CBE—Life Sciences Education
Vol. 10, 379–393, Winter 2011

Article

Applying Computerized-Scoring Models of Written Biological Explanations across Courses and Colleges: Prospects and Limitations

Minsu Ha,* Ross H. Nehm,* Mark Urban-Lurain,† and John E. Merrill‡

*The Ohio State University, School of Teaching and Learning, Columbus, OH 43210; †Michigan State University, 1428 Engineering, East Lansing, MI 48824; ‡Michigan State University, 6171 Biomedical Physical Sciences, East Lansing, MI 48824

Submitted August 27, 2011; Revised September 29, 2011; Accepted September 29, 2011
Monitoring Editor: Vivian Siegel

Our study explored the prospects and limitations of using machine-learning software to score introductory biology students' written explanations of evolutionary change. We investigated three research questions: 1) Do scoring models built using student responses at one university function effectively at another university? 2) How many human-scored student responses are needed to build scoring models suitable for cross-institutional application? 3) What factors limit computer-scoring efficacy, and how can these factors be mitigated? To answer these questions, two biology experts scored a corpus of 2556 short-answer explanations (from biology majors and nonmajors) at two universities for the presence or absence of five key concepts of evolution. Human- and computer-generated scores were compared using kappa agreement statistics. We found that machine-learning software was capable in most cases of accurately evaluating the degree of scientific sophistication in undergraduate majors' and nonmajors' written explanations of evolutionary change. In cases in which the software did not perform at the benchmark of "near-perfect" agreement (kappa > 0.80), we located the causes of poor performance and identified a series of strategies for their mitigation. Machine-learning software holds promise as an assessment tool for use in undergraduate biology education, but like most assessment tools, it is also characterized by limitations.

INTRODUCTION

In large introductory biology classes throughout the United States, multiple-choice (MC) formats typify both formative assessments (e.g., clicker questions, concept inventories) and summative tests (e.g., midterm and final exams; see Wood [2004] and Smith *et al.* [2008]). While there is little doubt among educators that MC formats in general are capable of providing cost-effective, reliable, and valid inferences about student knowledge and misconceptions in many content areas, not all types of student learning outcomes may be measured using MC formats (reviewed in American Association for the Advancement of Science [AAAS, 2011] and Nehm *et al.* [in press]). Moreover, despite generating useful assessment information, MC tests may also produce unintended, and rarely considered, negative consequences for learners, such as the generation of false knowledge (Mandler and Rabinowitz, 1981; Roediger and Marsh, 2005; Butler *et al.*, 2006; Kang *et al.*, 2007). Additionally, many MC tests are most conducive to detecting novice *or* expert (incorrect or correct) models of student thinking, whereas a large body of work in cognitive science indicates that many students construct mixed models of naïve and informed scientific information as they learn (e.g., Vosniadou [2008]; Opfer *et al*. [2011]); right *or* wrong options—the staple of MC tests—may limit the valid measurement of student learning gains (Nehm and Schonfeld, 2008; Nehm and Ha, 2011; Neumann *et al.*, 2011).

Collectively, these and many other limitations of MC formats should motivate biology educators to 1) develop and deploy a more diverse array of high-quality assessment methods and 2) measure a more expansive range of student

DOI: 10.1187/cbe.11-08-0081
Address correspondence to: Minsu Ha (ha.101@osu.edu).

knowledge, skills, and learning outcomes (e.g., AAAS [2011, p. 17]; Nehm *et al.* [in press]). The purpose of our study was to investigate the prospects and limitations of implementing new assessment methods in introductory biology—specifically, computerized scoring of short-answer scientific explanations. Can successful application of these innovative methods at one university be generalized, and if not, why not? What are the implications for adopting computerized-scoring programs as assessment tools in introductory biology and biology education research?

BACKGROUND

Open-Response Assessments in Biology Education

Educational researchers emphasize that assessments should be built upon and aligned with what we know about student learning and cognition (National Research Council [NRC], 2001). One major advance in our understanding of student learning is that learners do not progress directly from novice to expert levels; rather, the pathways of knowledge growth in biology (and other domains) are highly inefficient and involve integrating scientific ideas into naïve knowledge frameworks, generating heterogeneous mental models, or building coexisting models (Nehm and Schonfeld, 2008, 2010; Vosniadou, 2008; Kelemen and Rosset, 2009; Evans and Lane, 2011). These so-called mixed or synthetic models may persist for long periods of cognitive development (Vosniadou, 2008) and even through the years of college instruction (Nehm and Reilly, 2007).

If our biology assessments are intended to measure progress in student reasoning and build upon findings from educational research, our assessment items (whether formative or summative, MC or open response) must permit at least three general reasoning categories as options: 1) exclusively naïve answer choices; 2) assemblages of mixed or synthetic answer choices; and 3) exclusively scientific answer choices. Currently, many MC diagnostic and summative biology assessments (and concept inventories) contain option types 1 and 3 (novice and expert, respectively), despite mounting evidence in some biology domains that *most* students harbor "mixed models" of biological concepts (Nehm and Ha, 2011). Thus, MC formats (correct *or* incorrect options) appear to be discordant with what we know about how students learn science (NRC, 2001; Nehm and Schonfeld, 2008; Vosniadou, 2008). Thus, one advantage of open-response assessments is that they allow students to assemble heterogeneous knowledge elements and thereby reveal student thinking at a much more fine-grained level than "novice versus expert" assessments.

Perhaps the strongest argument for the inclusion of written assessments in introductory biology is that MC formats are not capable of measuring many desired learning outcomes for introductory biology courses (AAAS, 2011); a diversified assessment portfolio is needed to comprehensively capture students' learning progress (Corcoran *et al.*, 2009). Indeed, increasing emphasis has been directed at building assessments that mirror authentic, real-world tasks, not just those that are easily measured (NRC, 2001; Nehm *et al.*, in press). Many science education policy documents, for example, emphasize the importance of having students generate and evaluate scientific explanations (e.g., *Benchmarks for Science Literacy* [AAAS, 1994]; the *National Science Education Standards* [NRC, 1996]; *Taking Science to School* [Duschl *et al.*, 2007]; and *Vision and Change in Undergraduate Biology Education* [AAAS, 2011]; Braaten and Windschitl [2011]). The ability to generate scientific explanations can only be assessed using open-response formats.

One final argument for the inclusion of open-response assessments in introductory biology is that they better align with most real-world experiences than do MC formats. Increasingly, college graduates are expected to perform nonroutine tasks that cannot be automated, digitized, or outsourced (Nehm *et al.*, in press). From an educational point of view, deploying assessment tasks that model authentic problem-solving environments would help reinforce for students which types of performances are most highly valued by biology educators, and which types of evaluations they are likely to experience postgraduation (e.g., production vs. selection tasks). Overall, while MC assessments should remain in biology educators' assessment toolboxes, the many advantages of open-response formats call for their greater inclusion. Practical limitations have prevented the wider use of open-response assessments, but recent technological advances are beginning to change this situation.

Computer-Assisted Scoring Tools

The increasing use of computer-assisted scoring (CAS) in many educational contexts has been motivated by the numerous constraints that characterize human scoring of constructed-response (e.g., short-answer, essay) items. Some of the most obvious limitations are the large amounts of time, money, and expertise needed to score such responses, and the consequent delayed feedback to test takers (Nehm *et al.*, in press). A more serious issue with human scoring of written responses is the persistent problem of grading subjectivity and consequent reliability threats; such problems are often introduced by the need for many different human graders to score large data sets, such as those generated in undergraduate biology courses (Nehm and Haertig, in press). Moreover, differently trained graders often disagree about the scores that should be given to a response, requiring additional training time to equalize scoring among raters. Reliable and consistent scoring of constructed-response items cannot be solved by having one human grader score all of the responses; grading fatigue and changes in scoring precision are well-known limitations of human scoring (Nehm and Haertig, in press). Thus, many long-standing problems have limited the use of open-response formats.

Fortunately, the rapid pace of developments in computer technology and text analysis software has made CAS tools more economical and accessible to educators. Consequently, many of the aforementioned limitations of human scoring have been investigated empirically using a variety of different software tools. This work has demonstrated that computer software can be "trained" to score constructed-response items as accurately and reliably as human raters (Page, 1966; Yang *et al.*, 2002; Shermis and Burstein, 2003). Indeed, the Educational Testing Service and many other large companies now employ CAS methods in large-scale, high-stakes, standardized exams (Powers *et al.*, 2002). Examples of these CAS tools include C-rater (Sukkarieh and Bolge, 2008), E-rater (Burstein, 2003; Williamson, 2009), and Intelligent Essay

Assessor (Landauer *et al.*, 2001). Less work has focused on using these or similar tools for formative assessment purposes. Our research group has used two tools to perform CAS at a smaller scale, specifically grading short-answer responses within classes at individual universities: SPSS Text Analytics (SPSSTA; Urban-Lurain *et al.*, 2010; Nehm and Haertig, in press) and the Summarization Integrated Development Environment (SIDE; Nehm *et al.*, in press). SPSSTA is a commercial software package sold by a private company (IBM), whereas SIDE is a freely available software package distributed by Carnegie Mellon University (Mayfield and Rosé, 2010a,b). Although the performance efficacy of both computer programs has been demonstrated using samples of undergraduate science students (Ha and Nehm, 2011), the two programs differ in the ways in which they approach CAS, as well as in the methods used to perform scoring (for a review of these differences, see Haudek *et al*. [2011] and Nehm *et al.* [in press]). SIDE is capable of creating scoring algorithms automatically when provided with a sufficiently large set of human-scored data for training and validation of scoring algorithms. Because we had a large sample of human-scored, short-answer responses for this study, our work on CAS was ideally suited to using SIDE.

SIDE combines a natural language processing (NLP) engine for parsing text, along with a set of machine-learning algorithms for classifying text (see Witten and Frank [2005] for more details). Analyzing text responses using SIDE has two main steps: 1) defining the filters necessary for capturing the structure of the text and 2) specifying the summaries to be displayed and extracting the needed subsets (for details, see Mayfield and Rosé [2010a,b] and the supplemental material). Operationally, SIDE uses corpora of students' constructed responses that have been scored by humans to detect text patterns associated with the presence or absence of particular scientific concepts as measured by expert raters. For instance, terms such as "mutation," "genetic change," or "change in DNA" are indicative of the presence of variation, which is one of the key concepts necessary for explaining evolutionary change (Nehm and Reilly, 2007). Student responses, and the associated expert scores, are used as input to SIDE. Since we focused on five key concepts in this study, we needed to input a set of human-scoring information on whether or not the student text included a particular concept or not for each of the five key concepts of evolution that we investigated (see *Human Scoring of Explanations of Evolutionary Change*).

SIDE provides a number of interactive tools for refining and improving the accuracy of the predictions by allowing the user to examine cases where the machine-learning model misscored (either incorrectly classifying a response as containing the concept when it does not, or failing to classify a response as containing a concept when it in fact does). SIDE can save the scoring model and apply it to new text to predict human scoring (Mayfield and Rosé, 2010a,b).

In this study, we used initial data sets, which had been scored by two biology experts, to train SIDE. We then used new, expert-scored data to validate the accuracy of the SIDE-scoring models. If this cross-validation approach is successful, this provides evidence that we can predict human scoring of new sets of student responses to the same questions with as much confidence as we would have using human raters.

Evolution: A Core Concept in Undergraduate Biology

Our investigations of CAS were focused on a core idea in biology: evolutionary change (AAAS, 2011). A large body of work, spanning more than 30 yr, has revealed a diverse array of learning difficulties with evolution in general, and natural selection in particular (reviewed in Nehm and Schonfeld [2007] and Gregory [2009]). Much less research has focused on psychometric issues relating to the measurement of student knowledge, although some work suggests that open-response formats and clinical interviews more validly capture student thinking about evolution than currently available MC assessments (Nehm and Schonfeld, 2008; Batistta *et al.*, 2010; Nehm *et al.*, in press). Thus, our investigation of the efficacy of CAS methods is well suited to the topic of evolution. Improvement in the measurement of students' thinking regarding evolutionary change may help to generate a deeper understanding of student learning difficulties and inform improved instructional practice.

RESEARCH QUESTIONS

Our study explores three research questions:

1. Are scoring models built using machine learning generalizable across colleges and courses (majors and nonmajors at different universities)? In other words, do scoring models built using student responses at one school function effectively at other schools?
2. How many human-scored student responses are needed to effectively build scoring models? To what degree does sample size impact computer-scoring success?
3. What factors limit computer-scoring efficacy, and how can these factors be mitigated to enable scoring models to be used in introductory biology courses across universities?

METHODS

Sample

To answer our research questions, we utilized three samples of undergraduate students enrolled in biology coursework (Table 1): 1) nonmajors enrolled in introductory biology at Ohio State University (OSU; 264 students/1056 written explanations); 2) nonmajors enrolled in introductory biology at Michigan State University (MSU; 146 students/584 written explanations); and 3) biology majors enrolled in introductory biology at MSU (440 students/1760 written explanations). Student responses were gathered using two online survey systems (ACS and LONCAPA; for details, see www.evolutionassessment.org and www.lon-capa.org, respectively).

We only included responses from individuals who completed four survey items each with responses of more than five words. The number of participants in the OSU nonmajor sample who completed all four items (see following section) was 358 (77.7% of total participants). Given the significant labor involved in scoring open-response items, we randomly selected a subset of 264 students (1056 responses) from this sample. We sampled the MSU data using the same approach The number of participants in the MSU nonmajors sample who completed all four items with more than five words was

Table 1. Sample information

Institution[a]	Major	Participants (n)[b]	Ethnicity (%)			Gender (%)		Age
			White	Minority	None mentioned	Male	Female	
OSU	Nonmajor	264	79.1	14.4	6.5	42	58	20.1
MSU	Nonmajor	146	66.4	13.7	19.9	40	60	19.4
MSU	Major	440	79.1	11.8	9.1	42	58	19.6

[a]OSU = Ohio State University; MSU = Michigan State University.
[b]Note that n refers to subsampled data sets (see *Sample*).

146 (90.7% of total participants). Finally, the number of participants in the MSU biology major sample who completed all four items with more than five words was 440 (90.0% of total participants). Because of the time, money, and expertise required to score student responses, we randomly selected a subset of responses ($n = 500$) from each sample (OSU nonmajors, MSU nonmajors, and MSU majors; 1500 responses total) for the first research question. For our second research question, we scored more than twice as many responses from the OSU corpus ($n = 1056$).

Items Used to Generate Explanations of Evolutionary Change

Our response corpus was composed of student explanations from four open-response items about evolutionary change. This assessment format has been employed in biology education research for more than 25 yr (e.g., Clough and Driver [1986]) and has been shown to generate reliable and valid inferences about students' reasoning regarding evolutionary change (Bishop and Anderson, 1990; Nehm and Schonfeld, 2008; Nehm and Ha, 2011). Instrument items were isomorphic ("How would biologists explain how a living X species with/without small/large Y evolved from an ancestral X species with/without large/small Y?") but differed in specific taxa and traits (i.e., X and Y).

Assessment item features, such as trait functionality and organism type, are known to influence students' reasoning regarding evolutionary change (Nehm and Ha, 2011; Opfer *et al.*, 2011). Consequently, we standardized our prompts to include only animal examples and functional traits (e.g., fins, wings). Moreover, because the *familiarity* of taxa and traits is also known to influence students' reasoning regarding evolutionary change (Opfer *et al.*, 2011), item taxa and traits were constrained by their overall familiarity. To do so, we used the frequencies of "organism + trait" in Google rankings as a proxy for familiarity (Nehm, Beggrow, *et al.*, in press). Specifically, taxon/trait combinations included: shrew incisors, snail feet, fish fins, and fly wings.

Although students' short-answer explanations of evolutionary change varied in length, the average number of words did not differ among items (analysis of variance [ANOVA], $F = 3.04$, $P > 0.01$). Specific item lengths were 1) shrew: mean = 45.5, SD = 30.3, minimum = 6, maximum = 430; 2) snail: mean = 42.9, SD = 27.9, minimum = 6, maximum = 429; 3) fish: mean = 42.5, SD = 26.3, minimum = 6, maximum = 202; 4) fly: mean = 41.6, SD = 26.5, minimum = 6, maximum = 209.

Human Scoring of Explanations of Evolutionary Change

Undergraduate students are known to recruit a diverse array of cognitive resources to build explanations of evolutionary change and solve evolutionary problems (Nehm, 2010). These resources may include, for example, well-structured scientific schemas, such as natural selection; fragmented mental models built using mixtures of scientific and naïve knowledge elements; or naïve explanatory models (Nehm and Ha, 2011). Given such diversity, it is most practical for assessment purposes to capture the existence of constituent explanatory elements in students' explanatory models (cf. Nehm and Haertig [in press]). For our study of automated computer scoring, two trained human raters scored all student responses for five key concepts of natural selection that were outlined by Nehm and Reilly (2007), described by Nehm and Schonfeld (2008), and codified in the scoring rubrics by Nehm *et al.* (2010a). It is important to emphasize that these concepts are central to the construct of natural selection, and necessary for explaining the operation of natural selection (Nehm and Schonfeld, 2010). Thus, the elements selected for scoring are not trivial or superficial aspects of reasoning regarding evolutionary change, and are associated with explanatory competence, as measured by clinical oral interviews (Nehm and Schonfeld, 2008). These key concepts included: 1) the presence and causes of variation (mutation, recombination, sex), 2) the heritability of variation, 3) competition, 4) limited resources, and 5) differential survival. It is important to note that scoring of short-answer explanations was dominated by the recognition of collections of key terms and short phrases, rather than elaborate grammatical expressions (see Nehm *et al.* [2010a] for details). Nevertheless, scoring was performed such that only accurate expressions counted for the "presence" of a concept; students' faulty expressions about heredity, for example, would *not* count as the presence of the key concept of heredity.

A series of studies has demonstrated that the coding rubrics used to score the short-answer explanations of evolutionary change are sufficiently clear to produce high levels of human interrater scoring agreement with moderate training (Nehm and Reilly, 2007; Nehm and Schonfeld, 2008; Nehm *et al.*, 2009a; Nehm *et al.* 2010b; Nehm and Ha, 2011; Nehm and Haertig, in press; Nehm *et al.*, in press). In past studies, kappa agreement coefficients between human scorers with limited training ranged from 0.69 to 0.95, with an average of 0.86 (e.g., Nehm and Haertig, in press). In the present study, two biology experts (who have scored several thousand explanations of evolutionary change and have used the rubrics of

Nehm *et al.* [2010a] for more than 2 yr) evaluated all student responses with very high agreement levels. The kappa reliability coefficients ($n = 1056$) for the present study were: 0.995 for variability, 1.000 for heritability, 1.000 for competition, 1.000 for limited resources, and 0.988 for differential survival. In the rare cases of disagreements between the two human raters, consensus scores were reached via deliberation. These final consensus scores were used in all subsequent analyses of human–computer scoring correspondence.

Human–Computer Correspondence Measures

We used Cohen's kappa to quantify the magnitude of human–computer scoring correspondence (Bejar, 1991). Cohen's kappa values range from 0.0 to 1.0, and are commonly used to quantify levels of agreement among human raters, or between human and computer rating scores (Landis and Koch, 1977; Nehm and Haertig, in press). Landis and Koch (1977) introduced three general agreement levels for kappa statistics that we follow in our study: values between 0.41 and 0.60 are considered "moderate"; values between 0.61 and 0.80 are considered "substantial"; and those between 0.81 and 1.00 are considered "near perfect." We also report specific kappa values for all analyses, given the subjective nature of these categorical distinctions.

ANALYSES

Our first research question explores SIDE performance at detecting individual key concepts of natural selection in students' written responses. For each of our analyses, and for each key concept, two *categories* of agreement statistics are reported: 1) scoring-model *training* agreement values and 2) scoring-model *cross-validation* agreement values. The training agreement values are generated when SIDE first attempts to construct a scoring model using the corpus of human-scored student answers; that is, SIDE attempts to "learn" from the human-scoring patterns and builds a computational model that can account for these patterns. Then, SIDE examines the efficacy of this scoring model by calculating how well the model scores the *same* responses from which it learned. Kappa and percentage agreement values (which are automatically generated by the SIDE program) enable researchers to judge the strength of the machine-learning model and consider whether it is worthy of use on a new data set. In situations where the training kappa values are "substantial" (> 0.60), the SIDE-generated scoring model is then applied to a *new* corpus of human-scored responses to determine whether the scoring model functions effectively with a new response corpus (that is, the training model is *tested*). Even if a training model performs admirably, this does not necessarily mean that it will be effective at scoring a new response corpus; both training and cross-validation performances need to be evaluated. Model cross-validation efficacy (also measured using kappa and percentage agreement statistics) must be performed manually (in our case, using SPSS, version 19.0). Cross-validation kappa values and percentage agreement values enable us to determine whether the SIDE-generated scoring models are likely to effectively score additional student responses.

In addition to exploring training and cross-validation performance of SIDE for each individual key concept, we also explored composite measures of students' explanations of evolutionary change. Key concept score (KCS) is a composite measure of the number of scientific concepts employed in an explanatory context (Nehm and Reilly, 2007). Given that KCS has been used in prior research on learning gains (Nehm and Reilly, 2007), we examined how well SIDE performed relative to human expert scorers for KCS. Specifically, we used Pearson correlation statistics (in SPSS, version 19.0) to test for significant associations between human and computer scores of both variables (in contrast to the single-concept agreement statistics discussed above).

Our second research question examined to what degree sample size influences SIDE scoring–model performance. Specifically, we trained SIDE on two different corpora: 1) 500 responses from OSU students and 2) 1056 responses from OSU students. We then tested the two different scoring models on 1) the MSU nonmajor response corpus and 2) the MSU biology major response corpus. We calculated kappa and percentage agreement statistics to evaluate the influence of the training-sample size on SIDE scoring–model performance. All statistical tests were performed in SPSS, version 19.0.

Our third research question explored the factors that limit SIDE scoring–model performance; that is, why, in some cases, do scoring models fail to function at the desired near-perfect (kappa values >0.80) agreement levels? Are such disagreements the products of the students (majors, nonmajors) and how they explain evolutionary change; the universities (OSU, MSU); the sample sizes; the scoring models; or combinations of these factors? This research question required examining all of the instances in which SIDE-generated scores and human-generated scores did not match, and attempting to identify the factors that contributed to score mismatches. After locating the likely source of scoring disagreements, we explored whether there were ways to mitigate these performance limitations so that future work would be more effective.

RESULTS

Students' Explanations of Evolutionary Change

To provide readers with a sense of the types of explanations of evolutionary change that undergraduate students' generate, four unedited student responses were extracted from the response corpus (see Table 2). As is apparent, student explanations of evolutionary change vary in length (for details, see *Items Used to Generate Explanations of Evolutionary Change*), sophistication, scientific accuracy, and scientific complexity. Adjacent to the responses in Table 2 are two columns indicating the numbers and types of key concepts detected in each response (see scoring methods, in *Methods*, for details). Note that in the present study we investigated only the magnitudes of accurately expressed scientific concepts in student responses, not naïve ideas or misconceptions. Computer scoring of other explanatory elements is the focus of ongoing research.

Testing the Impact of Training Corpus on Scoring Success

Our first analyses explored whether training SIDE using different human-scored corpora had an impact upon

Table 2. Selected examples of students' written explanations of evolutionary change and corresponding human and computer scores

Taxon/trait/polarity	Student's explanation of evolutionary change	Human score (number of key concepts)	Computer score (number of key concepts)
Shrew incisors	"Incisors may have developed on shrews due to *a genetic mutation* [Variation]. An offspring of a normal shrew may have had a mutated baby that had incisors, or some earlier form of incisors. The incisors would have given the new shrew an advantage *in acquiring food* [Limited resources] and reproducing, so it would have *a higher fitness* [Differential survival] leading *the incisor trait to be passed on to other generations* [Heredity]. As the trait will then develop with each generation due to variation involving the trait and the levels of success attached to each variant."	4	4
Snail feet	"They would explain that once all the snails had small feet. Then one day there was a *mutation* [Variation] that produced a snail with a large foot. *The snail with a large foot was better able to produce more offspring* [Differential survival] in the environment *passing on his trait* [Heredity]."	3	3
Fish fins	"There was *a random change in the DNA sequence* [Variation] of the fish that coded for the production of the fin. Nonrandom mating could have occurred with females selecting males with fins as partners, which disrupts HW equilibrium and leads to the evolution of the fin because *the fish with fin are better able to produce viable offspring* [Differential survival]."	2	2
Fly wings	"The evolution of a fly species with a large wing from an ancestral fly with small wings could be through the process of natural selection or from *a random mutation* [Variation]."	1	1

scoring-model success (Figure 1). Six tests were performed (Figure 1, A–F) for *each* key concept of evolution (e.g., variation, heredity, etc.). For the majority of these tests, the scoring agreements reached or exceeded near-perfect kappa values (18/30 tests) and percentage agreements above 90% (24/30 tests). Three key concepts—variation, heredity, and limited resources—were detected at near-perfect levels regardless of the training or cross-validation samples used. In contrast, competition and differential survival were very sensitive to training and cross-validation samples; in only two of the 12 tests did they reach near-perfect kappa agreement levels (Figure 1, left). While raw percentage agreement values were robust for four of the five concepts (the exception being differential survival), these values do not take into account chance agreements. The dramatic difference between these two agreement statistics for competition suggests that sample size is contributing to these patterns, as we discuss in *Training Sample Sizes and Scoring Success*. Overall, the most significantfactor influencing the performance of the SIDE-generated scoring models was not the training or cross-validation corpus per se, but rather, concept-specific factors.

Concept Frequencies in Different Samples

As shown in Figure 2, human-identified frequencies of key concepts (blue bars) are in close alignment with computer-identified frequencies (red and green bars) for all samples (the different rows: OSU nonmajors, MSU majors, and MSU nonmajors). In addition, different SIDE training sets (i.e., MSU majors, MSU nonmajors) did not generate substantially different scoring frequencies of key concepts in comparison with human-generated scores (compare the different colored bars for each concept within each row). Small differences are apparent, however, among scores for differential survival in the OSU nonmajor sample (top row, right) and variation in the MSU major and nonmajor sample (middle and bottom rows, left).

One of the most striking patterns across all samples is that introductory biology students rarely used the concept of competition in their explanations of evolutionary change. In contrast, differential survival was used by a majority of students in all samples. Differences in the frequencies of key concept use were also apparent between samples (compare the rows in Figure 2): almost twice as many responses from the MSU samples employed the concept of variation, compared with the responses in the OSU samples (bottom two rows vs. top row, left). In addition, MSU majors used the concept of limited resources more often than the other groups. Overall, Figure 2 demonstrates that differences in the frequencies of particular concepts vary across samples and schools, but the SIDE-generated scoring model was able to detect these differences. The extremely rare occurrence of competition (Figure 2) was associated with poor model performance for this concept (Figure 1).

Training Sample Sizes and Scoring Success

Given that some key concepts were less common in the training and cross-validation data sets than other concepts (e.g., competition vs. variation), we investigated the impact of sample size on scoring-model efficacy. For *each* key concept (e.g., variation, differential survival) we performed two experiments. In the first, we trained SIDE using 500 human-scored responses from a sample of OSU nonmajors; and in the second, we trained SIDE using a sample more than twice as large: 1056 human-scored responses from OSU nonmajors.

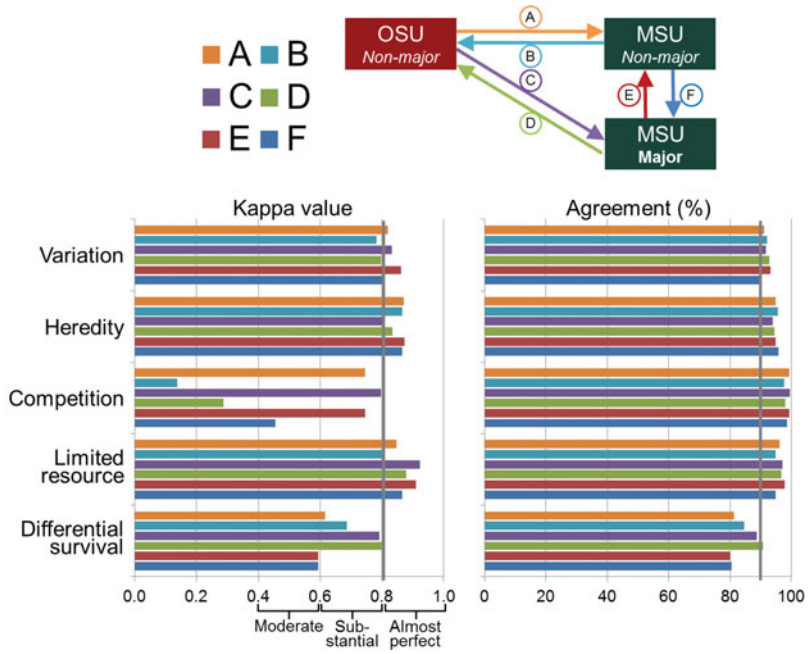

Figure 1. Magnitudes of agreement among human-scored and computer-scored explanations of evolutionary change from three samples (OSU, MSU nonmajor, and MSU major). For each of the three samples: $n = 500$ responses. Five key concepts of evolutionary change were examined separately (e.g., variation, heredity). Arrows indicate which sample was used to train the models and which sample was used to test the models. Kappa values compensate for chance agreements, whereas agreement values are raw percentages. (A) OSU sample model training and MSU sample nonmajor model cross-validation; (B) MSU nonmajor sample model training and OSU sample model cross-validation. (C) OSU sample model training and MSU nonmajor model cross-validation. (D) MSU major sample model training and OSU sample model cross-validation. (E) MSU major sample model training and MSU nonmajor model cross-validation. (F) MSU nonmajor model training and MSU major sample cross-validation.

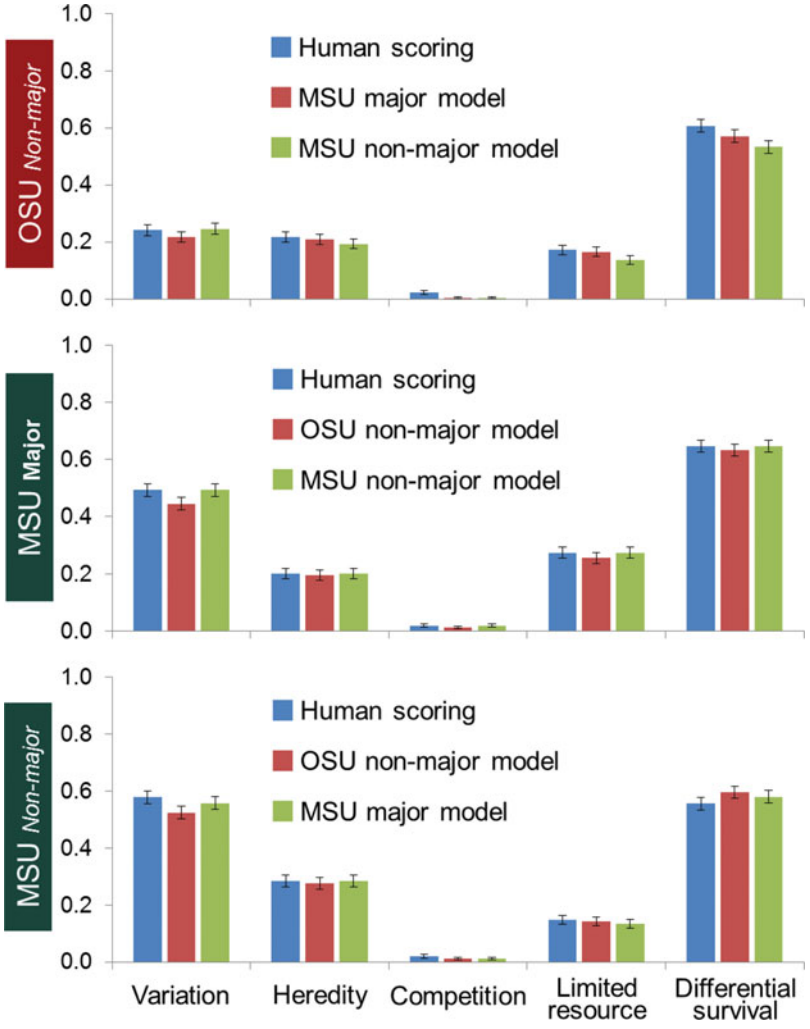

Figure 2. Frequencies (0–100%) of key concepts among samples and between human- and computer-generated scores. Blue bars = human-detected frequencies; red bars = frequencies detected using the MSU major computer-generated scoring model; and green bars = the frequencies detected using the MSU nonmajor computer-generated scoring model. In each of the three samples (OSU nonmajor; MSU major; MSU nonmajor), 500 responses were used. Error bars represent the SEM.

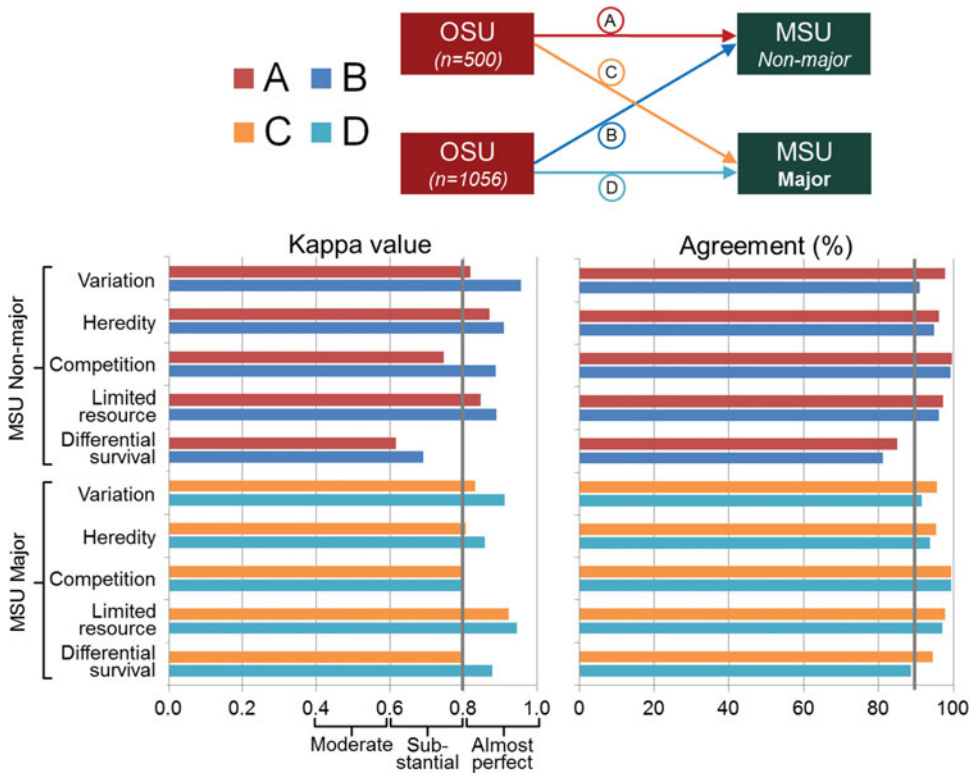

Figure 3. Cross-validation of the impact of training sample size on model performance. Four samples were used in the analysis (OSU nonmajors: $n = 500$; OSU nonmajors: $n = 1056$; MSU nonmajors: $n = 500$; and MSU majors: $n = 500$). Five key concepts of evolutionary change were examined separately (e.g., variation, heredity). Arrows indicate which sample was used to train the models and which sample was used to test the models. Kappa values compensate for chance agreements, whereas agreement values are raw percentages. (A) OSU sample ($n = 500$) training and MSU sample nonmajor cross-validation. (B) OSU sample ($n = 1056$) training and MSU sample nonmajor cross-validation. (C) OSU sample ($n = 500$) training and MSU sample major cross-validation. (D) OSU sample ($n = 1056$) training and MSU sample major cross-validation.

The two resulting SIDE-generated scoring models were applied separately to: 1) a corpus of MSU nonmajors' written explanations and 2) a corpus of MSU biology majors' written explanations (see *Methods*). As above, kappa agreement statistics and raw agreement percentages were calculated for all comparisons. Figure 3 illustrates the results of both experiments.

The scoring models built from the larger corpus produced higher correspondences with expert human raters in nine out of 10 tests; the exception was competition in the MSU biology major sample (Figure 3). Given that the smaller training corpus ($n = 500$) produced near-perfect correspondence with human raters in most tests, doubling the training size bumped only one concept—competition in the MSU nonmajor sample—to the desired benchmark (kappa > 0.80). Although the larger corpus did improve model performance for differential survival (Figure 3), in many cases it did *not* meet our benchmark using either training corpus ($n = 500$ or 1056). In terms of raw percentage agreements, the larger corpus did not always produce improved results; in fact, the smaller corpus in many cases produced slightly higher agreement percentages. Nevertheless, in tests of model performance, 17 out of 20 comparisons of computer and human agreement reached or exceeded 90%. Additionally, using the large training data set, 9/10 analyses produced results that were detected at or above the kappa benchmark of 0.80.

Overall, in nearly all cases, doubling the training corpus improved model performance, but not substantially. The most dramatic improvement was seen in the detection of competition in the MSU nonmajor sample. Thus, the frequencies of particular concepts in the training corpus must be considered, not just overall sample size (see Figure 2 "Human scoring").

Explanatory Structures

In addition to comparing individual key concept detection between human- and computer-scored explanations, it is useful to examine how students collectively assemble these concepts into explanatory structures (Figure 4). One approach for representing these explanatory interrelationships is to use concept-association diagrams (Nehm and Ha, 2011). Figure 4 illustrates both the frequency (the size of the circles) and co-occurrences of concepts (the thickness of the gray lines) in students' explanations of evolutionary change. For instance, approximately 20% of OSU nonmajors used both concepts of variation and differential survival in their responses (see connecting lines in Figure 4). Each row in Figure 4 compares explanatory structures between human expert raters (left) and SIDE-scoring patterns (right) for a particular student sample (e.g., top: OSU nonmajors; bottom: MSU biology majors). As is apparent from the figure, results for students' knowledge networks are remarkably similar, regardless of whether they were scored by humans or computers.

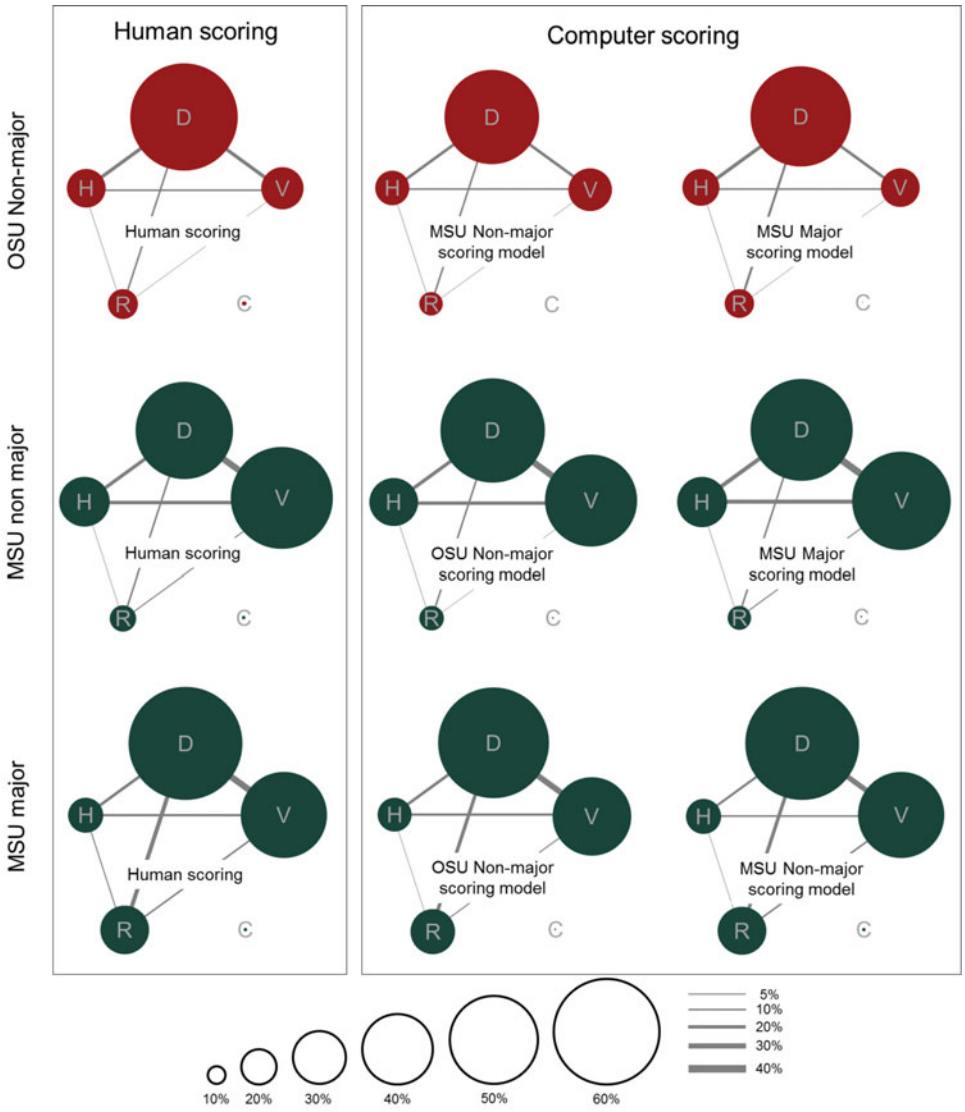

Figure 4. Holistic patterns of human–computer scoring correspondence (each row), taking into account all five key concepts. Circle sizes represent the frequencies of concepts; gray bars indicate the percentages of concept co-occurrence. D = differential survival; V = variation; H = heredity; R = limited resources; C = competition.

Comparing the columns in Figure 4 also reveals that SIDE-generated scoring models can detect different explanatory structures across student samples, and these patterns closely mirror the findings reported in Figure 2. The human-generated scores, for example, demonstrate that MSU majors and nonmajors used the concept of variation much more frequently than OSU nonmajors. SIDE produced the same patterns. Interestingly, human scorers also determined that MSU nonmajors used the concepts of variation and heredity more frequently than MSU majors; SIDE detected these patterns. MSU biology majors used the concept of limited resources much more frequently than MSU nonmajors, and this is also indicated in the SIDE-generated scores. While differences in the explanatory structures among majors, nonmajors, and institutions are interesting, the important point we wish to emphasize in Figure 4 is that SIDE-generated scores, which took minutes to generate, are in remarkable alignment with the patterns generated by humans during weeks of painstaking grading.

In addition to examining patterns of correspondence between human- and computer-generated visual representations of explanatory structure, Nehm and Reilly (2007) used KCS to quantify the number of different scientifically accurate evolutionary concepts that students use to explain evolutionary change in a prompt. Table 3 illustrates statistically significant ($P < 0.001$) and robust ($r = 0.79$ to 0.87) associations between human- and computer-generated KCS for all comparisons. Thus, using approaches for measuring student knowledge of evolution previously established in the literature (Nehm and Reilly, 2007; Nehm and Ha, 2011), SIDE-generated scoring models produce patterns equivalent to those derived from human raters.

Table 3. Correlation coefficients between human-scored and SIDE-scored student explanations for KCS[a]

Training sample	Testing sample	Human vs. SIDE KCD correlation ($^{**}P < 0.001$)
OSU nonmajor ($n = 500$)	MSU nonmajor ($n = 500$)	0.79**
MSU nonmajor ($n = 500$)	OSU nonmajor ($n = 500$)	0.80**
OSU nonmajor ($n = 500$)	MSU major ($n = 500$)	0.87**
MSU major ($n = 500$)	OSU nonmajor ($n = 500$)	0.85**
MSU major ($n = 500$)	MSU nonmajor ($n = 500$)	0.82**
MSU nonmajor ($n = 500$)	MSU major ($n = 500$)	0.82**

[a]In all cases, associations were strong and significant ($P < 0.001$). KCS represents the number of different scientific concepts in a prompt. For details, see *Methods* and Nehm and Reilly (2007).

Factors Limiting Computer-Scoring Success

Although SIDE and its machine-learning algorithms were shown to be highly effective at scoring the accuracy and complexity of undergraduates' explanations of evolutionary change, our studies revealed several limitations, which are summarized in Table 4. The key factors that limited the efficacy of computer scoring included: misspellings; nonadjacent key terms; very uncommon concept frequencies; and the diversity of expressions that students used to represent particular concepts.

Spelling and spacing errors produce human–computer score disagreements. While our human raters easily understood what students were attempting to explain when they wrote "preditor" [predator], the computer was not able to do so. Misspelled words such as "servive" [survive], "springoffs" [offspring], and "foodso" [food so] also produced misclassifications in our study (see Table 4).

We also found that when student responses included key terms suggestive of a concept, but the words constructing the concept were scattered throughout the written response

Table 4. Examples of the types of disagreements between human-scored and computer-scored explanations[a]

Scoring pattern	Category	Examples 1 to 5	Solution
Positive computer score but negative human score	Many key terms used, but important aspects were missing	(1) "The original Shrew, who didn't have incisors, may have not been a *fit* species. Meaning, they may have not been *reproducing* enough to pass on their traits. Another reason would be a mutation that survived. What I mean is that a few shrews may have developed an allele mutation that resulted in the shrews developing incisors. After the mutation, it was passed on and probably *survived* because of natural *selection*, sexual *selection*, or artificial *selection* [*more* fit, survive and reproduce *more*]."	Put a weight on core terms
	Key terms not adjacent, but scattered throughout the response	(2) "For the fly species with wings to *survive* through natural selection it had to evolve to a species without wings. The wings of the fly were no *longer* needed so over time they grow smaller and weaker and eventually they no *longer* appeared in the offspring [*survive longer*]."	Human augmentation of SIDE-scoring models
Negative computer score but positive human score	Very uncommonly used expression	(3) "The fish was filling a niche in an area that required a fish with smaller fins. Generations passed and *a mutant gene* for a fish with smaller fins did well and its offspring did well through time a new species was born."	Increase concept frequencies in training sample; human augmentation of SIDE-scoring models
	Complex expressions	(4) "Variation of living fish species may leads [sic] to random mutation. It creates new sequences of DNA that will code for new or different protein. *This protein help*[sic] *in the creation of a new living fly species with wings in this situation. This species then reproduce* [sic] *and evolve through some time*. It may help to explain how a new living fly species with wings evolve [sic] even though it is originally from an ancestral fly species that lacked wings."	Human augmentation of SIDE-scoring models
	Spelling errors and spacing errors	(5) preditor [predator], servive [survive], springoffs [offspring], foodso [food so]	Incorporate a spell-check program during data collection

[a]Categories: types of scoring problems; examples: specific student responses; solutions: approaches used to correct the computer–human disagreement.

(conveying a very different meaning), the SIDE-scoring models often mistakenly linked these words together and considered the concept to be present. For example, in Table 4, example 2, the scoring model identified text elements characteristic of the concept of differential survival ("survive longer") to be present, even though the response included the words "survive" and "longer" in separate sentences.

Complex expressions, or very long sentences, also posed problems for the software. For example, one student (correctly) explained differential survival in the following way: "the fish with large fins is not suitable for the living environment in that specific area, so the fish with smaller fins survive and reproduce." Because expressions like this were rare in the student response corpora, and did not contain explicit language about differential survival, the scoring models failed to detect them. Fortunately, expressions like this were uncommon in the samples.

Low concept frequencies also prevented the machine-learning algorithm from building successful scoring models; without enough positive cases to analyze, the computer failed to annotate new cases appropriately. Competition serves as an example of a concept that was very rarely used by students to explain evolutionary change; only about 10 instances of competition were found among 500 written explanations. In addition to low concept frequencies, unusual expressions also lead to misclassifications. The term "mutant gene," while clearly identifiable as a "cause of variation" by a biologist, was too rare to be incorporated into the scoring models (Table 4).

We found that the concept of differential survival was influenced by the frequencies with which particular samples used particular terms. While the kappa values between scoring models built using the OSU nonmajor and the MSU major sample were nearly 0.80 (near-perfect), the kappa values between scoring models built by the OSU nonmajor and MSU nonmajor samples and between the MSU major and MSU nonmajor samples did not meet this benchmark (0.60). The cause of this pattern appears to be that the language patterns that MSU nonmajors used were somewhat different from the language patterns that the OSU nonmajor and MSU major samples used to describe differential survival. For example, the term "differential" (such as "differential reproduction rate" or "differential survival success") was observed 12 times among 500 responses in MSU nonmajor sample, whereas the term "differential" was observed only three times among 500 responses in the OSU nonmajor sample and only once among 500 responses in the MSU major sample. Consequently, the program incorporated "differential" as a diagnostic term for the MSU nonmajor scoring model, but not for the other samples.

DISCUSSION

While CAS is becoming increasingly common throughout the educational hierarchy (Nehm *et al.*, in press), biologists have been slow to make use of this technological innovation. Two recent studies by Nehm and Haertig (in press) and Nehm *et al.* (in press) tested the efficacy, respectively, of SPSSTA, version 3.0 (Galt, 2008) and SIDE (Mayfield and Rosé, 2010a,b). Using large samples of undergraduate biology students in single classes at one university, they demonstrated that both of these analytical tools are capable of generating assessment scores equal in precision to those by trained, expert raters (biologists with PhDs). Overall, Nehm *et al.* (in press) suggested that when clear scoring rubrics have been developed, and student ideas on a particular topic are well established, SIDE is much more powerful and cost effective than SPSSTA (Haudek *et al.*, 2011; Nehm *et al.*, in press). Since both of these factors apply to our present study, we chose SIDE as our CAS tool. For biologists who have not developed robust grading rubrics, or who have not comprehensively investigated student thinking about a topic, SPSSTA will be a more appropriate starting point (Haudek *et al.*, 2011).

Prior studies of SIDE did not investigate several questions that arise when biologists apply scoring models beyond a single instructor, course, or college. First, are scoring models generalizable across colleges and courses (e.g., major vs. nonmajor)? That is, will successful scoring models built at one university work at another? Second, how much human scoring is needed to build a robust scoring model, and can human scoring of additional student responses compensate for scoring-model limitations across courses and colleges? Finally, what factors might account for scoring models that function effectively in a class at one university but fail at a similar class in another? Can these failures be fixed?

It is important to emphasize that CAS tools—including machine learning—are not capable of comprehending the meanings of students' lexical responses. Programs such as SIDE simply note the presence or absence of particular words (or word pairs) in response corpora, build large matrices of word combinations, and apply sophisticated machine-learning algorithms to predict human-scoring patterns (Mayfield and Rosé, 2010a,b). Consequently, machine-learning tools are very sensitive to language, but not its meaning(s). Expert human raters, in contrast, can effortlessly comprehend diverse linguistic expressions and understand their equivalence (e.g., "some live and some die" is equivalent to "differential survival"); in contrast, computers view different text as indicative of different information. For this reason, mundane text differences, such as spelling (color vs. colour; fecundity vs. feduncti[sic]) impact scoring-model success.

Depending upon the scientific key concept for which a scoring model is built (e.g., variation, differential survival, etc.), different lexical expressions are used in different frequencies. Indeed, different populations of students—such as biology majors and nonmajors—may use characteristically different linguistic expressions to represent biological concepts. Some word combinations in some samples will be more diagnostic and predictive of key concepts than in others. Because of these concept-specific and sample-specific issues, we discuss our specific results relating to sample source (university; majors vs. nonmajors) and sample size (500 responses vs. 1056 responses) separately for each concept for which a scoring model was developed: variation, heredity, competition, and differential survival.

For key concept 1, variation, we found that SIDE scoring–model success was not sensitive to sample source (Figure 1). That is, regardless of which response corpus was used to train SIDE (i.e., OSU vs. MSU students; majors vs. nonmajors), the scoring models generated excellent agreement with trained expert raters and near-perfect kappa values (> 0.80). However, we did find that scoring models for variation were

somewhat sensitive to sample size (that is, whether 500 or 1056 responses were used to build the scoring models). In comparison with the key concept of heredity, for example, in which a doubling of the training-sample size had almost no impact upon kappa values (adding 0.04 to 0.05), a doubling of the sample size for variation produced meaningful increases in kappa values (adding 0.14 to 0.80).

The explanation for the increase in kappa values with increasing training sample size for variation (but not heredity) appears to be related to the diversity and frequency of linguistic expressions that students used to represent these biological concepts. Although the most common term used by students to represent variation was "mutation," various other terms were also used, such as "different alleles," "genetic change," or "error in DNA." If only a few students used particular written expressions when linguistically representing the concept of variation (such as "genetic makeup"), then such expressions would be unlikely to be included in the machine-learning model, and downstream disagreements between human and computer scores would result. The frequencies of particular expressions, and their associations with other terms, influence scoring-model success. Indeed, we found that doubling the training sample for variation increased the frequencies of particular terms to a threshold at which they were included in the scoring models, producing improved kappa agreement statistics. For example, the matrix included 268 words for the $n = 500$ sample, while the matrix included 386 words for the $n = 1056$ sample. Matrix size is associated with differences in scoring-model performances.

For the concept of heredity, computer-scoring success was very stable and very successful regardless of sample size or source (Figures 1 and 3). Biology majors and nonmajors from different colleges and classes appear to use a consistent and detectable array of linguistic expressions to represent heredity concepts (e.g., Table 1).

The third concept we investigated was competition. Unlike the previous results, computer-scoring success for competition was sensitive to both sample source and sample size (Figures 1 and 3). Given our findings for variation and heredity, this result is surprising; the Nehm *et al.* (2010a) scoring rubric indicates that a very small set of terms is typically used to detect competition (e.g., compete, competition, competes). When we examine the frequency of students who used this concept, we find that only 1–2% of students used competition in their explanations of evolutionary change. Indeed, only 10 to 20 responses (out of 1000) included linguistic expressions relating to competition. Statistically, the probability that the algorithm will include such rare occurrences is low. Two solutions may be used to tackle the problem of rare responses: first, to amass a larger corpus of responses; or second, to use a special function in SIDE that allows users to augment the model and weight particular terms (see Nehm *et al.* [in press] for details). It is difficult for SIDE to build scoring models for extremely rare concepts.

The next concept we studied was limited resources. We found that the scoring models for this concept were stable in relation to both sample source and sample size (Figure 3). Kappa values were near-perfect (> 0.80) for the small data sets ($n = 500$) across samples, although there were some minor deviations. Overall, regardless of course or college, it appears that students commonly use consistent language patterns to represent this evolutionary concept, and scoring models for this concept work very well.

The final concept that we studied was differential survival. Similar to our findings for competition, differential survival was sensitive to both sample source and sample size. The comparatively weak performance of the differential survival scoring models was not a result of low response frequencies (as we observed with competition); large percentages of students utilized this idea in their explanations of evolutionary change (e.g., 60.3%; Figure 2). Scoring problems in this case were a product of students' highly variable language use. This is in line with the scoring rubrics of Nehm *et al.* (2010a), which were built using different student samples and also note the diverse expressions with which students represent this evolutionary concept (e.g., "increase their survival," "survived better," "the species dies while others survive"). Because we also found that SIDE-scoring models were sensitive to sample source, different linguistic expressions may have been related to instructor discourse patterns. If, for example, students are imitating instructors' language (cf. Nehm *et al.* [2010b]), and different instructors use different phrases to represent biological ideas, then the sample source will impact scoring-model efficacy (as we found). Although the scoring model built using the largest sample ($n = 1056$) demonstrated relatively good kappa values (e.g., 0.69, 0.89; see Figure 3), the highly variable ways of communicating the concept of differential survival appears to have limited scoring-model performance.

Generalizing Our Findings to Other Samples and Populations

Very few studies in biology education have examined the similarities and differences between different student populations' short-answer explanations of biological phenomena, including evolutionary change. In two studies of primarily underrepresented biology students (many of whom were English-language learners) from a minority-serving institution in the eastern United States, Nehm and Reilly (2007) and Nehm and Schonfeld (2008) used short-answer, constructed-response assessments similar to those in the present study to reveal students' thinking patterns regarding evolutionary concepts. Nehm and Schonfeld (2008) reported that their findings were generally similar to those of primarily white student populations documented in the literature. Our current findings—from primarily white, midwestern undergraduates in large, public, research universities—are also very similar to those documented in these prior studies (see Table 1). This suggests that undergraduate biology students, regardless of racial and ethnic background, may utilize a large but relatively constrained set of concepts when conceptualizing evolutionary change. Nevertheless, such conjecture should be tested using diverse student samples from different geographic regions of the country. Until such work is completed, we cannot with confidence argue that machine-learning tools will be effective for assessing *all* introductory biology students.

Implications for Introductory Biology Faculty

Our study has produced robust, automated, and generalizable scoring models capable of detecting most (but not all) of

the core evolutionary concepts emphasized in standards documents, curricula, and textbooks (Nehm *et al.*, 2009b). Biology educators can make use of our work by downloading the free software package SIDE (see Mayfield and Rosé, 2010a,b) and incorporating our scoring models (freely available from the senior authors) to evaluate their students' written explanations of evolutionary change. Using a PC computer with an i7 processor, scoring 1000 written responses takes seconds to a few minutes (depending on the concept) and produces high levels of accuracy that are comparable with consensus scores generated by two trained biologists (see Figure 4).

In addition to a user's manual (Mayfield and Rosé, 2010a), and details on the workings of SIDE (Mayfield and Rosé, 2010b), learning how to use SIDE is illustrated in a series of video tutorials (freely available at http://evolutionassessment.org). Given that this emerging form of assessment research is new, it is important to emphasize that the software is not packaged in a user-friendly format, and like other technological tools (e.g., clickers, new operating systems, new software), effort is required to learn to use it.

Our research to date has only validated a small set of biological concepts, and introductory biology instructors are likely to want to know how their students interpret a broader array of concepts in evolution (and other content areas). We are continuing to build scoring models for other concepts, such as naïve ideas or misconceptions of evolution (Ha and Nehm, unpublished results). We speculate that improved technology and advanced research on machine-learning assessment will enable more and more concepts to be detected in students' written responses.

National partnerships among introductory biology educators could make future work on machine learning more efficient and cost effective. Indeed, all biology educators, regardless of whether they view automated scoring as beneficial or not, could help move the field forward by collecting large corpora of students' written responses to different prompts across subject areas (genetics, matter and energy transformation, cell biology; Haudek *et al.*, 2011); this would help those researchers interested in using and refining machine-learning methods. Additionally, faculty from minority-serving institutions, or those teaching large English language–learning populations, are needed to expand our knowledge base on how scientific language is used to communicate core concepts in biology.

Perhaps the most significant implication of our work for introductory biology educators is that evaluating students' written work, especially in large classes, is not impossible. This is significant from an assessment standpoint, as we contend that the process of asking students to communicate their understanding of scientific phenomena is a worthwhile activity, regardless of whether automated methods will be employed to assess these responses (e.g., Chi *et al.* [1994]). When analyzing students' written responses, we have been surprised by students' limited capacity to communicate and explain core scientific concepts (such as evolution)—particularly those students who perform admirably on MC assessments (cf. Nehm and Schonfeld [2008]). Without providing students practice and feedback in communicating their scientific understanding, we cannot expect this situation to improve.

Future work is needed to expand our concept of what constitutes a sound explanation of evolutionary change. Quantifying students' use of necessary and sufficient scientific elements (key concepts) as a benchmark for competency, as we have done, captures only one facet of short-answer scientific explanations (cf. Braaten and Windschitl [2011]). Logic, persuasion, and argumentation skills are also important dimensions of scientific explanation, but they were not investigated in our study. Expanding our assessment framework will likely stimulate discussions about what facets of scientific explanation are most important for fostering scientific literacy.

Implications for Biology Education Researchers

Research in the use of machine learning (and text analysis in general) in biology education is only beginning (Haudek *et al.*, 2011; Nehm and Haertig, in press; Nehm *et al.*, in press); much remains to be learned. A community of practice on text analysis in STEM education has recently been established (see Haudek *et al.* [2011] and http://aacr.crcstl.msu.edu), providing a forum for researchers interested in learning more about these innovative assessment methods. Our current study has uncovered several findings likely to be of interest to researchers motivated to pursue this line of work.

First, even though we collected large response corpora, some concepts were nevertheless quite rare, limiting model performance. A large sample ($n = 500$) does not guarantee high concept frequency. In many instances, we were surprised by which concepts were used by students (and which were not). Second, we documented several factors that caused problems for machine-learning methods (e.g., misspellings; linguistic diversity) that nevertheless can be addressed by using a spell-checker during data gathering and weighting text expressions prior to analysis. Third, the diversity of linguistic expressions associated with concepts was highly variable (and generally unpredictable a priori), impacting scoring success. Some concepts were easily detected by the software, whereas others were not. Overall, the process of building automated scoring models is effortful and requires clear scoring rubrics and thousands of carefully evaluated responses.

ACKNOWLEDGMENTS

We thank Kristen Smock and Judith Ridgway, OSU, and Donna Koslowsky and Tammy Long, MSU, for assistance with data gathering; Luanna Prevost for comments on the manuscript; and the Automated Analysis of Constructed Response (AACR) group for discussions. R.H.N. thanks Elijah Mayfield and Caroline Rosé and the faculty of the Carnegie Mellon Pittsburgh Science of Learning Center summer school for help with machine-learning methods. We also thank two reviewers for helpful suggestions for improving the manuscript. We thank the National Science Foundation (NSF; REESE grant 090999 to principal investigator [PI] R.H.N.) for funding M.H. and collaborative NSF CCLI 1022653 to PIs Jenny Knight, R.H.N., and M.U.-L. for support of the AACR group. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the NSF. The research was conducted under OSU IRB Protocol # 2008B0080 (R.H.N., PI).

REFERENCES

American Association for the Advancement of Science (1994). Benchmarks for Science Literacy, New York, NY: Oxford University Press.

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education, Washington, DC: http://visionandchange.org.

Batistta 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.

Bejar II (1991). A methodology for scoring open-ended architectural design problems. J Appl Psychol *76*, 522–532.

Bishop B, Anderson C (1990). Student conceptions of natural selection and its role in evolution. J Res Sci Teach *27*, 415–427.

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

Burstein J (2003). The e-rater scoring engine: automated essay scoring with natural language processing. In: Automated Essay Scoring: A Cross-Disciplinary Perspective, ed. MD Shermis and J Burstein, Mahwah, NJ: Lawrence Erlbaum, 113–122.

Butler AC, Marsh EJ, Goode MK, Roediger HL (2006). When additional multiple-choice lures aid versus hinder later memory. Appl Cogn Psychol *20*, 941–956.

Chi MTH, Slotta JD, de Leeuw N (1994). From things to processes: a theory of conceptual change for learning science concepts. Learn Instr *4*, 27–43.

Clough EE, Driver R (1986). A study of consistency in the use of students' conceptual frameworks across different task contexts. Sci Educ *70*, 473–496.

Corcoran TB, Mosher FA, Rogat A (2009). Learning Progressions in Science: An Evidence-Based Approach to Reform, New York: Center on Continuous Instructional Improvement.

Duschl RA, Schweingruber HA, Shouse AW (2007). Taking Science to School: Learning and Teaching Science in Grades K-8, Washington, DC: National Academy Press.

Evans EM, Lane JD (2011). Contradictory or complementary? Creationist and evolutionist explanations of the origin(s) of species. Hum Dev *54*, 144–159.

Galt K (2008). SPSS text analysis for surveys 2.1 and qualitative and mixed methods analysis. J Mixed Meth Res *2*, 284–286.

Gregory TR (2009). Understanding natural selection: essential concepts and common misconceptions. Evol Educ Outreach *2*, 156–175.

Ha M, Nehm RH (2011). Comparative efficacy of two computer-assisted scoring tools for evolution assessment. Paper presented at the National Association of Research in Science Teaching, Orlando, FL, April 2011.

Haudek KC, Kaplan JJ, Knight J, Long T, Merrill JE, 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*, 149–155.

Kang SHK, McDermott KB, Roediger HL (2007). Test format and corrective feedback modify the effect of cross-validation on long-term retention. Eur J Cogn Psychol *19*, 528–558.

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

Landauer TK, Laham D, Foltz PW (2001). The intelligent essay assessor: putting knowledge to the test. Paper presented at the Association of Test Publishers Computer-Based Cross-validation: Emerging Technologies and Opportunities for Diverse Applications conference, Tucson, AZ, February 2001.

Landis JR, Koch GG (1977). The measurement of observer agreement for categorical data. Biometrics *33*, 159–174.

Mandler G, Rabinowitz JC (1981). Appearance and reality: does a recognition test really improve subsequent recall and recognition?. J Exp Psychol Learn *7*, 79–90.

Mayfield E, Rosé C (2010a). An interactive tool for supporting error analysis for text mining. Proceedings of the North American Association for Computational Linguistics (NAACL) HLT 2010: Demonstration Session, Los Angeles, CA, June 2010, 25–28.

Mayfield E, Rosé C (2010b). SIDE: The Summarization IDE (User's Manual). www.cs.cmu.edu/~cprose/SIDE.html (accessed 13 October 2011).

National Research Council (1996). National Science Education Standards, Washington, DC: National Academy Press.

National Research Council (2001). Knowing What Students Know: The Science and Design of Educational Assessment, Washington, DC: National Academies Press.

Nehm RH (2010). Understanding undergraduates' problem solving processes. J Biol Microbiol Educ *11*, 119–122.

Nehm RH, Beggrow E, Opfer J, Ha M. Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. Am Bio Teacher (*in press*).

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

Nehm RH, Ha M, Mayfield E. Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. J Sci Educ Technol *(in press)*.

Nehm RH, Ha M, Rector M, Opfer J, Perrin L, Ridgway J, Mollohan K (2010a). Scoring Guide for the Open Response Instrument (ORI) and Evolutionary Gain and Loss Test (EGALT). Technical Report of National Science Foundation REESE Project 0909999 http://evolutionassessment.org (accessed 10 January 2011).

Nehm RH, Haertig H Human vs. computer diagnosis of students' natural selection knowledge: cross-validation the efficacy of text analytic software. J Sci Educ Technol (*in press*).

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

Nehm RH, Poole TM, Lyford ME, Hoskins SG, Carruth L, Ewers BE, Colberg PJS (2009b). Does the segregation of evolution in biology textbooks and introductory courses reinforce students' faulty mental models of biology and evolution? Evol Educ Outreach 2, 527–532.

Nehm RH, Rector M, Ha M (2010b). "Force-talk" in evolutionary explanation: metaphors and misconceptions. Evol Educ Outreach *3*, 605–613.

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

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

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*, 1131–1160.

Nehm RH, Schonfeld IS (2010). The future of natural selection knowledge measurement: a reply to Anderson *et al*. J Res Sci Teach *47*, 358–362.

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

Opfer J, Ridgway J, Perrin L, Mollohan K, Nehm RH (2011). Applying cognitive science to assessment of evolution education. Paper

presented at the National Association of Research in Science Teaching, Orlando, FL, April 2011.

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

Powers DE, Burstein JC, Chodorow M, Fowles ME, Kukich K (2002). Stumping E-rater: challenging the validity of automated essay scoring. Comput Hum Behav *18*, 103–134.

Roediger HL, Marsh EJ (2005). The positive and negative consequences of multiple-choice cross-validation. J Exp Psychol Learn *31*, 1155–1159.

Shermis MD, Burstein J (2003). Automated Essay Scoring: A Cross-Disciplinary Perspective, Mahwah, NJ: Lawrence Erlbaum.

Smith MK, Wood WB, Knight JK (2008). The genetics concept assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci Educ *7*, 422–430.

Sukkarieh J, Bolge E (2008). Leveraging C-rater's automated scoring capability for providing instructional feedback for short constructed responses. In: Lecture Notes in Computer Science, vol. 5091, ed. BP Woolf, E Aimeur, R Nkambou, and S Lajoie, New York: Springer, 779–783.

Urban-Lurain M, Moscarella RA, Haudek KC, Giese E, Merrill JE, Sibley DF (2010). Insight into student thinking in STEM: lessons learned from lexical analysis of student writing. Paper presented at the National Association for Research in Science Teaching Annual International Conference, Philadelphia, PA, March 2010.

Vosniadou S (2008). The framework theory approach to the problem of conceptual change. In: International Handbook of Research on Conceptual Change, ed. S Vosniadou, New York: Routledge, 3–34.

Williamson DM (2009). A framework for implementing automated scoring. Paper presented at the American Educational Research Association, San Diego, CA, April 2009.

Witten IH, Frank E (2005). Data Mining, 2nd ed., Amsterdam: Elsevier.

Wood WB (2004). Clickers: a teaching gimmick that works. Dev Cell *7*, 796–798.

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