J Sci Educ Technol
DOI 10.1007/s10956-011-9300-9

Transforming Biology Assessment with Machine Learning: Automated Scoring of Written Evolutionary Explanations

Ross H. Nehm · Minsu Ha · Elijah Mayfield

Abstract This study explored the use of machine learning to automatically evaluate the accuracy of students' written explanations of evolutionary change. Performance of the Summarization Integrated Development Environment (SIDE) program was compared to human expert scoring using a corpus of 2,260 evolutionary explanations written by 565 undergraduate students in response to two different evolution instruments (the EGALT-F and EGALT-P) that contained prompts that differed in various surface features (such as species and traits). We tested human-SIDE scoring correspondence under a series of different training and testing conditions, using Kappa inter-rater agreement values of greater than 0.80 as a performance benchmark. In addition, we examined the effects of response length on scoring success; that is, whether SIDE scoring models functioned with comparable success on short and long responses. We found that SIDE performance was most effective when scoring models were built and tested at the individual item level and that performance degraded when suites of items or entire instruments were used to build and test scoring models. Overall, SIDE was found to be a powerful and cost-effective tool for assessing student knowledge and performance in a complex science domain.

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

M. Ha
e-mail: ha.101@osu.edu

E. Mayfield
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: elijah@cmu.edu

Published online: 24 April 2011

Keywords Machine learning · SIDE · Text analysis · Assessment · Computers · Evolution · Explanation

Introduction

Formative and summative assessments are increasingly recognized as essential components of effective teaching and learning (NRC 2001, 2007). High-quality assessments focusing on core competencies and performances have shown great promise in helping to foster meaningful learning gains in students throughout the educational hierarchy (NRC 2001). Such findings have catalyzed efforts to develop new assessment tools and practices that more closely mirror the complex dimensions of authentic, real-world problem solving (NRC 2001, 2007; Wagner 2008; Gitomer and Duschl 2007). Coupled with educational reform, new tools are needed to assess the types of skills that the twentyfirst century American workforce needs; that is, skills that cannot be easily automated, digitized, or outsourced (Wagner 2008; Gitomer and Duschl 2007; NRC 2008). Internet search engines and artificial intelligence machines, for example, are currently capable of rapidly and accurately answering many well structured questions, and yet these are the types of problems often emphasized in K-16 multiple-choice assessments (Wagner 2008; Markoff 2011). Simply put, many assessments are not measuring the types of skills or performances that are most highly valued by employers or educational reformers (NRC 2001, 2007; The Conference Board, Corporate Voices for Working Families, the Partnership for 21st Century Skills, and the Society for Human Resource Management 2007).

Real-world biological problems are often ill-structured, requiring performances such as task framing, weighing the value and relevance of information, and assembling

disparate knowledge elements into clear, logical, coherent and complex explanatory structures (Nehm 2010). Most high-stakes, multiple-choice assessments are only capable of exposing a small set of the problem-solving processes central to authentic scientific practice (NRC 2007). For these reasons, multiple-choice assessments are often poorly suited to assessing the most valuable skills and performances in biology (and other domains). Current multiple-choice assessments are also severely limited in their ability to measure communication skills essential to success in real world problem-solving environments. Indeed, written communication, critical thinking, and problem solving form a constellation of skills that employers agreed were most needed for success in the twentyfirst century workplace (Wagner 2008: 91).

While many students may be able to access and know large amounts of information, they may nevertheless be limited in their ability to organize and express such understanding in a clear, logical, and persuasive manner. Implementing assessments that help to reveal to students their progress toward competence in this regard is essential (NRC 2001). Such goals were highlighted by the NRC more than a decade ago (2001: 5): "[a]ssessments need to examine how well students engage in communicative practices appropriate to a domain of knowledge and skill, what they understand about those practices, and how well they use the tools appropriate to that domain" (NRC 2001, p. 5).

Increasing global competition for skilled workers has also thrown light on the United States frequent reliance on high-stakes, multiple-choice tests (Wagner 2008; NRC 2008). As noted by Andreas Schleicher, head of educational indicators at the OECD, "United States students tend to be rather good in multiple-choice tasks, when four choices are clearly laid out. They have a much harder time when they're given open ended tasks" (cited in Wagner 2008: 94–95). Given that performance-based assessments are increasingly being used in higher education and in international comparisons (e.g., the Collegiate Learning Assessment [CLA] and the Program for International Student Assessment [PISA], respectively), it is important for students to have opportunities for demonstrating their abilities in these formats. Indeed, twenty-five percent of the Trends in International Math and Science Study (TIMSS) assessment items, for example, now require students to demonstrate performance skills such as constructing explanations (Liu et al. 2008: 35). Reliance on high stakes, multiple-choice assessments may be sending the wrong message to students about what knowledge and performances are valued outside of schooling and in the workplace.

Ongoing national curriculum and instruction reform (e.g., NRC 2008) must be aligned with the development of innovative assessments that move away from inauthentic performance tasks, such as the selection of carefully packaged, discrete bits of pre-structured knowledge, and towards the construction and communication of more authentic tasks common to ill-structured, real-world problems (NRC 2001; Nehm and Haertig 2011). The question arises as to what tools may be used for such complex assessment tasks, and how well they function. To this end, our study involves one small step towards developing more innovative assessment practices in biology: the use of a new technological tool known as the Summarization Integrated Development Environment (SIDE) to automatically analyze and score written explanations of evolutionary change. We test the efficacy of this freely available software package and illustrate how it may be used in biology education more broadly.

Assessing Evolutionary Explanations

For the past 30 years, student-generated explanations of evolutionary change have been used on a small scale for assessing knowledge, revealing misconceptions, and measuring conceptual growth in secondary school and undergraduate populations (e.g., Clough and Driver 1986; Bishop and Anderson 1990; Demastes et al. 1995; Nehm and Reilly 2007; Nehm et al. 2010). Multiple-choice assessments—which by their very nature limit opportunities for assessing students' abilities to construct and communicate valid scientific explanations—have nevertheless gained in favor and frequency because of their implementation ease and scoring simplicity in large samples. Recent work in evolution assessment has highlighted numerous limitations of extant closed-response (e.g., multiple-choice) evolution instruments and has called for the broader adoption of open-response assessments (e.g., short answer or essay) to mitigate these limitations (Nehm and Schonfeld 2008, 2010; Nehm and Ha 2011).

Despite the many advantages of open-response evolution assessments (for reviews, see Nehm and Schonfeld 2008; Nehm and Haertig 2011), in practical terms they also carry with them a series of significant disadvantages. These include: (1) rubric development and validation costs; (2) scorer training costs; (3) grading time; (4) rater variability and associated reliability threats; (5) grading fatigue; and (6) interpretation difficulty. Fortunately, new tools and technologies, collectively known as Computer Assisted Scoring (CAS), are capable of addressing many of the aforementioned disadvantages (Page 1966; Yang et al. 2002; Shermis and Burstein 2003). Several commercial CAS tools for large-scale assessment, including C-rater (Sukkarieh and Bolge 2008), E-rater (Burstein 2003), Intelligent Essay Assessor (Landauer et al. 2001), and SPSS Text Analysis (Galt 2008), are being employed with increasing frequency in large-scale educational contexts.

Despite the growing use of CAS in national-level assessment projects, only one study to our knowledge has explored the use of CAS tools at a smaller scale, and in the complex but educationally important domain of evolutionary biology (Nehm and Haertig 2011). Specifically, Nehm and Haertig used SPSS Text Analysis 3.0 (SPSSTA) to detect so-called Key Concepts of natural selection in short answer responses to three prompts from the ORI instrument of Nehm and Reilly (2007). Their study revealed that the text analysis functions (or extraction rules) developed and deployed in SPSS Text Analysis to detect individual Key Concepts (KCs) produced "almost perfect" agreement (Kappas 0.81–1.00) with expert human raters in the majority of analyses (cf. Landis and Koch 1977). These promising findings were dampened by several disadvantages of using SPSS Text Analysis: most notably, the initial cost of the commercial product, and the immense amount of time and expertise required to develop appropriate term libraries and to build text extraction rules capable of performing the text analyses (Nehm and Haertig 2011).

SPSS Text Analysis is not the only program or method available for CAS of evolutionary explanations. One relatively new and freely available tool is the Summarization Integrated Development Environment (SIDE) (Mayfield and Rosé 2010). Unlike SPSS Text Analysis, the SIDE toolkit utilizes machine learning to perform text analysis (Witten and Frank 2005). SIDE is able to identify notable aspects of a text, and use them to perform analyses of text. Among other uses, this can include assigning a text to one of a set of categories (e.g., classifying a movie review as "thumbs up" vs. "thumbs down"), assigning a score to a text based on some metric (e.g., automated essay grading on an A–F scale), or producing a short summary of the most relevant or notable facets that the text contains (e.g., extracting assigned tasks from a meeting transcript).

In the domain of Computer-Supported Collaborative Learning (CSCL), SIDE and its predecessor TagHelper Tools (Donmez et al. 2005) have been used in a variety of ways: to automate or assist in coding student skills (Rose et al. 2005), to assist human moderators in online student discussions (McLaren et al. 2007), to trigger automated support when needed in a group discussion (Kumar et al. 2007), and to identify methods of argumentation in student collaborations (Rose et al. 2008). In contrast to these prior studies, the goal of our work with SIDE is to replace costly human scoring of evolutionary explanations with scoring models built using machine learning.

The machine-learning algorithm of SIDE requires the input of a set of features about an evolutionary explanation, and produces as output a decision about whether that explanation contains the key concepts of interest. The algorithm does this by automatically finding patterns in the text responses, and using these patterns to generate a map linking input features and output judgment. In order to find these patterns, a set of "training examples" must be provided; a human coder needs to evaluate each of the training examples for the presence or absence of each concept. Subsequently, the machine-learning model extracts patterns that are most likely to function effectively. The set of input features must be sufficiently expressive, and machine learning must be able to make use of those features in a general way so that they may extend to new data.

SIDE uses a simple representation, known as a "bag of words" model, to build features. Specifically, a vocabulary list is compiled of each word that appears in the training examples, excluding extremely common words, such as "the" or "and," which are not likely to carry useful information about response content. It also excludes very rare words (in our case, those which occur in fewer than five of our 2,260 training examples). When a new response is assessed, the model marks the presence or absence of each word in the vocabulary list, along with a feature indicating the number of words in the response. The features are then analyzed with a Support Vector Machine (SVM) model, which is considered state of the art in text classification (for more information about this model, see Witten and Frank 2005).

A core difference between SIDE and SPSSTA is that SIDE was designed for what may be termed *confirmatory* text analyses; that is, it functions superbly at elucidating patterns that differentiate sets of previously categorized text responses (e.g., democratic and republican speeches; student papers with high grades and low grades, etc.) and building scoring models based upon those categories. SPSSTA, in contrast, it designed for what may be termed *exploratory* text analysis; that is, when clear categories or dimensions of text are not well established (when a scoring rubric is absent, for example), the program may be used to begin identifying possible terms, categories, and/or themes in the corpus. While machine-learning approaches to this type of analysis exist, under the domain of unsupervised learning, they are not presently incorporated into SIDE. Thus, the two tools approach text analysis quite differently.

In order to utilize SIDE for machine scoring, the only materials that are needed are a corpus of scored responses and the free program. SPSSTA, which costs >$1,000.00 for an individual license, requires much more effort prior to beginning the scoring process. First, the software lacks the scientific vocabulary and verbs necessary for analyzing evolutionary responses; these must be added manually or obtained from another user who has done this work (Nehm and Haertig 2011). Considerable content expertise may be needed to define and build the term and type libraries necessary for text extraction. For example, a library of more than 450 biology terms and types needed to be

manually built in Nehm and Haertig's study using SPSSTA.

After the term and type library is created in SPSSTA, "rules" must also be built that assist the program in identifying which text combinations to tag in student responses. The performance of these rules must be evaluated relative to expert human scoring, and refined accordingly. This process is iterative, and can take considerable time and effort (hundreds of hours). The sophistication of the rules is primarily constrained by human ingenuity, as rules may be quite elaborate. Once rules are obtained and demonstrated to function effectively (i.e., in comparison to expert human scoring) they may be applied to new data sets.

SIDE, in contrast, uses terms, types, and rules that differentiate responses that are discovered automatically using machine learning (Witten and Frank 2005). SIDE attempts to build successful scoring models by discovering patterns in human-coded responses. In so doing, SIDE performs much of the difficult "rule-building" work that is done manually in SPSSTA. There are advantages and disadvantages to this time saving step, however. It is quite difficult to understand the scoring models that SIDE generates. While an interface is provided for exploring model performance within SIDE, it requires considerable background knowledge in machine learning. In comparison, the extensive labor involved in developing SPSSTA types and rules ensures that the researcher is able to directly trace the path from extraction rule to output score. If SIDE could be shown to perform effectively in assessing students' written explanations of evolutionary change, it would offer major financial and time advantages over other tools, such as SPSSTA.

Research Questions

A key assumption when using machine learning tools (such as SIDE) is that the training examples being employed are representative of the responses to which the model will subsequently be applied. A model that is built using responses to one prompt will perform best when evaluating other responses to the same prompt, of approximately the same length and of the same style. It is not always possible, however, to gather a large number of training examples for every prompt that one wishes to assess. Consequently, it is necessary to examine the generalizability of scoring models. One approach to examining model generalizability is to build a scoring model using an entire instrument, and examining the performance of the model on responses to a different instrument. A more fine-grained approach is to focus on individual items within an instrument that are used to build a scoring model—for instance, by building a model based on responses to an item about the evolution of roses, but applying the model to responses to an item about the evolution of snails. In this case, the underlying concept being tested (e.g., the heritability of variation) remains the same but the item context has changed. We used the instrument-focused and item-focused approaches in our study of the efficacy of SIDE.

Our study tests the efficacy of SIDE relative to human expert scoring using student responses to two different evolution instruments (the EGALT-F and EGALT-P) that contain prompts (items) that differ in various surface features (such as species and traits). We test human-SIDE scoring correspondence under four different conditions: Same Prompt, Same Instrument (SPSI); Same Prompt, Different Instrument (SPDI); Different Prompt, Same Instrument (DPSI); and Different Prompt, Different Instrument (DPDI). Additionally, we examine the efficacy of SIDE scoring of responses that differ in length and we calculate the number of responses that are needed to establish "near perfect" Kappa agreement levels (above 0.80) with human expert raters. Finally, we discuss our findings relative to Nehm and Haertig's (2011) related work using SPSS Text Analysis and make recommendations for future work on automated scoring of students' written evolutionary explanations and other performance tasks.

Sample and Methods

We used a corpus of 2,260 evolutionary explanations written by a sample of 565 undergraduates in our analyses. The students who generated these explanations had varying levels of evolution knowledge (specifically, non-majors taking their first college biology course and first-year majors completing a course employing evolution as a core theme). Responses were gathered using an online response system built within our university course management system. The evolutionary explanations that we analyzed were produced in response to a series of prompts about evolutionary change contained in two instruments: the Evolutionary Gain and Loss Test (EGALT) version F (Familiarity) and the EGALT version P (Progression) (For item details, see Nehm et al. 2010; Nehm and Ha 2011).

While the stems of the EGALT instrument items were nearly identical, the species and traits in the items were different both within and between instruments. Specifically, the EGALT-P items prompted explanations of trait gains and losses between the following species/trait/change combinations: Elm/seeds/gain; Rose/thorns/loss; Snail/poison/gain; Penguin/flight/loss. The EGALT-F items requested explanations of trait gains between the following species/trait/change combinations: Snail/poison/gain; Elm/

seeds/gain; Prosimian/tarsi/gain; Labiatae/pulegone/gain. Thus, students' written evolutionary explanations were produced in response to a diverse array of biological surface features.

For the EGALT F, responses were gathered from 320 undergraduate students enrolled in an introductory biology course for majors. Demographically, the sample was 78% White (non-Hispanic; $n = 251$) and 22% minority (African American, $n = 14$; Asian, $n = 28$; Hispanic, $n = 13$; Native American, $n = 1$; Other and non-disclosed, $n = 13$), 55% female, and had an average age of 21 years (SD = 2.3). Each student was administered four prompts (see above) resulting in a total of 1280 responses. For the EGALT P, responses were gathered from 245 undergraduate students also enrolled in an introductory biology course for majors. Demographically, this sample was 75.1% White (non-Hispanic; $n = 184$) and 24.9% minority (African American, $n = 19$; Asian, $n = 28$; Hispanic, $n = 5$; Native American, $n = 1$; Other and non-disclosed, $n = 7$), 57.6% female, and had an average age of 20.7 years. Each student was administered four prompts from the EGALT-P, resulting in a total of 980 responses.

Human Scoring of Evolutionary Explanations

Students' evolutionary explanations are known to contain a diverse array of explanatory elements, ranging from naïve to scientific, and assembled in various combinations and permutations (Nehm and Ha 2011). Our analyses using SIDE employed a construct-grounded approach, in which we sought to identify the elements of scientific explanation considered necessary and sufficient to account for evolutionary change via natural selection (Nehm and Schonfeld 2010). Specifically, three so-called "core concepts" of natural selection are considered necessary and sufficient to explain natural selection: (1) the presence and causes of variation, (2) the heritability of variation, and (3) the differential reproduction and survival of individuals (Patterson 1978; Endler 1992).

Student explanations of evolutionary change were graded by two expert human raters using the scoring rubrics of Nehm et al. (2010). The first expert had a Ph.D. in evolutionary biology, had published in this discipline, and had taught biology for more than a decade. The second expert had a master's degree in biology and had published extensively in the field of evolution education. In terms of scoring, the presence or absence of the three core concepts of natural selection (see above) was established via consensus in all 2,260 student responses between these two human raters. Overall, our study of machine scoring using SIDE examined the detection and measurement of core elements of content (the construct of natural selection) that have been recognized as such by evolution experts (Lewontin 1978; Pigliucci and Kaplan 2006:14; Patterson 1978; Endler 1992). Core concept scores were tallied separately for each item, and collectively for all four items, both pre- and post-course. In addition, the number of different core concepts used among all four items (hereafter: Core Concept Diversity) was scored for each participant.

Measures of Score Correspondence

Inter-rater agreement between human raters is a common metric for evaluating score comparability (Chung and Baker 2003: 28; Krippendorff 2004: 246–249). This approach may also be used to test for human–computer correspondence. Agreement may be quantified using the percentage of exact or adjacent agreements between SIDE scores and human expert scores. Percentage agreement statistics are problematic, however, as they are sensitive to the number of cases analyzed (Yang et al. 2002). Cohen's Kappa, values of which range from 0.0 to 1.0, has consequently been used to quantify levels of agreement between raters because it compensates for chance inter-rater agreements (Bejar 1991). Several different inter-rater agreement benchmarks have been established using the Kappa statistic: Kappa values between 0.61 and 0.80 were considered by Landis and Koch (1977) to be "substantial" and those between 0.81 and 1.00 to be "almost perfect." Krippendorff (1980) likewise followed these benchmarks in his bestselling guide to content analysis. In line with these studies, we consider Cohen's Kappa values between 0.41 and 0.60 to be "moderate", those between 0.61 and 0.80 to be "substantial", and those between 0.81 and 1.00 to be "almost perfect."

Analyses

There are many possible approaches for testing the efficacy of SIDE scoring models relative to expert human raters. Different analyses will reveal different facets of the efficacy and generalizability of a scoring model. Since our goal is to explore the utility of SIDE for automated scoring of evolutionary responses using many different item surface features (i.e., different taxa, traits, and change polarities), we perform several different evaluations comparing expert-generated scores and SIDE scores: (1) Same Prompt, Same Instrument (SPSI) (e.g., SIDE is trained and evaluated on examples from the Rose item from the EGALT-P); (2) Same Prompt, Different Instrument (SPDI) (e.g., SIDE is trained on examples from Elm items from the EGALT-P, and evaluated on Elm items from the EGALT-F); (3) Different Prompt, Same Instrument (DPSI)

(e.g., SIDE is trained on examples for the Elm items and evaluated on examples for Rose items, with both Elm and Rose items coming from *within* the EGALT-P); and (4) Different Prompt, Different Instrument (DPDI) (e.g., SIDE is trained on the Prosimian item within the EGALT-P and evaluated on Rose items within the EGALT-F).

To evaluate the performance of these models, we use a technique known as cross-validation. For each test condition, we partition our training sample into a series of separate subsamples. Then, we build a series of scoring models, each one using all but one of the subsamples. We then measure the accuracy of each model on the held-out subsample, and average the resulting performance values across all of the models. This simulates the performance that we might expect to see on new responses in a real-world scenario. Importantly, we do not change any settings of the types of features used or the model being applied in cross validation. Our goal is to explore the utility of SIDE for automated scoring of many different instruments; therefore we wish to evaluate its performance without the need for human fine-tuning of each model.

We performed one additional analysis of likely importance to others interested in using SIDE for science assessment: the effects of response length on scoring success. Specifically, we examined whether SIDE scoring models function with comparable success on short and long responses.

We used PASW 18.0 to perform several calculations: (1) Kappa agreement statistics between human and SIDE scores of core concepts of natural selection in the student responses; (2) Pearson correlation coefficients between Core Concept Diversity (CCD) measures generated using human experts and SIDE; and (3) regression calculations for examining the relationship between Kappa values and response length. The exact SIDE program settings used in our analyses are included in the Appendix.

Results

Student Explanations of Evolutionary Change

In order to provide readers with a sense of the evolutionary explanations that the students in our sample constructed, we include a series of examples along with human and computer score results (Table 1). As is evident in Table 1, students' explanations included a diverse array of explanatory elements, both scientifically accurate and conceptually naive. For the present study, our scoring of the > 2000 essays focused on the types of scientifically accurate

Table 1 Examples of students' written evolutionary explanations to four different prompts and respective human and computer scores

Prompt	Student's explanation of evolutionary changes	Human key concept score	SIDE key concept score
Familiar animal/trait gain (Snail/poison)	"A random mutation could have occured causing the snail to produce some sort of toxin that is poisonous. This chance mutation then could have increased the animals fitness making it less suseptable to predation. Since it was not eaten it was able to reproduce and pass along the ability to produce poison to its offspring"	4	4
Unfamiliar animal/trait gain (Prosimian/torsi)	"The ancestral prosimian species with short tails may have used the tails in mating rituals. This could mean that the females picked the males with the longer tails and as each generation after that the same process would occur. The choosing of the genetic trait made each generation of the species have longer and longer tails"	3	2
Familiar plant/trait gain (Elm/winged seed)	"A species of elm with winged seeds allows for seed dispersal thus spreading of potential offspring over a vast area. As a result more offspring are potentially scattered and the winged are the median of reproduction dispersal. This is very similar to the coconut floating in water to a distant island then implanting in the soil and maturing into a coconut/palm tree"	1	1
Unfamiliar plant/trait gain (Labiatae/pulegone)	"I do not know what pulegone is. However I'm sure that evolutionary horticulturist would. They would probably explain that pulegone give this particular species of labiatae an competitive edge. Other species of labiatae may not live in the proper setting or interact with the organisms that surround this species of labiatae. Hence this particular species of labiatae is the only one to employ pulegone"	2	2

See methods for complete items and scoring methods

elements in student responses—that is, the so-called "Key Concepts" (KC) of natural selection (ongoing research is exploring misconception detection).

Testing the Effects of Instruments and Items on SIDE Performance

Our presentation of results begins at the coarsest grain size of comparison between human expert scores and SIDE scores—that is, with between-instrument comparisons. Training SIDE using 980 human-scored responses to all four EGALT-P items for the variables KC1, KC2, and KC3 generated scoring models with Kappas of 0.87, 0.81, and 0.88, respectively (Fig. 1). Applying these scoring models to all EGALT-F instrument responses produced Kappas of 0.83 for KC1, 0.68 for KC2, and 0.64 for KC3. In all cases, percentage agreement values exceeded 80%. Reversing the analysis—performing SIDE training on 1280 EGALT-F responses and applying the scoring models for each KC to EGALT-P—produced similar findings (Fig. 1). In both analyses, KC1 agreement values were greater than those for KC2 and KC3 for both the training and the testing phases of the analysis. Overall, the training models were "near perfect" in most instances (Kappa scores above 0.80), but agreement degraded when the scoring models built with one instrument were applied to the another instrument.

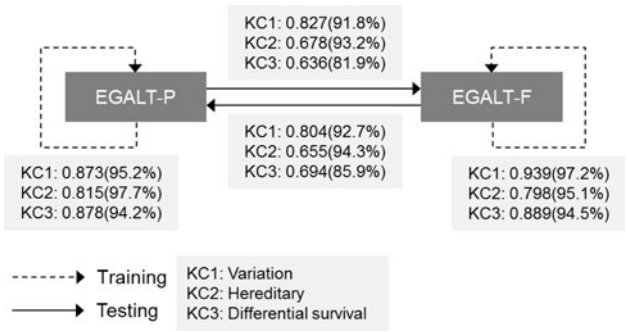

Fig. 1 SIDE performance at the 'whole instrument' scale. Here, training on the EGALT P is applied and tested on the EGALT F, and vice versa. Kappa values and agreement percentages are listed for each of the three Key Concepts (KC) of natural selection (see text for descriptions). Note that *dashed lines* represent model building, and *solid lines* represent model testing

In contrast to our previous analyses, our next set of tests explored correspondence patterns between SIDE scoring models built using all four items within the *same* instrument; that is, the 1,280 responses for EGALT-F were split in two, model building was performed on 50% of the sample ($n = 640$), and model testing was performed on the other 50% of the sample ($n = 640$). This procedure was performed eight independent times in order to compensate for possible sampling effects. For the model-building phase of the analysis, mean Kappa values for the three KCs from EGALT-F were consistently greater than those for EGALT-P, although for both instruments all values were above 0.75 (Fig. 2). Additionally, percentage agreement values always exceeded 90% for all KCs for both instruments. When these scoring models were applied to the other half of the responses, they produced comparable magnitudes of agreement (Fig. 2). Mean Kappa values for EGALT-F for KC1, KC2, and KC3 were 0.93, 0.75, and 0.83, respectively. Mean Kappa values for EGALT-P for KC1, KC2, and KC3 were very similar (0.80, 0.85, and 0.81, respectively). For all KCs, the percentage agreement values exceeded 90% (Fig. 2). Overall, five out of the six tests of the SIDE models exceeded "near perfect" Kappa scores (above 0.80, cf. Landis and Koch 1977).

In addition to analyses of each KC, we measured what we consider to be the most useful and valid indicator of natural selection understanding: Key Concept Diversity (KCD; Nehm and Reilly 2007; Nehm and Schonfeld 2008).

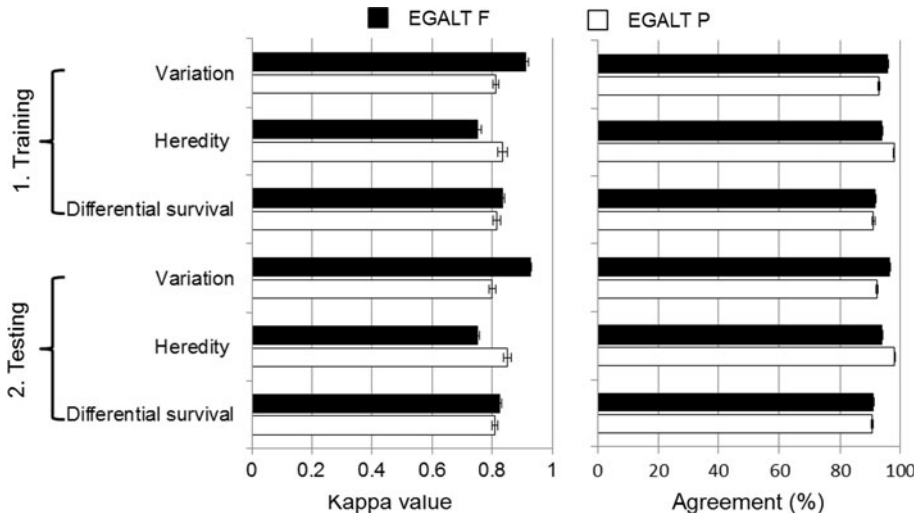

Fig. 2 Scoring model training/building and scoring model testing results for the three core concepts of natural selection for the EGALT F and EGALT P instruments. All four items within each instrument were used in separate analyses. A random selection of one half of the responses within each instrument was selected eight (independent) times to train/build the scoring models, and these models were subsequently tested on the other half of the responses. Displayed are mean values and standard errors, using all four items from each instrument

KCD is a broad measure of evolutionary knowledge: it captures student use of accurate scientific elements across an array of problems differing in surface features (Nehm and Ha 2011). For this reason, the efficacy of SIDE in evaluating KCD is of particular importance. Diversity refers to the number of necessary and sufficient explanatory concepts that are employed in an evolutionary explanation. Notably, KC1, KC2, and KC3 must be used together to produce a complete scientific answer. Using a training/testing sample split as described above, we compared human-expert and SIDE KCD scores using Pearson correlation coefficients. Mean KCD correlation coefficients for EGALT-F ranged from 0.87 to 0.92 (mean = 0.90) and those for EGALT-P ranged from 0.83 to 0.89 (mean = 0.86). All correlations were significant at $p < 0.01$. Thus, for what we consider to be the most important measure of evolutionary explanations, SIDE performance was outstanding.

Our next analyses examined the performance of SIDE at a finer grain size: we tested how SIDE scoring models that were built on sets of items within an instrument (e.g., elm + rose + penguin) functioned when applied to *different* items within the *same* instrument (e.g., Snail). In the case of EGALT-P, 735 responses were used to build scoring models (Fig. 3). These models were subsequently tested on a different single item from the same instrument ($n = 245$ in all analyses). In the case of EGALT-F, 960 responses were used to build scoring models based on the item sets. These models were subsequently tested on a different, single item from the EGALT-*F* ($n = 320$ in all cases). Scoring models built using item sets from EGALT-P displayed Kappa values ranging from 0.81 to 0.84 for KC1, 0.83–0.88 for KC2, and 0.81–0.88 for KC3 (Fig. 3, left column). Scoring models built using item sets from EGALT-F displayed Kappa values ranging from 0.93 to 0.94 for KC1, 0.77–0.81 for KC2, and 0.84–0.88 for KC3 (Fig. 3, right column). Tests of the scoring models (separately applied to the snail, elm, rose, and penguin items of EGALT-P) produced Kappa values ranging from 0.80 to 0.87 for KC1, 0.85–0.91 for KC2, and 0.68–0.77 for KC3. Tests of the scoring models (separately applied to the snail, elm, prosimian, and labiatae items of EGALT-F) produced Kappa values ranging from 0.88 to 0.94 for KC1, 0.62–0.86 for KC2, and 0.76–0.88 for KC3. In general, the lowest levels of correspondence between human and SIDE scores occurred in KC3 for the EGALT-P, but KC2 for EGALT-F. Overall, model building and model testing were influenced to some extent by items and instruments (Fig. 3). In the majority of model building and model testing analyses, however, "near perfect" agreement levels were reached.

Our final analyses focused on using individual items to build scoring models and to perform model tests on different items (both within and between instruments). For items in the EGALT-P instrument, 245 responses were used and for EGALT-F 320 responses were used. The results of this analysis are quite complex, given the number of permutations possible with two instruments, eight items, and three concepts (Fig. 4). In order to capture a broad picture of the findings, we first draw the reader's attention to the cell shadings illustrated in Fig. 4. The darkness of the shadings reflects the level of score agreement, with the darkest cells representative of the highest Kappa values (which are also shown numerically within each cell). We

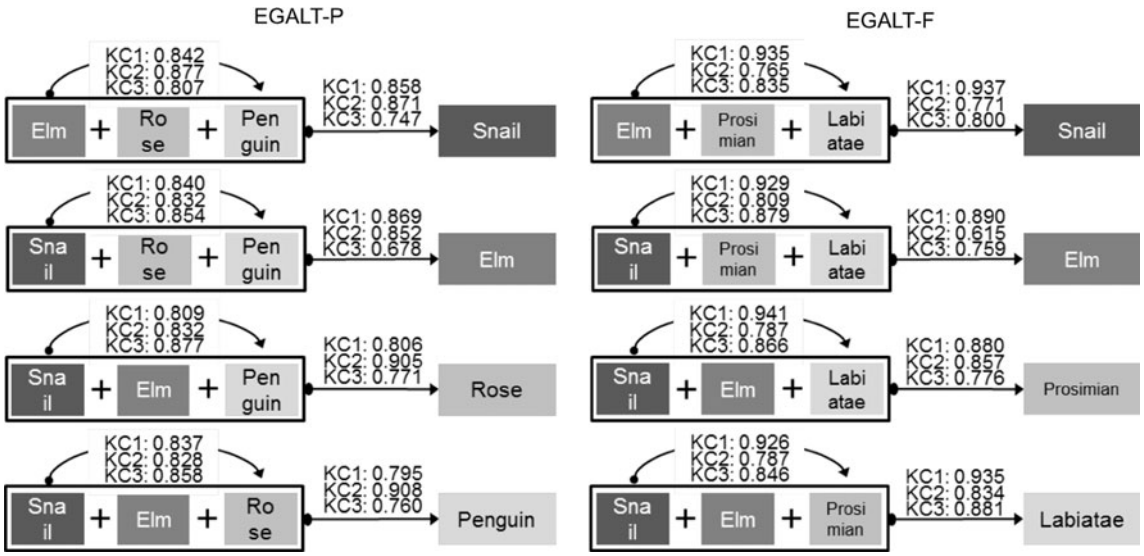

Fig. 3 Within instrument, different prompt comparisons. **a** EGALT P elm + rose + penguin ($n = 735$) trained and tested on snail ($n = 245$). **b** egalt F. Elm + prosimian + labiatae ($n = 960$) trained and tested on snail, etc. ($n = 320$)

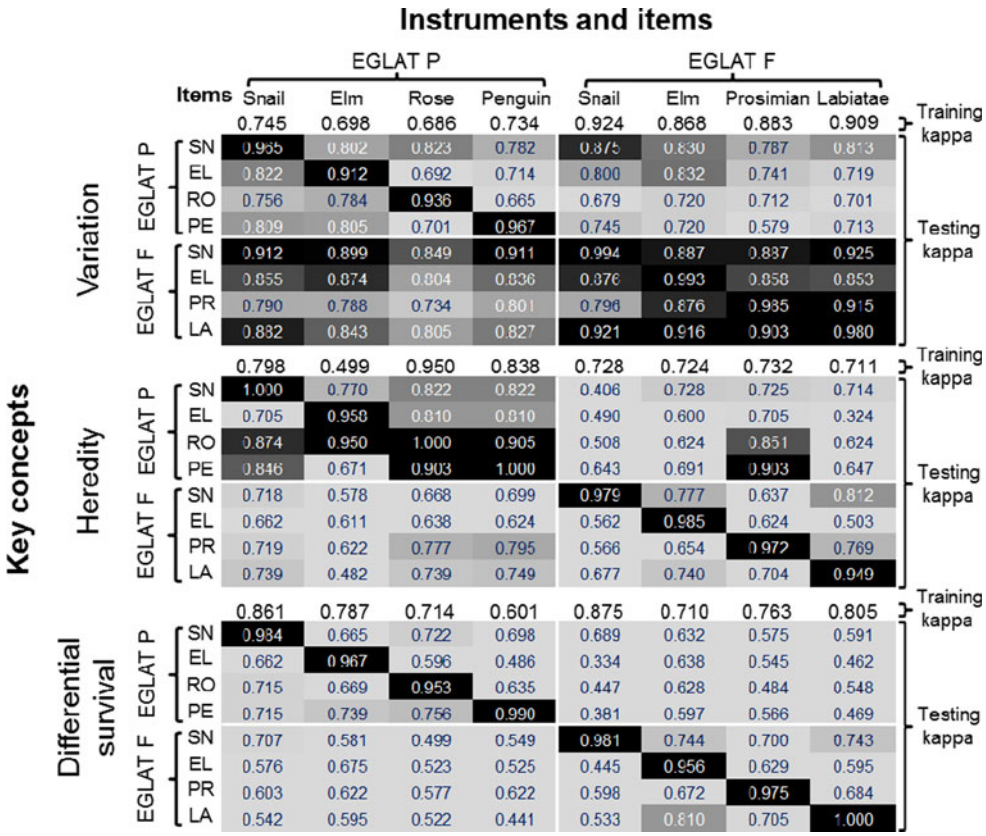

Fig. 4 Comparisons among instruments, items, and key concepts (variation, heredity, differential survival). EGALT P (all items, $n = 245$); EGALT F ($n = 320$). Cell darkness reflects Kappa scores, with darker cells having higher Kappas

note several patterns of interest in Fig. 4. First, dark diagonal "steps" may be observed in all three of the key concept panels (upper, middle, and lower). This pattern is a product of cases in which model testing was performed on the *same* item type from the *same* instrument (i.e., training on the EGALT-P snail item and testing on the EGALT-P snail item; recall that the same responses were not used because of sample splitting). This result is not surprising, as training and testing were performed on the most similar data.

The second pattern of note in Fig. 4 is the decreasing number of darkly shaded cells as one compares the top, middle, and bottom panels; KC1 (Variation) has many darkly shaded cells whereas KC3 (Differential survival) has few darkly shaded cells. This pattern indicates that KC3 scoring models are the most sensitive to item surface feature and instrument changes. In contrast, KC1 scoring models are the least sensitive to these changes; here, the vast majority of model tests produced high Kappa agreement levels between SIDE and human scorers.

The third pattern that we note in Fig. 4 is the effect of changing instruments on scoring model performance. Examining the upper panel in Fig. 4, it is apparent that items on the EGALT-F have greater agreement correspondences than items on EGALT-P for KC1. This is indicated by the darker shading in the lower right hand corner of the KC1 panel relative to the shading in the lower left hand corner of the KC1 panel. Instrument effects may also be noted for KC2 (Heredity). In this case, EGALT-P items outperform EGALT-F items. In the case of KC3 (differential survival), model performance is less sensitive to which instrument is used than for the other KCs. However, it is highly item specific.

Overall, these analyses revealed two patterns: First, instruments and their constituent item surface features in some cases significantly influence the performance of SIDE in scoring evolutionary explanations. Second, training SIDE using specific items (e.g., "Rose") and applying the model to the same items produces the highest degree of correspondence with human expert scores.

Response Length and SIDE Performance

Our final analysis explored whether SIDE scoring models functioned with comparable efficacy when applied to responses varying in length. SIDE scoring models were built for KC1, KC2, and KC3 using all 1,280 responses to EGALT-F. Similarly, scoring models were built for the three KCs using all 980 responses to EGALT-P. These scoring models were subsequently tested on subsamples (from the corresponding instrument) that differed in length. Specifically, the responses to each instrument were

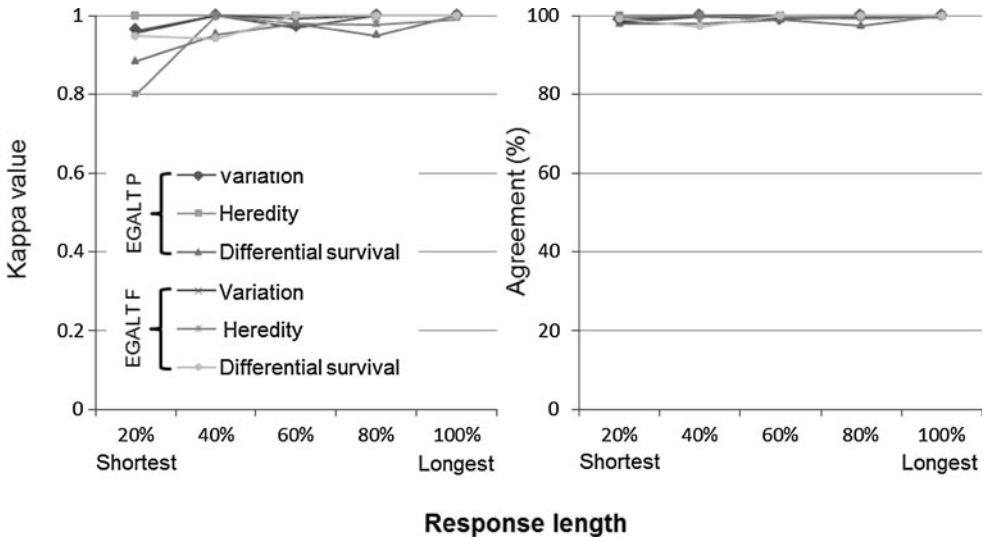

Fig. 5 Scoring models built using the entire sample of each instrument were applied to subsamples sorted by response length. While the scoring models were least effective in the shortest response subsample, all tests exceeded Kappas greater than 0.80

partitioned into five groups, ranging from the shortest 20% to the longest 20% in the sample (Fig. 5). Scoring models built with the EGALT-F were tested on various EGALT-F subsamples, as were those for the EGALT-P. As Fig. 5 illustrates, scoring model performance (measured using Kappa and percent agreement) did not differ substantially among the five subsamples: in all cases, Kappa values exceeded 0.80 and percentage agreement values exceeded 95%. While scoring models built using the entire sample performed well in all length conditions, the scoring models were least effective in the shortest response subsample (Fig. 5).

Discussion

We begin our discussion of SIDE scoring efficacy at the broadest scale of analysis: whole-instrument comparisons. At this scale, SIDE was trained using responses to the EGALT-P instrument, and the resulting scoring model was tested using the responses to EGALT-F (and vice versa) (Fig. 1). Despite outstanding model-building Kappa scores (Fig. 1), SIDE performance tests did not meet the Kappa benchmark (scores above 0.80) in four of the six comparisons with human expert scores. Thus, scoring models built at the whole-instrument level did not perform well using our response corpus. This is not surprising, as the prompts and their surface features across instruments are quite diverse.

Given the moderate performance of scoring models built at the whole-instrument level, our second set of analyses examined SIDE scoring model efficacy *within* each instrument (EGALT-P or F) for particular Key Concepts (KC). Scoring performance within instruments was superior to the whole-instrument comparisons, with 83% (5/6) of KC tests meeting or exceeding the Kappa performance benchmark of above 0.80 (Fig. 2). Interestingly, model performance across KCs was not consistent between instruments; the KC1 scoring model demonstrated the best performance in the EGALT-F but the worst performance in EGALT-P. Nevertheless, within-instrument scoring models performed very well in most cases.

Our third set of analyses focused on SIDE scoring of individual items (e.g., rose) using models built from combinations of different items (e.g., snail + elm + penguin). In these analyses we found that 66% (16/24) of scoring model tests met our Kappa benchmark (Fig. 3). While KC1 met our benchmark in all comparisons at this scale, KC2 and KC3 performance was variable. Thus, in many cases, SIDE scoring models built from item combinations failed to match our performance target.

Our final analyses focused on scoring model performance at the scale of individual items. SIDE performance was outstanding when training models were built and tested on the same items, with 100% of tests (24/24) exceeding our Kappa benchmark of above 0.80 (Fig. 4, see the diagonal cells in the upper, middle, and lower panels). It is unlikely that additional human scoring of these items will be necessary given their robust scoring models. Our analysis also revealed that KC1 training models were most effective across items, but this was not the case for KC2 or KC3. Overall, it appears that in the case of student-generated evolutionary explanations, SIDE is most effective when trained on the same items that it will subsequently be asked to score.

We have established robust scoring models for six different evolution items. Given that SIDE can score large response sets (>1,000) in less than 5 min, larger-scale testing of the generalizability of these scoring models should begin. Specifically, samples from different

geographic regions, more racially diverse samples, and varying levels of content preparation should be examined.

Use of these items and their associated SIDE scoring models should save considerable time and money. Following the scoring time and cost estimates of Nehm and Haertig (2011), employing and training a human rater to score 3,000 responses is estimated to take 100 h and to cost $2,000.00. Given that many introductory biology programs, including our own, educate thousands of students per year, the financial benefits of using SIDE to score open response evolution assessments is clear. But perhaps the greatest positive effect of using SIDE and related computerized assessment tools is that science assessments can begin to measure more complex scientific reasoning processes, such as the construction of evolutionary explanations, and shift away from measuring students' abilities to select isolated fragments of knowledge in multiple choice tests (NRC 2001, 2007; Alberts 2010; Nehm and Haertig 2011). Thus, SIDE may be used to leverage reform in the teaching and learning of biology.

Improving SIDE Scoring Models with Human-Built Feature Spaces

It is possible to augment the scoring models automatically built using SIDE, and thereby enhance scoring performance. The machine-learning methods in SIDE make inferences based upon the training examples that are provided, without any prior "knowledge" about the key words or phrases that might be helpful in distinguishing response types. Human experts may be able to detect rare but related text elements, and group them together into sets of key terms, so that the scoring model can recognize a pattern that would otherwise be obscured or missed. A second approach is to have domain experts identify—prior to computer analysis—those sections of the text that are most important for detecting the presence or absence of a concept. Prior work in other content areas has demonstrated that this interactive approach can improve scoring model performance (Arora and Nyberg 2009).

A drawback of human expert augmentation is that it requires expert time and resources to build each new machine-learning model. While all of the scoring models that we presented were produced automatically (with no human 'tuning' of feature spaces or models), an interactive approach would require expert analysis of many training examples prior to beginning machine learning. Preliminary augmentation attempts using SIDE demonstrated that while kappas can be increased a small amount through brief episodes of expert 'tuning' (often enough to nudge model performance from under to over our "near perfect" agreement benchmark level above 0.80, cf. Landis and Koch 1977). Dramatic increases in kappas (e.g., +0.2) were less easily obtained, however. Nevertheless, additional work is clearly needed to empirically investigate approaches to optimization of machine learning using human expert augmentation.

Advantages and Disadvantages of SIDE Relative to SPSS Text Analysis

In the only other study of computerized scoring of students' written evolutionary explanations, Nehm and Haertig (2011) explored the utility of SPSS Text Analysis for Surveys 3.0 (SPSSTA). For science education researchers interested in using text analysis to score students' written assessment responses, a general discussion of the advantages and disadvantages of the SPSSTA and SIDE programs may be helpful. Nehm and Haertig (2011) found that their text extraction libraries and rules built in SPSSTA were able to detect Key Concepts (KC) of natural selection in students' responses at comparable magnitudes to those from SIDE (i.e., Kappa scores above 0.80 in a majority of cases). Even though these two studies used different samples of student responses generated from different instruments (EGALT vs. ORI), both software packages functioned quite well at detecting KCs of natural selection. Both studies demonstrate the utility of using text analysis programs for scoring written explanations in biology.

There are costs and benefits of each program, however, that researchers will want to consider in choosing between these two programs (Table 1). Overall, for researchers who have clear and robust scoring rubrics and scored responses, SIDE appears to be much more cost and time effective. SPSSTA, on the other hand, appears to be the more appropriate tool for researchers who have not built rubrics for scoring responses, or who have vague models of the text features that define excellent or poor responses, for example. These broad generalizations should be interpreted with caution, however, as our experiences in the complex domain of evolutionary biology may not apply to other content areas (Table 2).

Implications for Assessments Beyond Evolution

The outstanding performance of SIDE in scoring evolutionary explanations suggests that it may have broad applicability in other domains, contexts, and educational levels, both within and outside of the biological sciences. In particular, assessment tasks in which clear criteria (i.e., rubrics) have been established, and corpora of scored responses are available, would be ideal candidates for testing the efficacy of SIDE relative to human scoring. Given that many assessments include rubrics that have been used on actual responses, and that SIDE is relatively

Table 2 Attributes of SIDE in comparison to SPSS text analysis for surveys 3.0

Features	SPSS text analysis	SIDE
Primary purpose	Exploration of text features and confirmation of text features	Confirmation of text features
Initial product cost	>$1,000.00 (personal license)	Free
Training time	Similar effort required	
Term library and rule creation	>100 h	Not needed
Human scoring time	Similar effort required	
Scoring rubric creation	Similar effort required	
Human/computer agreement statistics (e.g., Kappa)	Not provided	Provided
Key concept detection performance	Similar: Majority of cases "near perfect" (Kappas > 0.80)	

easy to use (and free), we see machine learning as a potentially transformative tool in advancing the assessment of more authentic, ill-structured problem solving tasks in many contexts (cf. NRC 2001).

In many respects, SIDE is unique in that it may solve the problem of developing innovative assessments at an "intermediate" scale and budget. In contrast, large companies, such as Educational Testing Service, will continue to develop large-scale, big budget, norm-referenced tests, and small schools will continue to hand-score performance assessments. But those stakeholders at an intermediate scale—public universities, academic departments, or school districts—may not have the budgets or expertise to build national-level, large-scale innovative assessments. SIDE may effectively fill this gap, as has been the case at our university; it may be easily adapted to 'home-grown' assessments in which rubrics and scored responses are available. We are hopeful that others will take advantage of this innovative assessment tool and explore its efficacy in other domains.

Conclusions

Many high-stakes, multiple-choice assessments are severely constrained in their ability to measure thinking and communication skills essential to success in real world problem-solving environments. Consequently, new assessment types and methods are needed that are capable of measuring more complex performances (Wagner 2008; Gitomer and Duschl 2007). Our study examined the efficacy of a new software tool (SIDE) in automatically scoring students' written explanations of evolutionary change. SIDE performance was most effective when scoring models were built and tested at the individual item level. Performance degraded when suites of items or entire instruments were used to build scoring models. SIDE performance was outstanding for what we consider to be the most important measure of evolutionary explanations: Key Concept Diversity (KCD). KCD measures the number of necessary and sufficient explanatory concepts that are employed in a written evolutionary explanation. When using SIDE for confirmatory text detection, as we did, it offers many advantages compared to commercial text analysis programs such as SPSS Text Analysis for Surveys. In the case of evolutionary explanations, SIDE scoring performance was found to be equivalent or superior to that of SPSSTA and required less time and financial investment. Technological tools such as SIDE have great potential in shifting the focus of assessments to more authentic, real world problem-solving tasks.

Acknowledgments We thank the faculty and participants of the 2010 PSLC (NSF Pittsburgh Science of Learning Center) summer school for financial and intellectual support; Prof. Carolyn Penstein Rosé for introducing us to the SIDE program; NSF REESE grant 0909999 for financial support.

Appendix

The SIDE program and user's guide may be downloaded at: http://www.cs.cmu.edu/~cpRosé/SIDE.html. Specific SIDE settings for performing the analyses in this study include: The machine-learning algorithm was selected as "weka-classifiers-functions-SMO"; options included: (1) "unigrams"; (2) "treat above features as binary"; (3) "line length"; (4) "remove stopwords"; and (5) "stemming" (details of these features may be found in Mayfield and Rosé 2010, p. 6). We also used the feature extractor plugin option "plugin.sample.fce.TagHelperExtractor"; This default option creates a feature table based upon the NLP extractions mentioned above. We also selected the "Remove rare features" option and set the value of 5, and, as noted above, chose the machine-learning algorithm "weka-classifiers-functions-SMO." We selected Cross-validation and set the value at 10. For the Default segmenter option, we selected "plugin.sample.segmenter. DocumentSegmenter" (Mayfield and Rosé 2010).

References

Alberts B (2010) Reframing science standards. Science 329(5991): 491

Arora S, Nyberg E (2009) Interactive annotation learning with indirect feature voting. In: Paper in the proceedings of student research symposium at NAACL-HLT 2009, Boulder, Colorado, USA. Accessed online at: http://www.cs.cmu.edu/%7Eshilpaa/NAACL_SRW_IAL.pdf

Bejar II (1991) A methodology for scoring open-ended architectural design problems. J Appl Psychol 76(4):522–532

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

Burstein J (2003) The e-rater scoring engine: automated essay scoring with natural language processing. In: Shermis MD, Burstein J (eds) Automated essay scoring: a cross-disciplinary perspective. Lawrence Erlbaum Associates, Inc, Mahwah, pp 113–122

Chung GKWK, Baker EL (2003) Issues in the reliability and validity of automated scoring of constructed responses. In: Shermis MD, Burstein J (eds) Automated essay scoring: a cross-disciplinary perspective. Erlbaum, Mahwah, pp 23–40

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

Demastes SS, Good RG, Peebles P (1995) Students' conceptual ecologies and the process of conceptual change in evolution. Sci Educ 79(6):637–666

Donmez P, Rosé C, Stegmann K, Weinberger A, Fischer F (2005) Supporting CSCL with automatic corpus analysis technology. In: Paper in proceedings of the international conference on computer support for collaborative learning (CSCL), Taipei, Taiwan

Endler JA (1992) Natural selection: current usages. In: Keller EF, Lloyd EA (eds) Keywords in evolutionary biology. Harvard, Cambridge, pp 220–224

Galt K (2008) SPSS text analysis for surveys 2.1 and qualitative and mixed methods analysis. J Mixed Meth Res 2(3):284–286

Gitomer DH, Duschl RA (2007) Establishing multilevel coherence in assessment. In: Moss PA (ed) Evidence and decision making. The 106th yearbook of the National Society for the Study of Education, Part I. National Society for the Study of Education, Chicago, pp 288–320

Krippendorff K (1980) Content analysis: an introduction to its methodology, 1st edn. Sage Publications, Thousand Oaks

Krippendorff K (2004) Content analysis: an introduction to its methodology, 2nd edn. Sage Publications, Thousand Oaks, London

Kumar R, Rosé C, Wang YC, Joshi M, Robinson A (2007) Tutorial dialogue as adaptive collaborative learning support. In: Paper in proceedings of the international conference on artificial intelligence in education, Los Angeles, USA

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

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

Lewontin R (1978) Adaptation. Sci Am 239:212–228

Liu OL, Lee HS, Hofstetter C, Linn MC (2008) Assessing knowledge integration in science: construct, measures, and evidence. Educ Assess 13(1):33–55

Markoff J (2011) Computer wins on 'jeopardy!': trivial, it's not. New York Times, 16 Feb

Mayfield E, Rosé C (2010) An interactive tool for supporting error analysis for text mining. In: Paper in proceedings of the demonstration session at the international conference of the North American Association for Computational Linguistics (NAACL), Los Angeles, USA

McLaren B, Scheuer O, de Laat M, Hever R, de Groot R, Rosé C (2007) Using machine learning techniques to analyze and support mediation of student e-discussions. In: Paper in proceedings of the international conference on artificial intelligence in education, Los Angeles, USA

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

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

National Research Council (2008) Rising above the gathering storm: energizing and employing America for a brighter economic future. National Academy Press, Washington, D.C.

Nehm RH (2010) Understanding undergraduates' problem solving processes. J Biol Microbiol Educ 11(2):119–122

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

Nehm RH, Haertig H (2011) Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. J Sci Educ Technol. doi:10.1007/s10956-011-9282-7

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

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, Schonfeld IS (2010) The future of natural selection knowledge measurement: a reply to Anderson et al. J Res Sci Teach 47(3):358–362

Nehm RH, Ha M, Rector M, Opfer J, Perrin L, Ridgway J, Mollohan K (2010) Scoring guide for the open response instrument (ORI) and evolutionary gain and loss test (EGALT). Technical Report of National Science Foundation REESE Project 0909999. Accessed online 10 Jan 2011 at: http://evolutionassessment.org

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

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

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

Rose C, Donmez P, Gweon G, Knight A, Junker B, Cohen W, Koedinger K, Heffernan N (2005) Automatic and semi-automatic skill coding with a view towards supporting on-line assessment. In: Paper in proceedings of the international conference on artificial intelligence in education, Amsterdam, The Netherlands

Rose CP, Wang YC, Cui Y, Arguello J, Stegmann K, Weinberger A, Fischer F (2008) Analyzing collaborative learning processes automatically: exploiting the advances of computational linguistics in computer-supported collaborative learning. Int J Comput Support Collab Learn 3(3):237–271

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

Sukkarieh J, Bolge E (2008) Leveraging c-rater's automated scoring capability for providing instructional feedback for short constructed responses. In: Woolf BP, Aimeur E, Nkambou R, Lajoie S (eds) Lecture notes in computer science: vol. 5091. Proceedings of the 9th international conference on intelligent tutoring systems, ITS 2008, Montreal, Canada, June 23–27, 2008. Springer, New York, pp 779–783

The Conference Board, Corporate Voices for Working Families, the Partnership for 21st Century Skills, and the Society for Human

Resource Management (2007) Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century workforce. Accessed online 22 Mar 2011 at: http://www.p21.org/index.php?option=com_content&task=view&id=250&Itemid=64

Wagner T (2008) The global achievement gap. Basic Books, New York

Witten IH, Frank E (2005) Data mining, 2nd edn. Elsevier, Amsterdam

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