JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 49, NO. 6, PP. 744–777 (2012)

Research Article

Cognitive Foundations for Science Assessment Design: Knowing What Students Know About Evolution

John E. Opfer,[1] Ross H. Nehm,[2] and Minsu Ha[2]

[1]*Department of Psychology, The Ohio State University, 1835 Neil Avenue, Psychology Building 245, Columbus, Ohio 43210*
[2]*School of Teaching and Learning, The Ohio State University, Columbus, Ohio*

Received 16 January 2012; Accepted 4 June 2012

Abstract: To improve assessments of academic achievement, test developers have been urged to use an ''assessment triangle'' that starts with research-based models of cognition and learning [NRC (2001) *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press]. This approach has been successful in designing high-quality reading and math assessments, but less progress has been made for assessments in content-rich sciences such as biology. To rectify this situation, we applied the ''assessment triangle'' to design and evaluate new items for an instrument (ACORNS, Assessing Contextual Reasoning about Natural Selection) that had been proposed to assess students' use of natural selection to explain evolutionary change. Design and scoring of items was explicitly guided by a cognitive model that reflected four psychological principles: with development of expertise, (1) core concepts facilitate long-term recall, (2) causally-central features become weighted more strongly in explaining phenomena, (3) normative ideas co-exist but increasingly outcompete naive ideas in reasoning, and (4) knowledge becomes more abstract and less specific to the learning situation. We conducted an evaluation study with 320 students to examine whether scores from our new ACORNS items could detect gradations of expertise, provide insight into thinking about evolutionary change, and predict teachers' assessments of student achievement. Findings were consistent with our cognitive model, and ACORNS was revealing about undergraduates' thinking about evolutionary change. Results indicated that (1) causally-central concepts of evolution by natural selection typically co-existed and competed with the presence of naïve ideas in all students' explanations, with naïve ideas being especially prevalent in low-performers' explanations; (2) causally-central concepts were elicited most frequently when students were asked to explain evolution of animals and familiar plants, with influence of superficial features being strongest for low-performers; and (3) ACORNS scores accurately predicted students' later achievement in a college-level evolution course. Together, findings illustrate usefulness of cognitive models in designing instruments intended to capture students' developing expertise. © 2012 Wiley Periodicals, Inc. J Res Sci Teach 49: 744–777, 2012

Keywords: assessment; cognition; constructed response; natural selection; misconceptions; explanations

Assessments of student knowledge and reasoning patterns play a central role in research on science teaching. At their most effective, assessment instruments provide valid and reliable inferences about student conceptual progress, thereby facilitating guidance in targeting instruction and evaluating instructional efficacy (NRC, 2001). Despite their high potential,

Additional Supporting Information may be found in the online version of this article.

Correspondence to: John E. Opfer; E-mail: opfer.7@osu.edu

DOI 10.1002/tea.21028
Published online 11 July 2012 in Wiley Online Library (wileyonlinelibrary.com).

however, assessment instruments for content-rich domains, such as biology, often lack validity in even the narrow sense described by Linn, Baker, and Dunbar, (1991)—that is, the ability to independently predict outcomes on real-world assessments (e.g., teacher-developed achievement tests). At their least effective, instruments may yield contradictory or false inferences about student knowledge, misconceptions, or reasoning processes (Nehm & Schonfeld, 2008). For some content areas—such as students' understanding of evolutionary processes—there are still remarkably few tools available for validly assessing students' progress (Nehm, 2006).

One suggestion for strengthening validity arguments for science assessment instruments comes from recent reform documents. The NRC (2001) report highlighted the need to develop measurement tools that are guided by cognitive models of progression towards competence. Indeed, as noted by the NRC (2001, p. 6) ''One of the main features that distinguishes the committee's proposed approach to assessment design from current approaches is the central role of a model of cognition and learning.'' The rationale for this recommendation is largely built from research on cognitive differences between novices and experts. Across a wide-range of subjects (e.g., chess, physics, mathematics), the gradual progression from novice to expert involves significant changes in what—and, more importantly, *how*—information is stored and retrieved from long-term memory, such as when solving problems (Chi, Feltovich, & Glaser, 1981; Ericsson & Smith, 1991; Sabella & Redish, 2007). Because science assessments necessarily require learners to access their long-term memory, these differences between novices and experts have strong implications for assessment design. Specifically, differences between experts and novices highlight the critical features of proficiency that should be the targets for assessment (NRC, 2001). This strategy has proven highly useful in designing assessments of mathematics and physics (e.g., Hunt & Minstrell, 1994; Marshall, 1995; White & Frederiksen, 1998), but it has never been applied to improve the validity of evolution assessment instruments.

Constructing a validity argument for science assessment instruments may be guided by an ''assessment triangle'' (NRC, 2001; see Figure 1). This heuristic emphasizes the relationships among *cognitive models*, *assessment methods*, and *inferences from assessment scores* (Kane, 2001; NRC, 2001). Within this framework, a cognitive model of the novice-to-expert progression in student thinking is crucial because it explains varying levels of performance and thus guides assessment design and interpretation. Valid assessment methods, in turn, are ''a set of specifications for assessment tasks that will elicit illuminating responses from students'' (NRC, 2001, p. 42)—that is, ones that will allow students to meaningfully express their scientific understanding and reasoning processes. Finally, valid inferences about student thinking and understanding are possible from assessment scores by using quantitative methods designed to detect ''patterns one would expect to see in the data given varying levels of student competency'' (NRC, 2001, p. 43). Together, these three vertices of the assessment triangle have proven useful for organizing, building, and evaluating integrative validity arguments (Kane, 2001; Marion & Pellegrino, 2006). A similar approach seems likely to help design and evaluate whether an evolution assessment captures how students' reasoning processes produce models of evolutionary causation.

In this paper, we report on an interdisciplinary effort aimed at developing and evaluating an assessment instrument that could measure *students' use of the core concepts of natural selection when explaining evolutionary change*. From the perspective of evolutionary biology, this construct is essential to assess because the actual work of evolutionary biologists is to use the core concepts of natural selection (variation, heritability, and differential survival) to explain what causes changes in phenotypic frequency to occur over time. Also, for more than

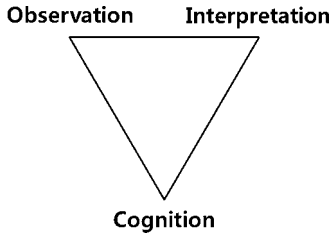

Figure 1. The assessment triangle from the NRC (2001).

30 years, constructed-response explanation tasks have been used to assess student reasoning about evolution (e.g., Bishop & Anderson, 1990; Clough & Driver, 1986). Recent work has provided evidence that inferences about students' evolutionary reasoning derived from these constructed-response explanation tasks: (1) are reliable and differentiate beginning biology students from evolutionary biologists (Beggrow & Nehm, 2012; Nehm & Ridgway, 2011; Nehm & Schonfeld, 2008) and (2) outlined some item characteristics that impact student reasoning processes (Nehm & Ha, 2011). This research led to a proposed set of items—ACORNS [Assessing COntextual Reasoning about Natural Selection]—for teachers to use as classroom case studies. Empirical evidence about the validity inferences for two of the items provided.

Here, we address several important issues about the proposed ACORNS items. The most important issue we sought to address was that ACORNS (like every other assessment in biology) had not been linked to a research-based cognitive model of student reasoning, despite NRC recommendations. Specifically, the novel contributions of the study reported in this paper include: (1) developing new items based on a cognitive model of student reasoning about evolutionary change, (2) defining characteristics to differentiate students' quality of explanations, (3) defining item characteristics that can be manipulated to assess students' consistency in the quality of explanation, and (4) providing empirical evidence to support interpretations about the quality of the explanations. In short, 30 years of prior work with evolution explanation tasks had not been linked to core findings from cognitive science, despite the rich potential of these findings for designing items that would allow us to know what students know about evolutionary change.

Following the ''assessment triangle'', our paper is organized into three main parts: (1) A research-based cognitive model of students' explanations of evolutionary change, (2) observation and inference, and (3) item evaluation. We start at the ''Cognition'' vertex of the triangle and review basic findings from cognitive science on the organization of information in memory and the gradual development of expertise (and science expertise in particular). Then, we turn to the ''Observation and Inference'' vertices of the triangle and describe a general set of specifications for science assessment tasks that are likely to reveal thinking in a manner that can illuminate gradations of competence (cf. Alonzo & Gotwals, 2012). Here, we show how four new ACORNS items illustrate these principles. Finally, in ''Item Evaluation,'' we report on an empirical study of undergraduate biology students that tested whether our new items were capable of capturing the graded sophistication in the explanations predicted by our cognitive model. Finally, we end our paper by discussing how the cognitive framework we developed may be applied to develop assessments of students' understanding in other science domains.

A Research-Based Cognitive Model of Students' Explanations of Evolutionary Change

In this section, we review basic findings from cognitive science on the organization of information in memory and the gradual development of expertise (and science expertise in particular). The purpose of this literature review is to illustrate how a research-based cognitive model can inform the design of an assessment that will measure students' use of the concepts of natural selection to explain evolutionary change. In this literature, four large themes were evident. As we will see, each theme has unique implications for how assessments should be designed, and how scores from such assessments can be used to infer cognitive competencies. The first two themes concern how information is represented and retrieved from long-term memory; these two themes are important because assessment instruments (at a minimum) require students to retrieve information from memory. The second two themes concern how expertise on evolutionary change gradually develops over the lifespan (cf. NRC, 2007). These next two themes are important because they provide assessment instruments with targets for graded competence.

Theme 1: In the Development of Expertise, Core Concepts Facilitate Long-Term Recall. From the perspective of cognitive science, much of the work of educators can be described as an effort to improve (directly or indirectly) how students store information in long-term memory so that information can be recalled for use in appropriate contexts, such as when explaining evolutionary change (Anderson, Reder, & Simon, 2000). Through the last century of research on memory, two major findings stand out as being particularly relevant for designing high-quality assessments. The first finding is that knowledge that is acquired with understanding (e.g., through ''elaborative rehearsal''; Craik & Lockhart, 1972) is retained better and transferred better than that which is acquired by rote (Katona, 1940). For example, by mentally converting, or ''encoding'', meaningless doodles as meaningful scenes (Bower, Karlin, & Dueck, 1975), or by encoding a meaningless string of sentences as having a *causal* interpretation (Bransford & Johnson, 1972; Krascum & Andrews, 1998), learners can improve their ability to recall previously studied information in appropriate contexts. One of the primary differences between expert and novices is the manner in which they organize information; this, in turn, affects how it is used (Chi, 2006). The second major finding is that when information is encoded into a meaningful ''gist'', the meaning of information can be retained for a longer period of time, but literal details of the information studied are no longer recognized (Brainerd & Gordon, 1994). For example, when learning about the biological properties of individual animals that could be encoded as ''true of birds'' versus ''true of bears'', recognition of the specific animals studied declines even as learning of the biological facts improves (Sloutsky & Fisher, 2004).

An important implication of these findings about long-term memory is that students' recall of STEM information (e.g., on exams and standardized tests) is likely facilitated by encoding information using the ''core concepts'' of the field (Griffin, Siegler, & Case, 1994; Inagaki & Hatano, 2004; Vosniadou & Brewer, 1992). For example, the biological concepts *plant* and *animal* are core concepts in taxonomy: they allow one to classify organisms, to make inferences on the basis of shared category-membership, and to make sense of newly discovered facts about individual plants and animals. Despite knowing the words ''plants'' and ''animals'', however, young children do not appear to share biologists' (and adults') concept *plant* and *animal* (e.g., children typically claim that trees are not plants and humans are not animals; Anglin, 1977; Carey, 1985). Lacking these core concepts, children instead encode new information about the biological properties of organisms (e.g., whether it breathes) by comparing the organism to people (Inagaki & Hatano, 2004). How novice

children encode biological information has an impact on their ability to recall biological facts. When asked about animals that resemble people, children recall much more than when asked about non-human animals (e.g., goldfish) that do not closely resemble humans—unless given a ''hint'' reminding them of the ways that the non-human animal is like people (Inagaki & Hatano, 2004). The importance of ''core concepts'' for encoding information is not unique to biology, but can be observed in other STEM disciplines as well, including mathematics (Case & Okamoto, 1996; Opfer & Siegler, 2012), astronomy (Vosniadou & Brewer, 1992), and physics (McCloskey, 1983). Moreover, this finding is educationally important because student understanding of core concepts typically predicts later memory for curriculum-based content (Au et al., 2008; Booth & Siegler, 2006), thereby making assessment of core concepts particularly important (see also, AAAS, 1993; NRC, 1996).

Within evolutionary, ecological, and organismal biology, the core concepts of natural selection, *variation*, *heritability, and differential survival*, are particularly important, allowing myriad biological facts to be acquired with understanding (Dobzhansky, 1973)—and thus (theoretically) more likely to be recalled. What makes these three concepts ''core'' in evolution is that they provide the necessary and sufficient conditions for natural selection (e.g., Endler, 1992; Lewontin, 1978; Nehm & Schonfeld, 2010; Patterson, 1978; Pigliucci & Kaplan, 2006). Further, in line with our cognitive model, use of these core concepts in explanations of evolutionary change is associated with increasing expertise in the field. That is, beginning biology majors use these core concepts very infrequently compared to more advanced majors (Beggrow & Nehm, 2012), and advanced majors use them less frequently in their explanations than practicing evolutionary biologists (Nehm & Ridgway, 2011).

Theme 2: In the Development of Expertise, Causally-Central Information is Weighted More Heavily in Student's Explanations. The second cognitive principle guiding our assessment development is the finding that experts are much more likely than non-experts to represent causally relevant features as explaining other information (Ahn, Gelman, Amsterlaw, Hohenstein, & Kalish, 2000; see Keil, 2006; Lombrozo, 2012, for reviews). For example, young children often understand that illness is a biological rather than social event, yet only older children grasp the causal role of germs in illness and what germs need to exist (Au, Romo, & DeWitt, 1999). Grasping causally relevant features is important, not only because they are useful markers of advanced understanding, but because they also enable students to appreciate how phenomena (e.g., the spread of illness) can be controlled, reduce attention to non-causal variables (e.g., sharing a swimming pool) that might interfere with understanding, and improve the quality of students' explanations (Keil, 2006).

The fact that experts are more likely to use causally-central information in their explanations than non-experts also helps to explain the previously described finding that experts can recall more information in their field of expertise than non-experts. That is, causal information can serve to organize other information in a way that makes it easier to recall (Fenker, Waldmann, & Holyoak, 2005; Krascum & Andrews, 1998; Woods, Brooks, & Norman, 2007). For example, rather than memorizing four features about birds—that birds build nests in trees, fly, have wings, and have bird DNA, an expert can understand these four disconnected features as just three causal relations—that is, birds build nests in trees *because* they can fly, they fly *because* they have wings, and they have wings *because* they have bird DNA. As might be expected, experiments indicate that fostering understanding of such causal relations helps novices both understand biological facts and to recall the information over a longer period of time (Krascum & Andrews, 1998; Woods et al., 2007).

To understand how evolutionary expertise might be associated with changes in explanation quality, research on the psychology of explanation is highly useful. Specifically,

psychologists (with roots in Aristotle) have identified three types of explanations (Lombrozo & Carey, 2006; Opfer & Gelman, 2001; Prasada & Dillingham, 2006). *Mechanistic* explanations (cf. Aristotle's ''efficient cause'') explain by appeal to parts and processes. *Essentialist* explanations (cf. Aristotle's ''formal cause'') explain by appeal to category membership. *Teleological* explanations (cf. Arisotle's ''final cause'') explain by appeal to functions or goals. To illustrate this taxonomy, Lombrozo (2012) considers explanations for why a particular tire is round. Explaining the roundness by appeal to the tire's manufacturing process (''it's round because the rubber is pressed together in a round mold'') would qualify as mechanistic; explaining by appeal to the objects' category membership (''it's round because it's a tire'') would qualify as essentialist; explaining by appeal to its function in generating efficient movement (''it's round so that it can roll'') would qualify as teleological.

Psychological research on explanation has direct implications for assessing the quality of students' explanations of evolutionary change. Some types of explanations are scientifically normative, whereas others are not. *Scientifically normative explanations* are ones using the core concepts of natural selection (discussed above), which appeal to purely mechanistic causes. In contrast, teleological explanations (e.g., ''winged seeds evolved so that they would disperse'') posit causes that are not normative: even if winged seeds had the effect of dispersal, this fact alone could not cause the trait to originate and be passed on to offspring. Similarly, essentialist explanations (e.g., ''the elm evolved winged seeds because it's an angiosperm'') are also not normative because they imply that categories of living things are fixed in nature, an idea that implicitly contradicts the very possibility of evolution by natural selection. Thus, a theoretically important element in the progression toward expertise in evolutionary biology is increasing appeal to the mechanistic causes that are central to evolution (i.e., the concepts of natural selection) and decreasing appeal to final and formal causes.

The idea that final and formal causes would play a large role in novice explanations of evolutionary change has received some support in research on cognitive development. Specifically, three ''cognitive biases'' have been identified in children's biological reasoning: (1) *essentialist biases*, the tendency to treat category members as having an enduring, heritable, ''true nature'' (Evans, 2001; Gelman, 2004; Shtulman, 2006; Shtulman & Schulz, 2008); (2) *teleological biases*, the tendency to treat features and actions of category members as existing for some purpose (Kelemen, 2003; Kelemen & Rosset, 2009; Opfer, 2002; Opfer & Gelman, 2001); and (3) *intentionality biases*, the tendency to assume that events are directed by some intelligence (Bartsch & Wellman, 1995; Inagaki & Hatano, 2002; Sinatra, Southerland, McConaughy, & Demastes, 2003; Sinatra, Brem, & Evans, 2008). Previous research with children and museum visitors had indicated that these three naïve beliefs tend to co-exist when *explaining* biological events (Evans et al., 2010; Poling & Evans, 2002). From previous assessment instruments (e.g., Bishop & Anderson, 1990; Brumby, 1984; Clough & Wood-Robinson, 1985; Dagher & BouJaoude, 1997), however, it is again unclear whether these same cognitive biases would affect biology majors' reasoning about evolution and whether overcoming these cognitive biases would be associated with increasing competency. Furthermore, it is unclear whether these previous instruments allow students to reflect such cognitive biases.

Theme 3: Throughout the Development of Expertise, Normative, and Naive Ideas Continue to Co-Exist. The third theme guiding our assessment development process is that growth of expertise is not all-or-none (Alonzo & Gotwals, 2012; Chi, 2006; Nehm & Ridgway, 2011; Vosniadou & Brewer, 1992). Rather, over the course of learning, expert and non-expert approaches often co-exist and compete for use in any given context or assessment event. A classic example of this co-existence of normative and naive ideas comes from

research by Vosniadou and Brewer (1992), who found that children who initially believed the earth to be flat spent many years after instruction to hold the contradictory beliefs that the earth is both flat and spherical.

This classic example of the co-existence of normative and naive ideas is certainly not unique to thinking about astronomy. For example, on the way to learning that plants and animals are living things, young children often make one of two errors—either maintaining that only animals are alive or maintaining that everything is alive (Opfer & Siegler, 2004). As children gain experience, these two errors become much less prevalent, but under time pressure, older children—and even college professors of biology!—display such errors, either by making them overtly, or by taking more time to recognize that a plant is alive than to recognize that an animal is alive (Goldberg & Thompson-Schill, 2009). Thus, even for a basic concept in biology (the life status of plants), naïve ideas can persist for many years and with much education. For the same reason, we might expect teleological and essentialist explanations to co-exist with use of core concepts as students progress in evolutionary biology.

Theme 4: In the Development of Expertise, Knowledge Becomes More Abstract and Less Specific to the Learning Situation. The final theme guiding our assessment is the finding that novices are much more heavily influenced than experts are by the superficial features of the problem-solving context (Chi et al., 1981; Kirsh, 2009; Nehm & Ridgway, 2011). Thus, young learners typically fail to generalize what they have learned because they attend to too much information or to the wrong information (Bulloch & Opfer, 2009; Gentner, 1988; Hartshorn et al., 1998; Keil & Batterman, 1984). These same deficiencies also mark adult performance in domains in which they have had little experience. Like children, adult novice generalizations overly rely on immediately perceptible and salient features while experts are able to use abstract information that has been most reliable in their past experience (Barnett & Ceci, 2002; Gentner & Markman, 1997; Gick & Holyoak, 1983; Newell & Simon, 1972; Opfer & Bulloch, 2007). For example, when solving biology problems that apply equally to plants and animals, early learners—who typically have much more experience with animals than plants—often succeed when the problems are about animals, but not when they are about plants (Opfer & Gelman, 2010).

The finding that knowledge becomes more abstract and less specific to the learning situation has an immediate implication for the quality of student explanations of evolutionary change. Specifically, experts would be expected to use the core concepts of natural selection in a way that abstracts over the superficial features of problems (e.g., whether explaining evolutionary changes in a plant versus an animal), whereas novices would be expected to use these concepts most often for problems that were highly similar to those they encountered in their lessons.

Observations and Inferences

Observations: The Assessment Tasks

The four themes from cognitive research on the progression of expertise provide central insights into the design of assessment instruments (NRC, 2001). Nevertheless, these cognitive principles must be carefully and deliberately linked to assessment tasks. As emphasized by the NRC (2001, p. 47), ''The tasks to which students are asked to respond on an assessment are not arbitrary. They must be carefully designed to provide evidence that is linked to the cognitive model of learning and to support the kinds of inferences and decisions that will be based on the assessment results.'' In this section, we discuss recent findings from science assessment research that illuminate how inferences about student cognition are constrained by

the types of tasks used to assess students. We focus of four key ideas stemming from our cognitive models: (1) prioritizing recall over recognition; (2) detecting students' use of causally central information; (3) permitting co-existence of scientific and naïve ideas; and (4) attending to task surface features. We link this work to the observations and inferences we seek to derive from our assessment tasks.

Prioritizing Recall Over Recognition. Previous research on natural selection assessment methods has revealed that knowledge recognition tasks (e.g., recognition of discrete ''pieces'' of information shown in multiple-choice tests) produced weaker associations with clinical interview scores than knowledge recall tasks (open-ended writing formats; Beggrow & Nehm, 2012; Nehm & Schonfeld, 2008). From a cognitive standpoint, these high correlations between clinical interviews and open-ended writing formats and low correlations between multiple-choice assessments and open-ended writing formats are to be expected. Both clinical interviews and open-ended writing formats require free recall rather than recognition, and free recall tasks provide a more robust test of whether students have converted the details of their lessons into a meaningful ''gist'' (see ''Core Concepts Facilitate Long-Term Recall'' above). For this reason, our assessments of student reasoning about natural selection require free recall, specifically the production of written explanations, rather than recognition.

Detecting Students' Use of Causally-Central Information. The second issue to be considered in assessment task design is whether the task allows students to reflect their understanding about the causally central features of natural selection and how this understanding is related to their understanding of non-causal features of the evolutionary process. From an assessment standpoint, eliciting *explanations* of evolutionary change—rather than descriptions—are highly useful for probing student's causal understanding of evolutionary change because students typically expect that good explanations should mention causally central information (Callanan & Oakes, 1992; Keil, 2006; Lombrozo, 2006). Moreover, the *types* of concepts that students use to explain evolutionary change are also highly revealing about their thinking about the *causes* of evolution.

Permitting Co-Existence of Normative and Naïve Ideas. The third issue to be considered in the design of assessment tasks is permitting the coexistence of normative (scientific) and non-normative ideas (cognitive biases) (Legare, Evans, Rosengren, & Harris, 2012; Nehm & Ha, 2011; Nehm, Ha, & Mayfield, 2012). To observe co-existence patterns, open-ended formats have been empirically demonstrated to be particularly useful (Ha, Nehm, Urban-Lurain, & Merrill, 2011). Multiple true-false formats (Frisbie, 1992), and ordered multiple choice formats (Briggs, Alonzo, Schwab, & Wilson, 2006) are also capable of detecting co-existence of normative and naive ideas but no such instruments exist for the topic of evolution. Instead, all extant multiple-choice evolution assessments force students to choose between normative scientific and non-normative naive explanations of evolutionary change (e.g., Conceptual Inventory of Natural Selection (CINS); Anderson, Fisher, & Norman, 2002) and therefore fail to uncover when students believe *both* (or neither) types of concepts to be correct.

For our purposes, normative scientific ideas refer to all key concepts (core and other) of natural selection outlined in the evolution education literature (e.g., Bishop & Anderson, 1990; Nehm & Reilly, 2007) and non-normative ideas refer to the cognitive biases outlined in the cognitive science literature (Sinatra et al., 2008).

Attending to Task Surface Features. A final consideration of assessment design relates to the surface features of the open-response tasks (Alonzo & Steedle, 2009; Nehm & Ha, 2011).

That is, assessment tasks need to be designed to assess students' concept use not only for features that are highly similar to the ones that they encountered in their coursework (e.g., Darwin's finches, antibiotic resistance), but novel contexts as well (e.g., prosimian tarsi). Item surface feature similarity is important because it should differentiate students who have succeeded in memorizing previous explanations of evolutionary change from those who possess an abstract and principled understanding of the mechanistic basis of evolution. Extant natural selection instruments like the CINS only assess student reasoning about familiar animals; plants and other lineages are absent from the assessment.

Which superficial features of items ought to be varied for students? A content analysis of popular textbooks in biology indicated that the most frequent exemplars of evolutionary change involve familiar items (e.g., *elms* rather than *labiatae*) and animals (e.g., *snails* rather than *elms*). As a result, we should be concerned about the impact of these two types of superficial features—familiarity of items and taxon—on students' explanations of evolutionary change. From a purely biological standpoint, of course, these superficial features should not affect students' explanations: that is, the key concepts of evolution hold equally for the familiar *elm* and the unfamiliar *labiatae*, as well as for *elms* and *snails*.

Similarly, the use of naïve ideas (such as cognitive biases) is equally invalid for familiar and unfamiliar organisms, as well as for plants and animals. Presumably for this reason, no previous assessments of evolution understanding attempt to systematically vary these superficial properties. However, from a cognitive standpoint, superficial characteristics can play a large role in students' reasoning and detecting conceptual progress, with non-experts applying better understanding to familiar than to unfamiliar items, as well as showing different patterns of reasoning for animals versus plants. Consequently, familiarity and taxon might also affect low-competence students' use of key concepts and cognitive biases about evolution. Indeed, a recent study of evolution experts and novices demonstrated that novices had greater difficulty reasoning across diverse suites of surface features (Nehm & Ha, 2011; Nehm & Ridgway, 2011).

Table 1 presents the assessment tasks developed in consideration of the four principles stemming from our cognitive model. The table reflects the manipulation of the taxa and the traits. To systematically vary animacy and familiarity, the taxa and traits used in the items

Table 1
Assessment items

	Plant	Animal
High frequency/familiarity	A species of elm (plants) produces winged seeds. How would biologists explain how an elm tree species with winged seeds evolved from an ancestral elm species that did not produce winged seeds?	A species of snail (animals) is poisonous. How would biologists explain how this poisonous snail species evolved from an ancestral snail species that was not poisonous?
Low frequency/familiarity	A species of labiatae (plants) is known to have pulegone. How would biologists explain how this labiatae species with pulegone evolved from an ancestral labiatae species that had no pulegone?	A species of prosimian (animals) has long tarsi. How would biologists explain how this prosimian species with long tarsi evolved from an ancestral prosimian species that had short tarsi?

were selected using PageRank (Page, Brin, Motwani, & Winograd, 1998), a key component of the Google search engine, and a highly-useful and broadly-tested proxy for the frequency of text that students encounter and their familiarity with that text (Griffiths, Steyvers, & Firl, 2007). PageRank makes it possible to approximate an equalization of familiarity over plants and animals (e.g., *snails* and *elms* are more similar in PageRank than *fish* and *elms*) and to vary familiarity within taxa (e.g., *elms* has fewer PageRanks than *labiatae*) on the basis of objective metrics rather than researcher intuition (please see Supporting Information Figure 1).

In summary, assessment tasks must be designed in a way that they are capable of revealing the cognitive processes central to detecting evolutionary reasoning progressions (cf. Beggrow & Nehm, 2012; Nehm & Ridgway, 2011). Specifically, these assessment tasks must be capable of: (1) prioritizing recall over recognition; (2) detecting students' use of causally central concepts; (3) permitting the co-existence of scientific and naïve concepts; and (4) attending to task surface features. Open-response assessment formats in general (and the items in Table 1 in particular) are capable of addressing all of these task features, and extensive scoring rubrics have been developed (for a review of the strengths and limitations of open-response assessments, see Nehm & Schonfeld, 2008). These are explained in the next section.

Coding Students' Explanations: Capturing Core Concepts and Cognitive Biases in Student Explanations of Evolutionary Change

To detect use of causally central versus non-causally central information in students' explanations, it is important to distinguish between students' use of two sets of scientifically normative ideas (core key concepts vs. other key concepts). As illustrated in Table 2, explanations containing core key concepts appeal to variation, heritability, and differential survival, rather than non-causal (but scientifically normative) concepts, such as competition, hyperfecundity, limited resources, and changes in population distributions.

Table 3 illustrates how students' use of key concepts and cognitive biases were scored, and Table 4 presents examples of scored explanations (for full details on scoring key concepts and cognitive biases, see Nehm et al., 2010; for previous method of scoring key concepts in ACORNS items, see Nehm, Beggrow, Opfer, & Ha, 2012). For example, to be credited with using the concept of *variation* as a cause of evolutionary change, a student had to explicitly refer to mutation or the random change of genetic information as causing the evolutionary change, rather than simply mentioning that organisms vary in some trait. Coding of cognitive biases also required explicit language linking final and formal causes to evolutionary change. To be coded as providing a teleological explanation, students had to provide explicit teleological language, such as *so that* or *in order to*, linking the goal or function of some trait with its origin.

Our method for inferring competence differs from that of other assessment tools. Existing evolution assessment instruments (e.g., CINS) give *equal* weight to students' knowledge of causally central and causally peripheral ideas. For example, some extant tests give equal weight to the concept of *differential survival* and *competition* despite the fact that these two concepts have different causal status (e.g., Lewontin, 1970). Such equal weighting is discordant with how evolutionary biologists conceptualize explanations of evolutionary change (Nehm & Ridgway, 2011; see also Table 4 for expert examples). Further, science educators have noted that increasing competence within a domain is often associated with increasing use of causally central concepts (Berland & McNeill, 2010; Nehm & Ridgway, 2011; Perkins & Grotzer, 2005).

Table 2
Description of the student explanation coding (italics indicate language corresponding to concept)

Concept Type	Concept Scored	Aspect of Explanation	Concept Definition	Example of Explanation Illustrating Each Concept
Key concepts: Core	Variation	Normative causal idea	The presence and causes of variation (mutation, recombination, sex)	*An ancestral labiatae plant's DNA mutated, giving it pulegone. This mutation did not harm the species, rather it made it equally or more successful from the labiatae without pulegones*. This allowed the mutation to be passed on via reproduction and eventually formed a new species. (Student #24207)
	Heritability	Normative causal idea	The heritability of variation (The degree to which a trait is transmitted from parents to offspring)	Having a poisonous gene for a snail would increase their survival rate. If the snail with the poison is the only one surviving *this trait is going to be passed on to offspring* and the non poisonous gene will become recessive or even non existent. (Student #15576)
	Differential survival/ reproduction	Normative causal idea	The differential reproduction and/or survival of individuals	An ancestral labiatae species had a mutation that caused a pulegone. *This pulegone was more beneficial to survival and reproduction in their environment than no pulegone. The labiatae with a pulegone produced more offspring and lived longer,* which kept their genes in the species. (Student # 23834)
Key concepts: Other	Competition	Normative non-causal idea	A situation in which two or more individuals struggle to get resources that are not available for everyone	Winged seeds could have come by chance mutation, allowing a new benefit to the tree. The seeds would be able to be more widely dispersed by the wind, enabling for them to go further away from the parent. *This would result in less competition for resources*, and the added ability to survive and reproduce. (Student # 28481)
	Hyper-fecundity	Normative non-causal idea	Hyper-fecundity or "overproduction" of offspring	*As more individuals of each species in this case the snails are born than can possibly survive* consequently there is a frequent recurring struggle for existence. Therefore if it evolves into something that can help it survive then the snails will have a better chance of living. And through inheritance any species can grow into its new and modified form. (Student # 27979)

(*Continued*)

Table 2
(*Continued*)

Concept Type	Concept Scored	Aspect of Explanation	Concept Definition	Example of Explanation Illustrating Each Concept
	Limited resources	Normative non-causal idea	Limited resources related to their survival such as food and predator and reproduction such as pollinator.	Animals with longer tarsi survived better, by *getting more food, escaping predators, or finding more mates*. This selected for animals with longer tarsi producing more offspring, and lead to the new long tarsi species. (Student # 19707)
	Change of population	Normative non-causal idea	A change in the distribution of produced phenotypic/ genotypic variation in the next generation	A genetic mutation allowed long tarsi to develop. The long tarsi might have given the prosimian a greater advantage in its environment, allowing it to survive longer and produce more offspring. The offspring then also had the long tarsi, and the same advantages. Due to the increased survivability, *the frequency of long tarsi would increase*, until all prosimians in the species had it. (Student # 28460)
Cognitive biases	Essentialism	Non-normative causal idea	Individuals and groups have an essential nature that allows for them to be sorted into ''natural'' categories, thereby overestimating between-category differences and underestimating within-category variability.	*Elm tree species* survived and reproduced at a more successful rate when they produced seeds with wings. *The trees* evolved from non-winged seeds to winged seeds because it allowed the species to survive and reproduce more efficiently and successfully. (Student # 28670)
	Teleology	Non-normative causal idea	The characteristics and actions of entities or groups to have a goal or to be inevitable	The plant may have evolved from a species which did not have pulegone because *it needed to attract certain insects*. The plant could have developed some way to produce *pulegone in order to attract certain animals or insects to spread its seeds and reproduce*. (Student # 23969)
	Intentionality	Non-normative causal idea	Events are directed by a mental agent.	The winged seeds with the help of wind create dispersal. The ancestral elm species adapted to their environment and *realized that* in order to produce more and more elm plants through dispersal, their seeds must be winged in order to disperse better. (Student # 23856)

Table 3
*Scoring guide for core concepts (normative causal) and cognitive biases (non-normative causal) in students' explanations of evolutionary change**

Concept Type	Concept	Coding Description	Example	Score
Core concept	Variation	Mutation or the random change of genetic information; may produce different phenotypes from parent's traits	*There was a mutation in the elm tree that caused it to have winged seeds.* This is advantageous because it allows the seeds to disperse farther than it would without winged seeds. Because this is advantageous, the elm tree with winged seeds survived and reproduced, and passed its genes for the winged seeds on.	1
		Mentioning "trait" but no mentioning about the 'cause' of the trait	Longer tarsi were "desired" by the environment that the species of prosimian lived in. The tarsi "became" longer and longer *as the trait for longer tarsi* was the one that was being passed down from generation to generation.	0
	Heritability	The transmission of the gene; possibility of "passing" the gene	The survival rate of labiatae with pulegone was higher in a population and thus the genes for pulegone accumulated and *were passed on from parents to offspring.*	1
		No mentioning the transmission of the particular 'gene'	One labiatae developed a pulegone (a mutation). This mutation was beneficial for survival and reproduction and enabled the plant to survive long enough *to produce viable offspring*.	0
	Differential survival/ reproduction	An individual survives/reproduces more or better	One snail underwent a mutation that produced poison. *This poison likely acted as a better defense mechanism and allowed the snail to reproduce more offspring.*	1
		No mentioning about the comparison to other individuals	In order to ward off predators, these snails developed poison *so that they could survive and pass on their genes.*	0
Cognitive bias	Essentialism	Response explains change at a level higher than the individual and fails to mention within species variability	Predation is major selective pressure acting on animals. There was most likely a gene mutation that caused *the snails* to become poisonous and as a result it allows them to be able to defend *themselves* more efficiently.	1
		Response explains change at a level higher than the individual but mentions within species variability.	For some reason, the plant needed pulegone to survive and eventually the plant evolved to contain pulegone. *The plants with pulegone survived better* and reproduced to produce plants with pulegone.	0

(*Continued*)

Table 3
(*Continued*)

Concept Type	Concept	Coding Description	Example	Score
	Teleology	Response contains teleological language without mentioning preexisting variation in "needed" trait; need causes trait to occur	The living snail *species needed to eat other organisms in order to survive, so over time they evolved over time to meet their needs.*	1
		Response does not contain teleological language	Winged seeds aid in dispersal increasing fecundity and fitness and the pressures of natural selection worked and this character to influence evolution of a new winged-seed species	0
	Intentionality	Explanation contains mental verb; agent of mental verb is evolving species or nature; mental verb causes evolutionary change	Since plants ever evolved, *they were looking for a way* to pollinate and disperse, and they couldn't *find* a way to help with that so with natural selection coming in handy, it produced winged seeds, to be easy for the plant to disperse and spread to different areas.	1
		Explanation contains mental verb, but agent of verb is not the evolving species or nature	The ancestral snail was probably dying off because its predators were eating it, but over evolution the snail became poisonous and its predators learned not to eat it or the predator will not feel well and even die.	0

*See Nehm et al. (2010) for the more detailed coding descriptions and scoring guide for other key concepts. Score 1 refers to the presence of the concept and score 0 refers to the absence of the concept. Italics indicate language corresponding to concept scored.

Table 4
Scoring of student and expert explanations of evolutionary change

Item	ID	Explanation	Causal Aspects of Explanation	Key Concepts in Explanation	Score	Naïve Ideas in Explanation	Score
A species of Elm (plants) produces winged seeds. How would biologists explain how an Elm tree species with winged seeds evolved from an ancestral Elm species that did not produce winged seeds?	Student # 17613	*It is necessary for plants to disperse seeds. Dispersal of seed is essential for plants fitness. An environmental pressure must of affected the plant that caused the favoring for winged seeds.*	Non-normative causal	None	0	Pressure as cause of change, Need/goal	2
	Student # 23087	*The elm needed a mode of dispersal for its seeds so some pressure for this dispersal caused the elm to develop winged seeds so that it could better spread its seed.*	Non-normative causal	None	0	Pressure as cause of change, Need/goal	2
	Student # 28501	*A mutation caused the seeds of an individual to have wings. Because this was good for seed dispersal, those genes were passed on and over time, became their own population.*	Normative causal	Variation, Heritability, Change of population	3	None	0
	Expert # 114	*There was genetic variation in the ancestral population in the tendency to produce winged seeds. Those trees that had winged seeds left more offspring than those without, so the trait spread by natural selection. Here the fitness advantage could be caused by dispersal of seeds, allowing greater spread of offspring.*	Normative causal	Variation, Differential survival/ reproduction	2	None	0

(*Continued*)

Table 4
(*Continued*)

Item	ID	Explanation	Causal Aspects of Explanation	Key Concepts in Explanation	Score	Naïve Ideas in Explanation	Score
A species of snail (animals) is poisonous. How would biologists explain how this poisonous snail species evolved from an ancestral snail species that was not poisonous?	Student # 22929	*Biologists would say that the external forces such as predators have caused a need for the snail to come up with some form of defense mechanism, in this case poison. This way the predators will learn that eating the snails will harm them, causing them to find new prey.*	Non-normative causal	None	0	Pressure as cause of change, Need/ goal, Intentionality	3
	Student # 10037	*The species of snail has evolved from not poisonous to poisonous through natural selection evolution. Outside pressures such as predation has caused it to develop into a poisonous creature in order to ensure survival and successful reproduction.*	Non-normative causal	None	0	Pressure as cause of change, Need/ goal	2
	Student # 7024	*There was a mutation in the snail that caused it to be poisonous. This is advantageous because it protects the snail from predators, and allows it to survive and reproduce and pass on the gene.*	Normative causal	Variation, Differential survival/ reproduction, limited resource, heritability	4	None	0
	Expert # 63	*A mutation occurred that caused a phenotypic change in the snail. This effect conferred some benefit to the snail that allowed its offspring to outcompete individuals without the mutations. Over time, the mutation became fixed and the snail was isolated enough to be a species*	Normative causal	Variation, Differential survival/ reproduction, Change of population, competition	4	None	0

Item Evaluation: Methods

Participants

Undergraduate participants ($n = 320$) were recruited from a larger sample of 431 students enrolled in a two-course introductory biology sequence for biology majors. Demographically, the sample was 78% White (non-Hispanic; $n = 251$) and 22% minority (African American, $n = 14$; Asian, $n = 28$; Hispanic, $n = 13$; Native American, $n = 1$; Other and non-disclosed, $n = 13$), 55% female, and with an average age of 21 years ($SD = 2.3$). For 93.8% of students in our sample, English was their first language, and for 6.2% of students in our sample English was their second language.

Assessment Tasks

Our assessment tasks were carefully designed to elicit aspects of cognition outlined above. Student explanations were elicited using our four new ACORNS items (Table 1) that presented students with carefully designed evolutionary change scenarios isomorphic with previously studied ACORNS items. Each item prompted students to write an explanation for how evolutionary change occurred (Tables 3 and 4).

The ACORNS items asked students to explain evolutionary change in taxa and traits likely to be familiar (i.e., elm, snail, poison, seeds) as well as unfamiliar to them (i.e., labiatae, prosimian, pulegone, tarsi). Additionally, unlike other evolution assessment instruments (e.g., CINS), ACORNS controlled for order effects by randomizing presentation of each item using a Latin-square design. Thus, by controlling for familiarity and order effects, the ACORNS controls for surface-level assessment features using a standard protocol in experimental cognitive psychology (Cozby, 1997). Responses were gathered using an online assessment system familiar to students at our university. The average number of words in the student explanations was 29.2 ($SD = 25.8$).

Coding of Written Explanations

For each of the four new items, two expert scorers with graduate degrees in the biological sciences (and trained by a cognitive scientist) independently coded student explanations for the presence or absence of the key and core concepts (KC) and cognitive biases (CB) using the scoring rubrics discussed above. Independent scoring of all KCs for the four new items easily exceeded inter-rater agreement scores of 0.81 (Kappa), whereas independent scoring of CBs revealed more variable scores (Kappas: Essentialism, 0.80; Teleology, 0.70; Intentionality, 0.52). Discordant scores were resolved via deliberation between the raters and the project leaders, leading to a set of consensus scores for all concepts. Scores for all KCs and CBs were used in our analyses. Overall, a matrix of 10 concepts $\times$ 4 items served as the primary data for our analyses, with myriad concept permutations evident in the sample.

Concept Frequency, Diversity, and Coherence Calculations

We performed three analyses on the concept score matrix noted above: concept frequency, concept diversity, and concept coherence. *Concept frequency* was calculated by summing the total number of concepts that students used across all four items. For example, one student might use six KCs across items, whereas another student might only use two KCs across items. *Concept diversity* was calculated by summing the total number of different concept types students used across the items. For example, if a student used three different KCs (e.g., variation, heritability, and differential survival) across two different items, the diversity score would be three. Diversity scores are important because a student may use the same KC types

across four items, yielding a KC frequency score of four, whereas another student may use four different types of concepts across the four items, also yielding a frequency score of four. Thus, the diversity calculations help to capture additional aspects of student response patterns (Nehm & Reilly, 2007). Finally, *concept coherence* was calculated by summing the number of times a student used the same type of KC (e.g., differential survival) across all four items. Thus, higher coherence values reflect more consistent concept use across the four items, which is important because the coherence of concepts represents the stability of a concept across contexts (Kampourakis & Zogza, 2009). As we will see, concept frequency and concept diversity yield informative data distinguishing levels of biological knowledge (Theme 1), distinguishing causal versus non-causal normative ideas about evolutionary change (Theme 2), and effects of superficial item characteristics (Theme 4). In contrast, concept coherence is useful for examining the degree to which normative and non-normative ideas co-exist and compete for use from item to item (Theme 3).

Reliability

Previous work established the content validity, convergent validity, and internal consistency of the ACORNS instrument using the same biology classes as the present study. In prior work, the reliability values (Cronbach's alpha) of ACORNS items were typically higher for key and core concepts than for naïve ideas (e.g., 0.77 and 0.67, respectively). In the present study, the reliability values for the four new ACORNS items were 0.76 for KCs and 0.68 for CBs, mirroring past work. These values are robust particularly given that (1) the items differ in surface features and (2) the item set is small (four constructed-response explanations per each student).

Item Evaluation: Results

Results are organized by the four themes of our cognitive model discussed above. Under Theme 1, we examine whether use of core concepts and cognitive biases on ACORNS successfully predicted future teacher assessments of biological knowledge (i.e., grades in an evolutionary, ecological, and organismal biology (EEOB) course that was taken after ACORNS was administered). This issue is important for Theme 1 because our cognitive model depicts students' knowledge of the core concepts of natural selection as making biology course material more meaningful to students, thereby facilitating their recall of this material on university exams, and thus leading to a positive correlation between core concept use on ACORNS and future grades in the EEOB course. Under Theme 2, we examine whether use of core concepts and cognitive biases better distinguish among high-, medium- and low-performing EEOB students than use of other key concepts of evolution. Under Theme 3, we examine the extent to which core concepts and cognitive biases co-exist in the same students and the degree to which cognitive biases are negatively correlated with core concepts versus other evolutionary concepts. Finally, under Theme 4, we test our prediction that the effect of item characteristics on key concept use would be negatively associated with EEOB grades.

Theme 1: In the Development of Expertise, Core Concepts Facilitate Long-Term Recall

We first examined the relation between performance on the ACORNS items and teacher assessments of biological knowledge in an Evolution, Ecology, and Organismal Biology (EEOB) course (see Figure 2). This course included explicit coverage of evolutionary topics in biology (e.g., natural selection, organismal, and population biology, phylogeny) and furnished 318 final course grades (2 missing/incomplete). These final course grades were analyzed for associations with ACORNS scores (i.e., KC, CB scores).

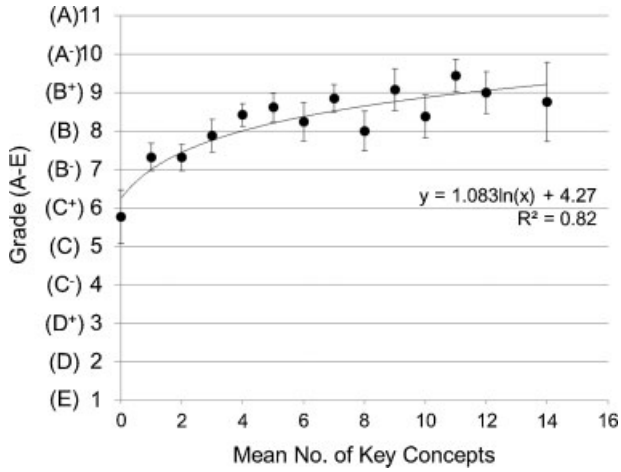

Figure 2. Relation between frequency of key concepts in explaining evolutionary change on ACORNS and course grades in a university biology class.

To examine the association, we converted students' final grades in the course into a numerical score ($E = 1, \ldots, A- = 10, A = 11$), and we examined whether these grades were predicted by the frequency and diversity of KCs and CBs contained in their responses to the four ACORNS items. To test for a relation between the *frequency* of KCs (0–28) and CBs (0–12) in students' explanations and their academic achievement in the course, we conducted a series of regressions.

To characterize the overall relation between concept frequency and grades, we first collapsed scores over subjects (Figure 2). In a series of regressions, we found that as the overall frequency of KCs mentioned in explanations of evolutionary change increased, grades in the course increased logarithmically ($F[1, 12] = 56.15$, $R^2 = 0.82$, $p < 0.0001$). Additionally, as the number of CBs increased, grades in the course decreased exponentially ($F[1, 7] = 8.92$, $R^2 = 0.56$, $p < 0.05$). To characterize the amount of variance in individual grades explained by the frequency of KCs and CBs in students' explanations, we performed the same analyses on individual scores. Again, we found that students who generated KCs the most frequently also tended to have the highest grades (logarithmic regression, $F[1, 316] = 39.74$, $R^2 = 0.11$, $p < 0.0001$), and students who generated CBs the most frequently tended to have the lowest grades (exponential regression, $F[1, 316] = 3.97$, $R^2 = 0.012$, $p < 0.05$).

To test whether using more *types* (diversity) of core concepts was associated with better biological knowledge, we then conducted a series of regressions to test for a relation between the diversity of KCs (0–7) and CBs (0–3) in students' explanations and their academic achievement in the EEOB course. When we collapsed scores over subjects, we found that as the overall diversity of KCs mentioned in explanations of evolutionary change increased, grades in the course increased logarithmically ($F[1, 5] = 156.20$, $R^2 = 0.97$, $p < 0.0001$). Additionally, as the diversity of CBs increased, grades in the course decreased linearly ($F[1, 2] = 23.12$, $R^2 = 0.92$, $p < 0.05$). The same relations held when we examined the amount of variance in individual grades explained by individual KC diversity and CB diversity. That is, students who generated responses with the most diverse array of KCs tended to have the

highest course grades (logarithmic regression, $F[1, 316] = 41.77$, $R^2 = 0.12$, $p < 0.0001$), and students who generated the most diverse array of CBs tended to have the lowest grades (linear regression, $F[1, 316] = 4.47$, $R^2 = 0.014$, $p < 0.05$). Overall, diversity and frequency measures reveal clear patterns of association with course grades.

The preceding findings suggest that understanding of evolution by natural selection—as measured by the frequency and diversity of KCs (core and other) and CBs on the four ACORNS items—provides a good predictor of students' understanding of an important branch of biology. An alternative interpretation of these findings is that overall academic proficiency impacts quality of students' explanations and their biology achievement independently, with no additional impact of student explanation quality on biology achievement. To test this alternative hypothesis, we controlled statistically for overall grade point average, and we examined the remaining partial correlations between KCs/CBs and biology grades. With overall GPA controlled, the remaining partial correlation between number of KCs and biology grade was still significant ($r[315] = 0.23$, $p < 0.0001$), as was the correlation between CBs and biology grade ($r[315] = -0.12$, $p < 0.05$). Thus, consistent with the idea that the causally central core concepts of natural selection help students to understand and later recall their biological knowledge, we found that the quality of student's explanations of evolutionary change successfully predicted overall grades in a college biology course. This relation held at several levels of analysis, whether collapsing over individual scores or not, whether examining number or diversity of CBs and KCs, and whether or not GPA had been controlled.

Theme 2: In the Development of Expertise, Causally-Central Information is Weighted More Heavily in Student's Explanations. To test whether core concepts and cognitive biases distinguished among high, medium, and low performers in the EEOB course, we used the EEOB grades to create three groups of students (''A'' students, ''B'' students, and ''C'' or below students). We also tested whether frequencies of core concepts and cognitive biases were more diagnostic of EEOB course performance than frequency of key concepts that are scientifically normative but are not causally central to explaining change (i.e., *competition*, *hyperfecundity*, *limited resources*, *change in allelic frequency*).

Table 5 presents the frequency with which each concept was used by all students, ''A'' students, ''B'' students, and ''C'' students, as well as the rank order of concept use by each of these groups. To test whether frequencies of core concepts and cognitive biases were more diagnostic of EEOB course performance than the frequencies of other KCs, we first assigned scores of 1 to As in the course, 2 to Bs, and 3 to lower grades, and we performed multiple Kruskal–Wallis tests to examine grade-differences for each of the KC and CB frequencies (Table 6). Among the core concepts, we found grade-differences in mentions of *causes of variation* (Chi-Square = 15.14, $df = 2$, $p < 0.001$, partial eta squared of rank = 0.05) and *differential survival* (Chi-Square = 24.70, $df = 2$, $p < 0.001$, partial eta squared of rank = 0.08), with a trend for grade-differences in mentions of *heritability* (Chi-Square = 7.03, $df = 2$, $p < 0.05$, partial eta squared of rank = 0.02). The Tukey *post-hoc* test illustrated that the ''A'' and ''B'' students' rank scores of *causes of variation* and *differential survival* conceptions were significantly higher than the rank scores of ''C'' students ($p < 0.05$). In addition, ''A'' students' rank score of the *heritability* conception was significantly higher than the rank score of ''C'' students ($p < 0.05$). Grade differences were also present among the three cognitive biases: *intentionality* (Chi-Square = 11.75, $df = 2$, $p < 0.01$, partial eta squared of rank = 0.04); *teleology* (Chi-Square = 12.28, $df = 2$, $p < 0.01$, partial eta squared of rank = 0.04); *essentialism* (Chi-Square = 6.77, $df = 2$, $p < 0.05$, partial eta squared of rank = 0.02). The Tukey *post-hoc* test indicated that ''A'' and ''B'' students' rank scores of the *intentionality* bias were significantly lower than the rank

Table 5
Percentage of concept use among students explaining evolutionary change (by future grades in biology course work)[*]

Concept Used	All Students % (Rank)	A Students % (Rank)	B Students % (Rank)	C Students % (Rank)
Differential survival	35.0 (1)	47.7 (1)	35.4 (1)	18.2 (2)
Variation	29.4 (2)	37.5 (2)	28.7 (3)	18.2 (3)
Essentialism	25.0 (3)	11.4 (4)	29.3 (2)	33.3 (1)
Heritability	8.8 (4)	14.8 (3)	7.3 (5)	4.5 (6)
Teleology	6.6 (5)	1.1 (7)	7.9 (4)	10.6 (4)
Limited resources	4.1 (6)	2.3 (6)	4.3 (6)	6.1 (5)
Change of population	1.9 (7)	3.4 (5)	1.8 (7)	0.0 (7)
Competition	0.3 (8)	0.0 (8)	0.6 (8)	0.0 (7)
Hyper-fecundity	0.0 (9)	0.0 (8)	0.0 (9)	0.0 (7)
Intentionality	0.0 (9)	0.0 (8)	0.0 (9)	0.0 (7)

[*]Numbers denote percentage of students in each group (all students, A students, B students, and C students) who used the concept more than three times across four items. Because a student could use more than one concept, columns do not total to 100.

scores of ''C'' students ($p < 0.05$). In addition, the ''A'' students' rank score of *teleology* was significantly lower than the rank score of ''C'' students ($p < 0.05$). For non-causal KCs, however, no grade differences were detected. Thus, core concepts and cognitive biases proved more useful in differentiating student achievement than many biological concepts typically stressed in other natural selection assessments (e.g., CINS).

Theme 3: Throughout the Development of Expertise, Normative and Naive Ideas Continue to Co-Exist and Compete for Use. Evolution assessments often include measures in which students are forced to choose between options that are and are not consistent with the construct of evolution by natural selection (defined above). Unlike these forced-choice assessments, like that of Bishop and Anderson (1990), our open-ended assessment can detect the frequency with which students express only the key concepts of natural selection, only

Table 6
Consistency of concept use among explanations of evolutionary change

	Percentage of Students Using the Concept on:			
Concept Used	1 Item	2 Items	3 Items	4 Items
Differential survival	73.1	49.4	35.0	13.8
Essentialism	70.3	43.1	25.0	7.8
Variation	60.3	45.6	29.4	14.4
Limited resources	59.7	14.7	4.1	0.0
Teleology	43.1	20.9	6.6	1.6
Heritability	30.0	15.3	8.8	1.9
Change of population	15.6	4.1	1.9	0.6
Intentionality	8.1	0.9	0.0	0.0
Competition	6.6	1.3	0.3	0.3
Hyper-fecundity	0.0	0.0	0.0	0.0

cognitive biases, both, and neither. In this section, we examined the extent to which KCs and CBs co-existed in the same students.

We first examined the co-existence of KCs and CBs among all students. Overall, we found that 22% of students provided one or more KCs and no CBs, 3% of students provided one or more CBs and no KCs, 73% of students provided both, and 2% of students provided neither.

To test the idea that KCs become more frequent and CBs become less frequent with increasing biological knowledge, we examined the relation between these four groups and EEOB course grades. Among these four groups (KC only, CB only, KC and CB, and no concepts), we performed a Kruskal–Wallis test to examine difference of numeric scale (1–11) of EEOB course grade (E to A). These four groups differed dramatically in their performance in the EEOB course (Chi-Square = 17.78, $df = 3$, $p < 0.001$), with grades of students offering only KCs (Mean Rank = 184.16, $n = 69$) and students offering both types of concepts (Mean Rank = 157.89, $n = 232$) being higher than those offering only CBs (Mean Rank = 78.68, $n = 11$) or neither concept type (Mean Rank = 86.33, $n = 6$). The partial eta squared of rank for EEOB course grades was 0.06 ($F[3, 314] = 6.22$, $p < 0.001$). The *post-hoc* test using mean rank (i.e., Tukey HSD test) indicted that the EEOB course grades for the ''exclusive key concept'' group were significantly higher than the EEOB course grades for the 'exclusive cognitive bias' group ($p < 0.05$). Thus, rather than most students exclusively using KCs or exclusively falling prey to CBs, KCs and CBs often co-existed *within* the same student, with use of KCs being associated with higher academic achievement. Our findings suggest that commonly used either-or item formats (e.g., CINS) are at odds with basic cognitive reasoning patterns.

Additional evidence for the co-existence of KCs and CBs came from our analysis of explanation consistency, which is ''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'' (Kampourakis & Zogza, 2009). We analyzed explanatory consistency across our four ACORNS items in a quantitative manner by measuring the consistency of element use (i.e., KCs and CBs) across the four items. For example, use of *heritability* by a participant across three or more prompts would be considered a consistent application of this concept, as would the application of a CB such as *essentialism* across three or more responses. In this way, our measure of consistency reveals the degree to which students view the same explanatory variables as relevant to problems differing in surface features. Because consistent application of causal core concepts is a hallmark of evolutionary expertise (Nehm & Ridgway, 2011), we would expect this measure to reveal gradations of evolutionary competency.

Table 6 presents the consistency that each concept was used across the four ACORNS items. Overall, we found that participants infrequently applied the same concepts across all four items. Rather, the composition of evolutionary explanations was highly contingent on the items (and their surface features), with fully half of the sample never using the same explanatory elements to solve the ACORNS items. Specifically, 0% of students consistently used *intentional* explanations across the four items differing in surface features, whereas 13.8% of students consistently applied the concept of *differential survival* across the four prompts. Thus, rather than biology students having a stable mental model for explaining all evolutionary change, it may be more accurate to describe students as having the propensity to display certain types of reasoning in some contexts but not others (cf. Nehm & Ha, 2011). These findings corroborate prior work suggesting that the unstable application of core concepts characterizes evolution novices (Nehm & Ridgway, 2011).

Given the importance of (causally central) core concepts and cognitive biases for student achievement in biology, we wanted to know the relations among them. Theoretically, our cognitive model holds that they should compete for use as expertise develops. To address this issue, we first examined the gross relation between KCs and CBs at the group and at the individual levels to determine if they were positively, negatively, or not correlated. When we collapsed scores over subjects, we found that as the number of CBs increased, the frequency of KCs decreased ($r[7] = -0.94$, $p < 0.0001$). The same negative relation between CBs and KCs held at the level of individual students, with students generating the most CBs also generating the fewest number of KCs ($r[318] = -0.34$, $p < 0.0001$). A similar pattern held for the *diversity* of KCs and CBs, with students generating the most diverse CBs also generating the least diverse KCs ($r[318] = -0.14$, $p < 0.02$). Thus, overcoming cognitive biases and providing key concepts of evolutionary change do not appear to be completely independent events.

We next examined the correlation structure among individual KCs and CBs to determine whether they were associated as we hypothesized. As shown in Figure 3, we found high

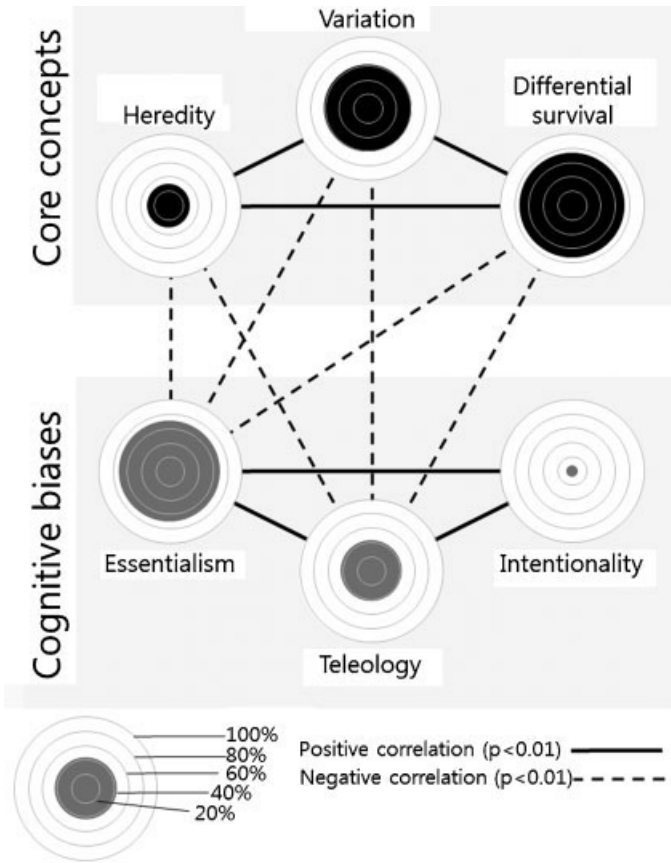

Figure 3. Relations among core concepts (black circles) and cognitive biases (gray circles) in students' explanations. Size of circles denotes percentage of students using a concept at least once across the four items.

positive correlations among each of the core concepts (*causes of variation*, *heritability*, and *differential survival*), which were also negatively correlated with the CBs *teleology* and *essentialism*. Additionally, there were high positive correlations among the CBs (*teleology*, *essentialism*, and *intentionality*), which had no or negative correlations with both core and non-core KCs. Thus, the correlation structure among KCs and CBs confirmed our prior conceptualization: core key concepts and cognitive biases were a relatively coherent group of cognitive structures (though with correlations being greater than zero and less than one, co-existence of KCs and CBs were also evident).

Theme 4: In the Development of Expertise, Knowledge Becomes More Abstract and Less Specific to the Learning Situation. The final issue that we examined was the influence of surface features of evolutionary problems (i.e., familiarity and animacy) on students' evolutionary reasoning. To examine this, we conducted a 2 (animal, plant) × 2 (familiar, unfamiliar) factorial repeated-measures ANOVA on the total number of KCs (0–7) and CBs (0–3) provided across student explanations to all four ACORNS prompts (see Figure 4). As hypothesized, more KCs were mentioned when explaining animal evolution ($M = 1.37$, $SD = 1.12$) than plant evolution ($M = 1.05$, $SD = 1.09$, $F[1, 319] = 51.5$, $p < 0.0001$, partial eta squared = 0.14). Additionally, more KCs were mentioned when explaining the evolution of familiar taxa and traits ($M = 1.37$, $SD = 1.17$) than unfamiliar taxa and traits ($M = 1.05$, $SD = 1.03$), $F[1, 319] = 46.7$, $p < 0.0001$, partial eta squared = 0.13). Finally, the effect of taxa interacted with familiarity ($F[1, 319] = 23.09$, $p < 0.0001$, partial eta squared = 0.13) with familiarity differences being much larger for animals (familiar, $M = 1.68$, $SD = 1.15$; unfamiliar, $M = 1.07$, $SD = 0.98$) than for plants (familiar, $M = 1.06$, $SD = 1.10$; unfamiliar, $M = 1.05$, $SD = 1.09$). On the other hand, we did not find significant differences in CBs between animals and plants (animal: $M = 0.55$, $SD = 0.72$, plant: $M = 0.58$, $SD = 0.77$, $F[1, 319] = 0.9$, $p > 0.05$), and between familiar and unfamiliar taxa (familiar: $M = 0.58$, $SD = 0.75$, unfamiliar: $M = 0.55$, $SD = 0.72$, $F[1, 319] = 0.8$, $p > 0.05$). In addition, we did not find an interaction effect for these factors ($F[1, 319] = 0.04$, $p > 0.05$). Thus, surface features of evolutionary problems generally had a large effect on students' use of key concepts in their evolutionary reasoning, whereas cognitive biases were present even for the least familiar organisms.

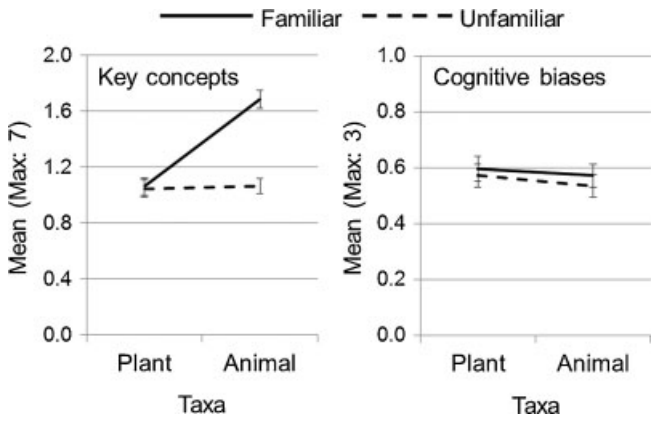

Figure 4. Mean key concepts and cognitive biases for items varying in familiarity and taxon (maximum number of KCs = 7, CBs = 3).

We next examined if more advanced students (i.e., those earning "A"s in the EEOB course) were less influenced by superficial item features than less advanced students (i.e., those earning "C"s or less). To test this issue, we first calculated an Animacy Bias and Familiarity Bias score for each student, where the scores reflected the ratio (expressed in difference of logarithms) of animal to plant KCs and CBs for the Animacy Bias and the ratio of familiar to unfamiliar KCs and CBs for the Familiarity Bias. Thus, positive Bias scores indicate a bias for animals or familiar items, whereas negative Bias scores indicate a bias for plants or unfamiliar items. As hypothesized, "C" students ($n = 68$) showed a greater Animacy Bias in their CBs ($M = 1.22$, $SD = 5.42$) than "A" students ($n = 88$; $M = -0.92$, $SD = 5.71$; $t[154] = 2.38$, $p < 0.05$, Cohen's $d = 2.62$), and the Animacy Bias also tended to be greater for the KCs of "C" students than "A" students ("C" students: $M = 2.19$, $SD = 4.9$; A-students, $M = 1.09$, $SD = 3.19$; $t[154] = 1.60$, $p = 0.11$, Cohen's $d = 0.27$). In contrast, Familiarity Biases did not differ between the two groups of students. Thus, as students advanced in their EEOB performance, some superficial features (animacy) had progressively less influence on the quality of students' explanations of evolutionary change.

General Discussion

Contemporary assessments of science understanding aim to situate students along a continuum of conceptual expertise (cf. Alonzo & Gotwals, 2012). Reform documents (e.g., NRC, 2001) have argued that cognitive research on the progressive development of expertise in a domain should therefore be used to ground and build assessment instruments capable of generating valid and robust inferences about students' conceptual development. In this paper, we used the NRC's (2001) assessment triangle to guide our research. This heuristic emphasizes the relationships among cognitive models, assessment methods, and inferences from assessment scores (NRC, 2001; Figure 1). While each vertex of the assessment triangle framed our work, it is important to emphasize that "[a] crucial point is that each of the three elements of the assessment triangle not only must make sense on its own, but also must connect to each of the other two elements in a meaningful way to lead to an effective assessment and sound inferences" (NRC, 2001, p. 49). This view resonates with contemporary theories of validity (e.g., Kane, 2001) in which construct validity elements require integration into a robust argumentative structure.

In line with the NRC (2001) policy document, our study used the assessment design triangle to frame an integrative cognitive model for natural selection assessment. First, an explicit and research-based *cognitive model* grounded the design of our assessment. Put briefly, this model posits that with the development of expertise, meaningful core concepts increasingly organize information in memory for long-term recall, causally central features become weighted more strongly in explaining data, normative scientific ideas co-exist but increasingly outcompete naive ideas in reasoning, and knowledge becomes more abstract and less specific to the learning situation. Our open-ended *assessment items* were specifically developed to align with these cognitive principles, and generate scores that prioritized recall over recognition, differentiated students' use of causally central and peripheral concepts; permitted the co-existence of normative scientific and non-normative naïve ideas in their responses; and manipulated item surface features (familiarity and animacy). Finally, we argued that *inferences from item scores* could be meaningfully linked back to the cognitive processes that span the continuum of expertise, including correspondence of task scores to clinical interviews, associations of task scores with expert performance patterns, and correspondence of task scores with independently developed natural selection knowledge measures.

As we will review below (''What We Learned about How Biology Students Think about Evolution''), our integrative cognitive model for natural selection assessment proved highly useful in characterizing the explanations of the typical biology undergraduate and in predicting students' real-world performance. In contrast, alternative evolution assessments have not been as effective at ''knowing what students know'' and perform poorly when attempting to predict performance on other assessments or real-world performance (e.g., an interview, teacher tests).

What We Learned About How Biology Students Think about Evolution

Core Concepts of Evolutionary Theory Facilitate Long-Term Recall of Biological Knowledge. In many domains of educational research (e.g., math learning), claims about ''core ideas'' are evaluated empirically by examining whether mastery of some part of the ''core'' is correlated with other parts of the ''core'' and whether degree of mastery of the ''core'' is correlated with overall achievement in the domain (e.g., grades in math classes, scores on math sections of standardized tests, etc.). We used a similar empirical strategy for examining whether the commonly accepted view, that evolution by natural selection is a core idea in biology (Dobzhansky, 1973), is empirically justified. Our basic hypothesis was that student's knowledge of the core concepts of natural selection would make biology course material more meaningful to them, facilitate their recall of this material on university exams, and thereby lead to a positive correlation between core concept use on ACORNS and future grades in biology.

Consistent with our idea that the core concepts of evolutionary theory facilitate recall of biological knowledge, we found that how students explained evolutionary change on ACORNS reliably predicted later grades in a course on ecological and organismal biology, even after controlling for overall GPA. Specifically, students who most often mentioned the *core concepts* of natural selection (heredity, variation, and differential survival; Lewontin, 1970) when explaining evolution later went on to achieve the highest grades in the biology course; a similar relation did *not* hold for non-central elements of natural selection (e.g., competition, hyperfecundity, etc.). Additionally, students who most often provided naïve essentialist and teleological explanations later went on to achieve the lowest grades in the biology course, suggesting that possession of these cognitive biases interferes with academic achievement. Again, these findings are consistent with the theoretical centrality of natural selection for making sense of the diversity of living things, as well as the broader educational importance of understanding the causes of evolution and of avoiding naïve biological ideas (cf. Dobzhansky, 1973). In contrast, alternative explanations of this correlation are much less plausible. Because course grades were earned *after* the ACORNS items were administered, for example, it is impossible for the causal direction to be the reverse. Further, many potential third variables that might explain this correlation can be ruled out, including overall academic proficiency (because GPA had been controlled) or biology-specific proficiency and/or motivation (because ACORNS scores were not found to correlate with performance in a previous cell and molecular biology course).

Core Concepts and Cognitive Biases Co-Exist and Compete in Students' Thinking. Although it is tempting to think of science learning as a simultaneous acquisition of scientific concepts and rejection of non-scientific concepts, our study of biology majors' understanding of evolution suggests a very different picture. That is, the vast majority of the biology students that we studied explained evolutionary change using *both* scientific concepts that cause evolutionary change (e.g., variation, heritability, and differential survival) *and* naïve cognitive biases (e.g., essentialism, teleology, and—to a lesser extent—intentionality). This finding is

important for two reasons. First, it suggests that previous assessments of evolutionary understanding (i.e., multiple choice tests like the CINS) can radically overestimate students' understanding by asking them to choose between a scientifically valid and invalid explanation of evolutionary change (see Nehm & Schonfeld, 2008 for empirical evidence of this pattern). In our study, 73% of students would have wanted to answer ''both'' for such a choice, and 2% would have wanted to answer ''neither''. Thus, forced-choice, *either-or assessments are at risk of misdiagnosing up to 75% of students' preferred explanations of evolutionary change*. Second, the co-existence of key concepts and cognitive biases strongly suggests that pedagogical efforts aimed only at introducing the necessary and sufficient causes of evolutionary change, or only at disabusing students of their naïve beliefs, are likely to prove insufficient. Indeed, our finding that continued use of cognitive biases—even with correct use of key concepts—was associated with lower grades in biology strongly suggests that educators should explicitly address cognitive biases when instructing students about the causal processes of evolutionary change. Unfortunately, few pedagogical and curricular tools are available for doing so.

Irrelevant Surface Features Impact Novice Students' Evolutionary Reasoning. An important recent advance in assessment of evolution has been the finding that the knowledge and naïve ideas that students employ depends very greatly on the specific contexts on which they are assessed (Nehm & Ha, 2011; see also Chi et al., 1981 for a classic study in physics). For example, some students use causally central key concepts to explain how a trait (such as the running speed of a cheetah) will *increase* in phenotypic frequency; however, the same students seldom mention these variables when explaining how traits (such as flight in birds) *decline* in phenotypic frequency. Indeed, understanding of one type of evolutionary change is typically a very poor predictor of understanding the other type. The present research reveals two *additional* surface features that impact the typical student's use of key concepts in evolutionary reasoning: overall familiarity with items and whether the item involves a plant or animal. Indeed, the surface feature of being an animal was particularly influential on the ''C'' students in the EEOB course. For these students, animal items were particularly likely to elicit cognitive biases and particularly unlikely to elicit key concepts.

These findings of context-dependent learning have a number of important implications for teaching evolution. First, curricula about evolution and natural selection may require much care in the choice of the ''cover stories'' (such as bacterial resistance to antibiotics) that are used to illustrate evolutionary change. Ideally, such examples would represent a diversity of evolutionary scenarios that could be systematically compared and contrasted. In a variety of subject areas (e.g., mathematics), choosing examples that allow systematic comparisons is known to help students identify the variables that are truly important for problem-solving (Gentner & Colhoun, 2010), and we think it quite likely that the same would be true in learning the important variables that cause evolutionary change via natural selection. Additionally, ''cover stories'' might be chosen to reveal and address the naïve ideas that plague student reasoning. Like students' understanding of the key variables in evolutionary change, misconceptions are also context-dependent, with misconceptions triggered by some contexts being rarely elicited by other contexts (Nehm & Ha, 2011).

Although superficial features had a marked effect on the typical students' use of key concepts in explaining evolutionary change, it is interesting to note that these features did not impact the probability that they used cognitive biases (such as essentialist or teleological explanations). One reason this may be the case is that essentialism and teleological thinking is so fundamental to the cognitive architecture that they pervade thinking about all living things, regardless of familiarity or animacy (Gelman, 2003). A similar idea was also

expressed by Dawkins (1996, p. 316), who wrote, ''It is almost as if the human brain were specifically designed to misunderstand Darwinism.'' More research is needed to evaluate these positions.

Knowing What Students Know About Evolution

Are other extant assessment instruments also capable of detecting the observed features of the typical undergraduate's thinking about evolution? Consider, for example, the Conceptual Inventory of Natural Selection (CINS) test, which comprises 20 multiple-choice questions that focus on common misconceptions about 10 components of natural selection (Anderson et al., 2002). Several features of CINS suggest that it cannot diagnose mental models of evolutionary change. First, by relying on multiple-choice items, the CINS chiefly tests recognition memory, which is sometimes highest among individuals with *lowest* competence (Sloutsky & Fisher, 2004), rather than requiring free recall, where task performance is more reliably correlated with expertise (Chi, 2006). By relying on such a low bar for remembering evolutionary facts, the CINS is predicted by our cognitive model to also show weaker correspondence to clinical interviews than ACORNS, an implication supported in prior work (Nehm & Schonfeld, 2008). Second, CINS scores do not differentially weight understanding of causally central versus peripheral evolutionary concepts, thereby threatening the ability of CINS to predict levels of expertise (an issue that has not yet been examined rigorously). Third, by asking students to choose between normative scientific ideas and misconceptions, the test precludes the ability to find the co-existence of both ideas in any given student, thereby threatening its ability to predict performance on other independently developed measures. Consistent with this idea, correlations between scores on the CINS and misconceptions on ACORNS are much lower than correlations between scores for key concepts on CINS and ACORNS. Finally, rather than varying levels of item familiarity, items from the CINS uniformly test performance on only one surface feature (familiar animals) that students are likely to have encountered in class. This is also unfortunate because items regarding unfamiliar animals, familiar plants, and unfamiliar plants typically fail to elicit causal core concepts among college students yet cause evolutionary biologists no difficulty.

One reason that extant evolution assessments fail to capture these central elements of evolutionary thinking is that these assessments are not based on general principles of cognition. Indeed, of all the extant assessments of evolutionary reasoning, none are designed with our four assessment features. This should not be surprising—none of the extant assessments were designed with an explicit, research-based cognitive model as a foundation.

In sum, we have demonstrated that the limitations of extant assessment instruments (e.g., CINS) may be traced to the lack of explicit, research-based cognitive models. As a result, these assessment tasks can generate faulty inferences about student thinking about natural selection. In many ways, our work demonstrates how ignoring the NRC (2001) framework may lead to problematic assessment tools and questionable evidence for improving science teaching. Put more positively, our findings also illustrate why core insights from cognitive science must be used in the design of assessment instruments intended to capture students' progress along a novice-expert continuum (cf. Alonzo & Gotwals, 2012). While our study has demonstrated how four cognitive principles—(1) prioritizing recall over recognition; (2) detecting students' use of *causally central* information; (3) permitting co-existence of scientific and naïve ideas; and (4) attending to task surface features—impact the assessment of natural selection, more work is needed to test the generality of these principles in other science domains.

Our results resonate with several recent findings in other science domains, and help to identify important directions for science education assessment research. In a study of students' reasoning about force and motion, for example, Alonzo and Steedle (2009) noted that item pairs prompting reasoning about the same phenomenon, but framed in different contexts, produced different reasoning patterns. The researchers' explanation for these patterns was that students performed better on items that used the same surface features as had been discussed in class (Alonzo & Steedle, 2009, p. 414). Thus, the performance differences that they documented may represent instances of memorized cases and/or may reflect the effects of familiarity on response processes (Bulloch & Opfer, 2009; Gentner, 1988; Hartshorn et al., 1998; Keil & Batterman, 1984). Alonzo and Steedle's (2009) results share many similarities with our findings about evolutionary reasoning. However, the experimental design of Alonzo and Steedle's (2009) study did not control for many variables, including order effects, and so their findings must be interpreted with caution. Nevertheless, these and related studies suggest that much more work is needed to pinpoint the types of surface features that impact student response processes on science assessments (Chi et al., 1981; Nehm & Ha, 2011).

Unknowns in What Students Know About Evolution

Although the current project was highly-revealing about what students know about evolution, several issues are in need of future investigation. First, many different approaches have been used to evaluate students' explanations in science education (Alonzo & Gotwals, 2012, chapters therein; diSessa, 1993; Russ, Scherr, Hammer, & Mikeska, 2008). Our focus on the causal components of explanations is but one method for measuring progressions of cognitive competency in relation to evolution. Alternative approaches, such as the analysis of causal language (diSessa, 1993), scoring the frequency of casual ''chaining'' (Russ et al., 2008), or evaluating explanations in the context of argumentative practices (Gotwals & Songer, 2010), may yield additional insights into student reasoning processes. Second, while written explanations were suitable assessment tasks for our sample of undergraduates, such formats risk under-representing measures of understanding in English language learners. Third, we operationalized students' progressions towards expertise using course grades in an evolutionary biology course. While this method confirmed the predictions of our model, course grades certainly can be influenced by factors other than student expertise. Additional studies using more robust measures would help to bolster our findings. Finally, more studies of how explanations are used by professional evolutionary biologists are needed (Nehm & Ridgway, 2011). Benchmarks of causal competency and sophistication should be grounded in empirical studies of how experts conceptualize and explain evolutionary change.

We thank the editors and reviewers of the special issue for providing helpful feedback on our manuscript. We wish to especially thank Dr. Ruiz-Primo for devoting so much time and effort to helping us improve the quality and clarity of our work. We also thank the National Science Foundation REESE Program (Award Number 0909999) for financial support of our study.

References

Ahn, W., Gelman, S. A., Amsterlaw, J. A., Hohenstein, J., & Kalish, C. W. (2000). Causal status effect in children's categorization. Cognition, 76, 35–43.

Alonzo, A. C., & Gotwals, A. W. (2012). Learning progressions in science: Current challenges and future directions. Rotterdam, The Netherlands: Sense Publishers.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. Science Education, 93, 389–421.

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

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

Anderson, J. R., Reder, L. M., & Simon, H. A. (2000). Applications and misapplications of cognitive psychology to mathematics education. Texas Education Review, 1, 29–49.

Anglin, J. M. (1977). Word, object, and conceptual development. New York: W.W. Norton & Company Inc.

Au, T. K., Chan, C. K. K., Chan, T., Cheung, M. W. L., Ho, J., & Ip, G. W. M. (2008). Folkbiology meets microbiology: A study of conceptual and behavioral change. Cognitive Psychology, 57, 1–19.

Au, T. K., Romo, L. F., & DeWitt, J. E. (1999). Considering children's folkbiology in health education. In M. Siegal & C. C. Peterson (Eds.), Children's understanding of biology and health (pp. 209–234). Cambridge, MA: MIT Press.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychological Bulletin, 128, 612–637.

Bartsch, K., & Wellman, H. M. (1995). Children talk about the mind. New York: Oxford University Press.

Beggrow, E. P., & Nehm, R. H., (2012). *Exploring the role of non-adaptive reasoning in students' evolutionary explanations*. Proceedings of the National Association for Research in Science Teaching (NARST) annual conference, Indianapolis, IN.

Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. Science Education, 94(5), 765–793.

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

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. Developmental Psychology, 41, 189–201.

Bower, G. H., Karlin, M. B., & Dueck, A. (1975). Comprehension and memory for pictures. Memory & Cognition, 3(2), 216–220.

Brainerd, C. J., & Gordon, L. (1994). Development of verbatim and gist memory for numbers. Developmental Psychology, 30(2), 163–177.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall1. Journal of Verbal Learning and Verbal Behavior, 11(6), 717–726.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. Educational Assessment, 11, 33–64.

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

Bulloch, M. J., & Opfer, J. E. (2009). What makes relational reasoning smart? Revisiting the relational shift in cognitive development. Developmental Science, 12, 114–122.

Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. Cognitive Development, 7, 213–233.

Carey, S. (1985). Conceptual change in childhood. Cambridge, MA: MIT Press.

Case, R., & Okamoto, Y. (1996). The role of central conceptual structures in the development of children's thought. Monographs of the Society for Research in Child Development, 61, 1–2.

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), The Cambridge handbook of expertise and expert performance (pp. 21–30). New York: Cambridge University Press.

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

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.

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

Cozby, P. C. (1997). Methods in behavioral research (6th ed.) Mountain View, CA: Mayfield Publishing Company.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing. A framework for memory research. Journal of Verbal Learning and Verbal Behaviour, 11, 671–684.

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

Dawkins, R. (1996). The blind watchmaker: Why the evidence of evolution reveals a universe without design. New York: W.W. Norton & Company.

diSessa, A. A. (1993). Toward an epistemology of physics. Cognition and Instruction, 10(2–3), 105–225.

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

Endler, J. A. (1992). Natural selection: Current uses. In E. F. Keller & E. A. Lloyd (Eds.), Keywords in evolutionary biology (pp. 220–224). Cambridge, MA: Harvard University Press.

Ericsson, K. A., & Smith, J. (1991). Toward a general theory of expertise: Prospects and limits. Cambridge, MA: Cambridge University Press.

Evans, E. M. (2001). Cognitive and contextual factors in the emergence of diverse belief systems: Creation versus evolution. Cognitive Psychology, 42, 217–266.

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

Fenker, D. B., Waldmann, M. R., & Holyoak, K. J. (2005). Accessing causal relations in semantic memory. Memory & Cognition, 33, 1036–1046.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. Educational Measurement: Issues and Practice, 11(4), 21–26.

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

Gelman, S. A. (2004). Psychological essentialism in children. Trends in Cognitive Sciences, 8, 404–409.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. Child Development, 59, 47–59.

Gentner, D., & Colhoun, J. (2010). Analogical processes in human thinking and learning. In A. von Müller & E. Pöppel (Series Eds.), B. Glatzeder, V. Goel, & A. von Müller (Vol. Eds.), On thinking: Vol. 2 towards a theory of thinking. Berlin, Heidelberg, New York: Springer-Verlag.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. American Psychologist, 52, 45–56.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15, 1–38.

Goldberg, R. F., & Thompson-Schill, S. L. (2009). Developmental "roots" in mature biological knowledge. Psychological Science, 20(4), 480–487.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. Science Education, 94, 259–281.

Griffin, S. A., Case, R., & Siegler, R. S. (1994). Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk for school failure. In K. McGilly (Ed.), Classroom lessons: Integrating cognitive theory and classroom practice (pp. 25–49). Cambridge, MA: MIT Press.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. Psychological Science, 18, 1069–1076.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized scoring models of written biological explanations across courses and colleges: Prospects and limitations. CBE-Life Science Education, 10, 379–393.

Hartshorn, K., Rovee-Collier, C., Gerhardstein, P., Bhatt, R. S., Klein, P. J., & Aaron, F. (1998). Developmental changes in the specificity of memory over the first year of life. Developmental Psychobiology, 33, 61–78.

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of physics. In K. McGilly (Ed.), Classroom lessons: Integrating cognitive theory and classroom practice (pp. 51–74). Cambridge, MA: MIT Press.

Inagaki, K., & Hatano, G. (2002). Young children's naive thinking about the biological world. New York: Psychology Press.

Inagaki, K., & Hatano, G. (2004). Vitalistic causality in young children's naive biology. Trends in Cognitive Sciences, 8, 356–362.

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

Kane, M. T. (2001). Current concerns in validity theory. Journal of Educational Measurement, 38, 319–342.

Katona, G. (1940). Organizing and memorizing: Studies in the psychology of learning and teaching. New York: Columbia University Press.

Keil, F. C. (2006). Explanation and understanding. Annual Review of Psychology, 57, 227–254.

Keil, F. C., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. Journal of Verbal Learning and Verbal Behavior, 23, 221–236.

Kelemen, D. (2003). British and American children's preferences for teleo-functional explanations of the natural world. Cognition, 88, 201–221.

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.

Krascum, R. M., & Andrews, S. (1998). The effects of theories on children's acquisition of family-resemblance categories. Child Development, 69(2), 333–346.

Legare, C. H., Evans, E. M., Rosengren, K. S., & Harris, P. L. (2012). The coexistence of natural and supernatural explanations across cultures and development. Child Development, 83, 779–793.

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

Lewontin, R. C. (1978). Adaptation. Scientific American, 239, 212–228.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20, 15–21.

Lombrozo, T. (2006). The structure and function of explanations. Trends in Cognitive Sciences, 10(10), 464–470.

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), Oxford handbook of thinking and reasoning (pp. 260–276), Oxford, UK: Oxford University Press.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. Cognition, 99, 167–204.

Marion, S. F., & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternate assessment. Educational Measurement: Issues and Practice, 25, 47–57.

Marshall, S. P. (1995). Schemas in problem-solving. New York: Cambridge University Press.

McCloskey, M. (1983). Intuitive physics. Scientific American, 248, 114–122.

National Research Council. (2007). Taking science to school: Learning and teaching science in grades K-8. Washington, DC: National Academy Press.

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

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press.

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

Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. The American Biology Teacher, 74(2), 92–98.

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

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. Journal of Science Education and Technology, 21(1), 183–196.

Nehm, R. H., Ha, M., Rector, M., Opfer, J. E., Perrin, L., Ridgway, J., & Mollohan, K., (2010). *Scoring guide for the open response instrument (ORI) and evolutionary gain and loss test (ACORNS).* Technical Report of National Science Foundation REESE Project 0909999.

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

Nehm, R. H., & Ridgway, J. (2011). What do experts and novices ''see'' in evolutionary problems? Evolution: Education and Outreach, 4(4), 666–679.

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.

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

Opfer, J. E. (2002). Identifying living and sentient kinds from dynamic information: The case of goal-directed versus aimless autonomous movement in conceptual change. Cognition, 86, 97–122.

Opfer, J. E., & Bulloch, M. J. (2007). Causal relations drive young children's induction, naming, and categorization. Cognition, 105, 206–217.

Opfer, J. E., & Gelman, S. A. (2001). Children's and adults' models for predicting teleological action: The development of a biology-based model. Child Development, 72, 1367–1381.

Opfer, J. E., & Gelman, S. A. (2010). Development of the animate-inanimate distinction. In U. Goswami (Ed.), The Wiley-Blackwell handbook of childhood cognitive development. Cambridge, UK: Blackwell.

Opfer, J. E., & Siegler, R. S. (2004). Revisiting preschoolers' living things concept: A microgenetic analysis of conceptual change in basic biology. Cognitive Psychology, 49, 301–332.

Opfer, J. E., & Siegler, R. S. (2012). Development of quantitative thinking. In K. Holyoak & R. Morrison (Eds.), Oxford handbook of thinking and reasoning. Cambridge, UK: Oxford University Press.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the web (Tech. Rep.). Stanford, CA: Stanford Digital Library Technologies Project.

Patterson, C. (1978). Evolution. Ithaca: Cornell University Press.

Perkins, D. N., & Grotzer, T. A. (2005). Dimensions of causal understanding: The role of complex causal models in students understanding of science. Studies in Science Education, 41(1), 117–165.

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

Poling, D. A., & Evans, E. M. (2002). Why do birds of a feather flock together? Developmental change in the use of multiple explanations: Intention, teleology and essentialism. British Journal of Developmental Psychology, 20, 89–112.

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. Cognition, 99, 73–112.

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. Science Education, 92(3), 499–525.

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

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

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

Sinatra, G. M., Brem, S. K., & Evans, E. M. (2008). Changing minds? Implications of conceptual change for teaching and learning about biological evolution. Evolution: Education Outreach, 1, 189–195.

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

Sloutsky, V. M., & Fisher, A. V. (2004). When learning and development decrease memory: Evidence against category-based induction. Psychological Science, 15, 553–558.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. Cognitive Psychology, 24, 535–585.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. Cognition & Instruction, 16, 3–118.

Woods, N. N., Brooks, L. R., & Norman, G. R. (2007). It all makes sense: Biomedical knowledge, causal connections and memory in the novice diagnostician. Advances in Health Sciences Education, 12, 405–415.