J Sci Educ Technol (2014) 23:160–182
DOI 10.1007/s10956-013-9461-9

Assessing Scientific Practices Using Machine-Learning Methods: How Closely Do They Match Clinical Interview Performance?

Elizabeth P. Beggrow · Minsu Ha · Ross H. Nehm · Dennis Pearl · William J. Boone

Published online: 16 July 2013

Abstract The landscape of science education is being transformed by the new *Framework for Science Education* (National Research Council, A framework for K-12 science education: practices, crosscutting concepts, and core ideas. The National Academies Press, Washington, DC, 2012), which emphasizes the centrality of scientific practices—such as explanation, argumentation, and communication—in science teaching, learning, and assessment. A major challenge facing the field of science education is developing assessment tools that are capable of validly and efficiently evaluating these practices. Our study examined the efficacy of a free, open-source machine-learning tool for evaluating the quality of students' written explanations of the causes of evolutionary change relative to three other approaches: (1) human-scored written explanations, (2) a multiple-choice test, and (3) clinical oral interviews. A large sample of undergraduates ($n = 104$) exposed to varying amounts of evolution content completed all three assessments: a clinical oral interview, a written open-response assessment, and a multiple-choice test. Rasch analysis was used to compute linear person measures and linear item measures on a single logit scale. We found that the multiple-choice test displayed poor person and item fit (mean square outfit >1.3), while both oral interview measures and computer-generated written response measures exhibited acceptable fit (average mean square outfit for interview: person 0.97, item 0.97; computer: person 1.03, item 1.06). Multiple-choice test measures were more weakly associated with interview measures ($r = 0.35$) than the computer-scored explanation measures ($r = 0.63$). Overall, Rasch analysis indicated that computer-scored written explanation measures (1) have the strongest correspondence to oral interview measures; (2) are capable of capturing students' normative scientific and naive ideas as accurately as human-scored explanations, and (3) more validly detect understanding than the multiple-choice assessment. These findings demonstrate the great potential of machine-learning tools for assessing key scientific practices highlighted in the new *Framework for Science Education*.

Keywords Applications in subject areas · Evaluation methodologies · Improving classroom teaching · Pedagogical issues · Teaching/learning strategies

Introduction

Science and engineering continue to play key roles in addressing social, economic, and political challenges facing the world (American Association for the Advancement of Science (AAAS) 2011). For this reason, quality science and engineering education remain high-priority goals, not only for the purpose of maintaining a high-skilled workforce, but also for fostering the development of a

E. P. Beggrow (✉) · M. Ha
Department of Teaching and Learning, The Ohio State University, 333 Arps Hall, 1945 N High Street, Columbus, OH 43210, USA
e-mail: beggrow.7@osu.edu

R. H. Nehm
Center for Science and Mathematics Education, Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794, USA

D. Pearl
Department of Statistics, The Ohio State University, 404 Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210, USA

W. J. Boone
Department of Educational Psychology, Miami University, 501 East High Street, Oxford, OH 45056, USA

scientifically literate citizenry that can engage in effective decision making about complex issues like global warming, energy security, disease prevention, and genetically modified organisms (National Research Council (NRC) 2012). In response to demands for a more scientifically literate citizenry and workforce, the landscape of science education is being transformed in the United States; science standards and instruction are continuing to move away from the presentation of science as a collection of facts to be memorized, and toward an inquiry-based framework that emphasizes the integration of practices (e.g., explanation, argumentation, and communication), crosscutting concepts (e.g., patterns, cause and effect, and system models), and core ideas (e.g., energy and evolution) into science teaching, learning, and assessment (National Research Council 2012).

Integrating practice-based skills into inquiry teaching is expected to help students learn how scientific understanding is generated, in contrast to lecture-based instruction that emphasizes the memorization of the *outputs* of the scientific enterprise (American Association for the Advancement of Science 2011), that is, students are expected to participate in authentic scientific practices by *doing* science. But embarking on this large-scale reform project will require a radical departure from everyday assessments dominated by multiple-choice and true–false formats. Indeed, developing and evaluating scientific arguments and explanations, and communicating such understanding to others, cannot be meaningfully assessed using forced-choice tests. A major challenge facing the field of science education is building assessment tools and systems that are capable of validly and efficiently evaluating authentic scientific practices (National Research Council 2001b, 2012).

The practice of explanation brings its own unique challenge to the task of assessment. Much of that challenge lies in the fact that several different perspectives have emerged on the practice of explanation, particularly about how to define an explanation and how to assess an explanation (e.g., Berland and McNeill 2012; Braaten and Windschitl 2011; Osborne and Patterson 2011; Russ et al. 2008; Songer and Gotwals 2012). Four of the more salient perspectives that have been proposed in the literature include defining explanations: (1) as mechanistic statements (Russ et al. 2008); (2) as components of argumentation (Berland and McNeill 2012); (3) as integrated packages consisting of scientific claims, evidence, and reasoning (Songer et al. 2009); and (4) as a means of making further sense of an agreed-upon fact or phenomenon (Osborne and Patterson 2011). Clearly, very different conceptualizations of the practice of explanation exist in the field of science education.

Despite the lack of consensus in the literature on the definition of a scientific explanation, researchers have found a variety of ways to introduce this practice to students at different age levels. For instance, Songer and Gotwals (2012) emphasized the importance of integrating inquiry with content and have explored how to scaffold the practice of explanation along with biology content learning (with both activities embedded within learning progressions). Their research focused on scaffolding elementary and middle school students' evidence-based explanation-building skills. Developing students' competencies in scientific practices is particularly challenging because younger students often lack science content knowledge *and* an understanding of the practice itself (McNeill et al. 2006).

Research has begun exploring the utility of computer-based learning environments for both developing and evaluating students' proficiency in constructing scientific explanations (e.g., Gobert et al. 2012; Woloshyn and Gallagher 2009). Specifically, Gobert and colleagues' work (both with the Calipers Project and *Science Assistments*) has created an intelligent tutoring system with the purpose of scaffolding the development of scientific process skills in middle school students. Along with assisting students with hypothesis generating and collecting and interpreting data, the intelligent tutoring system, *Science Assistments*, provides students with the opportunity to construct an explanation of a simulated experiment (Gobert et al. 2012). Yet, the authors acknowledge that the explanations were not autoscored, nor were larger conceptual issues like theory and content knowledge included; both the former and the latter were considered "beyond the scope of the project" (Gobert et al. 2012, p. 163). In addition to scaffolding scientific process skills in general, some research has specifically addressed how computer-based learning programs can guide students in self-explanation (see Woloshyn and Gallagher 2009 for a brief review), a skill that has been found to be useful in fostering learning in students (e.g., Chi et al. 1989; Lombrozo 2006, 2012). This work in computer-based learning is promising and has much to offer to the larger body of research in developing and assessing students' explanation skills.

Our study differs from prior work (e.g., Songer, McNeil, and Gobert and colleagues) in several respects. First, our study focuses on the assessment of students' content knowledge and scientific reasoning expressed in written explanations. Second, it does not explore how well students have mastered the practice of explanation itself (e.g., whether students express integrated packages of scientific claims, evidence, and reasoning). Third, our study does not examine the impact of having students engage in the practice of explanation on content learning (Chi et al. 1989; Lombrozo 2012). Fourth, it attempts to bring computer tools to bear on the 30-year history of using a particular type of written explanation to assess evolutionary understanding (e.g., Deadman and Kelly 1978). Given the long history of using written explanations in evolution assessment, we focus on this topic in detail.

Assessment of Core Ideas in Science: Evolution

Biological evolution is a topic with direct relevance to many current societal challenges emphasized by the National Research Council (2007): the evolution of antibiotic resistance (e.g., MRSA); the origins of new pathogens (e.g., new strains of Swine flu); the preservation of endangered species (e.g., Florida panther); and the development of genetically engineered crops, to name a few. Despite the NRC's (2012) recognition of evolution as one of four core ideas in science education (National Research Council 2012), a large body of work has shown that it is a particularly difficult subject for students to learn (Nehm and Reilly 2007). For this reason, evolution assessments are of considerable importance, not only for documenting student learning difficulties, but also for evaluating the comparative efficacy of alternative instructional approaches (e.g., Nehm and Reilly 2007).

Long before the current emphasis on the importance of scientific practices in teaching and learning, many science education research studies recognized the utility of written explanation tasks for gaining insights into students' mental models and reasoning about evolution and natural selection (e.g., Deadman and Kelly 1978; Bishop and Anderson 1990). Asking students to construct explanations enables researchers and teachers to glean important insights about the structure and composition of student thought, that is, the open-ended format permits many permutations of "right" *and* "wrong" knowledge elements, rather than either "right" *or* "wrong" multiple-choice formats (Nehm and Haertig 2012). Despite the long-standing interest in written explanation tasks in evolution education assessment, and the recent emphasis on practices such as explanation in science instruction, the usefulness of written response assessments is limited because of the time required for administering and scoring the tests [as compared to multiple-choice (MC) formats] (for a review, see Ha et al. 2011; Haudek et al. 2011; Nehm and Schonfeld 2008; Nehm et al. 2012b).

Although some inferences derived from commonly used MC instruments (e.g., the Conceptual Inventory of Natural Selection (CINS): Anderson et al. 2002) have been shown to be valid and reliable (Nehm and Schonfeld 2008), they are not without limitations. MC assessment items—because they typically contain numerous enticing naive idea distracters—may actually cause students to develop false knowledge (Roediger and Marsh 2005). Furthermore, MC items typically contain either normative, scientifically accurate answer options, such as key concepts (KCs) of natural selection (e.g., *variation, heritability, and competition*), *or* non-normative naive ideas (NIs) (e.g., needs *cause* evolutionary change). Yet, several empirical studies have shown that normative and non-normative ideas typically coexist in large percentages of students' mental models of evolution (Beggrow and Nehm 2012; Nehm and Schonfeld 2008). Simply put, many MC formats do not accurately reflect the structure of student thinking about the core idea of evolution (Opfer et al. 2012). While ordered MC (OMC) instruments can be developed to detect students' synthetic or mixed models (Briggs et al. 2006; Nehm and Ha 2011; Vosniadou et al. 2008), no such instruments have been developed for the topic of evolution (Opfer et al. 2012). More fundamentally, MC assessments simply cannot assess students' communication and explanation abilities.

Another widely used format for uncovering the structure of student thinking about core ideas is the clinical oral interview (Ginsburg 1981). Clinical oral interviews provide opportunities for educational researchers to gather student explanations in a way that is similar to the gathering of written responses and has been found to be as valid (Seddon and Pedrosa 1988), but with the advantage of possible clarification, follow-ups, and divergent conversational paths (Black 1999; Joughin 1998). In other words, because interviewers are in dialogue with participants, they can ask follow-up questions to clarify any points of ambiguity or confusion, or to determine whether or not students are using words or terms (such as the KCs of evolution) in a scientifically accurate way (Rector et al. 2013; Russ et al. 2012). This context provides interviewers with opportunities for gaining deeper and more valid insights into students' thinking processes and resolving instances of lexical ambiguity. With written assessments, in which there is no interaction between the student and the examiner, these opportunities for clarification and determining accurate use of terms are absent. Instead, the response must be interpreted "as-is" and evaluation relies upon the examiner's interpretation of meanings and intentions. For these reasons, the clinical interview is often considered to be a "gold standard" for assessing student reasoning and understanding within education research (Ginsburg 1981). Nonetheless, this format is far from ideal with large samples, as it is even more time-intensive than administering and scoring constructed-response assessments.

Recent research has shown that in the domain of evolution, written response scores and clinical oral interview scores tend to be strongly correlated (Nehm and Schonfeld 2008; Nehm et al. 2012a). Both of these assessment formats—interviews and written explanations—ask participants to recall information. In contrast, MC assessments ask students to *recognize* information and therefore may have weaker correspondence with clinical interviews relative to written explanation assessments (Nehm and

Schonfeld 2008; Opfer et al. 2012). In alignment with the NRC (2001b) goals of assessment design, assessments should prioritize recall tasks over recognition tasks, as information recall is a more robust indicator of meaningful learning (Opfer et al. 2012). This is where the CINS (and perhaps other MC instruments) fall short, as they rely on recognition over recall, and have unsurprisingly been found to have weaker associations with clinical oral interviews relative to recall task scores (Nehm and Schonfeld 2008; Nehm et al. 2012a). A written response assessment format thus has several advantages over MC assessments, despite its associated costs.

Automated Computer Scoring

Given the constraints of administering and scoring written assessments, recent work has turned to computer-scoring software as a possible solution (Graesser and McNamara 2012; Haudek et al. 2011; Moscarella et al. 2008; Page 1966; Shermis and Burstein 2003; Yang et al. 2002). In particular, the research on the efficacy of computer-assisted scoring has been extended from assessing writing skills (e.g., E-Rater) to assessing specific content such as math and science (e.g., C-rater, see Leacock and Chodorow 2003). Recent studies have also explored the role of computer-assisted scoring (CAS)—particularly machine learning—in biology education research (Ha et al. 2011; Haudek et al. 2011; Nehm et al. 2012b). Both C-rater and the CAS used in this study use automated computer-scoring systems to detect the presence or absence of the content knowledge in the written responses developed by natural language processing techniques (Leacock and Chodorow 2003). However, the major difference between these two CAS systems is that the CAS used in the present study is able to detect normative scientific ideas and non-normative naïve ideas (see the “right and wrong knowledge elements” section above in the “Assessment of core ideas in science: evolution” section). Consequently, our CAS system (like C-rater) is capable of reporting not only the richness or poorness of scientific ideas in written responses (e.g., total key concept score), but it can also detect four different reasoning patterns about evolution (e.g., scientific, mixed, naïve, and no models).

While computer-scoring models for written responses of evolutionary phenomena (and many other topics) have been found to be comparable to human scoring of written responses (Graesser and McNamara 2012; Ha et al. 2011), it remains to be determined how closely computer-generated scores of written responses compare to the “gold standard” of clinical oral interviews. The comparison of computer-scored responses to clinical interviews is crucially important, yet is lacking in prior studies of the efficacy of computer-scoring methods.

A second major limitation of previous studies investigating the correspondence of assessment methods has been the exclusive use of classical test theory (CTT) analyses (e.g., Nehm and Schonfeld 2008; Nehm et al. 2012a). CTT relies on many psychometric assumptions that may not always be met. For example, CTT assumes a normal distribution and scale consistency, although it uses raw scores, which can have greatly limiting consequences when the data are ordinal/categorical in nature, as differences between scores may not always be equivalent, as is typically assumed (Neumann et al. 2011). In other words, CTT uses raw scores even though on an instrument, the difference in meaning between a score of 85 and a score of 90, for example, is not necessarily equivalent to the difference in meaning between a score of 90 and of 95. The assumption of scale consistency is rarely met, and thus, CTT analyses tend to be less sound than other approaches, such as Rasch analysis (Boone and Scantlebury 2006).

The Rasch model transforms raw data into measures on a standardized equal interval scale (Boone and Scantlebury 2006). In addition to addressing the issue of equivalent scaling, Rasch analysis is also able to provide very rigorous assessment of reliability and validity through indices such as mean square (MNSQ) and standardized z-score (ZSTD) (see Neumann et al. 2011 for details) in order to judge the quality of items, which CTT is not able to provide. Researchers use the mean of fit indices for individual items generated by Rasch analysis and the number of misfit items to judge the quality of the entire instrument. Moreover, fit indices for individual persons can also be used to evaluate participant performance.

Despite the widely accepted advantages of Rasch analysis over CTT analyses, previous studies were not able to employ this more rigorous psychometric approach because of the small clinical oral interview sample sizes (e.g., $n = 18$; Nehm and Schonfeld 2008). Another limitation of Nehm and Schonfeld’s (2008) study was that it used a written assessment that (unlike the Assessment of Contextual Reasoning About Natural Selection test [ACORNS], see *Methods*) neither controlled for order effects (Rector et al. 2012) nor included items standardized by familiarity (Nehm and Ha 2011). In order to improve upon previous approaches, and to more rigorously examine the correspondence among assessment measures, we used Rasch analysis to compare computer-scored measures of ACORNS explanation tasks to MC assessment measures and clinical oral interview explanation tasks (Boone and Scantlebury 2006). Because Rasch analysis is robust with large sample sizes (e.g., >100), we carefully interviewed more than 100 students in order to use Rasch analysis to test the fidelity of computer-scored written explanations relative to these other metrics. We know of no other study that has amassed such a large corpus of data from clinical interviews to perform an evaluation of the quality of computer-scoring methods.

Research Questions

The central goal of this study was to examine the degree to which computer-scored explanations aligned with clinical oral interview scores. In pursuing this goal, we explored a series of related questions: (1) How well do measures derived from the MC CINS test, the computer-scored and human-scored ACORNS test, and the clinical interviews fit a Rasch model? (2) Using Rasch, to what extent do measures of explanation quality align across the assessment formats? and (3) How do the different assessment formats compare in terms of the insights they provide into students' conceptual frameworks about the core idea of evolution?

Methods

Sample

We gathered data from 104 undergraduate students enrolled at a large, public, Midwestern research university in the United States. Participants from four classes took part in our study: introductory biological anthropology ($n = 26$), introductory biology for majors ($n = 28$), advanced evolutionary biology ($n = 23$), and advanced mammalogy requiring a prior course in evolution ($n = 27$). The average age of participants was 20.9 years, 40.4 % were male, and the majority was white non-Hispanic. All 104 participants voluntarily took part in individual clinical oral interviews and were offered USD $20 for their participation. Details about the interviews and instruments are discussed below.

Instruments and Scoring Methods Overview

We used three approaches to gather information about participants' evolutionary reasoning patterns: (1) the multiple-choice Conceptual Inventory of Natural Selection (CINS) test (Anderson et al. 2002), (2) the open-response ACORNS instrument (Nehm et al. 2012a), and (3) clinical oral interviews (modeled after Nehm and Schonfeld 2008). All of these tools have been used previously in the literature (Anderson et al. 2002; Bishop and Anderson 1990; Nehm and Schonfeld 2008; Nehm et al. 2012a, b). We discuss the details of each approach in turn.

The CINS Test

The CINS is a multiple-choice distracter-driven test that has been shown to generate valid inferences about overall levels of students' evolutionary knowledge (specifically natural selection and speciation) (Nehm and Schonfeld 2008). Nevertheless, despite being widely used, the test has been shown to display psychometric problems at finer grain sizes (Battisti et al. 2010). The CINS consists of twenty items that are scored as correct/incorrect; thus, the total score of the instrument ranges from 0 to 20. While the CINS only allows for a student to choose a "right" (normative scientific idea) or a "wrong" (naive idea) answer option, other tests (e.g., see ACORNS below) allow participants to express both types of reasoning in an explanation. This constraint makes comparisons of concept-level scores from the CINS and other MC tests challenging. For example, when asked about the scientific accuracy of an evolutionary idea, if a student chooses a naive idea answer (but not a key concept answer), does this response indicate that the student (1) does not know the key concept, (2) prefers the naive idea but knows the key concept, or (3) prefers the naive idea but does not know the key concept? Given this complication, for our study, we followed the intentions of the CINS authors when interpreting the meaning of student responses, that is, if a student did not choose the "correct" (scientific) answer, they are considered to *not* know the scientific idea. While this assumption may not always be correct, it is in line with the design of this test.

A second complication arose in attempting to compare CINS scores with the other instruments. In order to standardize our comparisons among assessment methods (that is, only compare concepts that were assessed in all of the different tools), we necessarily had to omit some of the CINS items that assessed concepts that are almost never used by students to explain evolution (e.g., "biotic potential" and "population stability") or are typically part of larger concepts (e.g., "causes of variation" is part of the "variation" concept) (Table 1). However, we used the majority of CINS items (shown as the item number from the original published instrument, in parentheses) that corresponded to the following key concepts across instruments: *variation* (6, 9, 16, 19), *heritability* (7, 17)*, competition* (5, 15)*, limited resources* (2, 14)*, and differential survival* (10, 18; see Table 1 for a summary).

The ACORNS Test

New findings about how students think about evolution, such as the impact of biotic familiarity, plants versus animals, and gain versus loss, have called for more sophisticated assessments of student thinking about evolution (Nehm and Ha 2011; Opfer et al. 2012). The newly developed constructed-response ACORNS (Assessment of Contextual Reasoning About Natural Selection; Nehm et al. 2012a) addresses these concerns and allows students

Table 1 Concepts scored in the clinical interviews, the ACORNS explanation tasks, and the MC CINS assessment

Concepts scored in explanations (see Opfer et al. for details and rubrics)	Interview	ACORNS (human and computer)	CINS
Core concept			
Variation	Yes	Yes	Yes
Heredity	Yes	Yes	Yes
Differential survival	Yes	Yes	Yes
Key concept			
Competition	Yes	Yes	Yes
Limited resources	Yes	Yes	Yes
Non-adaptive reasoning	Yes	Yes	No
Naive idea			
Needs/goals	Yes	Yes	No
Use/disuse	Yes	Yes	No
Adapt/acclimation	Yes	Yes	No

Note that because in the CINS students can only choose a correct concept OR a naive idea, it is only possible to determine whether they know the correct concept. If students chose a naive idea, that does not mean that they do not know correct concept, only that they prefer the naive idea in the context of the particular item. See Nehm and Schonfeld (2008), Nehm et al. (2010), and Opfer et al. (2012) for details of many student responses and exact human-scoring protocols

to simultaneously incorporate a variety of NIs and KCs into their written explanations of evolutionary change across a variety of surface features known to impact reasoning processes (e.g., taxon/trait, loss/gain; Nehm and Ha 2011; Opfer et al. 2012).

The ACORNS has been shown to generate valid and reliable inferences about student reasoning about natural selection in different contexts (e.g., trait gain and loss in plants and animals; Beggrow and Nehm 2012; Nehm et al. 2012a; Opfer et al. 2012). The ACORNS prompts students to generate written explanations of evolutionary change. Though different philosophical perspectives on explanation exist in the literature (see Braaten and Windschitl 2011), our items ask students to demonstrate their understanding of the evolutionary process by describing how the *explanandum*—or phenomenon to be explained (in our case, the gain or loss of a trait)—came to be (Osborne and Patterson 2011). Thus, our task aims to elicit a theoretical explanation, unlike much of the other work that asks students to provide evidence to back up their claims (e.g., Berland and McNeill 2012; Sandoval and Millwood 2005; Songer et al. 2009). The prompts in the aforementioned studies often provide varying types of evidence (embedded within the item itself). In contrast, our items do not provide any evidence for students to include in their explanations. Our study also uses explanations as a tool for understanding student *reasoning about content*, not understanding student *reasoning using evidence* (e.g., Songer et al. 2009; Songer and Gotwals 2012), that is, we are not focused on student learning of the practice itself (as are Berland and McNeill 2012 and Songer and Gotwals 2012).

In our study, the explanation prompts consisted of four isomorphic items, carefully standardized by taxon and trait familiarity: How would biologists explain how a living X species with/without Y evolved from an ancestral X species that had/lacked Y? (X = Mouse/lily/snail/grape; Y = claws/petals/teeth/tendrils). The ACORNS responses were scored using the published rubrics of Nehm et al. (2010) for six KCs and three NIs of natural selection (KCs: *variation, heritability, competition, limited resources, differential survival, and non-adaptive ideas*; NIs: *needs/goals, use/disuse, and adapt/acclimation*). KC scores for each item ranged from 0 to 6 (0, 1, 2, 3, 4, 5, 6), and NI scores for each item ranged from 0 to 3 (0, 1, 2, 3). Participant responses were scored to consensus by two human raters (a PhD student in biology education and an evolutionary biologist) who demonstrated strong inter-rater agreement ($\kappa > 0.80$).

Computer-Assisted Scoring of the ACORNS

An automated computer-scoring model (ACSM) also scored students' written explanations produced in response to the ACORNS items. The ACSM used in our study was developed using machine-learning methodologies. Machine learning is a field of computer science that investigates how computers can be used to detect, analyze, and build models of patterns in data that humans have categorized and apply these models to new, uncategorized datasets (for details, see Abu-Mostafa 2012; Nehm et al. 2012b). In a general sense, the machine-learning software "learns" from human-scoring patterns and attempts to build models that are able to predict these patterns. In our study, we used the software package LightSIDE (see Mayfield and Rosé 2013), which is a free, open-source machine-learning software package, to build nine ACSMs for nine different evolutionary ideas (six scientific and three naive). The ACSMs were built using corpora of human-scored responses independent of the dataset used in our current study, that is, the response corpora used to build the ACSMs were different from the response corpora used to score the student responses in our current study.

Two aspects of machine learning are important to distinguish. The machine-learning software first extracts all of the features from students' explanations (e.g., single words or two-word combinations) and builds an optimal regression equation that predicts human graders' decision patterns (i.e., whether the concept was present or absent in a particular response). This step is informally known as *training* (Abu-Mostafa 2012). When the computer builds an optimal equation, it is important to check the validity of the equation (that is, does it really work?), which is accomplished by comparing the model performance on a subset of the human-scored dataset to a novel corpus of scored responses. This second step is informally known as *testing*. If good results emerge from both training and testing, then the ACSM can be used to predict scores for a novel response corpus. But in order to know whether the model performed well on the new corpus, humans also need to independently score this set of responses. In our study, we report training values as well as testing values for the ACSMs for nine different concepts.

Many different types of machine-learning parameters can be controlled using LightSide. These parameters include N-Gram selection (i.e., the number of contiguous feature sequences, such as unigrams or bigrams—such as "genetic," "change" for unigram, "genetic_change" for bigram), stemming (i.e., grouping words based on common stems—such as "adapt" for adapting and adaptation), and removing stop words (i.e., very common words such as "a," "an," "so," or "and"). Based on prior research (Ha and Nehm 2012; Ha et al. 2011; Nehm et al. 2012b), for each ACSM, we used different settings for different concepts: unigram for *variation*, *competition*, *limited sources*, *non-adaptive,* and *use/disuse concepts*, and bigram for *heredity*, *differential survival*, *needs/goals*, and *adapt/acclimation*. In LightSIDE, for feature extraction, we selected "stemming," "removing stop words," "line length," "treat above features as binary," "contains non-stop words," and threshold set at the value of 5 (see Mayfield and Rosé 2012, 2013, for details on the LightSIDE program). In addition, the scoring models were trained using tenfold validation. In LightSIDE, we used sequential minimal optimization (SMO) for training support vector (see Platt 1999 for details). Although it takes several minutes to prepare the datasets and select the particular settings for each concept, it typically takes fewer than 5 s for LightSIDE to build and apply the ACSMs to our corpus.

Our ACSMs for the nine natural selection concepts were built using a corpus of 6,232 written explanations. These explanations were written by (1) non-majors enrolled in an introductory biology class (274 students/1,096 written explanations), (2) STEM-related majors enrolled in an introductory biology (403 students/1,612 written explanations), (3) biology-related majors enrolled in an introductory biology class focusing on evolution (565 students/2,260 written explanations), (4) biology majors enrolled in an advanced evolution class (123 students/492 written explanations), and (5) experts in evolutionary biology (graduate students and professors) (193 experts/772 written explanations). The corpus of written explanations was produced in response to various ACORNS instrument taxon/trait combinations, such as: (1) snail-poison, (2) prosimian-tarsi, (3) elm-winged seed, (4) labiatae-pulegone, (5) penguin-flightless, and (6) rose-thornless. It is important to emphasize that the items for the student explanations that we used to *train* or build the ACSMs were different from the items in the current study.

Clinical Oral Interviews and Scoring

The clinical oral interview protocol was modified from the protocol of Nehm and Schonfeld (2008) and consisted of two ACORNS items, which were identical to those on the written instrument, and two novel items (to minimize any potential testing effects). All items were isomorphic and used taxa and traits that were standardized by familiarity (see Nehm et al. 2012a). The novel items were as follows: (a) "How would biologists explain how a living opossum species without a tail evolved from an ancestral opossum species that had a tail?" and (b) "How would biologists explain how a living cactus species with spines evolved from an ancestral cactus species that lacked spines?" During the interviews, participants were asked follow-up questions to the items that included prompts such as "Can you explain what you mean when you use the word X?" and "Can you tell me more about X?" These prompts provided participants with an opportunity to clarify their responses and the interviewer with an opportunity to determine whether participants understood the words and terms they were using and whether they were using them in an accurate manner. Interviews were scored 0/1 for the absence/presence of the six KCs and three NIs (KCs: *variation, heritability, competition, limited resources, differential survival, and non-adaptive ideas*; NIs: *need/goal, use/disuse, and adapt/acclimation*) using the published rubrics of Nehm et al. (2010). KC scores for each item ranged from 0 to 6 (0, 1, 2, 3, 4, 5, 6), and NI scores for each item ranged from 0 to 3 (0, 1, 2, 3). Transcripts are provided (see *Clinical Oral Interviews* below) to illustrate the types of responses given by participants and the methods used to score them. An evolutionary biologist and a biology education PhD student scored all interviews, and any scoring discrepancies were subsequently resolved via deliberation. Consensus scores were used in all subsequent analyses. All human scoring of written responses and clinical oral interviews met a cutoff value of inter-rater reliability of $\kappa > 0.80$ (see Ha et al. 2011 for details).

Comparing Student Ideas Across Assessments

Our first analysis calculated the percentages of participants who used each KC and NI for each item in all of the datasets (i.e., human-scored ACORNS, computer-scored ACORNS, clinical oral interviews, and CINS). Average frequencies of each KC and NI used across the items for each assessment were also calculated. This analysis provides a straightforward summary of the outputs of the different assessment methods.

Testing Rasch Model Fit

To answer our first research question, for each of the four datasets (i.e., human-scored ACORNS, computer-scored ACORNS, interviews, and CINS), we used Rasch analysis to calculate (1) the means of person and item MNSQ and ZSTD values, (2) person and item separation, and (3) person and item reliability in order to evaluate how well the data fit the Rasch model. Since the CINS assessment does not have items specifically designed for measuring particular naive ideas independent of those for capturing key concepts, two separate Rasch analyses were conducted. First, we compared the KC person measures of (1) the MC CINS, (2) the computer-scored ACORNS, (3) the human-scored ACORNS, and (4) oral interviews. Second, we compared the KC *and* NI person measures of (1) the computer-scored ACORNS, (2) human-scored ACORNS, and (3) oral interviews. Rasch analyses were conducted using the WINSTEPS program (Linacre 2006).

Quantifying Alignment with Clinical Oral Interviews Using Rasch Measures

In order to establish and compare agreement levels between measures derived from the different instruments and the "gold standard" oral interview, we performed a series of analyses. First, we performed regression and correlation analyses of the person measures generated by Rasch analysis to explore the extent to which the assessments predicted clinical oral interview measures. Pearson correlations between the Rasch measures (e.g., both KC and KC&NI) for the interview, the human-scored and computer-scored ACORNS, and the CINS were performed. In the next analysis, we performed a series of linear regressions of how well the Rasch measures for the different assessments predicted the Rasch measures of the interview. Pearson correlations and regression analyses were conducted using SPSS version 19.0.

Quantifying Alignment with Clinical Oral Interviews Using Ordered Participant Pairs

In addition to using regression to analyze the alignment of the different assessments with clinical interview scores, we also developed a more precise comparison method using ordered participant pairs. The same participant ability measures derived from Rasch analysis (KC and KC&NI) were used to examine how each assessment (i.e., the interview, the human-scored and computer-scored ACORNS, and the CINS) ordered all possible pairs of participants ($5{,}356 = 104 \times 103/2$). Thus, for each assessment, pairs of students were ordered based on the person measures derived from Rasch analysis. For example, for the interview measures, participant "A" could have been ordered above, below, or equal to participant "B."

To analyze alignment with the interview, the ordered participant pairs of each assessment (human-scored and computer-scored ACORNS, and the CINS) were compared to the ordered participant pairs of the interview. For example, consider the pair of participants, participant "A" and participant "B," for the computer-scored ACORNS and for the interview. The computer-scored ACORNS and the interview may agree on how to order this pair of participants (e.g., participant "A" is ordered above, below, or equal to participant "B" for both assessments), in which case the assessment measures would be in *full concordance* (Table 2). If one assessment orders the pair of participants as equals, but the other assessment orders participant "A" as above or below participant "B," then the assessment measures would be in *half discordance* (Table 2). Finally, should the computer-scored ACORNS order participant "A" above participant "B," and the interview order participant "A" below participant "B" (or vice versa), the assessment measures would be in *full discordance* (Table 2). For each student, we computed the average discordance level across pairings with the other 103 students for the computer-scored ACORNS measures versus the interview measures. This analysis was then repeated for the human-scored ACORNS measures versus the interview measures and for the CINS measures versus the interview measures. Overall, this analysis provides a much more precise comparison of instrument alignment than is possible with the regression analysis.

Quantifying Overall Evolutionary Reasoning Models

The previous analyses compared the performance of different assessment methods using tallies of concept scores. In order to gain better insight into participants' conceptual frameworks and answer our final research question, we categorized the responses from the ACORNS (human-scored and computer-scored) and the interviews into one of four model types (note that the CINS could not be analyzed in this way). These reasoning models provide a holistic snapshot—holistic in the sense of accounting for both KC and NI use—of the general approach that participants are using to tackle evolutionary problems (See Nehm et al. 2009, for the use of this approach with science teachers). A *scientific model* is defined by the exclusive use of normative scientific concepts (i.e., KCs) to explain evolutionary change (no naive ideas are used); a *mixed model* describes explanations that are composed of combinations of naive

Table 2 Possible outcomes for comparison of pair of participants A and B

Comparison of participants "A" and "B"	Interview measure of ability		
	A ordered above B	A and B equal	A ordered below B
Computer measure of ability			
A ordered above B	Full concordance	Half discordance	Full discordance
A and B equal	Half discordance	Full concordance	Half discordance
A ordered below B	Full discordance	Half discordance	Full concordance

ideas and key concepts; a *naive model* is defined by the exclusive use of naive, non-normative ideas (and no key concepts); finally, *no model* was used to describe cases in which participants did not use any of the nine concepts that we scored and simply repeated parts of the question or answered the prompt using irrelevant text.

Categorizing participants by reasoning models was performed using two different methods. The first method identified participants' model type for each ACORNS *item* and then a sample-level percentage was calculated for each model type per item (e.g., for the *snail* item, a percentage of no model, naive, mixed, and scientific models was calculated). The second method identified an overall reasoning model for each *participant* by averaging each individual participant's performance across all four items (e.g., participant "A" had a scientific model and participant "B" had no model). Then, a sample-level percentage was generated for each model. The purpose of quantifying these reasoning models was to gain insight into how different assessment methods reflected broader reasoning models and levels of expertise (Nehm and Ridgway 2011), and how closely particular test results aligned with clinical interview results.

Results

Prior to discussing the comparisons among assessment types, we first describe the results from each of the assessments.

Clinical Oral Interviews

The 104 interviews provided rich insights into participants' explanations of the causes of evolutionary change. The interviews lasted 13.78 min on average (minimum 6.13 min; maximum 29.4 min). Overall, a greater percentage of participants used KCs compared to NIs, and non-adaptive ideas were the least frequent concepts employed by participants (Table 3).

In addition to providing information about the composition of evolutionary explanations, the clinical interviews also provided an opportunity for participants to engage in dialogue with the interviewer, provide clarification about how they reasoned, and to reveal the intended meaning of the words that they chose to express their ideas (Black 1999). Such clarification is particularly important when explanations are being scored for the presence/absence of ideas and for evolutionary explanations in which the issue of lexical ambiguity commonly arises (Rector et al. 2013). Here, we provide examples when such clarification episodes meant the difference between being scored for the presence versus the absence of a naive idea. For example:

Participant C	…so, uh, the tail eventually became eliminated from the gene pool because it wasn't, uh, required, so not having a tail became an adaptation from a mutation.
Interviewer	And when you say adaptation, what do you mean by that?
Participant C	Um. An adaptation is something that started as a mutation or variation that was, uh, beneficial enough to the species, that, uh, eventually became prevalent, prevalent [sic].

In this example, when Participant C provided an initial explanation and used the term *adaptation*, it was not entirely clear whether she understood the meaning of this idea. However, upon responding to the follow-up question, it became apparent that the participant was using adaptation in a scientifically normative manner. This is not always the case, however, as participants will often use scientific terms without understanding the scientific meaning of the term that they use (Beggrow and Nehm 2012; Rector et al. 2013). For example,

Interviewer	…how would biologists explain how a living cactus species with spines evolved from an ancestral cactus species that lacked spines?
Participant E	I guess that could be similar with either, like a genetic mutation or maybe a genetic drift and, uh, just, could also have to do with, uh, like being an anti-predatory defense to, uh, protect it since it's in a harsh environment already, they kinda [sic]

Table 3 Percentage of participants using individual key concepts (KCs) and naive ideas (NIs) in each item across assessments

Concepts	Interview		Human		Computer		CINS	
	Item	%	Item	%	Item	%	Item	%
Variation (KC)	Snail	89.4	Snail	53.8	Snail	50.0	6	67.3
	Lily	83.7	Lily	37.5	Lily	38.5	9	71.2
	Opossum	86.5	Mouse	45.2	Mouse	44.2	16	96.1
	Cactus	88.5	Grape	47.1	Grape	43.3	19	79.4
	Average	87.0	Average	45.9	Average	44.0	Average	78.5
Heritability (KC)	Snail	59.6	Snail	19.4	Snail	18.3	7	79.8
	Lily	58.7	Lily	11.5	Lily	9.6	17	54.5
	Opossum	51.9	Mouse	15.4	Mouse	14.4		
	Cactus	53.8	Grape	14.4	Grape	15.4		
	Average	56.0	Average	15.2	Average	14.4	Average	67.1
Competition (KC)	Snail	24.0	Snail	2.9	Snail	2.9	5	90.4
	Lily	12.5	Lily	1.9	Lily	1.9	15	83.5
	Opossum	11.5	Mouse	1.0	Mouse	1.0		
	Cactus	14.4	Grape	1.0	Grape	1.0		
	Average	15.6	Average	1.7	Average	1.7	Average	86.9
Limited resources (KC)	Snail	76.9	Snail	25.2	Snail	24.0	2	95.2
	Lily	62.5	Lily	16.3	Lily	14.4	14	81.6
	Opossum	46.2	Mouse	8.7	Mouse	8.7		
	Cactus	87.5	Grape	9.6	Grape	8.7		
	Average	68.3	Average	15.0	Average	13.9	Average	88.4
Differential survival (KC)	Snail	69.2	Snail	54.8	Snail	52.9	10	81.7
	Lily	71.2	Lily	51.0	Lily	44.2	18	82.5
	Opossum	73.1	Mouse	51.9	Mouse	43.3		
	Cactus	75.0	Grape	46.2	Grape	40.4		
	Average	72.1	Average	51.0	Average	45.2	Average	82.1
Non-adaptive idea (KC)	Snail	1.9	Snail	1.9	Snail	1.9		
	Lily	1.9	Lily	1.9	Lily	1.9		
	Opossum	3.8	Mouse	4.8	Mouse	3.8		
	Cactus	1.0	Grape	1.9	Grape	1.9		
	Average	2.2	Average	2.6	Average	2.4		
Need/goal (NI)	Snail	33.7	Snail	19.2	Snail	19.2		
	Lily	26.9	Lily	18.3	Lily	24.0		
	Opossum	32.7	Mouse	29.8	Mouse	28.8		
	Cactus	15.4	Grape	15.4	Grape	14.4		
	Average	27.2	Average	20.7	Average	21.6		
Use/disuse (NI)	Snail	5.8	Snail	3.8	Snail	6.7		
	Lily	12.5	Lily	3.8	Lily	3.8		
	Opossum	12.5	Mouse	6.7	Mouse	8.7		
	Cactus	4.8	Grape	0.0	Grape	1.0		
	Average	8.9	Average	3.6	Average	5.0		
Adapt/acclimation (NI)	Snail	12.5	Snail	7.7	Snail	14.4		
	Lily	10.6	Lily	9.6	Lily	13.5		
	Opossum	18.3	Mouse	8.7	Mouse	11.5		
	Cactus	8.7	Grape	6.7	Grape	14.4		
	Average	12.5	Average	8.2	Average	13.5		

have to guard themselves from anything that's going to get water out of them.

Interviewer You said that it could be a genetic mutation or genetic drift, can you explain what you mean by genetic drift?

Participant E I always forget the definition of genetic drift. Um, it's kinda [sic] just like, uh, a swing towards one extreme instead of where it was before, but I guess, so I guess that kind of takes away from what I was going towards, I guess, kinda [sic] contradicted myself.

Interviewer What's swinging? What is that is swinging towards one extreme?

Participant E Uh, just kinda [sic] like, the genetic makeup or the, uh, actual structures that are present are kinda [sic] more of a shift from one to another based on the pressures that they're getting from the environment or from, from other species around them.

Interviewer When you say structures, do you mean, um, what do you mean by that?

Participant E Um, sorta [sic] like the actual spines, in this case, or uh, the the [sic] cells within, that are used to maintain the water and the moisture and hold it.

Interviewer And how, how does that process of swinging towards one of the extremes or another come about?

Participant E Um, I guess that would be more, type of a, kind of, more of natural selection, where they're, where if they don't change then they're just going to die out, so without, if the, if they didn't develop the spines, there'd be no way for them to completely protect themselves, they would, without spines they'd probably evolved in another way and came up with some sort of poison or something else that would have deterred predators and stuff from eating them.

Here, Participant E begins by using the key concept of *genetic drift*, which would be scored as a non-adaptive idea in the written ACORNS. However, upon further questioning, the participant demonstrates a lack of understanding of what genetic drift actually means and instead provides an adaptive explanation for this *non-adaptive* mechanism. This example illustrates how follow-up questions provide justification for using interviews as a "gold standard;" they are crucial for elucidating participants' knowledge and understanding.

Scores from different assessments were often in strong alignment (see "Methods," beginning p. 7, for details on how each type of score was derived). For example, Participant A had a Rasch person measure of 2.85 for the twelve CINS items (raw score of 18 out of 20 on the full CINS assessment), had a KC&NI Rasch person measure of 1.89 on the computer-scored ACORNS, and 1.8 for the human-scored ACORNS (average raw NI score of 0 and raw KC score of 3.25). Furthermore, Participant A was able to provide scientifically normative explanations for each interview item (including both adaptive and non-adaptive concepts) and had Rasch KC&NI person measures of 2.46 (average raw interview score of 4.5 KCs):

Interviewer …how would biologists explain how a living cactus species with spines evolved from an ancestral cactus species that lacked spines?

Participant A Well most biologists would suggest that you had your ancestral cactus species that did not have any spines and within that population of cacti you had the variation, whether that variation came from mutation, or introduced genes from another species, you know, like a weird cross-cactus pollination event or something, but gene flow in that way. And once the variation got into the population it somehow, the presence of spines became more prominent and then eventually fixed within the species. And the ways, the mechanisms that it would probably go from being a minority to a majority would either be genetic drift, where it's kinda [sic] random, happenstance, you know, maybe all the cactus got wiped out by a flood, except for the guy, you know, the four cactus over there that had spines so this new group of cactus all have spines because, uh, the parent stock all had spines, or maybe it was just kinda [sic] random and the gene went from being prominent to rare and back and forth until it reached fixture. Or you could have natural selection where you've got, you know, maybe a new predator came into the area and was feeding on the cactus, but it wasn't feeding on the spiny cactus because they don't like getting poked in the face. And when all of the non-spiny cactus got eaten, the only ones left were spiny ones and when they bred offspring, you know, some of them were probably spiny and some of them weren't but only the spiny ones survived and then eventually all of the cactus have spines. And that's more of a natural selection based explanation.

Interviewer Alright.

Participant A Because there was a reproductive advantage to have spines… you'd have a gradual switch of cacti.

In this particular explanation, Participant A includes four KCs and no NIs. Participant A's explanation demonstrates a clear understanding of evolution by both natural selection and genetic drift.

In contrast, Participant B performed poorly on the CINS and had a Rasch person measure score of 0.02 (raw score of 8/20), a KC&NI person measure of -3.03 for the computer-scored ACORNS, and -2.72 for the human-scored ACORNS (average raw scores: 1 NI and 0 KCs for each ACORNS item). Additionally, Participant B had a KC&NI person measure of -2.24 (raw interview scores demonstrated an average of 0.5 KCs and 1.5 NIs per item):

Participant B Um, (laughter), um basically changing over time, over a long long [sic] period of time.

Interviewer And what, exactly, is changing?

Participant B Hm (laughter), um, I guess just adapting, changing to its environment.

Interviewer You also use the term adapt, adapting, um, what do you mean when you use that term?

Participant B Um basically that it would see the cha- or see the environment around itself and, uh, you know the trait will, either stay or, change whether it's being used or not in that environment.

Interviewer So the trait will stay or change depending on, um, whether it's being used—

Participant B —surroundings, and yeah, if it's being used.

Interviewer Ok. (pause) And, how would the biologist explain that the trait came to-, or uh arrived in the first place?

Participant B Uh, basically that the need for a tail, uh, became necessary due to whatever environment it was living in.

The incorporation of both naive ideas and key concepts into an explanation, as Participant B did above, was common in our sample. While a handful of participants did not use KCs in one or two interview items, all participants managed to use a KC in at least half of their explanations

ACORNS Written Explanations

Overall, participants took an average of 9.88 min to complete the ACORNS assessment (minimum 1.67 min; maximum 36.98 min). A scientific reasoning model is exemplified by Participant G's response for the item, "How would a biologist explain how a living species of snail with teeth evolved from an ancestral species of snail that lacked teeth?"

Participant G In an ancestral population of snails lacking teeth the snails would have variety of heritable traits. In that initial population of toothless snails, some snails might have mutations that give them teeth-like structures in their mouths. Those snails in the population with teeth-like structures would somehow be favored by their environment, they might have access to a wider variety of food. Those snails would exhibit greater reproductive success among their population and the genes for teeth-like structures would become more frequent generation after generation. Some snails would have more developed teeth-like structures and be favored among the population for reproduction. After many generations of pressures favoring teeth the snail population would be full of toothed snails.

In contrast, a naive reasoning model is demonstrated by Participant H's response to the item involving a species of mouse. This response includes all three naive ideas but lacks key concepts:

Participant H That over time the mouse did not need the claws (they became useless to them) and they became [sic] adapted without them. The claws became a hinderance [sic] and cost the animal to keep the claws they did not use.

Human-Scored ACORNS

The highest score consisted of 0 NIs and 13 KCs for all four items ($n = 2$). The lowest score consisted of 4 NIs and 0 KCs for all four items ($n = 1$). When all of the human-scored ACORNS explanations were pooled together, there was a minimum of 0 NIs, a maximum of 8 NIs, a minimum of 0 KCs, and a maximum of 13 KCs. The average total amount of NIs used by participants was 1.30 (SD = 1.67). The average total amount of KCs used by participants was 5.25 (SD = 3.13).

Computer-Scored ACORNS

Scores were similar between the computer-scored explanations and the human-scored explanations. The highest computer-generated score consisted of 0 NIs and 13 KCs across all four items ($n = 2$). The lowest score consisted of 4 NIs and 0 KCs across all four items ($n = 2$). The computer-scored ACORNS had the same minimum and maximum of NIs and KCs used (0, 8; 0, 13, respectively). The average number of NIs used by participants was 1.61 (SD = 1.84), and the average number of KCs was 4.87 (SD = 3.18).

CINS Multiple-Choice Assessment

The CINS takes less than 30 min for participants to complete (Anderson et al. 2002; however, specific response times were not reported for their study). Overall, participants did well on the CINS, although a broad range of scores was obtained; the average score was 15.37 (out of 20 total). The highest score was 20 ($n = 11$) and the lowest score was 4 ($n = 1$).

Assessment Comparisons

There was some variation in participants' use of scientific and naive ideas across assessments (Table 3). A greater percentage of participants used scientific ideas in the interviews compared to written explanations and for some concepts, the differences between assessment scores were considerable (Table 3). For example, an average of 87 % of participants used *variation* in their clinical interview explanations, whereas less than 50 % of participants used *variation* in their ACORNS explanations. Slightly fewer than 80 % of participants selected *variation* in their CINS responses, which is more similar to the clinical interviews than to the ACORNS written explanations.

The percentage of participants selecting concepts on the CINS did not always align with the interview, however. In the clinical oral interviews, 15.6 % of participants used *competition,* compared to <2 % on the written ACORNS and 86.9 % of participants on the CINS. While the average percentage of participants using concepts differed between interview items and ACORNS items, the patterns of usage were similar: *variation* was used the most, followed by *differential survival*, *heritability,* and *limited resources*—which were roughly the same—and *competition* was the least common. The CINS results also differed from the other two assessments (ordered from most to least frequent: *Limited resources*, *competition, differential survival, variation,* and *heritability*), though the differences between concept frequencies within the CINS were relatively smaller (see Table 3 for details).

Differences in the percentages of participants using naive ideas in the interviews and ACORNS were not as large as those of participants demonstrating scientific ideas. For example, 27.2 % of participants used the NIs of *needs/goals* in the clinical interviews, whereas approximately 20 % of participants used *needs/goals* on the written ACORNS. Compared to the very large gaps observed between the concept frequencies of *variation* and *heritability* among items, the percentage gaps between assessments for participants using NIs were relatively smaller. The patterns for NIs were also similar for the interview and ACORNS explanations, with *needs/goals* being the most common, followed by *adapt/acclimation* and *use/disuse* (Table 3).

We used Rasch analysis to examine the degree to which our data fit a Rasch model and to appropriately compare score patterns among the instruments by transforming raw scores to person and item measures (which to our knowledge has not been done in previous instrument comparison work). Table 4 illustrates the person and item MNSQ and ZSTD values, separation values, and reliability statistics for each of the assessment methods. Outfit MNSQ revealed that the CINS did not meet this quality control benchmark (values should be <1.3; Bond and Fox 2001). Although item reliabilities for all four assessments met the benchmark (>0.9), the person reliability for the clinical interviews was the highest (0.70), followed by the human-scored ACORNS written assessment (0.61), computer-scored ACORNS written assessment (0.60), and finally, the CINS (0.41).

We explored how closely (1) the human-scored KC measures from the ACORNS, (2) the computer-scored KC measures from the ACORNS, and (3) CINS measures were associated with the KC measures from the clinical interviews (Fig. 1). The results showed that approximately 30.4 % of variance in the clinical interview measures was explained by human-scored KC measures from the ACORNS and 31.5 % of the variance in the clinical interviews was explained by the computer-scored KC measures from the ACORNS, while only 12.0 % of the variance in the clinical interview KC measures was explained by the CINS measures, that is, the correlation between the human-scored ACORNS KC measures and clinical interview KC measures was 0.551 and the correlation between the computer-scored ACORNS KC measures and clinical interview KC measures was 0.561, while the correlation between the CINS measures and clinical interview KC measures was 0.346. The correlation between the human-scored KC measures and the computer-scored KC measures from the ACORNS was 0.964.

We also examined how successfully the human-scored KC&NI measures and the computer-scored KC&NI measures from the ACORNS explained the variance seen in the KC&NI measures from the clinical interviews (Fig. 2). The correlation between the human-scored ACORNS KC&NI measures and the clinical interview measures was 0.611, and the correlation between the computer-scored ACORNS

Table 4 Item fit statistics derived from the Rasch analysis

	Outfit		Separation	Reliability
	MNSQ	ZSTD		
KC				
Interview				
Item	0.97	0.00	5.47	0.97
Person	0.97	0.00	1.53	0.70
Human				
Item	1.05	0.10	3.78	0.93
Person	1.04	0.20	1.24	0.61
Computer				
Item	1.06	0.10	3.62	0.93
Person	1.03	0.20	1.23	0.60
CINS				
Item	***1.36***	0.50	3.02	0.90
Person	***1.32***	0.10	0.83	0.41
KC&NI				
Interview				
Item	0.94	0.00	5.70	0.97
Person	0.94	0.00	1.72	0.75
Human				
Item	1.26	0.10	5.84	0.97
Person	1.07	0.10	1.63	0.73
Computer				
Item	1.25	0.20	5.68	0.97
Person	1.07	0.10	1.66	0.73

MNSQ above acceptable range bolded

KC&NI measures and clinical interview measures was 0.633. The correlation between the human-scored KC&NI measures and computer-scored KC&NI measures was 0.960. Our analyses also found that approximately 37.3 % of the variance in the clinical interview KC&NI measures was explained by the human-scored KC&NI measures, and 40.0 % of the variance in the clinical interview KC&NI measures was explained by the computer-scored KC&NI measures. The results illustrate that computer-scored written explanations have remarkably strong correspondence with oral interview measures, are capable of capturing participants' KCs and naive ideas as accurately as human-scored measures, and do not strongly align with the MC CINS assessment.

The ordered participant pair comparisons provided the most precise approach for examining our overarching research question: How closely did the different instruments align with the "gold standard" clinical oral interviews? In the ordered participant pairs analysis, the KC measures of the computer-scored written explanations had the highest concordance with the clinical interview (61.0 %; Table 5). The CINS assessment had the lowest concordance with the interview (52.9 %; Table 5), meaning that there were more pairs of participants ranked differently between the CINS and interview compared to the computer-scored and the human-scored written explanations (see Table 2 for possible outcomes). For the KC&NI measures, the computer-scored written explanations (65.7 %) had slightly greater concordance with the interviews than did the human-scored written explanations (64.3 %; Table 5).

Ordered participant pair comparisons were also plotted against one another to illustrate differences between the assessments in terms of their discordance with interview measures (Figs. 3, 4). Ordered participant pairs that were generated using KC person measures showed that the computer-scored ACORNS was in agreement with the interviews more often (71 participants) than was the CINS (30 participants) and three participants had equal agreement (see Fig. 3a). The human-scored ACORNS KC person measures were more often in agreement with the interview measures (69 participants) than were the CINS KC measures with interview measures (35 participants) (Fig. 3b). The computer-scored ACORNS KC measures had greater agreement with the interview measures (66 participants) than did the human-scored ACORNS KC measures with the interview measures (33 participants), and the measures had equal agreement for five participants (see Fig. 3c). For the KC&NI measures, the computer-scored ACORNS measures were more often in agreement with the interview measures (56 participants) compared to the human-scored ACORNS measures with the interview measures (42 participants) (six participants showed equal agreement; see Fig. 4). These results are in general alignment with the regression results (discussed above), although they more precisely illustrate the greater concordance of computer-scored written explanation measures with clinical oral interview measures for KC measures and KC&NI measures.

Evolutionary Reasoning Models Across Items

In addition to examining the components of participants' explanations, we also classified each participant into one of four different reasoning models (e.g., *scientific* model, *mixed* model, *naive* model, or *no* model) based on their interview and ACORNS responses (note that the CINS could not be analyzed in this way). This classification was carried out using two different approaches. In the first approach, participants were classified into reasoning patterns based on their explanations for each of the four items across the different assessments. Using this method, the average percentage of explanations demonstrating a *scientific* model in the clinical interviews and the written

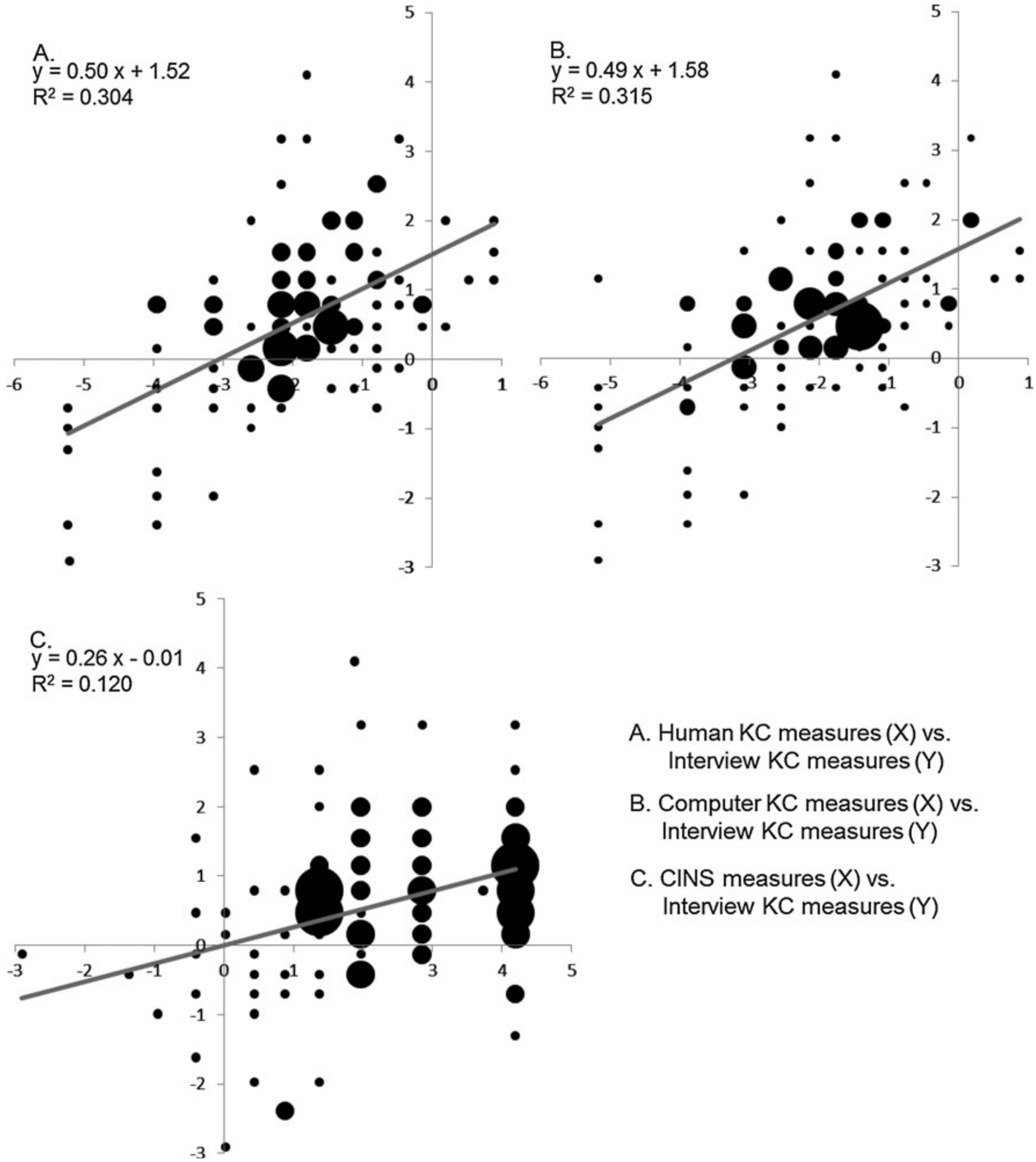

Fig. 1 The linear regressions of the person measures for KCs for human-scored and computer-scored ACORNS and the CINS. All *y*-axes reflect the interview KC person measures. R^2 represents the effect size

ACORNS was similar (interview: 63.2 %, human-scored ACORNS: 58.9 %, and computer-scored ACORNS: 51.9 %; Fig. 5), whereas the percentage of *mixed* models (interview: 34.6 %, human-scored ACORNS: 16.3 %, and computer-scored ACORNS: 18.0 %), *naive* models (interview: 1.9 %, human-scored ACORNS: 8.9 %, and computer-scored ACORNS: 13.2 %), and *no* model (interview: 0.2 %, human-scored ACORNS: 15.9 %, and computer-scored ACORNS: 16.8 %) shown in both interviews and ACORNS were different.

Evolutionary Reasoning Models Across Participants

Using this approach, participants' reasoning patterns were averaged across all of the items in order to provide an overall reasoning pattern for each *participant*. This second method yielded somewhat different results compared to those discussed above (the across-item analysis). The percentage of participants using a *scientific* model across assessment methods was relatively lower (interview: 32.7 %, human-scored ACORNS: 46.2 %, and computer-

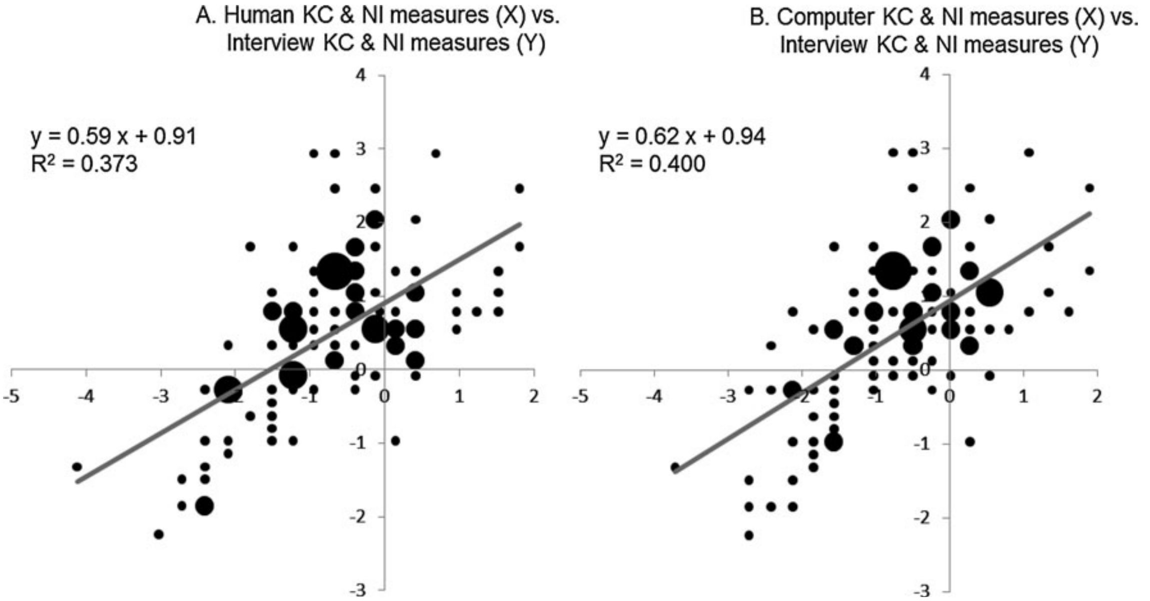

Fig. 2 The linear regressions of human and computer KC&NI person measures. All *y*-axes reflect the interview KC&NI person measures. R^2 represents the effect size

Table 5 Discordance levels for three instruments versus interviews for ordered participant pairs based on KC measures and on KC&NI measures

	Full discordance (%)	Half discordance (%)	Full concordance (%)	Mean ± SD ($n = 104$) (%)
KC				
Human-scored ACORNS versus interview	24.6	16.0	59.5	32.5 ± 15.4
Computer-scored ACORNS versus interview	23.7	15.3	61.0	31.3 ± 15.0
CINS versus interview	26.6	20.6	52.9	36.8 ± 15.5
KC&NI				
Human-scored ACORNS versus interview	23.5	12.2	64.3	29.6 ± 15.9
Computer-scored ACORNS versus interview	22.2	12.1	65.7	28.2 ± 14.3

scored ACORNS: 39.4 %), while the percentage of participants exhibiting *mixed* models was higher (interview: 67.3 %, human-scored ACORNS: 49.0 %, and computer-scored ACORNS: 53.8 %; Fig. 6). Using the second method of analyzing models, both the *naive* model and the *no* model categories were quite low, with no participants exhibiting *naive* models or *no* models in the interviews, and fewer than 6 % exhibiting *naive* or *no* models in the human-scored and computer-scored ACORNS (Fig. 6). Depending upon whether participants were categorized by individual-item analyses or by across-item analyses, differences in the percentages of reasoning patterns were apparent. Overall, however, the *across participants* method (method 2) displayed the greatest degree of similarity to the clinical oral interviews.

Discussion

General Discussion

The new *Framework for Science Education* (National Research Council 2012) emphasizes the centrality of scientific practices—such as explanation, argumentation, and communication—in science teaching, learning, and assessment. A major challenge facing the field of science education is developing assessment tools that are capable of validly and efficiently evaluating these practices and the ideas expressed within them. While many studies have compared the similarity of human-scored and computer-scored writing products (reviewed in Magliano and Graesser 2012), remarkably few studies have compared

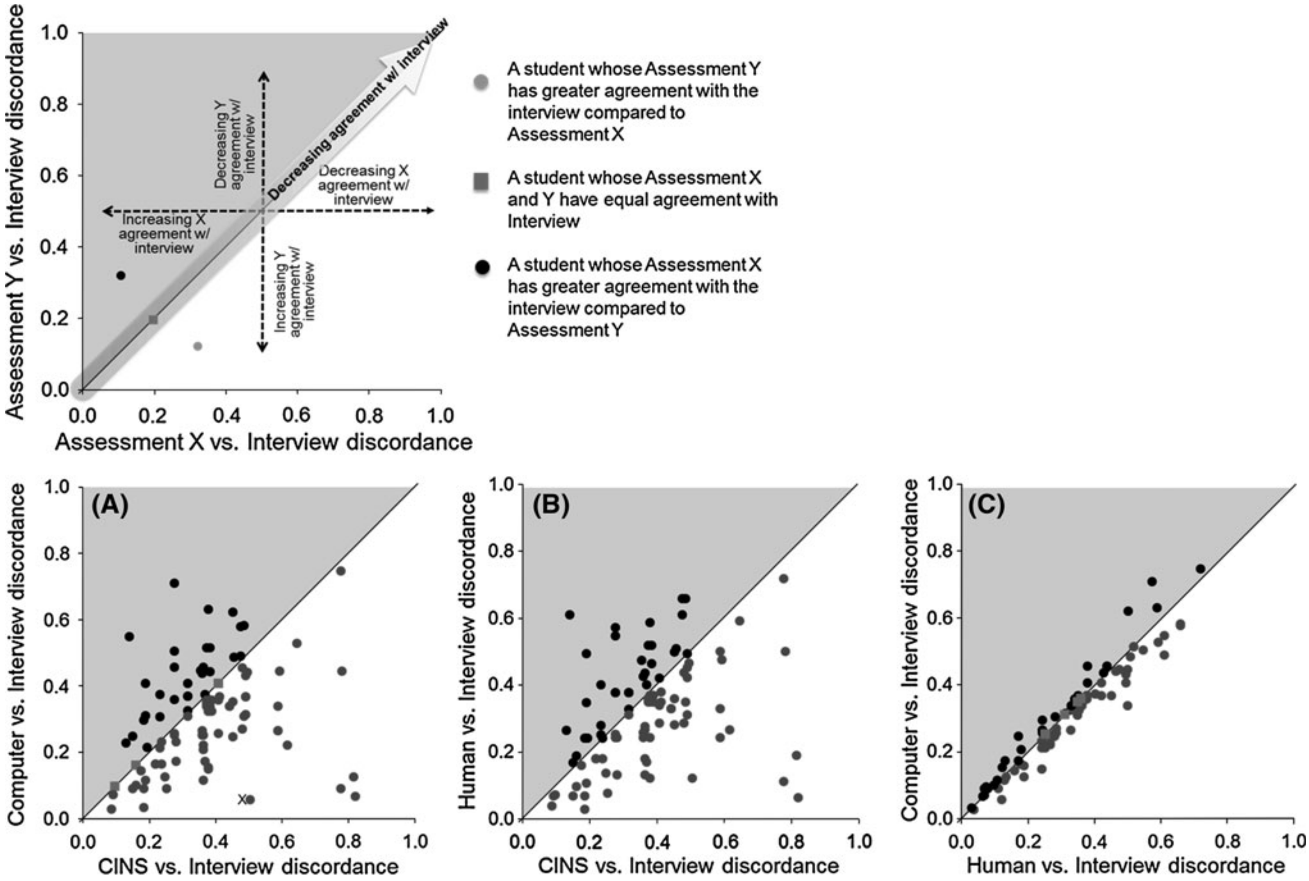

Fig. 3 A comparison of assessments using ordered participant pairs' KC measures. Each point in the plot shows the **a** computer-scored ACORNS versus interview and CINS versus interview, **b** human-scored ACORNS versus interview and CINS versus interview, and **c** computer-scored ACORNS versus interview and human-scored ACORNS versus interview discrepancies evaluating a participant's ability (in terms of KC measures) in comparison with other participants in the class. For example, consider the participant represented by the point marked with the "*X*" (Fig. 3a). The interview ordered that participant's ability differently than the computer-scored ACORNS in comparison with only 6 % of the other participants in the class but differently than the CINS score for about 50 % of the other participants. Participants, such as "*X,*" for whom the computer-scored ACORNS measure of ability was more often in agreement with the interview compared to the CINS measure, are shown as light gray (*points falling below the diagonal*) and account for 71 of the 104 members of the class. The computer measure of ability was less often in agreement with the interview compared to the CINS measure for 30 students (*black points falling above the diagonal*), while 3 participants (*dark gray squares*) show the computer-scored ACORNS measure and CINS having equal average agreement with the interview. For **b,** there are 35 *black dots* (human-scored ACORNS less often in agreement with interview), 69 *gray dots* (CINS less often in agreement with interview), and 0 *gray squares* (equal agreement). For **c,** there are 33 *black dots* (computer-scored ACORNS less often in agreement with interview), 66 *gray dots* (human-scored ACORNS less often in agreement with interview), and 5 *gray squares* (equal agreement)

computer-scored scientific explanations to the educational "gold standard" of clinical oral interviews. Indeed, that was the central question guiding our study: How well do computer-scored written explanation task scores align with clinical oral interview scores?

Written explanations are time-consuming for students to take and for instructors to grade. Education researchers have turned to technology in the hope that computer scoring is able to bridge the gap between the efficiency of MC tests and the richness of written explanations. While recent research with machine-learning software has shown that automated computer-scoring models (ACSM) can generate scores of evolutionary explanations (and several other topics) comparable to those generated by human experts (Ha et al. 2011; Nehm et al. 2012b), the question remains as to how closely computer-generated explanation scores approximate the gold standard of clinical oral interview scores, that is, it is entirely possible that while human-scored and computer-scored written assessments are in strong alignment, neither approach might approximate the communication and explanation skills exemplified by a clinical oral interview and emphasized in the new *Framework for Science Education* (National Research Council 2012).

Another significant limitation of prior empirical work comparing assessment formats is the near-exclusive use of measures derived from classical test theory (CTT) despite

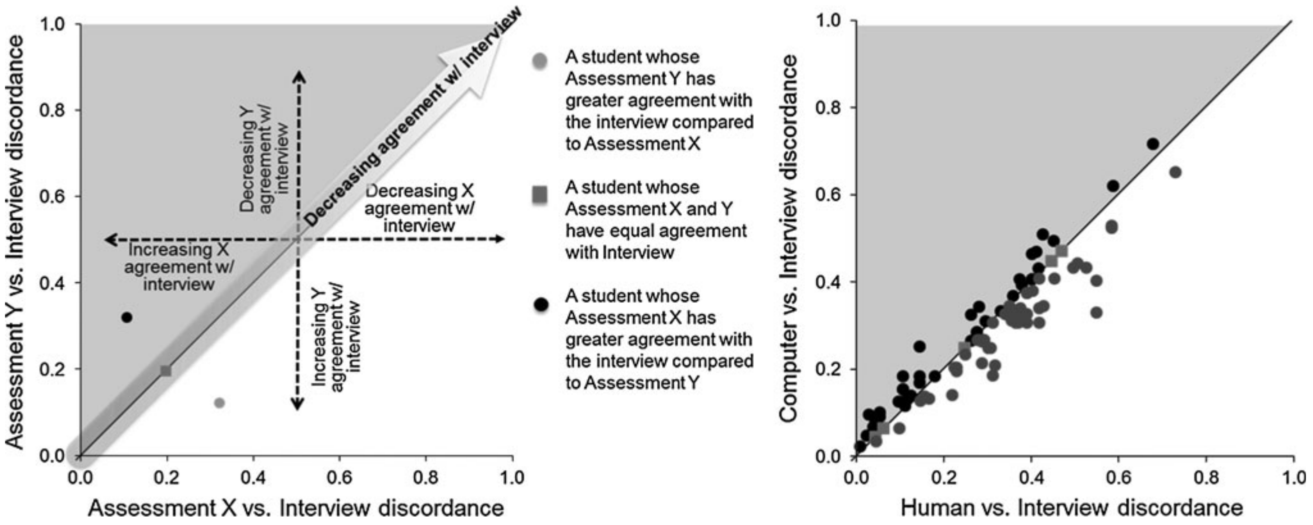

Fig. 4 A comparison of assessments using ordered participant pairs' KC&NI measures. Each point in the plot shows the computer-scored ACORNS versus interview and human-scored ACORNS versus interview discrepancies evaluating a participant's ability (in terms of KC&NI measures) in comparison with other participants in the class. Participants for whom the computer-scored ACORNS measure of ability was more often in agreement with the interview compared to the human-scored ACORNS measure are shown as light gray (*points falling below the diagonal*) and account for 56 of the 104 members of the class. The computer-scored ACORNS measure of ability was less often in agreement with the interview than the CINS measure for 42 students (*black points falling above the diagonal*), while 6 participants (*dark gray squares*) show the computer-scored ACORNS measure and human-scored ACORNS measure as having equal average agreement with the interview

the well-known advantages of using alternative approaches, such as IRT and Rasch (Boone and Scantlebury 2006). Rasch analysis ensures equivalent scaling and generates mean of fit indices that allow for rigorous assessment of reliability and validity. Rasch provides the means for evaluating not just the quality of an entire instrument, but participant performance as well. For these reasons, our study attempted to more rigorously evaluate the correspondence of different assessment methods using Rasch analyses.

The comparative framework for our study is also noteworthy. Many comparisons of human-scored and computer-scored text use only one comparative frame (e.g., agreement patterns in one type of score). Our study, in contrast, makes comparisons across several frameworks—within-item individual normative scientific key concepts and naive idea measures, across-item reasoning element measures, and holistic mental model portraits that are characteristic of different forms of reasoning (e.g., mixed models and scientific models). Comparing performance patterns across assessment formats using different comparative frames helps to more rigorously establish the generalizability of ACSM efficacy.

A final aspect of our study differs from prior work. While many studies, like our own, report general levels of raw agreement between methods, correlations between assessment scores, and Kappa statistics quantifying agreement (all using CTT-based assumptions), our Rasch-based comparisons were also conducted using a new participant pair ordering method that more precisely compares the performance of individual participants across assessment types. Collectively, our empirical approach attempts to rigorously examine the efficacy of ACSMs for scoring evolutionary understanding, which is recognized as one of four core content areas in the life sciences (National Research Council 2012).

Assessment Correspondence Patterns

The item fit statistics derived from Rasch analysis demonstrated that (1) the computer-derived scores from the written explanations and (2) the human-derived scores from the oral and written explanations conformed to the Rasch model but the CINS scores did not. Both item and person fit statistics for the CINS did not meet the Rasch model expectations, meaning that the ability levels of the students and difficulty levels of the items were poorly matched. More than expected by the model, low-ability students can thus perform well on high-difficulty items while high-ability students can perform poorly on low-difficulty items (Bond and Fox 2001). Thus, despite being a widely used test, in our sample of students, the CINS did not appear to accurately assess students' evolutionary knowledge (corroborating work by Battisti et al. 2010). Overall, the measures derived from the clinical oral interview best fit the Rasch model, which lends further support

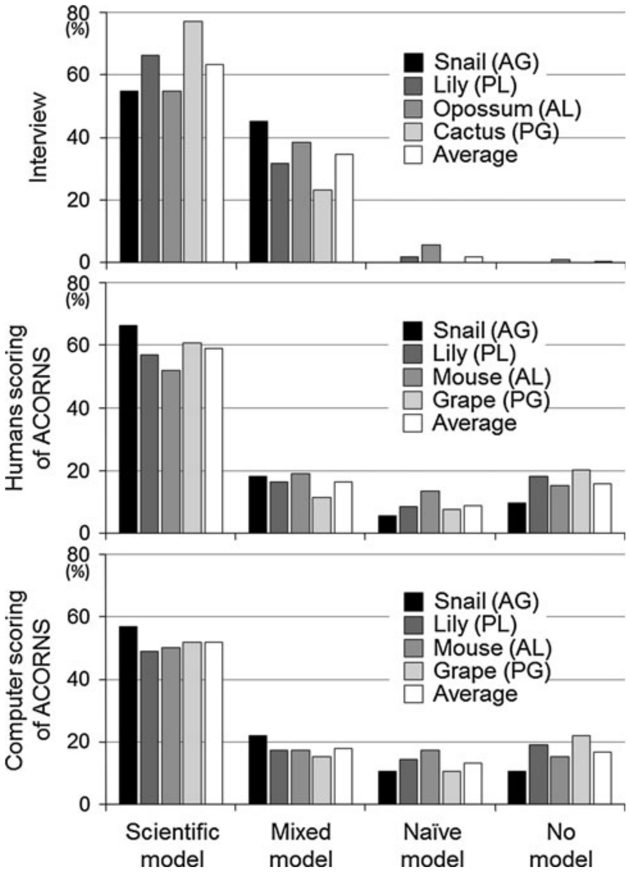

Fig. 5 Evolutionary reasoning models across items. The percentage of reasoning patterns captured by three different metrics and categorized by averaging participants' scores for each item and getting an average score/item for the sample. *Y*-axis represents percentage of participants. *Snail* and *lily* item are common across all three metrics. Contextual features are designated next to the item taxa; thus, *A* represents an animal taxa, *P* represents a plant taxa, *G* represents the gain of a trait, and *L* represents the loss of a trait (e.g., AG = an item in which an animal gains a trait)

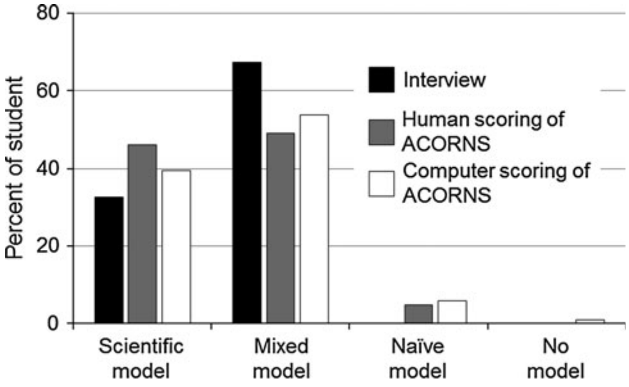

Fig. 6 Evolutionary reasoning models across participants. The percentage of reasoning patterns captured by three different metrics and categorized by averaging each individual participant's performance across items and getting an average model for each participant. *Y*-axis represents percentage of students

to its status as a "gold standard;" computer-generated measures also aligned very well with the best-fit standard set by the interviews.

The computer-scored measures were strongly aligned with interview measures. Pearson correlations demonstrated that computer-generated KC measures were more strongly associated with interview measures than were the CINS measures. While the computer-generated KC measures predicted the interview KC measures at a level comparable to human-generated measures, the CINS prediction levels were considerably lower, making it a poorer indicator of student knowledge. Therefore, although the CINS may be able to measure aspects of student knowledge (i.e., KCs), it may not be able to accurately measure the full extent of that knowledge (such as KCs *and* NIs).

While at first glance the 0.551 and 0.561 correlation coefficients between the KC measures of clinical interviews and the human-scored and computer-scored ACORNS assessments (respectively) are only moderate (0.50–0.60), it is important to keep in mind that the assessments themselves were different in terms of test format (interview vs. writing), that is, the correlation coefficients were not used to measure parallel forms equivalency. Moreover, judging the "highness" and "lowness" of correlation coefficients using absolute numerical values can be misleading; instead, the correlation coefficients are useful for comparing between two competing values. Our discussion is focused on the comparisons of interview measures and written assessment (ACORNS) measures on the one hand, and of the correlation between interview measures and multiple-choice assessment (CINS) measures on the other; our goal was not to examine the absolute magnitudes of the correlation coefficients. Indeed, it is not unexpected to find robust (but moderate) correlation magnitudes on tests of the same domain but focusing on different tasks (e.g., oral vs. written vs. selected response). The important point is that the coefficients of 0.551 (human-scored KC measures and interview KC measures) and 0.561 (computer-scored KC measures and interview KC measures) are relatively high compared to the CINS KC measures and interview KC measures coefficient of 0.346. In sum, we found that computer scoring not only successfully predicts interview KC measures, but that it also successfully predicts the KC&NI measures of the interviews as well as human scoring.

Our findings demonstrating the efficacy of computer scoring were also corroborated using the ordered-pair analyses. Specifically, Rasch-based, ordered-pair comparisons of students for both KC and KC&NI measures demonstrated that computer-generated measures had greater concordance with interview measures compared to the CINS measures or the human measures. In other words, the order of participant pairs that computer-scored measures generated had strong agreement with the order the interview measures generated more often than the CINS measures or the human-scored measures. This indicates that the computer scoring of written explanations will provide more consistent results with interview evaluations than the alternatives included in this study (i.e., human scoring of written explanations and the MC CINS test).

Overall, we found that computer-scored written explanation measures have stronger correspondence with clinical oral interview measures than did the multiple-choice CINS. Computer scoring is also comparable to human scoring of written explanations, and computer-scored written responses can exceed the performance of the most commonly used MC test for measuring student thinking about evolution. Computer scoring also overcomes the limitations of MC science tests in general. First, assessments of scientific practices, like the explanation-focused ACORNS, allow students to respond to prompts using a variety of reasoning models, consisting of many different assemblages of cognitive elements (e.g., mixed models and naive models). Second, the ACORNS takes into account recent findings from cognitive science, specifically examining student reasoning across contextual surface features (e.g., taxon and trait and loss or gain of trait), whereas the CINS items only addresses one reasoning context (familiar animals) (Opfer et al. 2012). Third, the ACORNS emphasizes recall of information, which is known to be a more valid indicator of understanding in comparison with MC recognition tests (Opfer et al. 2012). Considering that computer scoring of written explanations negates the time and cost constraints associated with constructed-response formats, and that the scoring models and software are free, the advantages of using the computer-scored ACORNS are clear. Overall, the open-source LightSIDE program may offer science educators in other domains (chemistry, earth science) a solution to the challenge of developing valid and efficient assessments of scientific practices and core ideas (National Research Council 2012).

Patterns of Student Thinking About Evolution Across Assessments

Despite differences in the fidelity of different methods to the clinical interviews, the majority of student explanations were categorized as either scientific models or mixed models (The CINS prevents the clear identification of mixed models). Thus, regardless of whether the students were questioned orally or in writing, their conceptual models were primarily built using purely normative ideas or different magnitudes and types of normative and non-normative ideas in combination. Importantly, purely naive models were rarely found in our 104 participants. These findings are in strong alignment with previous work on undergraduate evolutionary reasoning and call into question the cognitive validity of either–or MC evolution tests (Beggrow and Nehm 2012; Nehm and Ha 2011; Nehm and Schonfeld 2008).

Advantages of Oral Interviews Over Other Assessment Methods

Clinical oral interviews have become a "gold standard" in education research for assessing student knowledge and reasoning because of the many advantages that they offer (e.g., clarification and follow-up questioning) compared to other assessment formats such as MC and written explanations. Although MC and written formats typically prohibit scorers from determining how students interpreted the language used in the items, or what students meant by the language that they used to answer the item, interviews provide opportunities for resolving reasoning patterns and lexical ambiguity. By engaging in conversation with students, interviews also allow students to use unique combinations and arrangements of ideas that cannot be documented in most extant MC assessments. Moreover, interviews permit the evaluator to go beyond the composition of student thinking and detect the structure of that thinking.

As no other study to our knowledge has ever included such a large sample of clinical oral interviews on evolution, our sample of 104 participants provided a unique opportunity to paint a detailed portrait of undergraduate students' thinking about evolutionary change. Overall, we found that students tended to provide similar explanations in their interviews as they did in their ACORNS explanations, although interviews tended to generate much more information. In other words, follow-up questioning, along with the extra time associated with such questioning (e.g., oral interview items lasted about 3.45 min each while ACORNS items lasted only 2.44 min each), appears to have produced richer (though not necessarily more accurate) explanations in general, which translated into higher frequencies of key concepts and naive ideas (relative to written responses). Unsurprisingly, more time spent on an assessment task is associated with the documentation of more ideas (see Beggrow and Nehm 2012).

Analyses of the clinical interviews revealed that the majority of students in our sample had a very difficult time providing explanations of evolutionary change. Specifically, most students failed to provide a complete explanation in response to the oral interview prompts; indeed, several follow-up questions were often needed to elicit an explanation from the students. Nonetheless, interviews were informative and revealed that while the majority of students use normative key concepts of evolutionary change (e.g., *variation, heritability, and differential survival*), most students also included non-normative naive ideas. Furthermore, although many students used scientific "buzz" words in their explanations, in some cases follow-up questioning revealed that the students lacked an accurate understanding of those terms. This is a challenge inherent to scoring written explanations.

Our large sample of interviews with students demonstrated that many participants struggled to communicate their ideas to the interviewer (see *Clinical Oral Interviews* above). Participants displayed very limited science communication skills, which has an impact on assessment and research of student understanding. If students are not capable of communicating their ideas, then instructors and researchers' will be unable to generate a clear picture of the students' reasoning models. The limited scientific communication skills evident in our sample (see interview transcripts above) highlight the need for increased opportunities for students to practice these abilities in the classroom. Such communicative skills are in line with the NRC recommendations (National Research Council 1996, 2001b, 2012). Communicative skills play a role in "making students' thinking visible to…themselves" (National Research Council 2001b, p. 4), comprise a crucial component of scientific literacy and necessitate that students have the opportunity to practice expressing their thinking in different ways (orally, visually in graphs, diagrams, tables, etc.) (American Association for the Advancement of Science 2011; National Research Council 2001a, b, 2012). Unfortunately, commonly used MC assessments—used for formative or summative purposes—fail to help students hone their scientific communication skills (Nehm and Haertig 2012; National Research Council 2012). While interviews provide important opportunities for revealing this core aspect of science literacy, they are impractical to implement in the classroom. Written explanations provide one alternative for helping students communicate their ideas. Given that computer scoring of written explanations is a robust approach (at least in the context of evolution), it is a very promising teaching and learning tool for science education.

Alternative Approaches for Categorizing Student Evolutionary Reasoning Models

One of the purposes of assessment is to better understand students' conceptual frameworks so that instruction can be designed accordingly (National Research Council 2001a, b). In our study, we documented participants' reasoning patterns using two different methods—by averaging across items for each participant or by averaging across participants for each item. While the *across participants* method demonstrated greater alignment between the human-scored and computer-scored ACORNS and the interviews, each method of categorization offers instructors and researchers different views of students' reasoning patterns. For instance, if a researcher is interested in how students reason in particular *contexts* (e.g., trait loss in familiar animals), then averaging student scores for particular items would be most appropriate. Accordingly, instructors could adjust their lessons to focus on what they found (e.g., mixed reasoning models about trait loss in familiar animals). However, if a researcher is interested in how a particular *student* reasons about evolution *in general* (content mastery), then documenting student reasoning across contexts/items would be more appropriate. It is likely that the most effective methodological strategy will depend on the research or instructional questions being asked. Regardless of which approach was used, computer scoring provided a robust proxy for student reasoning models.

Study Limitations

One weakness of this study was the lack of exact alignment of the CINS items with those of the ACORNS and clinical oral interviews. Considering that the latter two assessments were identical in format, it is perhaps not surprising that their outputs were more strongly associated than was the CINS. The ideal (but impractical) comparison study would have been to build and validate an entirely new multiple-choice test that more closely aligned with the ACORNS and interviews. However, developing such an instrument would have taken several years and been very costly. Given this constraint, we employed the most widely used evolution instrument (the CINS) as a comparative benchmark. This study limitation prevents us from conclusively demonstrating that our results will apply in all MC formats, or that interviews will always be more strongly associated with computer-derived scores than with MC tests. In short, our results may not generalize to other samples, instruments, or content domains. Nevertheless, they do provide robust support for the similarity of clinical interviews and computer-scored explanations of evolutionary change, which to our knowledge has not been investigated previously.

Our study also specifically examined a subset of nine concepts. Currently, only three robust computer-scoring models for naive ideas and six for key concepts have been developed. Previous research has demonstrated that other naive ideas that we have not studied also occur in student explanations (e.g., Beggrow and Nehm 2012; Nehm et al. 2012a). Therefore, as we build additional computer-scoring models for additional naive ideas, the portrait of student reasoning that our study painted may change. More advanced machine-learning methods could also improve upon the identification of student ideas, and build much more elaborate computational models of how students reason about evolutionary change. The greater the diversity of concepts that are detected using machine scoring models, the closer they are likely to approximate clinical oral interviews.

Overall, our work indicates that machine-learning methods are one important solution for validly and efficiently evaluating core ideas (such as evolution) embedded in scientific practices (National Research Council 2012). However, further work is needed to begin assessing students' competencies in scientific practices themselves (e.g., Berland and McNeill 2012; Gobert et al. 2012; Songer et al. 2009; Songer and Gotwals 2012).

Acknowledgments We thank John Harder, Ian Hamilton, the Ohio State University's Center for Life Science Education and Department of Anthropology's Graduate Teaching Associate program for assistance with data collection, the National Science Foundation REESE program (DRL 0909999) for funding portions of this study, and Meghan Rector for helpful reviews of the manuscript. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the National Science Foundation.

References

Abu-Mostafa YS (2012) Machines that think for themselves. Sci Am 307(1):78–81

American Association for the Advancement of Science (2011) Vision and change in undergraduate biology education. AAAS, Washington

Anderson DL, Fisher KM, Norman GJ (2002) Development and evaluation of the conceptual inventory of natural selection. J Res Sci Teach 39(10):952–978

Battisti BT, Hanegan N, Sudweeks R, Cates R (2010) Using item response theory to conduct a distracter analysis on conceptual inventory of natural selection. Int J Sci Math Educ 8:845–868

Beggrow EP, Nehm RH (2012) Students' mental models of evolutionary causation: natural selection and genetic drift. Evolut Educ Outreach 5(3):429–444

Berland LK, McNeill KL (2012) For whom is argument and explanation a necessary distinction? A response to Osborne and Patterson. Sci Educ 96(5):808–813

Bishop BA, Anderson CW (1990) Student conceptions of natural selection and its role in evolution. J Res Sci Teach 27(5): 415–427

Black TR (1999) Doing quantitative research in the social sciences. Sage Publications, London

Bond TG, Fox CM (2001) Applying the Rasch model: fundamental measurement in the human sciences. Lawrence Erlbaum Associates, Mahwah

Boone WJ, Scantlebury K (2006) The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. Sci Educ 90(2):253–269

Braaten M, Windschitl M (2011) Working toward a stronger conceptualization of scientific explanation for science education. Sci Educ 95(4):639–669

Briggs DC, Alonzo AC, Schwab C, Wilson M (2006) Diagnostic assessment with ordered multiple-choice items. Educ Assess 11(1):33–63

Chi MTH, Bassok M, Lewis MW, Reimann P, Glaser R (1989) Self-explanations: how students study and use examples in learning to solve problems. Cogn Sci 13:145–182

Deadman JA, Kelly PJ (1978) What do secondary school boys understand about evolution and heredity before they are taught the topics? J Biol Educ 12(1):7–15

Ginsburg H (1981) The clinical interview in psychological research on mathematical thinking: aims, rationales, techniques. Learn Math 1(3):4–11

Gobert JD, Sao Pedro MA, Baker RSJD, Toto E, Montalvo O (2012) Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. J Educ Data Min 4(1):153–185

Graesser AC, McNamara DS (2012) Automated analysis of essays and open-ended verbal responses. In: Cooper H, Panter AT (eds) APA handbook of research methods in psychology. American Psychological Association, Washington

Ha M, Nehm RH (2012) Using machine-learning methods to detect key concepts and misconceptions of evolution in students' written explanations. Paper in proceedings of the National Association for Research in Science Teaching, Indianapolis, Indiana

Ha M, Nehm RH, Urban-Lurain M, Merrill JE (2011) Applying computerized scoring models of written biological explanations across courses and colleges: prospects and limitations. CBE Life Sci Educ 10:379–393

Haudek KC, Kaplan JJ, Knight J, Long T, Merrill J, Munn A, Nehm RH, Smith M, Urban-Lurain M (2011) Harnessing technology to improve formative assessment of student conceptions in STEM: forging a national network. CBE Life Sci Educ 10(2):149–155

Joughin G (1998) Dimensions of oral assessment. Assess Eval High Educ 23(4):367–378

Leacock C, Chodorow M (2003) C-rater: automated scoring of short-answer questions. Comput Humanit 37(4):389–405

Linacre JM (2006) WINSTEPS Rasch measurement software [Computer program]. WINSTEPS, Chicago

Lombrozo T (2006) The structure and function of explanations. Trends Cogn Sci 10:464–470

Lombrozo T (2012) Explanation and abductive inference. In: Holyoak KJ, Morrison RG (eds) Oxford handbook of thinking and reasoning. Oxford University Press, Oxford, pp 260–276

Magliano JP, Graesser AC (2012) Computer-based assessment of student-constructed responses. Behav Res Methods 44:608–621

Mayfield E, Rosé C (2012) LightSIDE: text mining and machine learning user's manual. Carnegie Mellon University, Pittsburgh

Mayfield E, Rosé C (2013) LightSIDE: open source machine learning for text. In: Shermis MD, Burstein J (eds) Handbook of automated essay evaluation. Routledge, New York

McNeill KL, Lizotte DJ, Krajcik J, Marx RW (2006) Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. J Learn Sci 15(2):153–191

Moscarella RA, Urban-Lurain M, Merritt B, Long T, Richmond G, Merrill J, Parker J, Patterson R, Wilson C (2008) Understanding undergraduate students' conceptions in science: using lexical analysis software to analyze students' constructed responses in biology. Proceedings of the National Association for Research in Science Teaching (NARST) annual conference, Baltimore, MD

National Research Council (1996) National science education standards. The National Academies Press, Washington, DC

National Research Council (2001a) Investigating the influence of standards: a framework for research in mathematics, science, and technology education. The National Academies Press, Washington, DC

National Research Council (2001b) Knowing what students know. The National Academies Press, Washington, DC

National Research Council (2007) Taking science to school. The National Academies Press, Washington, DC

National Research Council (2012) A framework for K-12 science education: practices, crosscutting concepts, and core ideas. The National Academies Press, Washington, DC

Nehm RH, Ha M (2011) Item feature effects in evolution assessment. J Res Sci Teach 48(3):237–256

Nehm RH, Haertig H (2012) Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. J Sci Educ Technol 21(1):56–73

Nehm RH, Reilly L (2007) Biology majors' knowledge and misconceptions of natural selection. Bioscience 57(3):263–272

Nehm RH, Ridgway J (2011) What do experts and novices "see" in evolutionary problems? Evol Educ Outreach 4(4):666–679

Nehm RH, Schonfeld IS (2008) Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. J Res Sci Teach 45(10):1131–1160

Nehm RH, Kim SY, Sheppard K (2009) Academic preparation in biology and advocacy for teaching evolution: biology versus non-biology teachers. Sci Educ 93(6):1122–1146

Nehm RH, Ha M, Rector M, Opfer JE, 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 RH, Beggrow EP, Opfer JE, Ha M (2012a) Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. Am Biol Teach 74(2):92–98

Nehm RH, Ha M, Mayfield E (2012b) Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. J Sci Educ Technol 21(1):183–196

Neumann I, Neumann K, Nehm R (2011) Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. Int J Sci Educ 33(10):1373–1405

Opfer JE, Nehm RH, Ha M (2012) Cognitive foundations for science assessment design: knowing what students know about evolution. J Res Sci Teach 49(6):744–777

Osborne JF, Patterson A (2011) Scientific argument and explanation: a necessary distinction? Sci Educ 95(4):627–638

Page EB (1966) The imminence of grading essays by computers. Phi Delta Kappan 47:238–243

Platt J (1999) Fast training of support vector machines using sequential minimal optimization. In: Schölkopf B, Burges CJC, Smola AJ (eds) Advances in Kernel methods—support vector learning. MIT Press, Cambridge, pp 185–208

Rector MA, Nehm RH, Pearl D (2012) Item sequencing effects on the measurement of students' biological knowledge. Paper in the proceeding of the National Association of Research in Science Teaching, Indianapolis, IN, 25–28 March 2012

Rector MA, Nehm RH, Pearl D (2013) Learning the language of evolution: lexical ambiguity and word meaning in student explanations. Res Sci Educ 43(3):1107–1133

Roediger HL III, Marsh EJ (2005) The positive and negative consequences of multiple-choice testing. J Exp Psychol Learn Mem Cogn 31(5):1155

Russ RS, Scherr RE, Hammer D, Mikeska J (2008) Recognizing mechanistic reasoning in student scientific inquiry: a framework for discourse analysis developed from philosophy of science. Sci Educ 92(3):499–525

Russ RS, Lee VR, Sherin BL (2012) Framing in cognitive clinical interviews about intuitive science knowledge: dynamic student understandings of the discourse interaction. Sci Educ 96(4):573–599

Sandoval WA, Millwood KA (2005) The quality of students' use of evidence in written scientific explanations. Cogn Instr 23(1):23–55

Seddon GM, Pedrosa MA (1988) A comparison of students' explanations derived from spoken and written methods of questioning and answering. Int J Sci Educ 10(3):337–342

Shermis MD, Burstein J (2003) Automated essay scoring: a cross-disciplinary perspective. Lawrence Erlbaum Associates, Inc., Mahwah

Songer NB, Gotwals AW (2012) Guiding explanation construction by children at the entry points of learning progressions. J Res Sci Teach 49(2):141–165

Songer NB, Kelcey B, Gotwals AW (2009) How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. J Res Sci Teach 46(6):610–631

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

Woloshyn V, Gallagher T (2009, December 23) Self-explanation. Retrieved from http://www.education.com/reference/article/self-explanation/

Yang Y, Buckendahl CW, Juszkiewicz PJ, Bhola DS (2002) A review of strategies for validating computer automated scoring. Appl Meas Educ 15(4):391–412