'' (conceptually equivalent) items on multiple-choice natural selection assessments. Unfortunately, like most of the previous studies that we reviewed, Nehm and coworkers failed to explicitly control for context effects in their assessment item design, evaluation, comparison, and interpretation.

Studies continue to be published that identify, but do not explicitly investigate, how evolution assessment item features may be controlling knowledge and misconception measurement. Kampourakis and Zogza (2009), for example, used five prompts (all of which included animals but differed in many other features) to investigate explanatory coherence in evolution, and found marginal response coherence. Evans

et al. (2010), using a small sample ($n < 30$) of museum visitors, employed a series of taxa (e.g., diatoms, flies, humans) in their assessment tasks and, like Clough and Driver (1986), Ha et al. (2006), Nehm and Schonfeld (2007, 2008) and Kampourakis and Zogza (2009), found significant differences in explanatory reasoning patterns—including naïve and scientific concept use—among prompts differing in contexts (in this case, taxon). None of these studies attempted to carefully control for item context effects, precluding identification of generalizations relevant to evolution assessment.

Theoretical Framework

Contextuality in science education has been broadly defined as how people respond and reason in different situations (e.g., diSessa et al., 2004). Such views are consonant with many facets of situated cognition, in which ''each problem is tied to a concrete setting and is resolved by reasoning in situation specific ways. . .'' (Kirsh, 2009, p. 264). The perspective that knowing is inseparable from context contrasts with aspects of classical problem solving theory (e.g., Newell & Simon, 1972). From the stance of situated cognition, tasks are not abstracted or internally represented by their fundamental structure; rather, the nuances of physical contexts matter because they interactively contribute to the framing, conceptualization, and enactment of the tasks at hand. Overall, in this view, problem solving is achieved using contextualized knowledge (Kirsh, 2009).

Importantly, classical problem solving theories do not discount the importance of task or prompt features (Newell & Simon, 1972). Indeed, internal representations and abstractions are recognized as influenced by task features through cuing and schema activation (Sabella & Redish, 2007). Empirical studies have demonstrated that isomorphic items testing for the same concept but differing in superficial features produce significantly different patterns of problem solving success; this is a result of differences in problem abstraction or internal problem representations (e.g., Chi et al., 1981). Furthermore, internal problem search spaces may be constrained by task cues, further supporting the claim that item contexts influence patterns of knowledge elicitation (Sabella & Redish, 2007). Thus, from the stance of both classical problem solving theory and situated cognition, contextuality is a significant contributor to how people perceive, use, internally represent, and solve problems.

Task contextuality is an issue of considerable importance for science assessment, perhaps more so in biology than in other science domains. Context effects in chemistry are highly predictable and empirically established: water, for example, is conceptualized as being composed of H_2O regardless of location and is predicted (and observed) to exist in a liquid state at particular pressures and temperatures; and Carbon is modeled as having six protons regardless of where it is and what other atoms are bonded to it. Likewise, in physics, forces are quantified in predictable relationships regardless of whether one launches a satellite from, say, the United States or Korea today or tomorrow; the relationship between acceleration and mass is not considered by scientists to be variable from place to place on the surface of the Earth.

Unlike all of the prior examples, the core units of biology—individuals and species—are not conceptualized by scientists as structurally or compositionally the same in different places or in different contexts/environments. Indeed, unlike the billions of carbon atoms on Earth, the nearly seven billion humans—or *Escherichia coli*, HIV, or *Maize*—are never modeled as the same across space or time. Likewise, a particular population within the same species is unlikely to be phenotypically or genetically identical to another population from a different place or time. Furthermore, the evolutionary trajectories of populations, species, and clades may be different depending on where and when they live. Indeed, different populations of the same species—subjected to similar biotic and abiotic factors—are known to display different evolutionary trajectories (e.g., see Endler, 1986). Thus, future states—unlike those in physics and chemistry—are generally not predictable in evolution (Gould, 1980). Thus, while context features are known to be important in many fields of science, a case may be made for the fundamental, unique, and complex nature of context in evolutionary thinking and reasoning. Consequently, evolution assessment tasks intended to measure knowledge and/or alternative conceptions may be characterized by heightened sensitivity to context effects, especially if students are reasoning with contextual knowledge, as suggested by theories of situated cognition (Kirsh, 2009).

Evolution Assessment Item Features and Contexts

A perusal of the vast evolution education literature reveals at least five major categories of assessment item features and contexts with which students have been asked to reason (e.g., Bishop & Anderson, 1990; Kampourakis & Zogza, 2009; Nieswandt & Bellomo, 2009) (See Supplementary Materials for many examples). First, comparisons have been framed across different evolutionary *units* or scales: within one species, between an ancestor and descendant species, between a living and an extinct species, between two similar living species, among many different living species, among many similar living species, or among combinations of populations and species (see Supplementary Materials).

Second, prompts have also contained different lineages or *taxa* (some of which are monophyletic, others that are paraphyletic, e.g., ''fish'') with fundamentally different attributes (e.g., ''bacteria,'' plants, animals, and fungi). Third, in addition to different *scales* and *taxa* (or no such specifications), evolutionary prompt features have differed in the types of *traits* (or character states): morphology (e.g., color, size, trait presence, or absence), sensory features (e.g., vision, hearing), and behavior (e.g., running speed).

Fourth, prompts have differed in character state polarity (i.e., gain or loss of a character state or trait): some prompts address evolutionary *gain* of a feature whereas others address the *loss* of a feature. Fifth, some prompts have elicited evolutionary *explanations* in which initial and final states have been specified, whereas others seek *predictions* from an initial state to a future state; and some do not clearly specify states. Overall, the diversity of assessment contexts employed in the literature is extensive, haphazard, and inconsistently balanced (see Supplementary Materials for examples). Studies are urgently needed that explore whether robust and nonrandom contextualization features cue predictable evolutionary reasoning patterns or whether such effects are too ''noisy'' to permit generalizations.

Hypotheses and Predictions

Testing hypotheses about evolutionary item context effects has the potential to: (1) enrich our understanding of the cognitive processes that underlie evolutionary task interpretation (NRC, 1996); (2) inform assessment designs so that significant context effects may be accounted for and appropriately balanced; and (3) establish conclusions regarding the cognitive coherence of evolutionary thinking and reasoning (diSessa, 2008; Vosniadou, 2008).

We test four predictions regarding a constrained set of major context variables on undergraduate biology majors' evolutionary explanations:

(1) Evolutionary tasks employing within-species (intraspecific) contexts elicit a significantly greater number of accurate explanatory elements (''key concepts of natural selection'') than do those employing between-species (interspecific) task contexts.
(2) Evolutionary tasks employing interspecific (between species) contexts elicit a significantly greater number of inaccurate explanatory elements (naïve ideas or ''misconceptions'') than do those employing intraspecific task contexts.
(3) Evolutionary tasks employing intraspecific trait gains elicit significantly greater numbers of accurate explanatory elements (''key concepts of natural selection'') than do those employing intraspecific feature loss contexts.
(4) Evolutionary tasks employing interspecific trait loss elicit a significantly greater number of inaccurate explanatory elements (''misconceptions'') than do those employing interspecific gain tasks.

Materials and Methods

A team of four biologists and biology educators collaboratively developed 12 open-ended assessment items that prompted participants to explain the causes of state changes in biological systems in different contexts and at different hierarchical scales (Table 1 and Supplementary Materials). These items were subsequently reviewed for clarity and readability by an English educator and revised prior to use on our sample. In addition to review by a panel of evolution experts, a pilot study of 37 students enrolled in an undergraduate course in biological evolution was used to examine the validity and reliability of these items. First, we found that the types of key concepts and naïve ideas produced by the new instrument overlapped completely with those documented in previous studies using the ORI instrument of Nehm and Reilly (2007).

Table 1
Item features used in the present study

Scale of Comparison	Evolutionary Direction	Lineage	Organism	Trait
Intraspecific (population)	Gain	Animal	Cheetah	Running speed
	Gain	Animal	Locust	Resistance
	Gain	Plant	Broken bush	Poison
	Loss	Animal	Birds	Flight
	Loss	Animal	Salamander	Eyes
	Loss	Plant	Rose	Thorns
Interspecific (species)	Gain	Animal	Cheetah	Running speed
	Gain	Animal	Locust	Resistance
	Gain	Plant	Broken bush	Poison
	Loss	Animal	Birds	Flight
	Loss	Animal	Salamander	Eyes
	Loss	Plant	Rose	Thorns

This suggests that the new items evoke responses characteristic of evolutionary reasoning. Second, we found a strong and significant correlation between the numbers of correct responses to our new items and overall scores on the CINS instrument ($r = 0.603$, $n = 37$). CINS scores are a proxy for knowledge of natural selection, and therefore these results provide convergent validity evidence (Nehm & Schonfeld, 2010). Third, Cronbach α values of 0.75 ($n = 37$) provide acceptable evidence of reliability, particularly given the small number of items. Overall, these three lines of evidence support the interpretation that the responses to our assessment items are valid and reliable indicators of student thinking about evolution.

Six of the items in our new instrument focused on the scale of populations (i.e., intraspecific or within species comparisons) and six items focused on the scale of species (i.e., interspecific or between species comparisons). In the six intraspecific items, three items dealt with the gain of an evolutionary trait (or character state), and three items dealt with the loss of an evolutionary trait.

The same was also true of the six interspecific items. Within the intraspecific item set, three items incorporated taxon/trait state changes as gains using both animals and plants: animal (cheetah → gain of running speed), animal (locust → gain of resistance), plant (broken bush → gain poison). Also within the intraspecific item set, three items incorporated taxon/trait state changes as loss (again using both animals and plants): animal (salamander → loss of eyes), animal (birds → loss of flight), and plant (roses → loss of thorns). The within and between species item sets (six items each, 12 total) were constructed in parallel form, with the same taxon/trait/state change patterns (e.g., birds/flight/loss for both intra and interspecific items). However, different hierarchical scales characterized the within and between-species item sets: for the intraspecific item set, all items prompted explanation of trait/state changes within one population, whereas for the interspecific item set, all items prompted explanation of trait/state changes between two species. The main units of comparison therefore were: within versus between species (standardized by taxon, trait, and state), and evolutionary gain and evolutionary loss.

Our sample of essay responses was gathered using an online response system built within the university course management system. Evolutionary explanations were extracted from undergraduate student participants enrolled in the introductory biology sequence for majors at a large public Midwestern university. This sample was randomized into two groups, with one group assigned the six intraspecific prompts and the second group receiving the six interspecific prompts. Students ($n > 200$) received extra course points for choosing to participate in the study, which involved responding to these six evolutionary prompts. Participation rates were $> 75\%$, and 1,200 sufficiently complete essays (600 for intraspecific and 600 for interspecific) were gathered and scored in our current study.

Students' evolutionary explanations were atomized into a series of units using a scoring rubric established in prior research and validated using extended clinical interviews (for details, see Nehm & Reilly, 2007; Nehm & Schonfeld, 2008). The first set of units extracted from participants' evolutionary explanations pertains to the scientifically established causal elements used to explain evolutionary change via natural selection.

The construct of natural selection—and its constituent elements—is generally well established and agreed upon (Nehm & Schonfeld, 2008). Nevertheless, it is important to note that there is some variance in the literature regarding the number of ''essential'' elements that comprise this construct (Nehm & Schonfeld, 2010). At a minimum, three key concepts (KCs) are considered necessary and sufficient to explain evolutionary patterns using the natural selection model: (1) The presence and causes of variation (mutation, recombination, sex); (2) The heritability of variation; (3) The differential reproduction and/or survival of individuals (Endler, 1986, p. 220; Lewontin, 1970; Patterson, 1978, p. 1; Pigliucci & Kaplan, 2006, p. 14). Many other authors also acknowledge: (4) Hyper-fecundity or ''overproduction'' of offspring; (5) Limited resources, (6) Competition, and (7) A change in the distribution of produced phenotypic/genotypic variation in the next generation (Endler, 1986; Patterson, 1978). Other authors extend the list of elements even further, and include ''population stability'' and ''speciation'' (Anderson, Fisher, & Norman, 2002). Perhaps the most debate exists as to whether ''speciation'' is a necessary element of natural selection—or whether a consensus exists in regard to this issue (Gould, 2002).

Our analyses of students' scientifically accurate explanatory elements straddle a middle ground between the above extremes, and code for seven of the most commonly accepted elements (1–7, above). Given the very infrequent invocation of the elements ''population stability'' and ''speciation'' by comparable samples in prior studies (notably, 0% of participants used them in Nehm & Schonfeld 2007, 2008 or Nehm, Kim, & Sheppard, 2009), we did not include them in our study. Our coding of explanatory elements thus encompasses the three ''essential'' elements—what we subsequently refer to as ''core concepts''—along with the most widely accepted additional elements denoted by evolutionary biologists (what we refer to as ''key concepts''; see also Endler, 1986; Gould, 2002; Patterson, 1978).

The first set of variables extracted from the 1,200 essays related to student knowledge of these seven concepts of natural selection (Mayr, 1982). The coding rubric was used to quantify the presence or absence of these seven key concepts in each of the students' 12 tasks (see above). Two scorers independently coded 1,200 essay responses for the presence or absence of these units (Mean inter-rater reliability values, measured using Kappa, were > 0.8 for all variables; see Ary, Jacobs, & Razavieh, 2002, pp. 265–266 for details). In cases of disagreement, all coding discrepancies were resolved via deliberation. These concept scores were tallied separately for each of the 12 item tasks, and collectively for each student. In addition, the number of different key and core concepts used among all items (hereafter: key concept diversity) was scored in this manner.

In addition to the scoring of scientifically established explanatory elements, we coded contextually inaccurate explanatory units (Nehm, Rector, & Ha, 2010). Regardless of whether these units are best described as misconceptions, alternative conceptions, phenomenological primitives, or cognitive resources (see diSessa, 2008), they are scientifically and contextually inaccurate explanatory elements. Rather than coin a new term for such elements, we will subsequently refer to them as naive ideas. Examples of naive ideas include the notion that ''needs and/or goals cause evolutionary change,'' that ''pressures'' applied to organisms can ''push'' them to change, and that the disuse of phenotypic features proximally produces evolutionary loss. All of the naive ideas that we identified have been discussed in the literature (e.g., Bishop & Anderson, 1990; Nehm & Schonfeld, 2008). Two scorers independently coded the 1,200 essay responses for these naive ideas and in cases of disagreement reconciled all coding discrepancies. Naive ideas were scored in a similar manner as key and core concepts of natural selection, where each naïve idea was given one point (see Nehm & Schonfeld 2007, 2008). Naive ideas were tallied for each prompt and collectively for each student. In addition, we calculated the number of different naive ideas used among the six items.

Core concept (CC), key concept (KC), and naïve idea scores were used in three two-way Analyses of Variance (ANOVA) (i.e., CC scores $\times$ scale [between vs. within species] $\times$ pattern [evolutionary gain vs. loss]; KC scores $\times$ scale $\times$ pattern; naïve idea scores $\times$ scale $\times$ pattern). Correlation coefficients were also calculated between CC and naïve idea scores within and between tasks.

Results

Student Explanations

Prior to presenting the frequencies of naïve and scientific elements relative to prompt contexts, we provide an example from one of the 1,200 explanations that we coded for scientific and naïve explanatory

elements. Several additional coded examples may be reviewed in the Supplementary Materials accompanying this article. These examples include the six taxon/trait combinations used in our study (see Materials and Methods Section) and illustrate the diversity of scientifically accurate and naïve elements that our tasks elicited (see Supplementary Materials, Appendix A, for complete items and Materials and Methods Section for explanatory element descriptions). Italicized text is indicative of a particular coding concept (e.g., a Key Concept). Numbers below the quotations refer to sample (S) and participant (P) codes.

> "The island probably lacks the large herbivores [KC5] which are the primary reason thorns are beneficial to mainland plants [Need/goal]. Upon arriving to the island, individuals with smaller thorns had more energy to spend elsewhere [Energy allocation] and had greater relative reproductive success [KC6] than those individuals that spent a lot of energy on thorns. So, over time, alleles for small or absent thorns [KC1] spread throughout the population. Also, the change could be due to a small founding population, the members of which randomly happened to have alleles [KC1] for no thorn." (S114, P40: Rose thorn item [between species]; Naïve ideas: "needs/goals" and "energy allocation", scientific concepts: KC1, KC5, KC6).

Item Context Effects: Evolutionary Scale and Trait Change

Two-way ANOVAs were used to explore context effects and their putative interactions with the constituent explanatory elements (naïve and scientific) from students' responses. Specifically, we compared the effects of evolutionary scale [between vs. within species] and trait polarity [loss vs. gain] on the frequencies of (1) key concept (KC) use, (2) core concepts of natural selection (CCNS) use, and (3) contextually inaccurate explanatory element—or naive idea-use (naïve ideas) (Figure 1).

For naïve ideas (Figure 1A), we found significant main effects for both evolutionary scale and the polarity of trait change (Scale: $F_{1,410} = 67.8, p < 0.01$; Trait: $F_{1,410} = 55.7, p < 0.01$), indicating greater naïve idea scores for trait loss scenarios at both evolutionary scales (within-species *and* between-species). We also found a significant interaction effect for naïve ideas between scale and trait change ($F_{1,410} = 8.2, p < 0.01$). Significant main effects were also found for evolutionary scale and trait change on key concept frequency (KC) (Figure 1B: Scale: $F_{1,410} = 60.3, p < 0.01$; Trait: $F_{1,410} = 555.1, p < 0.01$). As Figure 1B illustrates, trait gain items produce greater magnitudes of key concepts at both hierarchical scales (within and between species). As with naive ideas, we found significant interaction effects between scale and trait change for KC ($F_{1,410} = 6.1, p < 0.05$). Finally, and perhaps unsurprisingly, we found very similar patterns for core concepts of natural selection (CCNS) as we did for key concepts (KC). Specifically, we found significant main effects and interaction effects for CCNS (Scale: $F_{1,410} = 76.8$, $p < 0.01$; Trait: $F_{1,410} = 76.8$, $p < 0.01$; Interaction: $F_{1,410} = 31.4$, $p < 0.05$). These patterns mirrored those of key concepts (Figure 1B,C).

In summary, our two-way ANOVAs documented clear and significant effects and interactions among evolutionary scale (intra- vs. interspecific levels) and trait change polarity (loss vs. gain) on the types of students' explanatory elements. Explanations of evolutionary loss were associated with a significantly greater number of naïve biological elements in all contexts; explanations of evolutionary gain included a significantly greater number of core and key concepts of natural selection; within-species explanations produced significantly more core and key concepts of natural selection and less naïve ideas, whereas between-species contexts were associated with fewer core and key concepts of natural selection as well as significantly more naïve ideas.

Assessment Item Effects on Explanation Patterns

Although generally consistent patterns of evolutionary explanation emerged between task contexts (trait gain/loss and within/between species), variation in the frequencies of core concepts and naïve conceptions within these categories were also noted (Figure 2). In terms of core concept use in the trait gain items (Figure 2, top panel), the locust, broken bush, and cheetah items did not display significant differences at different scales (within/between species). Nevertheless, the locust resistance item always elicited a greater number of core concepts relative to the bush and cheetah items, which themselves produced similar levels of core concepts. In terms of core concept use in the trait loss items, the bird flight item produced the highest

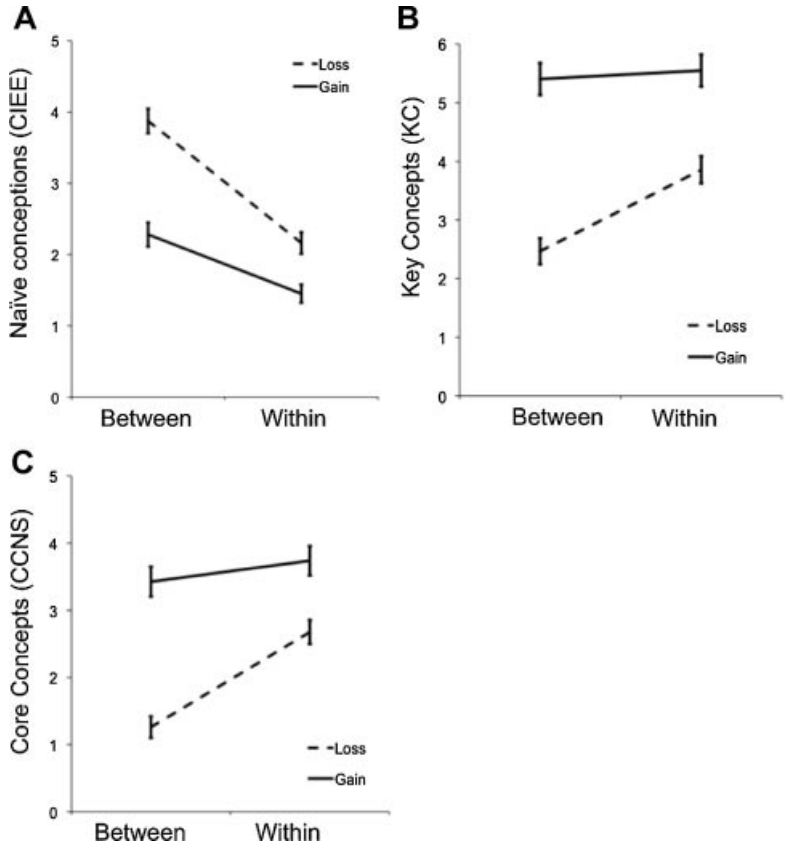

Figure 1. Two-way ANOVAS of context effects and their interactions. A: Naïve ideas (contextually inaccurate explanatory elements). B: Key concepts of natural selection. C: Core concepts of natural selection (KC1, KC2, and KC6). Between, between species; Within, within species; Gain, train gain items; Loss, trait loss items. Error bars represent 1 standard error about the mean.

average number of core concepts in the intraspecific context but the least number of core concepts in the interspecific context—this item was the most sensitive to the ''scale'' context. In contrast, the rose thorn and salamander vision items produced comparable magnitudes of core concepts in both the intra and interspecific contexts. Overall, as noted above, more core concepts were always elicited—regardless of taxon or trait—in the within species contexts.

Less consistent patterns of naive idea use emerged between task contexts (trait polarity and evolutionary scale) (Figure 2, bottom panel). The most consistent elicitation of naïve ideas occurred among the locust, bush, and cheetah intraspecific gain items: in these cases, no significant differences among items were noted (Figure 2, bottom panel). This was not the case, however, with the interspecific loss or gain items or the interspecific gain items. Here, in many comparisons, significant differences in naïve idea use are apparent. For example, the bush poison item elicited a significantly greater number of naïve ideas than the cheetah speed item in the between species context, as did the salamander vision item and the rose thorn item. Differences are also apparent among items within contexts (e.g., interspecific). The salamander vision item on average elicited twice as many naïve ideas as the cheetah speed item. However, as noted above, more naïve ideas were always elicited—regardless of taxon or trait—in the between species contexts relative to the within species contexts, and naïve idea use was more variable than core concept use in all contexts.

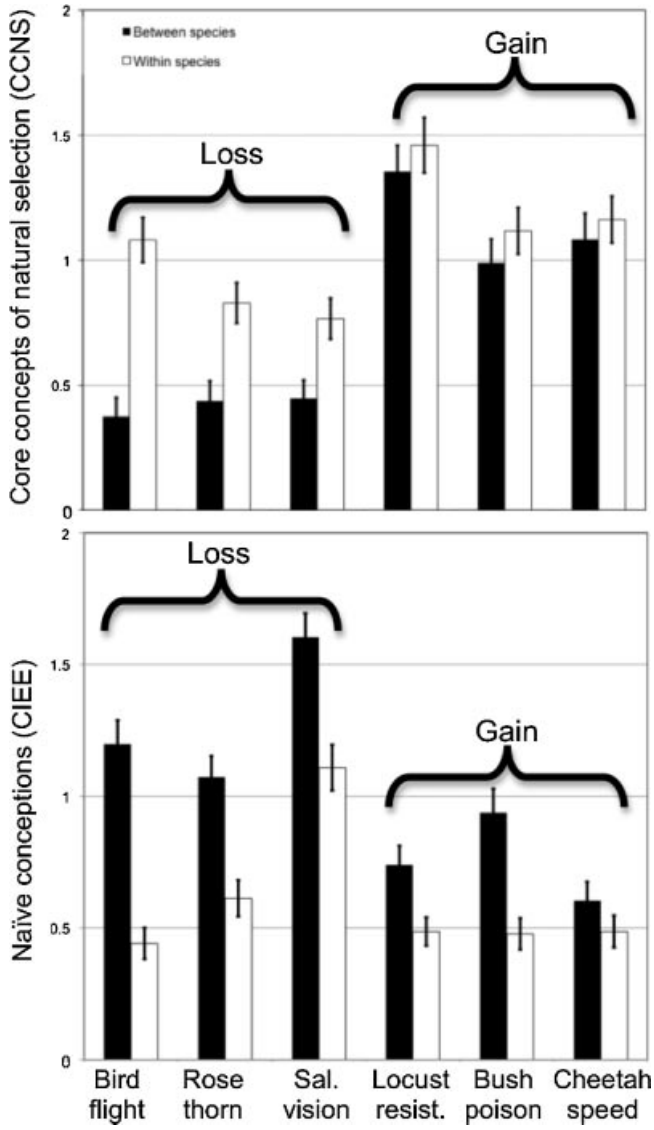

Figure 2. Item effects on evolutionary explanations. Upper panel: mean values (and standard errors) for core concepts of natural selection (CCNS). Lower panel: mean values (and standard errors) of naïve ideas (contextually inaccurate explanatory elements). See Appendix for complete item descriptions. Loss, items involving trait loss; Gain, items involving trait gain; Sal., salamander; Resist., resistance.

In summary, three major patterns were noted in the item comparisons. First, naive idea elicitation—regardless of context—is always more variable among items than core concept elicitation. Second, the *same* taxon/trait combinations (e.g., birds/flight) elicit *different* magnitudes of both scientifically accurate and naïve explanatory elements depending on scale (within/between species). Third, some assessment items, such as the locust resistance item and the salamander vision item, elicit high magnitudes of explanatory elements regardless of context.

Naïve Biological Ideas

Six categories of naïve ideas were identified among the 12 items: (1) needs and goals, (2) use and disuse, (3) intentionality, (4) adapting and acquiring traits, (5) deliberate energy allocation, and (6) pressure as a direct cause of change. Comparing the top and bottom panels of Figure 3, it is apparent that trait loss items and trait gain items produced different types and frequencies of naïve ideas; use/disuse and energy allocation are nearly absent from trait gain contexts but present in trait loss contexts. Additionally, the hierarchical scale of the items influenced the elicitation of naive ideas. Between-species contexts (illustrated in black) typically produced more naïve ideas than within-species contexts. For example, for trait loss items, needs and goals were used more than twice as often in between-species contexts as they were in within-species contexts. Likewise, pressure (for trait loss) and adapt/acquire (for trait gain) were used significantly more frequently in interspecific contexts. Some of the same naïve ideas were also used in different magnitudes depending on trait gain or loss patterns. Pressure as a cause of change, for example, was evoked much more frequently in trait loss contexts than in the trait gain contexts. Finally, some naïve ideas were evoked in comparable magnitudes despite different contexts. In gain items, pressure was used similarly both within and between species, as was adapt/acquire in the context of trait loss. Overall, trait loss and gain contexts typically elicited different types and frequencies of naïve biological ideas, and in most cases intra- and interspecific contexts elicited different frequencies of naïve ideas (Figure 3).

Scientific Elements: Core Concepts of Natural Selection

Three core concepts of natural selection (KC1: variation and its causes; KC2: hereditary variation; KC6: differential survival and reproduction) were identified in students' evolutionary explanations. While our prior analyses revealed clear patterns of collective core concept use in different contexts, they did not examine students' deployment of individual core concepts. As Figure 4 illustrates, students' use of the three core concepts differed significantly in relation to hierarchical level (within vs. between species) and trait change (gain vs. loss). In the between-species context, for example, the three core concepts were used nearly twice as often in trait gain contexts relative to trait loss contexts. The rank order of concept use is relatively consistent within and between species (i.e., KC6 $>$ KC1 $>$ KC2). Nevertheless, not all patterns are consistent among contexts. While KC6 was used most frequently in both intra- and interspecific contexts for trait gain, this was not the case for trait loss. Additionally, while KC2 was the most infrequently employed concept in both gain and loss contexts, it was not used to the same degree (Figure 4). While we do not illustrate parallel results for the key concepts (KC) of natural selection, they revealed comparable patterns (also indicated in our ANOVA results shown above and in Figure 1). Overall, although students did not employ the three core concepts of natural selection at comparable magnitudes, similar use patterns are apparent in (1) trait change contexts and (2) at different hierarchical levels.

Explanatory Element Interrelationships

Our previous analyses atomized evolutionary explanations in order to better understand and test individual concept or ''element'' use in assessment items differing in controlled context features. It is also important to explore how these numerous elements—both scientific and naïve—are assembled and packaged into larger explanatory compounds. As shown in Figure 5, and discussed in detail above, different item contexts displayed different frequencies of scientific and naïve elements. This is represented by circle size and shading. To these well-established patterns we now add the *associations* among elements, which are represented by solid and dashed lines. Solid black lines represent significant positive correlations ($p < 0.01$), solid gray lines represent significant negative correlations ($p < 0.01$), and dashed lines represent the same patterns but at weaker significance levels ($p < 0.05$) (Figure 5).

The first pattern to note is that core scientific elements (KC1, KC2, and KC6) are strongly and positively correlated in all four contexts (despite different frequencies). The second pattern of interest is the frequent positive and significant association among three of the naïve explanatory elements: intentionality, needs/goals, and adapt/acquire. This association cluster occurs in three of the four contexts (between-species, trait gain, and trait loss) but is notably broken in the within species context. The third pattern of interest in the complete absence of negative correlations in the between-species context; here, intentionality, needs/goals,

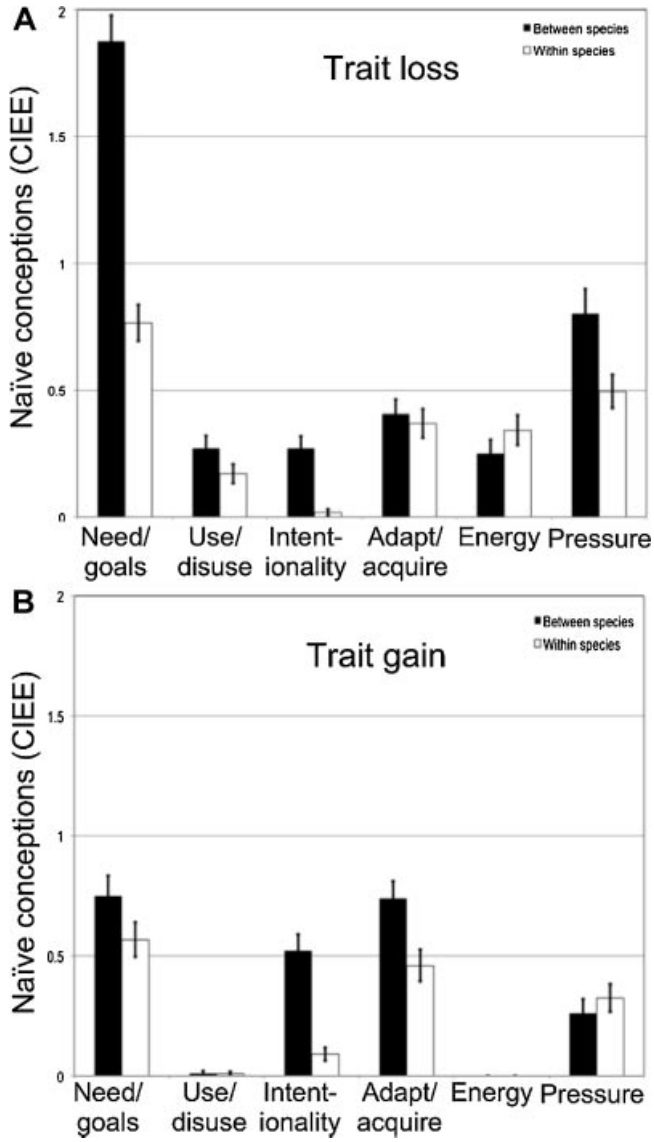

Figure 3. Naïve ideas (contextually inaccurate explanatory elements) in different contexts. A: Trait loss naïve idea patterns. B: Trait gain patterns. See text for detailed explanations of each naïve idea. Error bars represent one standard deviation about the mean.

and adapt/acquire are significantly associated. The fourth pattern is that the naïve idea of pressure is always negatively associated (or not significantly related) with scientific core concepts.

Perhaps the most interesting question is whether particular naive elements are associated with core scientific ideas (and the consistency of such patterns). In the within-species context (Figure 5A), all naïve ideas have negative or no significant correlations with core scientific concepts, but many naïve and scientific elements are associated with one another. In the between-species context (Figure 5B), only energy is significantly and positively correlated with KC1; otherwise, accurate and naïve ideas associate into independent groups notably lacking negative correlations between them. In the trait gain context (Figure 5C)

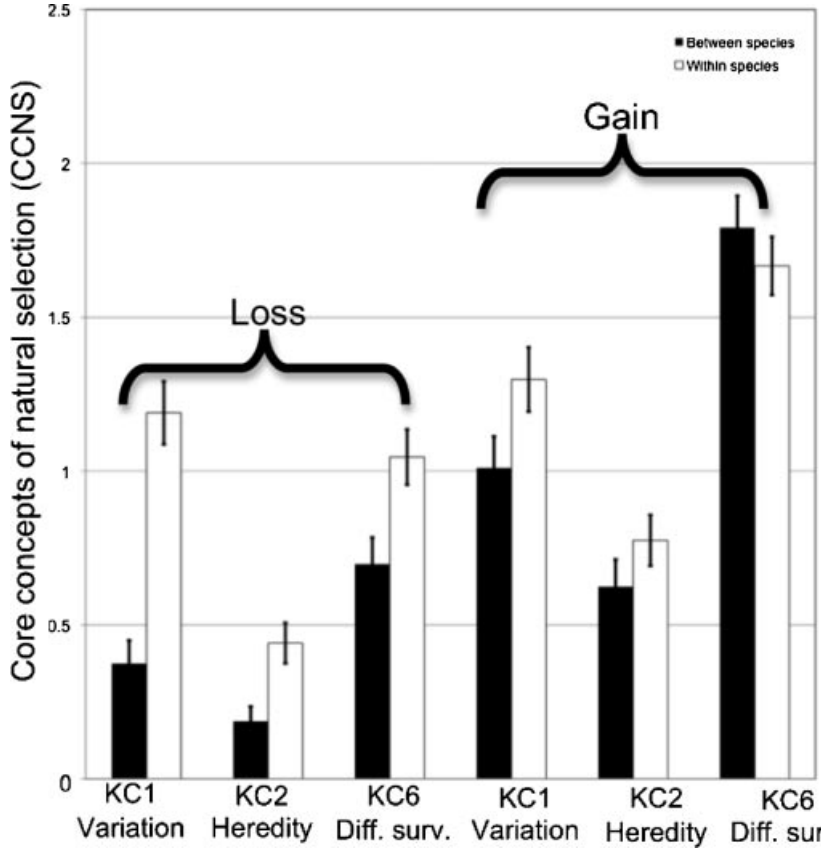

Figure 4. Core concepts of natural selection (CCNS) patterns in different contexts. Black bars, between species values; Loss, trait loss contexts; Gain, trait gain contexts; Diff. surv., differential survival. See text for detailed explanations of each core concept. Error bars represent one standard deviation about the mean.

we again note many significant negative associations between core scientific elements and naïve elements, but strong positive associations within the suite of core scientific ideas; the same is true for naïve ideas. Finally, in the trait loss context (Figure 5D), as in the previous cases we find similar association patterns among elements, with many core scientific elements significantly and positively associated with one another; many naïve elements significantly and positively associated with one another; and negative associations between many naïve and scientific elements. Overall, despite differences in the frequency of use of different elements, naïve elements and scientific elements display relatively distinct and consistent patterns of association in all four contexts.

In addition to examining correlations among explanatory elements, we may also examine the ''purity'' of cognitive resource use patterns; that is, we may ask: How many students used exclusively core or naïve conceptions in their explanations? How many participants used both naïve and scientific elements in their evolutionary explanations? We found that in no contexts (i.e., among the 12 items) did the percentage of biology majors using scientific elements alone reach a large majority (minimum 4.2%, maximum 52.3%) (Table 2). Additionally, many explanations were devoid of any identifiable scientifically accurate or naive concepts; that is, some students merely rephrased the question in statement form (minimum 5.2%, maximum 17.7%). Furthermore, many explanations were composed of exclusively naïve biological ideas (minimum 16.2%, maximum 60.4%). Finally, 16.2–39.6% of students packaged mixtures of naïve and scientifically accurate elements. Overall, it is clear that naïve ideas are often mixed with accurate elements in students'

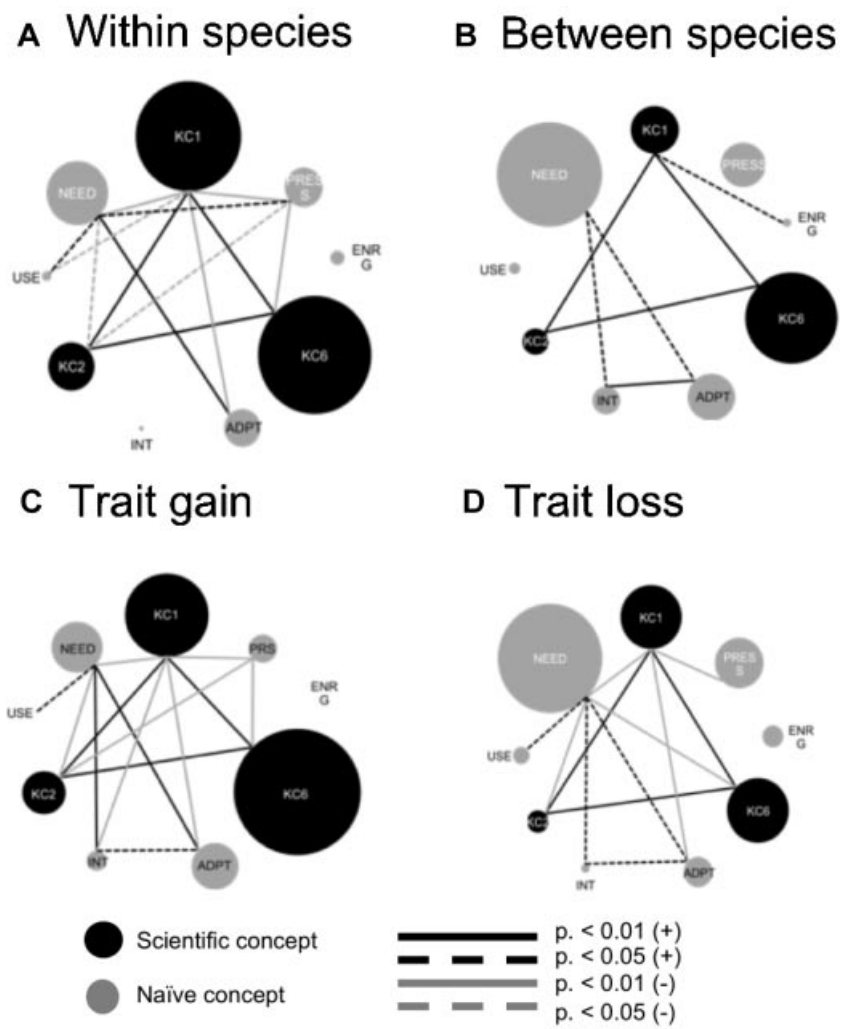

Figure 5. Context effects on the interrelationships among explanatory elements (scientific and naïve ideas). Circle size represents the frequency of explanatory element use, and circle color represents element class (scientific is black and naïve is gray). Lines represent statistically significant correlations among elements. Solid lines represent positive significant correlations at $p < 0.01$; gray lines represent negative significant correlations at $p < 0.01$. Dashed lines represent significant correlations at $p < 0.05$. NEED, needs and goals; PRESS, pressure; ENRG, energy allocation; INT, intentionality; USE, use and disuse; ADPT, adapt and acclimate. KC1, variation and its causes; KC2, heritable variation; KC6, differential survival and reproduction. A. within species; B. between species; C. trait gain; D. trait loss.

evolutionary explanations, although such ''mixed'' models never exceeded 36% of the sample. Disturbingly, among the 12 tasks, the percentage of biology majors using exclusively scientific explanatory elements never exceeded 53% of the sample.

Discussion

Thirty years of research studies in evolution education have failed to carefully consider or control for context effects in assessment tasks despite well-established findings in cognitive science, cognitive psychology, and other science domains that such effects occur (Chi et al., 1981; Clough & Driver, 1986;

Table 2
Percentages of responses in our sample of biology majors that were composed of "pure" (naïve ideas or scientific ideas only) and "mixed" (naïve + scientific ideas)

	Trait Loss			Trait Gain		
	Bird Flight	Rose Thorn	Sal. Vision	Locust Resist.	Bush Poison	Cheetah Speed
Within species						
None	14.4	13.5	3.6	8.1	7.2	8.1
Core concepts	48.6	39.6	27.0	46.8	52.3	51.4
Naïve ideas	16.2	29.7	44.1	18.9	24.3	19.8
Both	20.7	17.1	25.2	26.1	16.2	20.7
Between species						
None	17.7	12.5	5.2	5.2	9.4	10.4
Core concepts	5.2	8.3	4.2	35.4	29.2	41.7
Naïve ideas	56.3	57.3	60.4	19.8	26.0	24.0
Both	20.8	21.9	30.2	39.6	35.4	24.0

diSessa et al., 2004; Goldstone & Son, 2005; Greeno, 2009; Ha et al., 2006; Sabella & Redish, 2007; Silver, 1979; Son & Goldstone 2009). Consequently, we began our study by establishing a simple evolutionary context space and mapping variables that have been implicitly or explicitly employed in assessment tasks or interview prompts (see Supplementary Materials for examples). This context space likely represents a restrictive set of contexts, and further work should attend to its expansion. For the present study, we investigated two of these major contextual features and their impact on students' evolutionary explanations: hierarchical level (within vs. between-species) and the polarity of trait change (gain vs. loss).

Unlike any previous evolution assessment study, we carefully standardized our assessment item contexts by controlling for: (1) taxon and trait, (2) prompt format (two group comparisons), and (3) explanatory structure (initial and final states were specified). Within this framework, we tested four hypotheses: (1) Evolutionary tasks employing within-species (intraspecific) contexts elicit a significantly greater number of accurate explanatory elements (key concepts of natural selection) than between-species (interspecific) task contexts; (2) Evolutionary tasks employing interspecific contexts elicit a significantly greater number of naive explanatory elements (misconceptions) than intraspecific task contexts; (3) Evolutionary tasks employing intraspecific trait gains elicit significantly greater numbers of accurate explanatory elements (key concepts) than those employing intraspecific trait losses; and (4) Evolutionary tasks employing interspecific trait loss elicit a significantly greater number of naïve ideas than those employing interspecific gains.

All four hypotheses were supported. The ANOVAs documented clear and significant effects and interactions among evolutionary scale (intra- vs. interspecific levels) and trait change direction (loss vs. gain) on the types and magnitudes of students' explanatory elements (Figure 1). Explanations of evolutionary loss included a significantly greater number of naïve biological ideas in all contexts; explanations of evolutionary gain included a significantly greater number of core and key concepts of natural selection; within-species explanations produced significantly more core and key concepts of natural selection and less naïve ideas, whereas between-species explanations produced less core and key concepts of natural selection and more naïve ideas. To directly answer our research question, robust and nonrandom contextualization patterns *do* emerge in evolution assessment tasks—such effects are not too ''noisy'' to permit generalizations.

Clough and Driver (1986), in the first study to explicitly explore evolution item effects on so-called cognitive coherence, noted that scientific concept use was much more stable among sets of evolution tasks than naïve concept use. Similarly, Nehm and Schonfeld (2008) found that different assessment methods tended to elicit comparable magnitudes of their so-called key concepts (scientific elements of natural selection) while this was not found to be the case with what they termed alternative conceptions. Our analysis of 1,200 essays generally corroborated the findings from both of these studies: trait loss and gain contexts typically elicited different types and frequencies of naïve biological ideas, and in most cases intra- and interspecific contexts also elicited different frequencies of particular naïve ideas. In the case of core concepts

of natural selection, however, much more similar core concept use patterns were apparent in (1) trait change contexts and (2) at different hierarchical levels. Thus, while the collective frequencies of scientific and naïve elements were found to display very predictable patterns, individual naïve ideas were more sensitive to elicitation in particular task contexts than were scientific core concepts.

The evolution education literature is replete with haphazardly combined and applied contextual variables in assessment tasks, greatly limiting the meaningful measurement—or legitimate comparability—of knowledge and misconception measures. Furthermore, the number and diversity of context variables employed in a study will likely be associated with measures of response consistency (i.e., cognitive coherence). Thus, models of how stable or coherent biological ideas appear will be contingent upon the contexts used in assessment items and interview prompts. For many types of studies and research questions in biology education, explicit and careful control of contextual variables is essential. Thus, prior evaluations of the efficacy of instructional interventions in evolutionary biology must be interpreted with caution given our findings of unambiguous item context effects on knowledge measurement. Items focusing on between-species comparisons may be minimizing estimates of student evolutionary knowledge while maximizing estimates of misconceptions; assessments employing trait gain items will be poor proxies for understanding of trait loss items. The ratio of gain/loss and within/between species items in an instrument or interview will control measurement outcomes to a large degree. Thus, a large body of prior work must be re-evaluated relative to the efficacy of the instructional interventions (e.g., Bishop & Anderson, 1990; Crawford et al., 2005; Nehm & Reilly, 2007).

Prompt Contexts and Knowledge Coherence

The context dependency of assessment tasks is not only of significance for science assessment. In the of field science education—perhaps more so than in cognitive psychology—the role of context in knowledge elicitation has been conceptually tied to empirical studies, theoretical models, and debates about cognitive coherence (diSessa, 2008; Vosniadou, 2008). In brief, elemental theories (diSessa, 1988) characterize knowledge as existing in a multitude of quasi-independent and flexible pieces (of varying types, ontologies, and scales) that may be actively coalesced into different explanatory assemblages particular to a situation. In contrast, coherence theories (Ioannides & Vosniadou, 2002) characterize knowledge as much more structured, stable, and theory-like, akin to beliefs or presuppositions (Ozdemir & Clark, 2009). Methodologically, evidence against cognitive coherence models (such as stable, theory-like alternative conceptions) has often been derived from studies revealing a lack of explanatory consistency across assessment tasks or prompt contexts (e.g., employing Clough & Driver's, 1986 contingency coefficients) (see also diSessa et al., 2004; Ozdemir & Clark, 2009). In contrast, the persistence of highly stable and structured explanatory frameworks across assessment contexts has been used as evidence in support of coherence theories (Ioannides & Vosniadou, 2002). Thus, evidence for or against coherence models are tightly linked to assessment tasks. Nevertheless, this issue remains largely unexplored in biology education—a better understanding of how contexts control knowledge elicitation would inform theoretical models of cognitive coherence in biology.

If, as we found, different task contexts (evolutionary scale and trait polarity) elicit predictable—albeit different—explanatory element (naïve and scientific) use patterns, what do these findings imply about cognitive coherence in evolutionary thinking (cf. diSessa et al., 2004; Ioannides & Vosniadou, 2002; Ozdemir & Clark, 2009)? The answer depends of course on (1) how we conceptualize coherence; (2) which findings we draw upon to answer the question; and (3) at what scale we look (sample, item, context). We could, for example, operationalize ''coherence'' as the exclusive deployment of scientific *or* naïve frameworks across contexts (much like Clough & Driver, 1986, and somewhat similar to Kampourakis & Zogza, 2009). From this vantage point, our study found some evidence for coherence across prompts; 17.1–39.6% of students formed mixtures of naïve and scientifically accurate elements in their evolutionary explanations, indicating the majority of the sample used ''pure'' (i.e., coherent) models.

Using both naïve and scientific conceptions may not, of course, be easily justified as a lack of explanatory coherence; naïve and scientific elements could be packaged into consistent frameworks across contexts. Indeed, we found some evidence of this. In Figure 5, we illustrate stable association patterns among mixtures of naïve and key ideas (and well as some negative relationships). Nevertheless, at a finer granularity

such stability is much less apparent (e.g., Figure 4). Given that our study has delineated clear contextual patterns, perhaps examining coherence within these domains (e.g., within species, or within loss items) may also be justified as reasonable (e.g., Figure 5A). Indeed, in many respects these contexts shed light upon the bounds of coherence. Overall, in the science education community, cognitive coherence has been tied to context effects, but this framework is much too vague at present to permit clear comparative work across disciplines and no consensus has emerged as to what framework is most appropriate (see diSessa et al., 2004 for additional complexities and Ozdemir & Clark, 2009 for an important case study of within-domain replication).

Implications for Biology Assessment, Curriculum, and Instruction

Evolution Assessments Must be Re-Conceptualized. Several recent papers have noted that growing attention to the teaching and learning of evolution by scientific organizations and individual researchers has not led to concomitant efforts in evolution assessment development and evaluation (Nehm, 2006; Nehm & Schonfeld, 2010). The relatively small numbers of evolution education instruments that exist differ significantly in terms of the presence, frequency, and diversity of task features. Given that the contextual variables that our study investigated have significant effects on the elicitation of both scientific and naïve biological explanations, our findings must be added to a growing list of concerns with extant measures of evolution and natural selection knowledge (Nehm & Schonfeld, 2010). Evolutionary assessments that do not take context and scale into consideration—or carefully control for their effects—will be limited in their ability to accurately or comprehensively assess students' evolutionary reasoning patterns.

Specifically, any assessment tool that is exclusively composed of evolutionary trait gain items (like the Concept Inventory of Natural Selection test) may lack construct validity given that evolutionary trait loss is an equally important evolutionary phenomenon that may be explained using natural selection. It will also only reveal particular types of naïve ideas because of contextual constraints like the ones that we documented. Moreover, given that between-species explanatory tasks tend to elicit greater numbers of naïve biological ideas and fewer key concepts, prior research using *only* between-species comparisons should be interpreted with caution (Bishop & Anderson, 1990; Nehm & Reilly, 2007; Nehm & Schonfeld, 2008); it is likely that findings from such work may provide a *maximum* estimate of evolutionary ''misconceptions'' and a minimum estimate of accurate ''key concepts.'' A greater balance of contexts would provide useful range values for measuring students' abilities to explain evolutionary scenarios (Catley & Novick, 2009).

Biology Curricula Focus on Trait Gains Within Species, But Trait Loss Between Species Is the Major Conceptual Challenge. The hierarchical nature of evolutionary biology is one of its most interesting but challenging attributes (Gould, 1980, 2002; Lloyd, 2007). Establishing and describing the factors associated with explanatory failure, as we have done, may be used to guide instructional attention to the most challenging contexts in evolutionary biology (i.e., between-species trait loss). It is clear that the translation of within-species variation into between-species fixed traits is most difficult for students to understand; not only did we find that between-species contexts elicited significantly more naïve biological ideas, but they also produced fewer accurate scientific concepts.

Curricula about evolution and natural selection—such as the commonly used examples of Darwin's finches, Peppered Moths, and antibiotic resistance—deal with trait gain (largely within species) and may thus foster little if any understanding of the evolution of trait loss or between species change (e.g., Campbell & Reece, 2004). Naïve biological ideas targeted using evolutionary gain scenarios or in-class formative assessments associated with these subjects may offer little in regard to exposing or mitigating naïve ideas associated with trait loss. Thus, curricular revisions are sorely needed that build empirical case studies about trait loss between species that are coupled with experimental research that illustrates the causes of such observable patterns. Given that hundreds of extant and extinct species display trait loss (Nehm & Schonfeld, 2010; Nehm, 2001), there is a wealth of material to draw upon.

Instructionally, Cognitive Resource Management—Not "Adding" Accurate Knowledge—Must be Emphasized. As documented in our study, explanatory failure appears in many cases *not* to be caused by the intrinsic *absence* of accurate knowledge, but rather its failure to be recruited, activated, or reasoned with in

particular contexts (as conceptualized by both situated cognition and classical problem solving theory) (Newell & Simon, 1972; Kirsh, 2009). This finding motivates a very different perspective from previous studies that assume that students harbor either misconceptions *or* lack accurate knowledge (e.g., Anderson et al., 2002).

Specifically, within-species prompts clearly revealed that students do in fact harbor so-called key and core concepts of natural selection, but for whatever reason they do not activate or reason with such knowledge in the between-species contexts; alternatively, naïve ideas may block core concepts or overwhelm working memory as problem spaces are searched. Thus, helping students appropriately deploy their extant cognitive resources in different contexts and scenarios may be a worthy focus of future work in biology education. Notably, such instructional approaches are quite different from the misconception ''mitigation'' strategies that characterize much of evolution education (e.g., Nehm & Reilly, 2007).

Causes of Knowledge Elicitation Patterns

Finally, perhaps the most interesting question that our study raises is: How is task interpretation contributing to the findings that we document? While our study sheds little light on such causes, we consider the rich literatures in the historical development of evolutionary theory and cognitive psychology to be profitable avenues for future work (Barsalou, 2009; Bowler, 1983; Darwin, 1859; Evans, 2008, Evans et al., 2010; Gruber, 1981). Similar to our biology students of today, evolutionary trait loss presented a much more challenging explanatory scenario for biologists in the past than evolutionary trait gain. Indeed, both Darwin and Wallace—despite creating, disseminating, and defending the theory of natural selection—adopted and retained elements of Lamarckian models (particularly use and disuse) to explain trait loss throughout their careers (Darwin, 1859, p. 134, Wallace, 1912, p. 436). Such frameworks are remarkably similar to the explanatory patterns that we documented in our study. Thus, the history of science is likely to provide important insights into how scientists progressed through such conceptual quagmires, perhaps informing us as to why between-species patterns of evolutionary trait loss remain more difficult to understand and explain than within species evolutionary trait gain (cf. Gruber, 1981; Wiser & Carey, 1983).

Many findings in cognitive psychology are pertinent to the context effects that we documented (Evans, 2008; Gelman, 2003). In particular, essentialism—the notion that individual organisms or species have a hidden essence or causal power—has figured prominently in both the historical and cognitive literature on evolutionary theory (Evans, 2008; Gelman, 2003; Mayr, 1963, 1991; Sober, 2006). Darwin's recognition of the central importance of individual and population level variation was key to his rejection of essentialist reasoning (Mayr, 1991). Gelman (2003) has amassed a corpus of evidence supporting the ubiquity of essentialistic reasoning in children, and Shtulman and Schulz (2008) and Evans et al. (2010) and have demonstrated its role in adults' evolutionary explanations.

We consider our findings of greater naïve concept use in between-species evolutionary scenarios to be congruent with essentialistic cognitive biases. Specifically, like many pre-Darwinian naturalists, the students in our sample may be essentializing species-level properties and thus failing to recognize the meaning and significance of within-species variation. The downstream effects of such constraints (e.g., problem categorization effects, cf. Chi et al., 1981) may be the elicitation of fewer core concepts of natural selection and the activation of naïve biological schemas. Further work will of course be needed in order to corroborate such conjecture.

Although our results are congruent with cognitive bias explanations (particularly essentialism), it is unclear as to how comprehensively they may account for the reasoning patterns in our scientifically inclined adult sample (i.e., students intending to pursue science careers). For example, while goals, needs, and intentions were indeed used by students in their explanations (similar to the findings of Kelemen & Rosset, 2009), so too were other naïve ideas such as ''pressure as a force that pushes organisms to change'' and ''deliberate energy saving choices by organisms.'' Clearly, cognitive biases comprise a crucial but incomplete framework for understanding evolutionary cognition.

Study Limitations

Our study documented patterns of evolutionary explanation associated with assessment item features in a large sample and dataset (1,200 essays derived from 200 participants). The use of extra credit to entice

participation may have biased our sample and responses. While statistically significant and meaningful differences were noted, our study did not explore *how* participants' conceptualized or reasoned with the tasks or problems that they were asked to tackle. It is possible that the tasks as envisioned by the research panel did not in fact correspond to how participants' conceptualized these tasks (e.g., see discussions in Halldén, 2008). Thus, conclusions that item features (e.g., ''trait loss'') exclusively accounted for the response differences that we documented must be interpreted with caution. Furthermore, our work was firmly situated within domain-specific paradigms of knowledge elicitation derived from the fields of science assessment, situated cognition, and classical problem solving theory (e.g., Chi et al., 1981; Kirsh, 2009; Simon & Newell, 1972). Several other disciplinary frameworks, most notably semiotics, discourse analysis, and text comprehension have great potential in enriching and informing how assessment item context features were used to reason evolutionarily (e.g., Greeno, 2009; Hallden, 2008; Kirsh, 2009; Newell & Simon, 1972; Shapiro, 1998). Similarly, reasoning and argument construction are central features of essay prompts, and so these perspectives would be worthy foci for further investigations.

This material is based in part upon research supported by the National Science Foundation under REESE grant number 090999 (Nehm, P.I.). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the NSF. The authors gratefully acknowledge Judith Ridgway for help with data collection; John Opfer and David Haury for comments on earlier versions of the manuscript, and the helpful efforts made by the editors and anonymous reviewers to strengthen our work.

References

Anderson, D.L., Fisher, K.M., & Norman, G.J. (2002). Development and evaluation of the conceptual inventory of natural selection. Journal of Research in Science Teaching, 39(10), 952–978.

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

Barsalou, L.W. (2009). Simulation, situated conceptualization, and prediction. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1521), 1281–1289.

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

Bowler, P.J. (1983). The eclipse of Darwinism: Anti-Darwinian evolution theories in the decades around 1900. Baltimore, MD: The Johns Hopkins University Press.

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

Bryce, T.G.K., & MacMillan, K. (2009). Momentum and kinetic energy: Confusable concepts in secondary school physics. Journal of Research in Science Teaching, 46(7), 739–761.

Campbell, N.A., & Reece, J.B. (2004). Biology (7th ed.). San Francisco, CA: Benjamin-Cummings Publishing Company.

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

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5(2), 121–152.

Chu, H.E., Treagust, D.F., & Chandrasegaran, A.L. (2009). A stratified study of students' understanding of basic optics concepts in different contexts using two-tier multiple-choice items. Research in Science & Technological Education, 27(3), 253–265.

Clark, D.B. (2006). Longitudinal conceptual change in students' understanding of thermal equilibrium: An examination of the process of conceptual restructuring. Cognition and Instruction, 24(4), 467–563.

Clough, E.E., & Driver, R. (1986). A study of consistency in the use of students' conceptual frameworks across different task contexts. Science Education, 70, 473–496.

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(6), 613–637.

Darwin, C. (1859). On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. New York: D. Appleton.

Deadman, J.A., & Kelly, P.J. (1978). What do secondary school boys understand about evolution and heredity before they are taught the topics? Journal of Biological Education, 12(1), 7–15.

Demastes, S.S., Good, R.G., & Peebles, P. (1996). Patterns of conceptual change in evolution. Journal of Research in Science Teaching, 33(4), 407–431.

diSessa, A.A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), Constructivism in the computer age (pp. 49–70). Hillsdale, NJ: Erlbaum.

diSessa, A.A. (2008). A bird's-eye view of the ''pieces'' vs. ''coherence'' controversy. In S. Vosniadou (Ed.), International handbook of research on conceptual change (pp. 35–60). New York: Routledge.

diSessa, A.A., Gillespie, N.M., & Esterly, J.B. (2004). Coherence versus fragmentation in the development of the concept of force. Cognitive Science, 28(6), 843–900.

Endler, J.A. (1986). Natural selection in the wild. Princeton, NJ: Princeton University Press.

Evans, E.M. (2008). Conceptual change and evolutionary biology: A developmental analysis. In S. Vosniadou (Ed.), International handbook of research on conceptual change (pp. 263–294). New York, NY: Routledge.

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

Gelman, S.A. (2003). The essential child: Origins of essentialism in everyday thought. New York: Oxford University Press.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. Cognitive Science, 10(3), 277–300.

Goldston, M.J., & Kyzer, P. (2009). Teaching evolution: Narratives with a view from three southern biology teachers in the USA. Journal of Research in Science Teaching, 46(7), 762–790.

Goldstone, R.L., & Son, J.Y. (2005). The transfer of scientific principles using concrete and idealized simulations. The Journal of the Learning Sciences, 14(1), 69–110.

Gould, S.J. (1980). Is a new and general theory of evolution emerging? Paleobiology, 6(1), 119–130.

Gould, S.J. (2002). The structure of evolutionary theory. Cambridge, MA: Harvard University Press.

Greeno, J.G. (2009). A theory bite on contextualizing, framing, and positioning: A companion to son and goldstone. Cognition and Instruction, 27(3), 269–275.

Gruber, H.E. (1981). Darwin on man: A psychological study of scientific creativity. Chicago, IL: University of Chicago Press.

Ha, M., Lee, J.K., & Cha, H.Y. (2006). A cross-sectional study of students' conceptions on evolution and characteristics of conception formation about it in terms of the subjects: Human, animals and plants. Journal of Korean Association for Research in Science Education, 26(7), 813–825.

Hallden, O. (2008). The contextuality of knowledge. In S. Vosniadou (Ed.), International handbook of research on conceptual change (pp. 509–532). New York, NY: Routledge.

Ioannides, C., & Vosniadou, S. (2002). Exploring the changing meanings of force: From coherence to fragmentation. Cognitive Science Quarterly, 2(1), 5–61.

Jones, M.G., Carter, G., & Rua, M.J. (2000). Exploring the development of conceptual ecologies: Communities of concepts related to convection and heat. Journal of Research in Science Teaching, 37(2), 139–159.

Kampourakis, K., & Zogza, V. (2009). Preliminary evolutionary explanations: A basic framework for conceptual change and explanatory coherence in evolution. Science & Education, 18(10), 1313–1340.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. Cognition, 111, 138–143.

Kirsh, D. (2009). Problem solving and situated cognition. In P. Robbins & M. Aydede (Eds.), The Cambridge handbook of situated cognition (pp. 264–306) Cambridge, MA: Cambridge University Press.

Lewontin, R.C. (1970). The units of selection. Annual Review of Ecology and Systematics, 1(1), 1–18.

Lloyd, E. (2007). Units and levels of selection. In D.L. Hull & M. Ruse (Eds.), The Cambridge companion to the philosophy of biology (pp. 44–65). Cambridge: Cambridge University Press.

Mayr, E. (1963). Animal species and evolution. Cambridge, MA: Harvard University Press.

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

Mayr, E. (1991). One long argument: Charles Darwin and the genesis of modern evolutionary thought. Cambridge, MA: Harvard University Press.

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

Nehm, R.H. (2001). The developmental basis of morphological disarmament in Prunum (Neogastropoda; Marginellidae). In M. Zelditch (Ed.), Beyond heterochrony (pp. 1–26). New York, NY: John Wiley & Sons.

Nehm, R.H. (2006). Faith-based evolution education? BioScience, 56(8), 638–639.

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

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

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

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

Nehm, R.H., Kim, S.Y., & Sheppard, K. (2009). Academic preparation in biology and advocacy for teaching evolution: Biology versus non-biology teachers. Science Education, 93(6), 1122–1146.

Nehm, R.H., Rector, M., & Ha, M. (2010). "Force talk" in evolutionary explanation: Metaphors and misconceptions. Evolution Education and Outreach, (3), 605–613.

Newell, A., & Simon, H.A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.

Nieswandt, M., & Bellomo, K. (2009). Written extended-response questions as classroom assessment tools for meaningful understanding of evolutionary theory. Journal of Research in Science Teaching, 46(3), 333–356.

Ozdemir, G., & Clark, D. (2009). Knowledge structure coherence in Turkish students' understanding of force. Journal of Research in Science Teaching, 46, 570–596.

Palmer, D.H. (1999). Exploring the link between students' scientific and nonscientific conceptions. Science Education, 83(6), 639–653.

Patterson, C. (1978). Evolution. London: British museum (natural history) and Routledge & Kegan Paul.

Pigliucci, M., & Kaplan, J. (2006). Making sense of evolution: The conceptual foundations of evolutionary biology. Chicago, IL: University of Chicago Press.

Potari, D., & Spiliotopoulou, V. (1996). Children's approaches to the concept of volume. Science Education, 80(3), 341–360.

Sabella, M.S., & Redish, E.F. (2007). Knowledge organization and activation in physics problem solving. American Journal of Physics, 75, 1017–1029.

Samarapungavan, A., & Wiers, R.W. (1997). Children's thoughts on the origin of species: A study of explanatory coherence. Cognitive Science, 21(2), 147–177.

Settlage, J., & Jensen, M. (1996). Investigating the inconsistencies in college student responses to natural selection test questions. Electronic Journal of Science Education, 1(1), (Retrieved 1/14/04.) Available at: http:ejse.southwestern.edu/.

Shapiro, B. (1998). Reading the furniture: The semiotic interpretation of science learning environments. In B.J. Fraser & K.G. Tobin (Eds.), International handbook of science education (pp. 609–621). Dordrecht, The Netherlands: Kluwer.

Shtulman, A., & Schulz, L. (2008). The relation between essentialist beliefs and evolutionary reasoning. Cognitive Science: A Multidisciplinary Journal, 32(6), 1049–1062.

Silver, E.A. (1979). Student perceptions of relatedness among mathematical verbal problems. Journal for Research in Mathematics Education, 10(3), 195–210.

Sober, E. (2006). Conceptual issues in evolutionary biology. Cambridge, MA: MIT Press.

Son, J.Y., & Goldstone, R.L. (2009). Contextualization in perspective. Cognition and Instruction, 27(1), 51–89.

Taasoobshirazi, G., & Glynn, S.M. (2009). College students solving chemistry problems: A theoretical model of expertise. Journal of Research in Science Teaching, 46(10), 1070–1089.

Vosniadou, S. (2008). The framework theory approach to the problem of conceptual change. In S. Vosniadou (Ed.), International handbook of research on conceptual change (pp. 3–34). New York: Routledge.

Wallace, A.R. (1912). Darwinism. London: Macmillan and Company.

Wason, P.C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. Quarterly Journal of Experimental Psychology, 23, 63–71.

Wiser, M., & Carey, S. (1983). When heat and temperature were one. In D. Genter & A. Stevens (Eds.), Mental models (pp. 267–297). Hillsdale, NJ: Erlbaum.