ORIGINAL SCIENTIFIC ARTICLE

What Do Experts and Novices "See" in Evolutionary Problems?

Ross H. Nehm · Judith Ridgway

Abstract Considerable research has focused on differences in expert and novice problem representation and performance within physics, chemistry, and genetics. Here, we examine whether models of problem solving based on this work are useful within the domain of evolutionary biology. We utilized card sort tasks, interviews, and paper-and-pencil tests to: (1) delineate problem categorization rules, (2) quantify problem solving success, and (3) measure the relationships between the composition, structure, and coherence of problem solutions. We found that experts and novices perceived different item features to be of significance in card sort tasks, and that sensitivity to item surface features was adversely associated with problem solving success. As in other science domains, evolutionary problem representation and problem solving performance were tightly coupled. Explanatory coherence and the absence of cognitive biases were distinguishing features of evolutionary expertise. We discuss the implications of these findings for biology teaching and learning.

Keywords Evolution · Natural selection · Problem solving · Misconceptions · Expertise · Undergraduates

R. H. Nehm (✉)
School of Teaching and Learning and Evolution, Ecology, and Organismal Biology, The Ohio State University, 1945 N. High Street, Columbus, OH 43210, USA
e-mail: nehm.1@osu.edu

J. Ridgway
Center for Life Sciences Education, The Ohio State University, Jennings Hall, Columbus, OH 43210, USA

Introduction

Fostering effective and efficient problem solving skills is a longstanding and universally agreed-upon educational goal (Huxley 1898; Kuhn 2005). Beginning with the pioneering work of Dunker (1945), disciplines ranging from cognitive psychology to science education have contributed explanatory elements to an increasingly elaborate conceptualization of successful and unsuccessful problem solving in science and other domains (Dunker 1945; Polya 1957; Newell and Simon 1972; Larkin et al. 1980; Marshall 1995; Markman 1999; Novick and Bassok 2005; National Research Council 2005; Kuhn 2005; Reif 2008). Comparative studies of experts and novices in different subject areas have been central to growth in understanding domain-general and domain-specific elements of problem representation and performance (Chi et al. 1981; Smith 1983; Novick and Bassok 2005). Experts and novices perceive, conceptualize, and internally represent problems differently, and these differences are associated with varying degrees of problem solving success (although see Smith (1983) and Hardiman et al. (1989) for discussions of gradational characterizations of expertise). Specifically, experts typically are more adept at distinguishing relevant problem components (domain principles) from contextual details or peripheral surface features (objects, places, people), identifying the formal structure of a problem, and generalizing across similarly structured problem types (Silver 1979). Likewise, high- and low-performing novices have been shown to differ in their perceptions of problem elements and, as with expert and novice comparisons, these differences have been shown to be associated with problem solving competency (Krutetskii 1976; Silver 1979; Chi et al. 1981; Markman 1999; Smith 1983; Hardiman et al. 1989; Reif 2008).

The study of problem solving can be split into two interrelated processes: problem representation and problem performance. Problem representation refers to the solver's construction of a model of the underlying structure, essential nature, or categorization of the problem at hand (Silver 1979; Novick and Bassok 2005; Greeno 2009). Structurally equivalent (or isomorphic) problems have been found to vary in difficulty because of problem solvers' different mental representations of the problems (Hayes and Simon 1977). Different features trigger different background knowledge, prior problem categorizations, and schemas, which impact problem representation (Marshall 1995; Nehm and Ha 2011). Thus, item features and the solver's background knowledge (and "comfort/confidence" level with the material) interact to influence how the problem is represented in the solver's mind.

In addition to different internal representations of problems, many other factors explain disparities in problem solving performance between experts and novices. Experts are typically better problem solvers because they harbor greater funds of well-organized domain knowledge (Reif 2008). Furthermore, so-called chunking, or the grouping of deep domain principles, theories, or performances into manageable units, coupled with effortless activation of such chunks, facilitates rapid and successful navigation through a problem search space (Newell and Simon 1972; Chi 2006). Additionally, similar problems may activate different goals and consequent solution strategies in experts and novices (Chi et al. 1981). Behavioral and attitudinal differences characterize experts and novices. Experts are often more persistent and principled in their problem attack, and enlist motivation and confidence in such efforts (Taasoobshirazi and Glynn 2009). Moreover, experts solve problems for very different reasons than novices (e.g., employment vs. grades). Overall, problem representation and performance, coupled with facilitative affective states, are key research areas within the large body of work on problem solving.

Problem Solving in Evolutionary Biology

While considerable research has focused on problem representation and performance in the domains of physics, chemistry, and genetics (e.g., Chi et al. 1981; Gabel and Bunce 1994; Heyworth 1999), no published work to our knowledge has explicitly explored experts' evolutionary problems solving. Evolutionary problem solving, unlike that in many areas of physics, chemistry, and genetics, often requires more probabilistic explanations (Jonassen 2000); demands the consideration of a multitude of forces or causal factors acting on a system; requires close attention to contextual factors (e.g., place, time, history); requires

careful separation of chronologies and explanations (O'Hara 1988); necessitates the consideration of emergent properties (Lewontin 1991); and, perhaps most importantly, requires the distinction between proximate and ultimate causal explanations (Mayr 1991). For these reasons, we were motivated to explore whether existing models of expert-novice problem representation and problem solving in science built largely upon cognition research in physics also have explanatory power within the realm of evolutionary thinking.

Given the complexity of evolutionary explanations, models, and theories, and a lack of evolutionary laws equivalent to those of chemistry and physics (i.e., those with broad boundary conditions and unambiguous predictive validity), a significant challenge confronting researchers interested in evolutionary problem representation and problem solving is the delineation of a disciplinary topic in which the problem solving strategies are theoretically and operationally manageable. We believe natural selection (Darwin 1859) is the best such candidate. Regardless of lineage, environmental context, or timeframe, natural selection is considered by biologists to be the major (but not exclusive) driving mechanism of evolutionary change. Notably, it is not the only evolutionary mechanism; debates throughout the history of biology have focused on the relative importance of natural selection as a causal agent relative to genetic drift and species sorting, among others (Bowler 1983; Gould 2002). Importantly, no contemporary biologists have questioned that natural selection is a major cause explaining patterns of organismal diversity in time and space (Morris 2001; Gould 1996, 2002).

In short, natural selection is a theory that explains how the constant production of novel heritable variation—through mutation, genetic recombination, and sex—is differentially sorted by the environment; competition for limited resources, exacerbated by the constant overproduction of offspring, leads to frequent mortality and subsequent population distributions that are more closely aligned to the current state of the biotic and abiotic environment. Thus, if natural selection is employed as an explanation for biotic change through time, several components or concepts are typically acknowledged: (1) the causes of variation (mutation, recombination, sex), (2) heritability of variation, (3) hyper-fecundity or "overproduction" of offspring, (4) limited resources and competition, (5) differential survival and reproduction of individuals, (6) a change in the distribution of produced variation in the next generation. We follow Nehm and Reilly's (2007) terminology and refer to these components as key concepts of natural selection or simply as key concepts (KCs).

In addition to being a useful area of evolutionary biology for the exploration of problem representation and problem performance, natural selection is also a concept that has

been shown to be particularly difficult to grasp for high school students, undergraduate nonmajors, biology majors, and science teachers (Nehm and Reilly 2007; Nehm and Schonfeld 2007, 2008; Nehm et al. 2009a, b, 2010; Gregory 2009; Nehm and Ha 2011). A large body of work in science education has delineated a daunting compendium of student alternative conceptions about natural selection (Duit 2004; Nehm and Schonfeld 2007, 2008, 2010). Many of these alternative conceptions are quite resistant to innovative instruction (Nehm and Reilly 2007; Gregory 2009). Overall, evolutionary mechanisms such as natural selection may provide a useful context for the study of problem representation and problem performance in novices and experts, given its central importance to the discipline of biology and the extensive research base on student background knowledge and alternative conceptions. Research on expert and novice problem representation and performance may provide useful insights into the teaching and learning of natural selection.

Research Hypotheses and Predictions

Our study begins investigating evolutionary problem solving by examining the putative association between problem representation and problem solving performance in novices and evolution experts. We test three hypotheses:

1. Evolution experts and novices perceive different item features to be of significance when classifying evolutionary problems.
2. Perceived item feature significance is associated with the types of cognitive resources that are used to solve evolutionary problems.
3. Problem classification is associated with problem solving performance in evolution experts and novices.

Sample

Our sample consisted of 35 individuals characterized by varying levels of evolutionary expertise drawn from a large tier 1 Midwestern research university. Our novice sample comprised 25 undergraduate students who had successfully completed the second quarter biology course for majors. This course focused on concepts of organismal and evolutionary biology and ecology, including all of the topics mentioned in our card sort task and problem performance tasks (discussed below). Students voluntarily agreed to participate in the study in exchange for the opportunity to be entered in a random drawing for two prizes worth $150. Participating students were mostly female (76%) and White non-Hispanic (72%), with 8%

Asian, 8% African–American, 4% Hispanic, and 8% mixed race. Novice ages ranged from 18 to 37 years (mean 21.3 years) and had overall grade point averages ranging from 1.85 to 4.00 (mean, 3.29/4.00). Thirty-two percent of novices received final course grades of A, 44% B, and 20% C. Overall, the novice sample closely approximated the diversity of students taking the course from which participants were drawn.

The sample of ten experts comprised individuals who had completed at least a baccalaureate degree in the biological sciences (one B.S., two M.S., and seven Ph.D.) and whose research or teaching focused on organismal, evolutionary, and ecological biology (in contrast to molecular, cellular, or developmental biology). Experts volunteered to participate without compensation. Participating experts were mostly White non-Hispanic (100%), male (60%); and had, on average, 13 years of post-Ph.D. research or teaching experience (range, 0–33). Expert ages ranged from 28 to 61 years (mean 45.5 years). In the expert sample, six participants were of the rank of course coordinator, lecturer, or assistant professor; two were of the rank of associate professor; and two were of the rank of full professor.

Methods

Study 1: Expert and Novice Card Sorting Task

Ten questions (Appendix 1) were developed that focused on evolutionary concepts to varying degrees by a panel comprised of biologists, biology educators, and English educators. All of the items focused on macroscopic patterns of evolutionary change. Each item was placed on a uniquely colored 4×6 inch note card along with a small image that corresponded in some way to the item content (e.g., a shark photograph). Two of the items were only indirectly related to evolutionary biology in that they attempted to elicit explanations drawn from the biological disciplines of biomechanics/functional morphology (item 2) or physiology (item 1). These two items required what Mayr (1991) referred to as proximal explanations rather than evolutionary or ultimate (teleological, theistic) explanations. The remaining eight items included six open-response questions and two multiple-choice questions containing prompts explicitly evoking evolutionary principles (e.g., natural selection) but containing many different superficial item features.

A priori superficial item feature designations included the type of question (closed/open response), evolutionary scenarios involving sensory features (bat/turkey vulture vs. others), water conservation (cactus/apple tree vs. others), flying animals (bat/vulture vs. others), "resistance" (bacteria/locust vs. others), environment type (e.g., water, land),

human-related events (tuberculosis/DDT vs. others), and taxonomic groups (animals/plants). In contrast, deep item features included, for example, the types of selection involved in the event (directional, stabilizing, and disruptive), the phylogenetic scale of the evolutionary comparison (within vs. between species, ancestor vs. descendant), and the causal components of natural selection (KCs, as discussed above).

The 30-minute card sorting task (CST) involved three stages: (1) participants reading the items and preparing for the sorting task, (2) sorting items using a think-aloud procedure and explaining the sorting patterns, and (3) evaluating the meaning of a prepared item grouping of the same cards. After completing the teleology instrument (see below) in a waiting area, participants entered an office with two investigators (male and female), two empty chairs, and a sorting table. They were told that they would be asked a series of questions about biology items. Participants were provided the ten note card questions, asked to take them to the chair facing the window, take as much time as they needed to read and understand the questions, and indicate to the investigators when they were finished. At such a time, they were asked to move to the chair at the sorting table and bring the cards with them.

At the sorting table, participants were asked to think about the concepts or principles that the questions were about and then sort the items into groups that made sense to them. They were told that they could make as many or as few groups as they would like. Participants were also asked to think aloud as they performed the CST. This was described as "like talking to yourself, but we'll be listening." After the first sort, participants were asked to explain what concepts or principles united each stack of cards and how each stack differed from the other stacks. After follow-up questions to clarify participants' explanations, they were asked if they could think of any alternative grouping arrangement that would make sense to them and perform such a sort. Participants were allowed to engage in as many as five sorts before being asked about the final pre-prepared item grouping. Participants performed on average three sorts (maximum five sorts, minimum two sorts).

The sorting table contained a large white foam poster board on which the CST took place. On the back of this board was a prepared grouping of cards of an identical nature (size, color) to those used in the CST but fixed into two groups: (1) Appendix items 1 and 2, which focused on evolutionarily peripheral patterns and (2) the other eight items about evolutionary patterns. Participants were asked (even if they also constructed the prepared grouping) if they could explain why someone might have grouped the cards in this pattern and whether the grouping made sense to them.

In addition to audio-recording the above tasks, both investigators completed a detailed scoring rubric that permitted efficient specification of card sort groupings as well as participant explanations for their groupings (the rubric is available from the senior author upon request). Thus, the variables captured during each CST included the number of groups, group item composition, the explanations for each grouping, and the ability to recognize the prepared grouping. Finally, all novice and expert grouping factors were categorized and tallied into superficial, deep, or ambiguous categories based on three criteria:

1. Whether the sort rules were emblematic of superficial or literal aspects of the items (e.g., plants vs. animals or flying vs. nonflying animals); such rules often comprised what Chi et al. (1981:125) referred to as item surface features. Examples of surface features included: objects referred to in the problem (e.g., bacteria, water) or the literal biological terms contained in the problem (e.g., evolve, ancestor),
2. Whether the sort rules were naïve or scientific ideas. Naïve ideas included the notions that organismal needs may stimulate evolutionary change or that putting pressure on organisms will push them to evolve, and
3. Whether domain principles in evolutionary biology, such as the structure of evolutionary comparisons (e.g., ancestor–descendant vs. intraspecific comparisons) or the types of selection likely involved in an evolutionary scenario (e.g., directional, disruptive) were employed (Table 1). Two scorers independently rated and agreed upon on all sort rule classifications.

In addition to our exploration of individual card sort rules and their use by experts and novices, we quantified card sort performance by scoring categorization features as expert (+1) or novice (−1; as discussed above and illustrated in Table 1). We summed these values for each participant for all of their sort tasks. This quantitative measure was used to explore sort rule patterns at a coarser granularity and quantify the direction and magnitude of associations among card sort expertise levels and problem solving performances (discussed below). Finally, we quantified card sort patterns with problem solving performance (as discussed below). All calculations were performed in SPSS 16 (SPSS, Inc.).

Study 2. Expert and Novice Performance on Evolutionary Problems

Two paper-and-pencil instruments were used to measure expert and novice knowledge and naïve ideas about natural selection. The first instrument was a truncated version of the Open Response Instrument (ORI), which was developed using questions from Bishop and Anderson (1990) and Nehm and Reilly (2007). The instrument comprised four open-ended essay questions: (1) explain why some

Table 1 Item features used by experts and novices to classify evolutionary problems

Categorization feature	Feature level	Expert feature frequency (%)	Expert using feature (%)	Novice feature frequency (%)	Novices using feature (%)
Structure of evolutionary comparison (ancestor–descendant vs. intraspecific, etc.)	Expert[b]	*22.2*	*80*	*8.0*	*40*
Types and components of selection (directional, disruptive, stabilizing)	Expert[b]	*19.4*	*70*	*5.6*	*28*
Evolution vs. function/physiology	Expert[b]	*16.7*	*60*	*9.6*	*48*
Environment, weather, seasons, climate, place where organism live vs. other factors	Novice[a]	8.3	30	14.4	72
Sensory features vs. non-sensory features	Novice[a]	5.6	20	2.4	12
Artificial vs. natural selection	Expert[b]	*5.6*	*20*	*1.6*	*8*
Immunity or resistance vs. non-resistance	Novice[a]	2.8	10	9.6	48
Various taxonomic groupings (single vs. multicellular	Novice[a]	2.8	10	4.0	20
Predator-prey vs. other factors	Novice[a]	2.8	10	3.2	16
Niche vs. other factors	Novice[a]	2.8	10	1.6	8
Biotic vs. abiotic selective factors	Expert[b]	*2.8*	*10*	*0.8*	*4*
Experimental/hypothesis driven	Expert	*2.8*	*10*	*0.0*	*0*
Microevolution vs. macroevolution	Expert	*2.8*	*10*	*0.0*	*0*
Observable change vs. non-observable change	Novice[a]	2.8	10	0.0	0
Human involvement vs. non-human involvement	Novice	0.0	0	9.6	48
Body form, features, traits vs. other factors	Novice	0.0	0	6.4	32
Adapting to environment vs. other factors	Novice	0.0	0	4.8	24
Plant vs. animals (or other groups)	Novice	0.0	0	4.8	24
Pressure vs. other factors	Novice	0.0	0	4.8	24
Needs vs. other factors	Novice	0.0	0	3.2	16
Feeding or food-related vs. other factors	Novice	0.0	0	1.6	8
Flying vs. non-flying	Novice	0.0	0	1.6	8
Stronger survive vs. other factors	Novice	0.0	0	1.6	8
Behavioral traits vs. physical traits	Novice	0.0	0	0.8	4
Total		100.0		100.0	

%=percentage. See the "Methods" and "Terminology" sections for details

Italics expert categorization

[a] Novice categorization but used by some experts

[b] Expert categorization but used by some experts

bacteria have evolved a resistance to antibiotics (that is, the antibiotics no longer kill the bacteria); (2) 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? And (3) 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? And (4) if biologists wanted to speed up evolutionary change, how might they do it? Participants were asked to be as complete as they could and provided a large space to answer each question. For more details on the validity, reliability, coding, and interpretation of these items, see Nehm and Reilly (2007) and Nehm and Schonfeld (2008).

The second instrument was a modified version of Bartov's (1978) teleological reasoning test consisting of ten items (Appendix 2). This instrument was used to measure participants' recognition of the appropriateness of teleological expressions. The Bartov instrument measures were therefore somewhat different from the ORI measures in that the Bartov instrument measured students' perceptions of the accuracy of teleological language whereas the ORI measured students' use of teleological reasoning in their evolutionary explanations. Each item in the Bartov instrument was scored such that a correct response received two points, a "not sure" response received one point, and a wrong answer received no points. We scored responses such that 15–20 points represented a score indicating clear recognition of scientifically inappropriate teleological reasoning, 10–14 points represented uncertainty about the appropriateness of teleological expressions, and a score less than 10 represented an inability to recognize teleological expression as biologically incorrect.

Problem solving performance was measured using the ORI and teleology test. The first set of variables extracted from the ORI related to participant knowledge of seven KCs of natural selection (Mayr 1982; Nehm and Schonfeld 2008): (1) the causes of phenotypic variation (e.g., mutation, recombination, and 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. The coding rubric was used to quantify the presence or absence of these seven key concepts in each of the students' three essay questions (see above). These rubrics are available from the authors upon request. Key concept scores were tallied separately for each question and collectively for each novice or expert. In addition, the number of different key concepts used among all questions (hereafter, key concept diversity) was scored for each participant.

Misconceptions or naïve ideas were scored in a similar manner as key concepts: each naïve idea was given one point (see Nehm and Schonfeld 2007; 2008). Naïve idea scores (which we will abbreviate as MIS for misconceptions) were tallied for each question and collectively for each participant. In addition, we calculated the number of different naïve ideas used among the four items (hereafter, naïve idea diversity). Correlation coefficients were calculated among: (1) instrument KC scores, (2) instrument MIS scores, and (3) card sorting task (CST) variables from study 1 discussed above and, (4) the teleology instrument scores as discussed above.

Terminology

We provide a brief synopsis of the terminology that we use throughout our paper in order to facilitate clear understanding of our results and interpretations. Experts and novices refer to socially established categories emblematic of the degree of formal educational training and experience in the subject domain irrespective of measured knowledge or skills in the domain; expert- and novice-like refer to descriptions of behaviors, performances, or knowledge considered typical of the social categories "expert" and "novice"; sort rules refer to grouping and categorization features generated by participants in the card sort task; sort scores refer to quantitative measures that summarize the degree to which participants employed literal or conceptual sort rules during the card sorts; problem categorization features refer to the specific item features that were used as sorting rules; explanatory composition refers to the knowledge or naïve idea elements that comprised evolutionary

explanations; explanatory coherence refers to the consistency of explanatory elements across prompts or items; explanatory structure refers to a description of how cognitive resource elements are assembled together into a holistic model or cognitive network; naïve idea diversity refers to the number of different naïve biological ideas or resources employed across item prompts; and KC diversity refers to the number of different scientifically accurate resources employed across all prompts.

Results

Evolutionary Problem Characterization

Differences in problem categorization and representation are thought to have causal associations with problem solving success (Novick and Bassok 2005), and so our work began by exploring how experts and novices grouped and categorized evolutionary problems using a CST. During the CST, two scorers independently documented and subsequently verified 24 different evolutionary sort rules or problem categorization features that were employed by expert and novice participants (see Table 1 for a complete list). The categorization feature "flying vs. nonflying," for example, was used by some novices to describe two item groups (1) items about bats, birds, and locusts and (2) the remaining seven questions about nonflying organisms. The sort rule "evolution vs. function/physiology," for example, was used by some participants to distinguish functional/physiological items (the shark and apple questions) from evolution-focused items (the other eight questions).

In many instances, novices employed sort rules that were never used by experts (Table 1). Specifically, novices uniquely employed 10/24 categorization features; all of these categorizations were also classified as novice-like (see above). These novice-exclusive sort rules could be divided into two classes: (1) literal feature sort rules (e.g., plants vs. animals or behavioral vs. physical traits) or (2) naïve biological sort rules (e.g., needs vs. other factors or pressure causes change vs. other factors). The former sort class included six sort rules and the latter class included four sort rules (Table 1). The most common literal sort rule—employed by 48% of novices—was whether the item groupings involved humans or not. The most common naïve sort rule—employed by 24% of novices—was "adapting to the environment vs. other factors" (Table 1).

Expert participants employed considerably fewer sorting rules than novices (3/24 vs. 10/24, respectively). These unique-to-expert rules included "experimental/hypothesis driven or not", "microevolution vs. macroevolution," and "observable vs. non-observable change". These three expert sort rules were rarely used overall, however (less than 10%

of experts employed these three rules). While we categorized the former two sort rules as expert-like because of their conceptual nature, the latter rule was not because "observation" is a conceptually basic scientific concept (even though no novices used "observable" as a sort rule).

Interestingly, the evolutionary sort rules used by novices and experts overlapped in many (11/24) instances. Five of these overlapping sort rules were classified as expert-like and six as novice-like (Table 1). Despite overlap, a greater percentage of evolution experts employed expert-like sort rules than did novices. For example, while 80% of experts used the "structure of evolutionary explanation" as a sort rule, only 40% of novices did so. Likewise, in the vast majority of cases, a greater percentage of novices employed novice-like sort rules than experts. For example, while 72% of novices viewed "environment and weather" as a salient sort rule, only 30% of experts did so.

Two of the sort features that overlapped between experts and novices displayed comparable percentages (Table 1): sensory features and niche. Overall, experts and novices, when confronted with the same evolutionary problems, independently derived many of the same sorting rules. In such cases, experts tended to use expert-like rules more commonly than novices did, and vice versa. Finally, novices generated many unique sort rules characteristic of literal item features and naïve biological concepts.

Expert and novice sort scores were significantly different ($t=-4.7$, 33 df, $p<0.001$). Novice sort scores ranged from −6 to +2 (mean −2.48, SD=2.3) and expert sort scores ranged from −2 to +5 (mean=1.6, SD=2.3). While 80% of expert sort scores were positive, 80% of novice sort scores were negative, indicating that most novices employed novice-like sort rules and most experts employed expert-like sort rules. While no novices' sort scores approached the highest scores of experts, three novices achieved the average expert score (about +2). Likewise, two experts had novice-level card sort scores (<0). Overall, card sort scores derived from judgments of card sort rules were highly predictive of social expertise level (e.g., evolutionary biologist or student). Nevertheless, overlap was also documented between the groups, indicating that some undergraduate students employed expert-like sort rules and some experts used novice-like sort rules.

Problem Solving Performance

In order to quantify expert and novice evolutionary problem solving abilities, we employed a version of the ORI and used it to measure KCs of natural selection diversity and naïve idea diversity (see "Methods" section). Overall, KC diversity differed significantly between experts and novices ($t=-3.55$, 33 df, $p<0.001$). Likewise, experts and novices

differed significantly in KC use for each item ($p<0.01$; N. B., for the "speeding up evolution" item, $p=0.015$). Experts on average used 4.9 KCs per item whereas novices used 3.8 KCs per item.

Interestingly, the pattern of KC use was highly consistent in the expert group: (1) 100% of experts employed KC1 (the causes of variation: mutation, recombination, sex), KC2 (heritability of variation), and KC6 (differential survival of individuals); (2) not a single expert employed KC4 (hyper-fecundity or "overproduction" of offspring) in any of the items and only 10% used KC5 (resource limitation); (3) 90% of experts used KC3 (competition) and KC5 (a change in the distribution of variation in the next generation) in their responses. Interestingly, a similar pattern of KC use occurred in novices, with KC4 again nearly absent from all responses.

Naïve biological ideas were also tabulated within and among the four item responses for all novices and experts. Such ideas included the notion that "needs drive evolutionary change," that pressure applied to organisms can push them to change, and that the disuse of phenotypic features proximally produces evolutionary loss. As was the case for KCs, the diversity of naïve ideas differed significantly between novices and experts, with experts expectedly employing fewer of them ($t=4.53$, 33 df, $P<0.001$). Likewise, naïve ideas differed significantly between experts and novices for each of the individual items ($P<0.01$ for all items except for the salamander item, in which $P>0.05$). As a group, experts on average employed 0.2 naïve ideas among the four items whereas novices employed an average of 1.44 naïve ideas among items.

A qualitative review of the types of naïve ideas used by participants indicated that experts and novices employed different naïve ideas. Notably, only one of the experts appeared to use a naïve biological idea (the redistribution of energy from vision to other senses). Novices employed a wider array of naïve biological ideas. These included not only the energy redistribution idea mentioned by an expert, but also: (1) needs drive change, (2) pressure forces change, (3) use and disuse explain change, (4) acclimation is the same as adaptation, (5) inheritance of acquired characteristics, and (6) intentionality explains change. For example, novices referred to the disuse of eyes in salamanders as an explanation of loss of vision. In another example, novices stated that organisms changed in response to needs caused by abiotic or biotic factors, such as the need to become immune to antibiotics, to outcompete other organisms, or to avoid predation. None of the experts employed any of these naïve biological ideas.

Explanatory Coherence

In addition to exploring the explanatory composition and structure of expert and novice problem solutions, we

analyzed the coherence of explanations across instrument items. According to Kampourakis and Zogza (2009) "exhibiting explanatory coherence means providing the same type of explanation to all tasks; in other words, thinking of all processes in the same terms and explaining them by using the same type of explanation." A priori, we expected experts to employ all seven concepts of natural selection in all of their item responses (thus displaying explanatory coherence). We analyzed explanatory coherence in a quantitative manner by measuring the consistency of element use (i.e., KCs and naïve ideas) across the four items for novices and experts (Table 2). For example, use of KC 2 (heritability) by a participant across three or more prompts would be considered a coherent application of this concept, as would the application of a naïve biological concept, such as "needs drive evolutionary change" across three or more responses.

As Table 2 indicates, experts' and novices' item response consistency for KCs differed greatly. Experts used six of the KCs with varying degrees of consistency whereas novices used only three. Seventy percent of experts consistently used KC 6 (Differential survival) whereas only 36% of novices did so. While experts consistently employed more KCs than did novices in their problem solutions, some concepts (e.g., limited resources) were rarely used consistently, and one (KC 4, hyperfecundity) was never used (Table 2). Overall, it is clear that experts displayed significantly greater explanatory coherence, and did so using scientifically accurate resources, than novices. Nevertheless, experts' use of scientifically accurate concepts was much less coherent than expected. Novices did not display explanatory coherence using scientifically accurate elements of natural selection.

Because naïve ideas were relatively common in novice responses, we also explored whether their application displayed patterns of explanatory coherence. Recall that experts almost never employed naïve biological ideas, and so it was not possible to explore consistency in this regard. As Table 2 illustrates, novices never employed naïve ideas in a coherent manner. Each item typically elicited unique (and naïve) biological resources.

In summary, the four open-response evolutionary problems (1) elicited different types and magnitudes of accurate and naive cognitive resources in experts and adults; (2) produced different structural arrangements of ideas; and (3) produced different degrees of explanatory coherence, with the greatest degree of consistency in experts' use of KCs of natural selection.

Teleology Test Performance

In addition to the ORI, expert and novice participants were also clearly distinguished by their teleology test scores. Novices displayed average scores of 7.6/20 and experts exhibited average scores of 16.9/20; these differences were significant ($t=-7.6$, 33 df, $P<0.001$). Only five of the 25 novices scored above 50% on the teleology test, whereas all of the experts did so. It is apparent that most novices were not able to recognize teleological statements as biologically inappropriate. Finally, we explored the association of teleology scores with other problem solving measures (Table 3). Teleology scores were significantly positively correlated with card sort scores ($r=0.4$, $P<0.05$) and negatively correlated with naïve idea diversity measures ($r=-0.42$, $p<0.05$); teleology scores were only marginally related to KC diversity scores ($r=0.33$, $p>0.05$).

Table 2 Explanatory coherence patterns in novice and expert participants

Cognitive resource	Description	Consistency[a] of expert use (%)	Expert rank order	Consistency[a] of novice use (%)	Novice rank order
Key concept 6	Differential survival	70	1	36	1
Key concept 1	Causes of variation	60	2	4	3
Key concept 2	Heritability	60	3	8	2
Key concept 7	Frequency change	50	4	0	n/a
Key concept 3	Competition	20	5	0	n/a
Key concept 5	Resources	10	6	0	n/a
Key concept 4	Hyperfecundity	0	7	0	n/a
Naïve concept 1	Needs drive change	n/a	n/a	0	n/a
Naïve concept 2	Pressure forces change	n/a	n/a	0	n/a
Naïve concept 3	Use and disuse explain change	n/a	n/a	0	n/a
Naïve concept 4	Acclimation=adaptation	n/a	n/a	0	n/a
Naïve concept 5	Inheritance or acquired traits	n/a	n/a	0	n/a
Naïve concept 6	Intentionality explains change	n/a	n/a	0	n/a

[a] Consistent use (>75%) of concept across item prompts

Table 3 Correlation patterns among measures

	Card sort score	Teleology score	Key concept diversity score	Misconception diversity score
Card sort score	1	0.399*	0.448**	−0.545**
n		35	35	35
Teleology score		1	0.328	−0.423*
n			35	35
Key concept diversity score			1	−0.547**
n				35
Misconception diversity score				1
n				35

*$p < 0.05$
**$p < 0.01$

Relationships between Problem Categorization Patterns and Problem Solving Performances

Evolutionary problem categorization, problem solving, and expertise are linked, as indicated by the significant associations among representative measures (Fig. 1, Table 3). Specifically, composite card sort scores, representative of problem categorization patterns, were significantly associated with all problem solving measures: (1) positively associated with KC diversity scores ($r = 0.45$, $P < 0.01$); (2) negatively associated with naïve idea scores ($r = −0.54$, $P < 0.01$); and (3) positively associated with teleology test scores ($r = 0.40$, $P < 0.05$). Additionally, the significant differences between expert and novice participants in all measured variables indicate the degree to which experts and novices

Fig. 1 Summary of results from study 1 (problem categorization in a card sort task) and study 2 (problem solving performance). See "Methods" section for details of task score calculations. *Error bars* are 2 SE about the mean. **$P < 0.01$, *$P < 0.05$. Misconceptions refer to all naïve ideas, as defined in the "Methods" section

differed in their representations of and solutions to evolutionary problems (Fig. 1). While causal conclusions connecting categorization and solving are precluded by the nature of our study design, we note several patterns that shed additional light on such relationships.

Our data set reveals three different patterns of association among problem sort rules and problem solving performance. First, the recognition of superficial item features by novices was, as may be expected, directly linked to naïve problem solutions. Novices who identified "immunity/resistance" a salient sort feature, for example, typically solved evolutionary problems containing these literal components using naïve ideas relating to acclimation (cf. Bishop and Anderson 1990). Specifically, nearly half of novices characterized immunity/resistance as different from other types of evolution. This representation or conceptualization is associated with the use of naïve biological mechanisms in problem solutions (e.g., acclimation or adapting or "getting used to it"). Thus, we find a clear link between problem categorization and problem solving performance.

The second permutation links the recognition of deep item features with successful scientific problem solving. The sort rule of "evolution vs. function/physiology" (Table 1) is a deep conceptual feature that was recognized by 80% of experts and 40% of novices. These individuals recognized the difference between problems (a) taking place over many generations and requiring evolutionary explanations and (b) those occurring within a generation and requiring what may be termed proximal biological explanations (e.g., function, physiology; see Kampourakis and Zogza 2009). The individuals who recognized such distinctions in the card sort tasks were much more likely to successfully solve evolutionary problems.

The third permutation that we documented was the case of a superficial sort feature having little association with expertise or problem solving success. Sensory vs. non-sensory features are an example of this permutation; these two classes of phenotypic traits are in fact not subject to different evolutionary or selective processes, and yet some
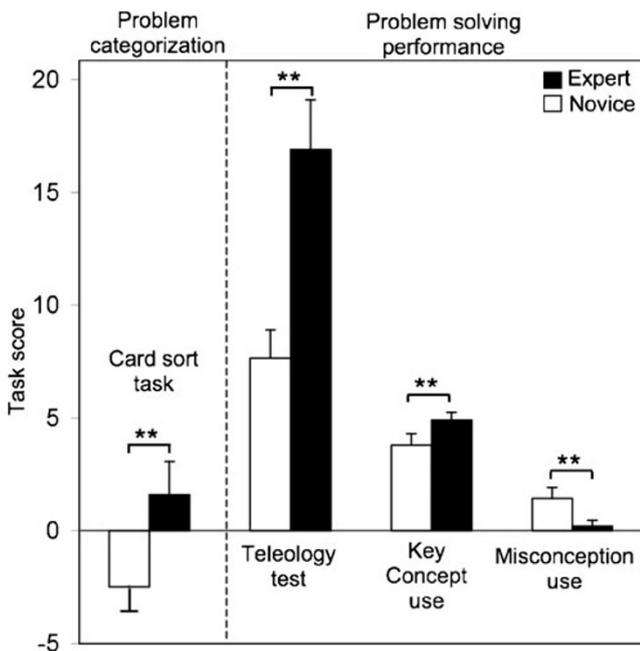
experts judged this item feature as salient and characteristic of different item groupings. Recognition of this superficial item feature was not, however, associated with problem solving success or failure; experts who recognized this superficial sort rule were good problem solvers, whereas novices who recognized this superficial feature as salient were typically poor problem solvers. Thus, in contrast to the previous two examples, the quality of problem characterization features or sort rules was not connected to evolutionary problem solving success.

Discussion

Crosscutting Themes in Scientific Problem Solving

The results of our study mirror several conclusions drawn from the problem solving literature in other content domains (e.g., physics, mathematics, chemistry, and genetics). First, we found that the novices in our study are "not a uniform crowd" (Hardiman et al. 1989: 633); biology majors with comparable academic preparation demonstrated significantly different problem categorization patterns and problem solving abilities along an expansive continuum (Smith 1983; Camacho and Good 1989). Some novices outperformed experts in their problem characterization richness and equaled experts in their problem solving success. At the other end of the spectrum, however, some novices exclusively identified literal or concrete item features (categories such as flying vs. nonflying, plants vs. animals, behavioral traits vs. physical traits) and employed teleological and intentional models characteristic of young children (see below). Overall, the construct of a novice evolutionary problem solver is problematic given the diversity of cognitive representations and strategies documented among individuals in our study.

Remarkably, some novices exhibited expert-like evolutionary problem categorization and solving abilities within the first year of their academic careers (see also Smith 1983, for similar findings in the domain of genetics). An important question for science educators that was asked (but not answered) by Camacho and Good (1989) in their study of chemistry students was what factors facilitated such precocious development in these novices? Which experiences, domain knowledge elements, metacognitive strategies, attitudes, and argumentation skills explained this apparently remarkable achievement (see also Shin et al. 2003)? Purposive sampling of such individuals and exploring the factors that may have contributed to their expertise would be a profitable endeavor.

As with the many studies we reviewed from other science content domains, evolution problem categorization is significantly associated with evolution problem solving success.

Novices employing expert item categorization groupings performed significantly better on evolution problem solving tasks (Fig. 1). Likewise, despite a diversity of novice explanation and card sort patterns (as noted above), experts displayed significantly higher evolution problem categorization scores and evolution problem solving scores. Furthermore, these scores were significantly associated with one another (Table 2). Thus, similar to the findings from research in physics education, we found that our participants' sensitivity to item surface features in problem categorization was adversely associated with problem solving success (see also Chi et al. 1981; Sabella and Redish 2007; Reif 2008; Nehm and Ha 2011). Indeed, those participants who employed more novice-like categorization features (Table 1) displayed lower problem solving scores. While our study design prohibits the attribution of causal conclusions, it has generated results suggestive of a relationship between problem characterization type and problem solving success emblematic of many other science disciplines.

Unsurprisingly, cognitive biases played a role in college-level evolutionary problem solving (Sinatra et al. 2008). We found that many undergraduate novices employed cognitive biases characteristic of children (e.g., intentionality and teleology) in both their problem categorizations and their problem solutions. Such cognitive biases were completely lacking in all expert responses. Some novices, for example, went so far as to use "needs" as a problem categorization feature and likewise explained the causes of evolutionary change in terms of organismal intentions and needs (see Southerland et al. 2001 for similar patterns in secondary school children). These results support the recent work of Kelemen and Rosset (2009), who demonstrate the ontogenetic persistence and masking (rather than replacement) of cognitive strategies characteristic of young children in college-age adults (see also Lombrozo et al. 2007 for comparable conclusions derived from adult samples afflicted with Alzheimer's disease). Overall, then, problem representation using primitive cognitive explanations was in some cases linked to explanations using the same elements. This result bolsters our conclusions that problem representation and problem solving performance are tightly linked in the domain of evolution.

Explanatory Coherence and Evolutionary Expertise

The consistency of explanations across items varied between experts and novices. As anticipated, expert problem solutions were almost exclusively composed of KCs of natural selection, whereas many novices' solutions comprised mixtures of key concepts, naïve ideas, and other contextually inappropriate cognitive resources (e.g., adaptation as acclimation and Southerland et al.'s 2001 P-prim "need as a rationale for change"). Thus, both the compo-

sition and structure of expert and novice evolutionary problem solutions differed in significant respects. Indeed, the greater use of key concepts in evolutionary explanations is but one empirical hallmark of evolutionary expertise.

Explanatory coherence may be considered another component of what may be termed evolutionary expertise (cf. Kampourakis and Zogza 2009). In our analysis of the structure and composition of problem solutions among experts and novices, we noted a diversity of ideas (scientific and naive) that were recruited and assembled into a variety of structural arrangements. We also delineated models of evolutionary explanations depending on whether the resources used by participants were primarily scientific KCs (expert), naïve biological concepts and/or item surface features (novice) or mixtures of the two (mixed models). Furthermore, we noted that social expertise levels (e.g., student, evolutionary biologist) were closely matched to the novice, mixed, and expert models. Finally, we noted that the consistency of explanatory model types across items (i.e., "coherence") was considerably greater for experts than novices. These patterns suggest that explanatory coherence is a useful measure of evolutionary expertise.

In a recent article exploring the coherence of evolutionary explanations in secondary students, Kampourakis and Zogza (2009) noted that "Students did not provide the same type of explanation to all tasks; rather this depended on the specific qualitative characteristics…in the tasks." Similarly, Nehm and Schonfeld (2008) and Nehm and Ha (2011) found that student explanations for evolutionary questions were context dependent in many cases; they particularly found that naïve idea measures were different depending on the measurement method and item features. Our results bring additional empirical evidence to bear upon such findings. It is clear that many superficial item features (e.g., locusts, sparrow wings, eye loss) activate different suites of cognitive resources in experts and novices (Hammer et al. 2005; Nehm 2010). Indeed, item features and contexts are perceived sensitively by novices and are associated with unique patterns of (1) prior knowledge cueing, (2) problem categorization and recognition, and (3) schema activation and solution pathways. Thus, biology educators must begin to attend to such context effects in their teaching, as well as better understand how they influence evolutionary learning. Given that explanatory coherence and insensitivity to such surface effects is a hallmark of evolutionary expertise, biology teachers must begin to tackle the issue of how to effectively teach students to represent and conceptualize evolutionary problems prior to attempting to solve them.

Study Limitations and Suggestions for Further Research

The conclusions generated from our evolution problem characterization tasks are constrained by the types of problems that we employed. Specifically, the problem types used in our card sort task were (a) ill-structured, (b) qualitative in nature, and (c) focused on the origin(s) and loss(es) of individual phenotypic features (e.g., sight, spines). A diverse array of problem types characterizes the domain of evolutionary biology, including quantitative and probabilistic problems relating to sub-cellular elements (e.g., alleles, base pairs, and other genetic elements); macroevolutionary scenarios such as adaptive radiations (Catley et al. 2011); and genomic and bioinformatic models (Nehm and Budd 2006). In order to better understand student thinking about evolution, which is prerequisite to designing effective instructional interventions and assessments (National Research Council 2001), further problem solving research is needed on a diversity of problem types that we did not attempt to investigate. Additionally, our expert sample was relatively small ($n=10$), racially homogeneous, and dominated by men. A larger and more diverse sample of experts could lead to different findings. Nevertheless, our sample does approximate the racial and gender distribution of the field of evolutionary biology.

Methodologically, defining, describing, and analyzing evolutionary solutions require a rich and empirically validated taxonomy of explanatory elements (reminiscent of di Sessa's 2008 "call to arms" in this regard). The categories of analysis presented here are woefully incomplete in terms of the ways in which solution elements could be characterized and modeled (e.g., causal structure, P-prims, coordination classes, etc.; see Hammer et al. 2005). A richer taxonomy of the elements that comprise evolutionary explanations would permit a more fine-grained analysis of the compositional and structural differences between experts' and novices' open-ended explanations. The task of documenting and analyzing these explanatory elements in written text may be enhanced using new technological tools (Nehm and Haertig 2011; Nehm et al. 2011).

Finally, visual representations have a longstanding association with biological thinking and problem solving (e.g., Darwin 1859; Kindfield 1991; Catley et al. in press). Indeed, diagrams and their representative structure have been linked to biological problem solving success in domains such as genetics (Kindfield 1991, 1993). However, none of the problems employed in our card sort task incorporated visual or conceptual diagrams. Exploring putative connections among diagrammatic representations, problem representation, and problem solving success has great potential in bolstering our understanding of evolutionary thinking and problem solving.

Appendix 1

1. Most apple trees typically drop their leaves in the autumn or winter. Scientists have found that at very cold temperatures leaves lose significant amounts of water. They have also noticed that apple trees lose their leaves during long summer droughts. How might they explain the loss of leaves in summer?

2. Hammerhead and Great White sharks have very different head shapes. Great White sharks can swim much faster than Hammerhead sharks. How would a scientist explain the faster swimming speed of the great white shark?

3. Research suggests that the ancestors of nocturnal (night-time) feeding bats fed during the day. Interestingly, nocturnal (night-time) feeding bats have a very advanced sense of hearing. In contrast, it appears that diurnal (daytime) feeding bats have a much poorer sense of hearing. Scientists recently published a paper explaining how they think nocturnal feeding bats developed their superior hearing. What do you think they discussed in their report?

4. Medical researchers are working to explain a troubling situation. In the 1950s, medical researchers developed antibiotics that successfully treated an illness known as tuberculosis. These antibiotics work by killing the bacteria that cause the illness. Today, in many hospitals in the USA, these antibiotics have been found to no longer kill the bacteria that cause tuberculosis. What explanations do you think medical researchers have used to explain this pattern?

5. Almost all birds in North America lack a sense of smell except for the turkey vulture. This species of bird flies high above the ground for most of the day and is able to locate food from this great distance using its highly developed sense of smell. Scientists have found that the closest relatives of the turkey vulture living in North America lack a sense of smell. How would a scientist explain how the turkey vulture's amazing sense of smell developed?

6. For many years farmers have controlled the spread of locusts (a type of insect that eats crop plants) with the chemical DDT. Recently, farmers have been complaining that DDT "doesn't work anymore". Scientists have studied this issue and confirmed what farmers' thought: DDT fails to kill many locusts. What reasons do you think scientists have used to explain this situation?

7. Scientists have recently completed a survey of the plant life in dry and hot (desert) regions and concluded that these plants typically have very small leaves or spines. They hypothesized that these leaf patterns are connected to the observation that large leaves tend to lose more water than small leaves or spines. They also discovered that the species from dry and hot regions have genetic relatives with large leaves that live in wet and cool regions. How would scientists explain the differences in leaf size between hot/dry and cool/wet regions?

8. Sparrows with average-sized wings survive severe storms better than those with longer or shorter wings, illustrating: (1) the bottleneck effect, (2) stabilizing selection, (3) neutral variation, and (4) disruptive selection

9. According to Seaworld.org, bullfrogs can jump 3–6 ft in a single hop. If scientists discovered that the bullfrogs alive today had ancestors that could not jump as far, how would they explain the frog's change in jumping ability?

10. The average length of jackrabbit ears decreases gradually with increasing latitude. This variation is an example of: (1) directional selection, (2) polymorphism, (3) genetic drift, and (4) disruptive selection

Appendix 2

Participants were given the following written instructions on a prepared sheet with boxes to check. Directions: "The following statements have been proposed for inclusion in a new biology textbook. Evaluate each statement by placing an "X" in the appropriate box." The box choices were: "Biologically correct statement", "Not sure", and "Biologically incorrect statement".

1. Some plants growing in arid regions have tried to develop adaptations to dry conditions.

2. A plant growing in a room tries to bend toward the window in order to get more light which is needed for photosynthesis.

3. Deciduous trees need to become dormant in winter in order to protect themselves from damages caused by low temperatures.

4. A male peacock opens its tail because he wants to impress the female.

5. Birds preparing for long flights try to store fat because they need it for their long journeys.

6. Light reaching the eye stimulates an impulse that travels to the brain.

7. Jose was running 3 km daily in order to strengthen his respiratory and circulatory functioning in preparation for a marathon.

8. Polar bears that live in snow-covered habitats developed white fur in order to protect themselves from predators.

9. Bacteria repeatedly exposed to antibiotics typically need to evolve or else they will likely become extinct.

10. Evergreen tree species have developed deep roots so that they will be able to survive in summer when there is no rain.

References

Bartov H. Can students be taught to distinguish between teleological and casual explanations? J Res Sci Teach. 1978;15:567–72.

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

Bowler PJ. The eclipse of Darwinism: anti-Darwinian evolution theories in the decades around 1900. Baltimore MD: The Johns Hopkins University Press; 1983.

Camacho M, Good R. Problem solving and chemical equilibrium: successful versus unsuccessful performance. J Res Sci Teach. 1989;26:251–72.

Catley KM, Novick LR, Shade CK (2011) Interpreting evolutionary diagrams: When topology and process conflict. J Res Sci Teach

Chi MTH. Laboratory methods for assessing experts' and novices' knowledge. In: Ericsson KA, Charness N, Feltovich P, Hoffman R, editors. The Cambridge handbook of expertise and expert performance. New York: Cambridge University Press; 2006. p. 167–84.

Chi MTH, Feltovich PJ, Glaser R. Categorization and representation of physics problems by experts and novices. Cognitive Sci. 1981;5(2):121–52.

Darwin C. 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; 1859.

di Sessa AA. A bird's-eye view of the "pieces" vs. "coherence" controversy. In: Vosniadou S, editor. International handbook of research on conceptual change. New York: Routledge; 2008. p. 35–60.

Duit R (2004) Bibliography: students' and teachers' conceptions and science education database. University of Kiel, Kiel, Germany. http://www.ipn.uni-kiel.de/aktuell/stcse/stcse.html. Accessed 1 March 2009.

Dunker K. On problem solving. Psychol Monogr. 1945;58(5):270.

Gabel DL, Bunce DM. Research on problem solving: chemistry. In: Gabel D, editor. Handbook of research on science teaching and learning. New York: Macmillan; 1994. p. 301–16.

Gould SJ. Full house: the spread of excellence from Plato to Darwin. New York: Harmony Books; 1996.

Gould SJ. The structure of evolutionary theory. Cambridge MA: Harvard University Press; 2002.

Greeno JG. A theory bite on contextualizing, framing, and positioning: a companion to son and goldstone. Cognition Instruct. 2009;27 (3):269–75.

Gregory TR. Understanding natural selection: essential concepts and common misconceptions. Evolution: Education and Outreach 2009;2:156–175.

Hammer D, Elby A, Scherr RE, Redish EF. Resources, framing, and transfer. In: Mestre J, editor. Transfer of learning from a modern multidisciplinary perspective. Greenwich, CT: Information Age; 2005. p. 89–120.

Hardiman PT, Dufresne R, Mestre JP. The relation between problem categorization and problem solving among experts and novices. Mem Cognition. 1989;17(5):627–38.

Hayes JR, Simon HA. Psychological differences among problem isomorphs. Cognitive Theory. 1977;2:21–41.

Heyworth RM. Procedural and conceptual knowledge of expert and novice students for the solving of a basic problem in chemistry. Int J Sci Educ. 1999;21(2):195–211.

Huxley TH. Method and results: collected essays, vol. I. London: Macmillan; 1898.

Jonassen DH. Toward a design theory of problem solving. ETR&D-Educ Tech Res. 2000;48(4):63–85.

Kampourakis K, Zogza V. Preliminary evolutionary explanations: a basic framework for conceptual change and explanatory coherence in evolution. Sci & Educ. 2009;18(10):1313–40.

Kelemen D, Rosset E. The human function compunction: teleological explanation in adults. Cognition. 2009;111(1):138–43.

Kindfield ACH. Confusing chromosome number and structure: a common student error. J Biol Educ. 1991;25:193–200.

Kindfield ACH. Biology diagrams: tools to think with. J Learn Sci. 1993;3(1):1–36.

Krutetskii VA. The psychology of mathematical abilities in school-children. Chicago, IL: University of Chicago Press; 1976.

Kuhn D. Education for thinking. Cambridge, MA: Harvard University Press; 2005.

Larkin JG, McDermott J, Simon DP, Simon HA. Models of competence involving physics problems. Cognitive Sci. 1980;4:317–45.

Lewontin RC. Biology as ideology: the doctrine of DNA. Concord, Ontario: House of Anansi Press; 1991.

Lombrozo T, Kelemen D, Zaitchik D. Inferring design: evidence of a preference for teleological explanations in patients with Alzheimer's disease. Psychol Sci. 2007;18(11):999.

Markman AB. Knowledge representation. Mahwah, NJ: Erlbaum; 1999.

Marshall SP. Schemas in problem solving. New York: Cambridge University Press; 1995.

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

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

Morris DW. Learning from the games animals play: using behavior to assess spatial structure and stochasticity in natural populations. Ann Zool Fenn. 2001;38:37–53.

National Research Council (2001) Knowing what students know: The science and design of educational assessment. Committee on the Foundations of Assessment, J.W. Pellegrino, N. Chudowsky, and R. Glaser (Eds.), Center for Education. Division on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press

National Research Council. How students learn: history, mathematics, and science in the classroom. Washington, DC: National Academies Press; 2005.

Nehm RH. Understanding undergraduates' problem solving processes. J Biol Microbiol Ed. 2010;1(2):119–21.

Nehm RH, Budd N. Missing 'links' in bioinformatics education: expanding students' conceptions of bioinformation and evolution using paleobiological databases. Am Bio Teacher. 2006;7(91):1–9.

Nehm RH, Ha M. Item feature effects in evolution assessment. J Res Sci Teach. 2011;48(3):237–56.

Nehm RH, Haertig H. Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of Text Analytic software. J Sci Ed Tech. 2011. doi:10.1007/s10956-011-9282-7.

Nehm RH, Reilly L. Biology majors' knowledge and misconceptions of natural selection. Bioscience. 2007;57(3):263–72.

Nehm RH, Schonfeld IS. Does increasing biology teacher knowledge of evolution and the nature of science lead to greater preference for the teaching of evolution in schools? J Sci Teac Educ. 2007;18(5):699–723.

Nehm RH, Schonfeld IS. Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. J Res Sci Teach. 2008;45(10):1131–60.

Nehm RH, Schonfeld IS. The future of natural selection knowledge measurement. A reply to Anderson et al. J Res Sci Teach. 2010;47(3):258–362.

Nehm RH, Kim SY, Sheppard K. Academic preparation in biology and advocacy for teaching evolution: biology versus non-biology teachers. Sci Educ. 2009a;93(6):1122–46.

Nehm RH, Poole TM, Lyford ME, Hoskins SG, Carruth L, Ewers BE, Colberg PJS. Does the segregation of evolution in biology textbooks and introductory courses reinforce students' faulty mental models of biology and evolution? Evol Educ Outreach. 2009b;2(3):527–32.

Nehm RH, Rector M, Ha M. "Force Talk" in evolutionary explanation: metaphors and misconceptions. Evol Ed Outreach. 2010;3:605–13.

Nehm RH, Ha M, Mayfield E. transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. J Sci Ed Technology. 2011. doi:10.1007/s10956-011-9300-9.

Newell A, Simon HA. Human problem solving. Englewood Cliffs, NJ: Prentice-Hall; 1972.

Novick LR, Bassok M. Problem solving. In: Holyoak KJ, Morrison RG, editors. Cambridge handbook of thinking and reasoning. New York: Cambridge University Press; 2005. p. 321–49.

O'Hara RJ. Homage to Clio, or, toward an historical philosophy for evolutionary biology. Syst Biol. 1988;37(2):142.

Polya G. How to solve it: a new aspect of mathematical method. 2nd ed. New York: Doubleday Anchor Books; 1957.

Reif F. Applying cognitive science to education: thinking and learning in scientific and other complex domains. Cambridge, MA: MIT Press; 2008.

Sabella MS, Redish EF. Knowledge organization and activation in physics problem solving. Am J Phys. 2007;75:1017–29.

Shin N, Jonassen D, McGee S. Predictors of well-structured and ill-structured problem solving in an astronomy simulation. J Res Sci Teach. 2003;40(1):6–33.

Silver EA. Student perceptions of relatedness among mathematical verbal problems. J Res Math Educ. 1979;10(3):195–210.

Sinatra GM, Brem SK, Evans EM. Changing minds? Implications of conceptual change for teaching and learning about biological evolution. Evol Educ Outreach. 2008;1(2):189–95.

Smith MU (1983) A comparative analysis of the performance of experts and novices while solving selected classical genetics problems. Unpublished doctoral dissertation, Florida State University.

Southerland SA, Abrams E, Cummins CL, Anzelmo J. Understanding students' explanations of biological phenomena: conceptual frameworks or p-prims? Sci Educ. 2001;85(4):328–48.

Taasoobshirazi G, Glynn SM. College students solving chemistry problems: a theoretical model of expertise. J Res Sci Teach. 2009;46(10):1070–89.