J Sci Educ Technol (2012) 21:56–73
DOI 10.1007/s10956-011-9282-7

Human vs. Computer Diagnosis of Students' Natural Selection Knowledge: Testing the Efficacy of Text Analytic Software

Ross H. Nehm · Hendrik Haertig

Published online: 28 January 2011
© Springer Science+Business Media, LLC 2011

Abstract Our study examines the efficacy of Computer Assisted Scoring (CAS) of open-response text relative to expert human scoring within the complex domain of evolutionary biology. Specifically, we explored whether CAS can diagnose the explanatory elements (or Key Concepts) that comprise undergraduate students' explanatory models of natural selection with equal fidelity as expert human scorers in a sample of >1,000 essays. We used SPSS Text Analysis 3.0 to perform our CAS and measure Kappa values (inter-rater reliability) of KC detection (i.e., computer–human rating correspondence). Our first analysis indicated that the text analysis functions (or extraction rules) developed and deployed in SPSSTA to extract individual Key Concepts (KCs) from three different items differing in several surface features (e.g., taxon, trait, type of evolutionary change) produced "substantial" (Kappa 0.61–0.80) or "almost perfect" (0.81–1.00) agreement. The second analysis explored the measurement of human–computer correspondence for KC diversity (the number of different accurate knowledge elements) in the combined sample of all 827 essays. Here we found outstanding correspondence; extraction rules generated using one prompt type are broadly applicable to other evolutionary scenarios (e.g., bacterial resistance, cheetah running speed, etc.). This result is encouraging, as it suggests that the development of new item sets may not necessitate the development of new text analysis rules. Overall, our findings suggest that CAS tools such as SPSS Text Analysis may compensate for some of the intrinsic limitations of currently used multiple-choice Concept Inventories designed to measure student knowledge of natural selection.

Keywords Computer scoring · Text analysis · Assessment · Knowledge measurement · Concept inventories · Evolution · Natural selection

Introduction

Recent reform documents emphasize that STEM education must place greater emphasis on the teaching, learning, and assessment of critical content or "core ideas" (NRC 2001, 2007). These reform documents also highlight the urgent need for assessment tools that harness advances in technology and are guided by cognitive models of progression towards competence (NRC 2007). Development of such assessment tools is particularly important for (1) revealing critical junctures in the development of student conceptual understanding and (2) measuring the instructional efficacy of teaching such core ideas. In line with these reform documents, our study explores a central problem in STEM education—assessing students' cognitive models of natural selection—that must be addressed in order for substantial progress to be made in the teaching and learning of this extremely important but greatly misunderstood "core idea" in biology.

Despite the unequivocal recognition that natural selection is a "core idea" in the biological sciences, learners at all levels of the educational hierarchy throughout the world—from high school to medical school, from America to New Zealand—are characterized by low levels of

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

H. Haertig
Institute of Physics Education, University of Duisburg-Essen,
Essen, Germany
e-mail: hendrik.haertig@uni-due.de

www.manaraa.com

understanding of natural selection, as well as myriad misconceptions (e.g., Grose and Simpson 1982; Brumby 1984; Clough and Wood-Robinson 1985; Zimmerman 1987; Bishop and Anderson 1990; Demastes et al. 1995; Dagher and BouJaoude 1997; Sinatra et al. 2003; Newport 2004; Nehm and Reilly 2007; Nehm and Schonfeld 2008, 2010 Nehm et al. 2010a, b). The increasing importance of natural selection within many fields of science—not just biology—is paradoxically coupled with persistent public confusion about it. Consequently, much work has focused on the teaching and learning—but not assessment—of evolutionary ideas (Nehm 2006; Donnelly and Boone 2007).

The development and rigorous evaluation of instruments that measure knowledge of, and misconceptions about, natural selection in learners of different ages and educational backgrounds remains a comparatively peripheral focus of evolution education research (Nehm 2006). This dearth of attention to assessment makes evaluation of the effectiveness of national and state standards, as well as particular pedagogical strategies used to teach natural selection, difficult if not impossible. The little work that has been done in recent years relating to evolution knowledge measurement has been directed at the development of multiple-choice Concept Inventories (CI) (for a review, see D'Avanzo et al. 2008). CIs have been increasingly used to measure student knowledge in many content areas, not just natural selection (see Liu 2010 for many examples). Given their practical nature (e.g., small item sets that are easy to score) CI development and use has become popular in STEM education. Nevertheless, many CIs have limitations that may be compensated for using new technological tools.

Concept Inventories in Science Education

The growing popularity of CI development and use in the sciences may at first glance call into question the need for other formats and methods, such as open-response Computer Assisted Scoring (D'Avanzo et al. 2008). In at least some instances, however, so-called Concept Inventories are not designed to measure in any meaningful sense student understanding of a *concept* (such as natural selection; Nehm and Schonfeld 2010). Rather, many CIs are checklists of disarticulated fragments of student thinking about a range of concepts and alternative conceptions across very different problem types. Moreover, such problem types are often characterized by an amalgamation of very different contexts and surface features (Nehm and Schonfeld 2010; Chi et al. 1981). Furthermore, CIs nearly always offer concise menus of very limited answer choices. But just as perusing a restaurant menu may reveal that your favorite entre is excluded, CIs may not offer students' true preferences. Likewise, even if students' two "favorite dishes" *are* on the menu—a healthy entre and a unhealthy dessert—typically they are only permitted to choose one option, despite the well established finding that many students—in some cases majorities—harbor both accurate *and* inaccurate ideas about a particular concept or problem (Nehm and Reilly 2007; Ha and Cha 2009; Nehm and Ha 2011; Nehm et al. 2009). Consequently, CIs in many cases are incapable of providing a holistic or authentic snapshot of student understanding of a concept; that is, what food choices *they* would select to compose a meal. Rather, CIs often reveal a hodgepodge of accurate and contextually inaccurate knowledge elements constrained by the options provided (Nehm and Schonfeld 2010).

But there is a more serious concern with some CIs that *do* attempt to target a particular concept or causal theory (e.g., natural selection). Even if such CIs were able to validly measure the elements of student thinking in comparable contexts, evidence indicative of knowing all of the "pieces" or elements of a causal theory does not necessarily provide evidence about how (or *if*) students think these elements work together (Resnick and Resnick 1992). That is, CIs may not reveal students' abilities in regard to the degree to which they can assemble the pieces of a concept into a coherent and functional explanatory structure.

Moreover, even if such CIs could measure students' abilities to integrate fragmented knowledge elements, such abilities would not necessarily be indicative of whether students can actually *apply* their knowledge of a concept to a particular context or problem. Thus, as noted by Nehm and Schonfeld (2010), science educators should be less interested in CIs that can only reveal isolated fragments of student thinking and be more interested in instruments that can reveal how students choose to assemble and employ these elements in explanatory models across different contexts (e.g., in the classroom and in the "real world").

Finally, the newly proposed *Framework for Science Education* (Alberts 2010) has identified several so-called "critical strands" that have been missing from science education and are in urgent need of curricular integration. These include: "generating and evaluating scientific evidence and explanations, understanding the nature and development of scientific knowledge, and participating in scientific practices and discourse." (Alberts 2010:491). Notably, extant multiple-choice CIs are unable to assess these core aspects of scientific practice, particularly *generating scientific explanations*. Constructed response assessments, however, may be suitable to this task.

Constructed Response Solves Some CI Constraints

Constructed response instruments may help to solve some (but not all) of the intrinsic constraints of extant CIs in

المنارة للاستشارات
www.manaraa.com

evolutionary biology. When confronted with a prompt designed to elicit an explanation of how evolutionary change has occurred, interviews with students indicate that they do indeed consider such problems to require a causal explanation, which they subsequently provide (Nehm and Schonfeld 2008). Furthermore, such interviews demonstrate strong associations with written constructed-response answers (Nehm and Schonfeld 2008). Such written responses therefore provide insights into: (1) which knowledge elements students consider to be important to explain a phenomenon (as opposed to the selection of a prescribed "either-or" menu of options); (2) how students assemble these elements into an explanation; and (3) how the explanation is applied to the context or situation represented in the item (Kirsh 2009; Nehm and Ha 2011). Consequently, some open-response instruments have been shown to display greater correspondence to clinical interviews than multiple-choice CIs (Nehm and Schonfeld 2008, 2010).

Bridgeman (1992:253) outlined several additional reasons why constructed-response formats may minimize the "false positives" noted in some multiple-choice assessments (e.g., Nehm and Schonfeld 2008): (1) they reduce measurement error associated with random guessing; (2) they eliminate unintended corrective feedback, that is, if an expected incorrect answer is not present in the menu options, the student knows that a change in strategy is required to correctly solve the problem; and (3) they prevent students from working backwards from the answers. Consequently, a large body of psychometric research argues for the inclusion of constructed response items in knowledge measurement (e.g., Traub and MacRury 1990; Morgan and Maneckshana 1996; Kuechler and Simkin 2004).

While many science educators undoubtedly recognize some of the intrinsic benefits of open-response instruments in knowledge assessment, many reasons account for their uncommon use with large populations. These include: (1) grading time and cost; (2) scorer training costs; (3) the complexity of rubric development and evaluation; (4) inconsistent scores among raters due to differences in scorer expertise and subjectivity; (5) grading fatigue; and (6) responses that may be difficult to interpret. Fortunately, new tools and technologies, collectively known as Computer Assisted Scoring, are capable of solving many of the aforementioned problems.

Computer Assisted Scoring

Announcements of the impending revolution in computer assisted scoring (CAS)—begun in the 1960s—are justified at long last (Page 1966; Yang et al. 2002; Shermis and Burstein 2003). Several CAS tools, notably C-rater (Sukkarieh and Bolge 2008), E-rater (Burstein 2003), Intelligent Essay Assessor (Landauer et al. 2001), and SPSS Text Analysis (SPSS Inc 2006), are being employed with increasing frequency in educational contexts. Moreover, CAS and related tutoring systems are beginning to administer, capture, and analyze more advanced performance skills in large populations, particularly in medical fields and in higher education (Clauser et al. 2000; Mislevy et al. 2002; Braun et al. 1990).

The growing use of CAS tools in many academic disciplines is driven in part by the numerous disadvantages that characterize human scoring of constructed response items, most notably the high costs (in terms of scoring time and expert training) and delayed feedback to test takers. Furthermore, human scoring is problematic for many reasons, including grading fatigue, inconsistent training and/or background knowledge of graders, and the intrinsic subjectivity associated with interpretation (Yang et al. 2002). Consequently, the development CAS has been justified by its purported ability to compensate for the weaknesses of human scoring by producing greater reproducibility, objectivity, reliability, and efficiency (Williamson et al. 1999; Powers et al. 2002a, b). Finally, the repeatedly documented comparability of computer-administered and paper and pencil administered test scores provides further justification for making use of such readily available electronically formatted responses (Keith 2003; Kingston 2009).

The persistent question asked of CAS systems is whether they can measure written responses as accurately as human scorers. The validation of CAS methods has been approached from many angles. The most straightforward approach quantifies levels of agreement between CAS scores and the scores generated by trained experts. Agreement may be quantified using the percentage of exact or adjacent agreement between CAS scores and human expert-generated scores. These measures have their disadvantages, however; most notably, they are influenced by the number of cases analyzed and score distributions (Yang et al. 2002). Consequently, Cohen's Kappa has been employed to quantify levels of agreement between CAS scores and expert scores because it compensates for chance inter-rater agreements (Bejar 1991). Other measures have also been used, such as the creation of average judgment scores (as estimates of "true scores") or consensus scores (among experts). Correlations among rating scores from many experts (or methods) have also been employed in the literature (Yang et al. 2002). Overall, many approaches have been used to test the efficacy of CAS systems.

Regardless of the validation method, CAS system-generated scores have been repeatedly found to display robust agreement patterns with expert raters. Beginning with the Project Essay Grade (PEG) system (Page 2003),

www.manaraa.com

researchers have found agreement levels between human and computer scoring >0.80. Specifically, Page found that "…the PEG program predicted human judgment well—better even than three human judges" (Page 2003). Work with Intelligent Essay Assessor has likewise demonstrated outstanding correspondence with human raters: "IEA-generated scores agreed better with ratings given by people with higher rather than lower expertise." (Yang et al. 2002, p. 402). Indeed, across exams "…IEA score[s] agreed with single readers as well as single readers agreed with each other" (Landauer et al. 2003). Such promising findings continue with more widely used commercial projects, such as Educational Testing Service's C-rater (Sukkarieh and Bolge 2008).

Despite such promising findings, it is important to note that statistical agreement does not necessarily indicate that what has been measured is meaningful; that is, high levels of agreement on a superficial or peripheral learning target would be of minor significance to educators (Landauer et al. 2000). Furthermore, agreement metrics are sensitive to the "grain size" of analysis; fine-grained scoring is much less likely to display high levels of agreement relative to whole-essay score agreements. Thus, both construct attributes and scale emerge as important considerations that must be attended to in the interpretation of computer–human correspondence scores Shermis and Burstein (2003).

The validation of computer assisted scoring systems has received increasing attention given their expanding role in educational evaluation. In brief, contemporary conceptualizations of validity involve the "representativeness and relevance of the test scores to the construct intended to be measured" (Yang et al. 2002:404). Validation evidence for CAS, according to Yang et al. (2002), may be gathered using empirical data, expert judgments, relevant literature, and logical analysis. Methodologically, we employ all of these approaches in our study of the efficacy of natural selection knowledge measurement using a CAS system (see below). Overall, however, the establishment of human–human agreement, coupled with corresponding human–computer score agreement, remains as one of the core approaches for validity evidence for CAS metrics (Yang et al. 2002).

The purpose of a test and the intended uses of its scores must also be considered as a component of validity (AERA, APA, and NCME 1999). High-stakes test scores—e.g., determining whether a biology teacher should be certified to teach or not—based on CAS scores alone would be unlikely to be considered valid despite their very high levels of agreement with human scorers (e.g., Kappas 0.85); in contrast, comparable agreement values on a CAS scored test designed for formative purposes—such as guiding instruction—may be interpreted as valid despite similar Kappa values as above. Thus, levels of agreement must be considered in light of the purpose for which the test scores will be put. Our work on evolutionary knowledge measurement is anchored in the pursuit of formative diagnosis of student thinking about natural selection, and this purpose must frame our data, evidence, and interpretations.

Measuring Knowledge of Natural Selection

The construct of natural selection—and its constituent elements (or "Key Concepts")—is generally well established (Nehm and Schonfeld 2008). Nevertheless, there is some variance in the evolutionary literature regarding the number of "essential" elements of this construct (Nehm and Schonfeld 2010). At a minimum, three Key Concepts (KCs) are considered necessary and sufficient to explain natural selection: (a) the presence and causes of variation (mutation, recombination, sex); (b) the heritability of variation; (c) the differential reproduction and survival of individuals (Lewontin 1978:220; Pigliucci and Kaplan 2006:14; Patterson 1978:1; Endler 1992:220). Many other authors acknowledge the importance of: (d) hyper-fecundity or 'overproduction' of offspring; (e) limited resources, (f) competition, and (g) a change in the distribution of produced phenotypic/genotypic variation in the next generation (Patterson 1978; Endler 1992). One group of authors goes even further, and expands this list of "essential" elements to include "population stability" and "speciation" (Anderson et al. 2002). Some debate also exists as to whether "speciation" is a necessary element of natural selection (e.g., Gould 2002).

Our analyses consider these opposing viewpoints regarding the content validity of natural selection by studying what we term Key Concepts and Core Concepts of natural selection. Key concepts include all seven of the most commonly accepted elements (1–7, above) whereas Core Concepts of natural selection include the three Key Concepts considered necessary and sufficient to explain natural selection (i.e., the presence and causes of variation; the heritability of variation; and the differential reproduction and survival of individuals). We leave "population stability" and "speciation" out of our analyses, although this omission may have little relevance given that in previous published studies, 0% of student samples [n > 100] ever used these elements in their explanations (see Nehm and Reilly 2007; Nehm and Schonfeld 2007, 2008). Suffice it to say that our analyses encompass the three "essential" elements of natural selection along with the most widely accepted additional elements denoted by evolutionary biologists (e.g., Patterson 1978; Endler 1992; Gould 2002). Overall, then, our study of the efficacy of a CAS system emphasizes the measurement of core elements of content

المنارة للاستشارات
www.manaraa.com

(the construct of natural selection) recognized by evolution experts (Lewontin 1978; Pigliucci and Kaplan 2006:14; Patterson 1978; Endler 1992).

Methods

Sample and Data

Our sample of constructed responses was gathered using an online response system built within our university course management system. Responses were captured from undergraduate student participants enrolled in the introductory biology sequence for majors. Demographically, the sample was approximately 80% White (non-Hispanic) and 20% minority (African American, Asian, Hispanic, Native American), 60% female, and with an average age of 20 years. The sample includes responses from 2 years: 2008 and 2009. Students received extra course points for choosing to complete an instrument (pre- and post-course) containing three evolutionary prompts about bacterial resistance, cheetah running speed, and salamander sight that have been widely used in the literature (see Bishop and Anderson 1990; Nehm and Reilly 2007). Participation rates were >75%; 812 sufficiently complete student responses were gathered from the 2008 sample and 428 from the 2009 sample.

Human Scoring

Students' evolutionary explanations were atomized into a series of units using a scoring rubric established in prior research and validated using extended clinical interviews (for details, see Nehm and Schonfeld 2007; 2008). The elements extracted from participants' evolutionary explanations pertain to the scientifically established causal elements used to explain evolutionary change via natural selection (see above). The coding rubric was used to identify the presence or absence of seven Key Concepts (KC) of natural selection in each of the students' three essay responses (see above). Recall that three of these KCs, because of their special importance to the theory of natural selection, are denoted as Core Concepts. Overall, then, a matrix of 7 concepts × 3 items was constructed.

Two expert scorers independently coded the essay responses for the presence or absence of the KCs using the scoring rubrics. Rater 1 holds a Ph.D. degree in evolutionary biology, has conducted evolutionary research, published articles in the primary scientific literature on evolution, and has taught biology for >10 years. Rater 2 holds a M.S. degree in Zoology and a Ph.D. in science education and has taught biology for >10 years. Both scorers discussed, modified, and finalized the rubrics after several episodes of practice scoring. Subsequently, all scoring was performed independently (these scores are the focus of our study).

The CAS System: SPSS Text Analysis

While many CAS systems are now available for use (see above), we focused on one of the least expensive commercially available products: SPSS Text Analysis 3.0 (subsequently: SPSSTA). Because a complete user's guide for SPSSTA is available (SPSS Inc. 2006), in the interest of space we only provide a very condensed description of how the program works. Readers interested in the nuances of program function are encouraged to consult the user's guide. In brief, SPSSTA uses linguistic-based techniques to identify, extract, and classify text. Such classifications are based upon semantic networks and text co-occurrence patterns. SPSSTA was designed to analyze short, open-ended responses. The best results have been obtained with single words or a few sentences, although responses may be as long as 4,000 characters. Analyses of these responses are based on a combination of linguistic and statistical extraction techniques. Text analysis involves the identification of equivalent classes of terms; locating synonyms of such term classes; indexing and grouping terms; and finding distribution patterns in the responses (SPSS Inc 2006). Improving the efficacy of automated extraction can be achieved by changing the programming rules, which we discuss below.

Programming SPSS Text Analysis

Some researchers using SPSSTA report successful automatic text extraction using the large term library that is provided with the program (Galt 2008). However, in the present study such "automatic" extractions could not be performed for two reasons: First, the term libraries provided with the program did not include verbs or most biology terms relating to evolutionary biology (e.g., mutations, recombination in meiosis, etc.). Second, automatic categorization was not able to differentiate between some KCs and evolutionary misconceptions (e.g., the term "adapt" may refer to either a correct or incorrect conceptualization of evolutionary processes). We therefore used a dataset of 812 essays to manually develop more expansive term libraries, program extraction rules, and calibrate the software.

Methodologically, Rater 1 (see above) initially used a previously developed rubric (Nehm et al. 2010a, b) to identify and manually extract text from student responses that were considered representative of each of the seven KCs of natural selection. He then copied and pasted this text into the corresponding rubric cell for each of the three

المنارة للاستشارات
www.manaraa.com

essay items about evolutionary change using Microsoft Excel. Rater 2 (see above) scored a subset ($n = 100$) of these essays. Inter-rater reliabilities (Kappa values) between the two human expert raters were >0.80 for all seven KCs, indicating that the first rater's results were appropriate and could be replicated. At the completion of scoring 812 responses, the Excel spreadsheet contained columns of text characteristic of each of the seven KCs. KC1, for example, included words and phrases culled from many responses, such as:

> "genetic mutation, mutation, mutations, random mutation, mutation for a faster running speed, fast genes, gene for becoming quicker, genes that make them faster, genetic variation provided for faster running, some that were slightly different, variation within the population…"

The key words and phrases identified by Rater 1 were subsequently used to build term libraries and relationship functions in SPSSTA. The first step involved building a library in SPSSTA that included all of the appropriate scientific and common language terms used by students that were lacking in the default SPSSTA libraries (i.e., all the text that Rater 1 extracted and that could be helpful in creating extraction rules). Including the inflections of verbs, 433 terms needed to be added to the SPSSTA library. These *terms* were then grouped into so-called *types*. One *type* may contain one or more *terms*. An example of a type is "abilities" which includes the terms e.g. "fast", "blind", "fight antibiotics", "fitness" or "resistance", etc. Using more types is useful, as it may assist in the creation of more specific and elaborate text extraction rules (see below).

After embedding the terms and types into the SPSSTA library, it was then possible to extract such terms and types from the 812 student responses that were imported into the SPSSTA program. Once extraction was performed, all of the terms that were included in the library were highlighted in different colors in all of the student responses in the sample. At this point, the library had been augmented and term extractions had been completed. The next step was to define and build text *categories*.

For our study, *each* of the KCs was represented as a separate category in SPSSTA. There are three possible approaches for creating a *category* in SPSSTA: (1) One or more individual *terms* may be used to define the category; (2) One or more *types* may be used to define the category; or (3) Combinations of different *types* and/or *terms* may define the category. Using HR1's extracted text for each key concept, it was in some cases obvious that single terms were sufficient for representing a KC (e.g. "mutation" was indicative of KC1). In other cases, a group of very similar terms was used to indicate a specific Key Concept (e.g. different synonyms for "reproduce," such as "multiply," "propagate," etc.). In such cases, the category included a *type*. Differentiating types and terms is of practical importance; in the library the types are used to organize, condense, and group terms. If a type is included in a category, then if a new term is added into this type SPSSTA will automatically extract it and add it to the category.

In most instances, combinations of types and terms were used to create categories. Equation 1 illustrates a very simple example in which one term was not sufficient to extract a KC; thus, combinations of different terms and types were needed. These combinations are called *rules* in SPSSTA. A *rule* is built using terms, types (indicated through < >), and the operators AND [&], OR [|], NOT [∼], and BRACKETS [()] (see Eq. 1, below).

$$\begin{aligned}&(<\text{kc2\_1\_spreading}>|<\text{kc6\_1\_reproduce}>|\text{offspring})\\&\&<\text{abilities}>\&((\text{will}\ \&\ \text{also}\ \&\ \text{be})|(\text{can}\ \&\ \text{do}\ \&\ \text{same})|\\&((\text{same}\ |\text{with this})\&(\text{ability}\ |\ \text{trait}\ |\ <\text{abilities}>))|\text{already})\end{aligned}\tag{1}$$

In order to further clarify rule functioning, we provide several examples to illustrate how they work. We use Eq. 1 as the example. The text in the student answers that is relevant to rule application is shown in bold. After each answer, the parts of the formula that relate to the bold text are explained in italicized text. The formulas should be compared with Eq. 1 to see how the AND and OR operators, as well as the brackets, are used to constrain searches for specific text combinations within students' answers. The student answers we use as examples are from the three different items. This explains why types have been used: the same rules may be used in very different prompt responses.

Answer 1 (Bacteria item): "There are some bacteria that survive the antibiotics which split off to make more bacteria that will also be resistant to the antibiotic."*("make more" is a term in the type < kc2_1_spreading >) & "resistant is a term in the type < abilities > & ((will & also & be)).*

Answer 2 (Cheetah item): "Over the generations the cheetahs evolved to better respond and live in their environment. In order to survive and catch prey, the heritable trait of speed was better suited for the environment. The cheetahs with this trait reproduced and were more fit than cheetahs that were not as fast."*("reproduce" is a term in the type < kc6_1_reproduce >) & "speed" and "fast" are terms in the type < abilities > & ((with this) & (trait)).*

Answer 3 (Salamander item):"When their ancestors changed their environment to the cave, perhaps the ones that already had poor vision (but whose other senses like hearing were heightened) could survive better. Those reproduced with the genes for poor vision but better hearing until today, where cave salamanders can no longer see."*("reproduce" is a term in the type < kc6_1_reproduce >) & "see" is a term in the type < abilities > & (already).*

المنارة للاستشارات
www.manaraa.com

We created our text extraction rules a posteriori; that is, they were identified using the corpus of text built by Rater 1. (It is also possible to develop rules a priori, but we did not do so). Formulation and finalization of extraction rules is an iterative process. After examining HR1's text, possible rules were built and text was extracted accordingly. After the extraction, the results were compared with HR1's scores. Similarities and differences were examined, rules were refined, and correspondence was examined. It is important to note that the *categories* were the same for the three assessment items (bacteria, cheetah, salamander); that is, the category KC1 included exactly the *same* terms, types and rules for the bacteria, salamander, and cheetah items. Rule-building was quite complex; preparing the rules required examining correspondences among three items and seven KCs. Overall, the SPSSTA default libraries are not sufficient for automatic text extraction; terms, types, and categories must be identified, and rules must be built using terms and types. Once complete, the efficacy of these "training" rules may be tested against human scores. After sufficient levels of agreement have been reached, the program has been "trained" and is ready to be tested using new data sets.

Measures of correspondence among human and computer scores

Measures of inter-rater agreement between human raters are typical metrics for testing score comparability (Chung and Baker 2003:28; Krippendorff 2004: 246–249). The same approach may also be used to test for human–computer correspondence given that it too attempts to identify the presence or absence of a particular knowledge element (i.e., a Key Concept). Agreement may be quantified using the percentage of exact or adjacent agreements between CAS scores and human expert-generated scores. As noted above, 'percentage agreement' statistics are problematic because they are sensitive to the number of cases analyzed (Yang et al. 2002). Consequently, Cohen's Kappa commonly has been employed to quantify levels of agreement between CAS scores and expert scores because it compensates for chance inter-rater agreements (Bejar 1991). Kappa values range from 0.0 to 1.0.

A review of the literature revealed that several different inter-rater agreement benchmarks have been established using the Kappa statistic. Landis and Koch (1977), for example, considered inter-rater agreement values between 0.61 and 0.80 to be "substantial" and those between 0.81 and 1.00 to be "almost perfect." Krippendorff (1980) likewise followed this latter benchmark in his well-known guide to content analysis. In contrast, other authors considered inter-rater values greater than 0.60 to be indicative of acceptable and meaningful agreement (e.g., Shermis and Burstein 2003; Altman 1991). On the other hand, Fleiss, Levin, and Paik (2003) describe Kappa values higher than 0.74 as "excellent" and values lower than 0.41 as "poor." Given the diversity of benchmarks established in the literature, and the lack of normative or theoretical justifications for such benchmarks, there is no clear consensus on the matter. For our study, we will follow the benchmarks introduced by Landis and Koch (1977), and echoed by Krippendorff (1980): Cohen's Kappa values between 0.41 and 0.60 are seen as moderate, between 0.61 and 0.80 are considered substantial, and between 0.81 and 1.00 as almost perfect.

In some cases, the marginal score totals were insufficient for calculating Kappa values; in these cases, Yule's Y is reported (Spitznagel and Helzer 1985). Yules Y values are similar to Kappa values in that 1.0 represents the highest possible agreement. Given that Yule's Y has a similar scale as Kappa, we use the same benchmarks as those introduced by Landis and Koch (1977, see above). In cases of multiple agreement comparisons, we employed interclass correlation coefficients (ICCs). The ICC is used as a test for inter-rater reliability between (or among) two or more variables (Field 2009); therefore, it allows analyses of the agreement among multiple raters. Other indices only permit pair-wise comparisons. The same benchmark categories were used as in Kappa and Yules's Y, although it is important to note that reaching an ICC of 0.80 is slightly more difficult than reaching a kappa of 0.80 (Field 2009). We used SPSS 16.0 and Microsoft Excel to calculate all Kappa values, Yule's Y values, ICCs, and correlation coefficients.

Human–Computer Agreement and Analysis Grain Size

As noted above, both construct attributes (e.g., the particular key concepts of natural selection that we identify) and scale (e.g., individual or total key concepts within or among items) are important considerations that must be attended to in the interpretation of computer–human correspondence scores. Given that we used three different items (bacteria, cheetah, salamander) in a pre-post-test design, and seven Key Concepts (and three Core Concepts) were scored for their presence or absence, there are several different grain sizes at which we may analyze the correspondence between human and computer scores.

Key concept scores were tallied separately for each item, and collectively for all three items, both pre- and post-course. In addition, the number of *different* key concepts used *among* all three items (hereafter: Key Concept Diversity) was scored for each participant. Given these items and response patterns, many different analyses of human–computer score agreement are possible. Consequently, we perform five different tests of human–computer score agreements (referred to as analyses 1–5).

المنارة للاستشارات
www.manaraa.com

Analysis 1

The first analysis is a fine-grained study comparing inter-rater (human–computer) agreement for *each* of the seven KCs for *each* of the three different essay prompts (Bacteria, Cheetah, and Salamander) pre- and post-course. This analysis tested whether the SPSSTA extraction rules (see above) could detect KCs 1–7 with equal fidelity as the expert human rater. First, pre and post responses for each item were analyzed separately. Second, pre and post answers for each item were analyzed together. As noted above, Kappa was used to statistically test correspondence.

Analysis 2

The second analysis explored the measurement of inter-rater agreement (human–computer) for each KC in a sample combining all 812 essays. The extraction rules used in this analysis were identical for all three prompts (Bacteria, Cheetah, and Salamander). This analysis is informative because while human raters may use distinctive criteria for the scoring of KCs for each prompt (e.g., bacteria, cheetah), the SPSSTA rules in this case do not. Thus, this analysis indirectly tests whether the superficial item features (e.g., type of organism and its physical/environmental and temporal context) are relevant factors in the scoring of evolutionary explanations. This analysis also tackles a problem identified in the first analysis (above), namely the low marginal totals for some KCs (i.e., particular KCs were mentioned very infrequently by the students in the sample). Thus, this analysis bolsters sample sizes for our statistical tests.

Analysis 3

The third analysis explores the measurement of the inter-rater agreement (human–computer) for Key Concept Diversity (KCD). Recall that KCD is a measure of the number of *different* Key Concepts (KC) of natural selection employed by students *among* the three essay prompts (pre- or post-course). Thus, this measure quantifies the number of *different* accurate explanatory elements that an individual student uses. This measure may be thought of as the complexity of students' accurate natural selection models (within the confines of the instrument tasks). KCD is also the coarsest measure of natural selection knowledge; even if, for example, we were to find that the inter-rater reliabilities for single KCs (e.g., Analyses 1 and 2 above) do *not* approach "near perfect" Kappa values, KCD—because it is not as fine-grained of a measure—may in fact meet this benchmark. If all *individual* KC inter-rater agreements match this Kappa benchmark, then of course KCD measures will also display high values. In other words, if analyses 1 and 2 fail to match our human rating benchmark, they may nevertheless function effectively for KCD measurement.

Analysis 4

The fourth analysis examines human–computer agreement for the three Core Concepts (CC) of natural selection. Note that these three particular Key Concepts (KCs) are considered by many evolutionary biologists to represent the necessary and sufficient elements of an explanation using natural selection (e.g., Nehm et al. 2010a, b; Lewontin 2010). Other evolutionary biologists, however, often subscribe to a much more expansive definition of natural selection that encompasses all or most of the Key Concepts (KCs). Given the uncertainty regarding the exact boundaries of the construct of natural selection (that is, whether KCs or CCs best describe the content we are attempting to measure), we analyze whether the measurement of student knowledge by our human experts and computer software functions with equal fidelity under the two different construct definitions.

Analysis 5

Our final analysis expands upon our previous *compositional* analyses and explores how individual KCs are packaged into explanatory *structures*. That is, we explore KC association patterns among items and among human and computer raters. These analyses may be considered holistic snapshots of students' assemblies of KCs into meaningful explanations. We measure KC co-occurrence patterns using Spearman correlation coefficients and represent these patterns visually in a new type of visual display.

Results

Software Training Analysis 1

Analysis 1 revealed that the functions (or extraction rules) developed and deployed in Text Analysis 3.0 (see above) matched or exceeded inter-rater agreement values of 0.81 ("almost perfect") in the vast majority of cases for individual Key Concepts (KCs) for all three item prompts both pre- and post-intervention (see Fig. 1). The SPSSTA program was able to detect the presence of KCs in a comparable manner as the expert human rater; the program also agreed with the expert human rater regarding the *absence* of KCs in the vast majority of cases. The weakest correspondence between SPSSTA and the human rater was noted for KC7 (A shift in the generational genotype/phenotype distribution).

المنارة للاستشارات
www.manaraa.com

Software Training Analysis 2

The second analysis explored the measurement of human–computer correspondence for each Key Concept (KC) in the combined sample of all 812 essays. Here we found outstanding correspondence, with the exception (as also noted in the first analysis, above) of KC7. Cohen's Kappas were: KC1 ($n = 199$, k = 0.939, $p < 0.01$), KC2 ($n = 256$, k = 0.882, $p < 0.01$), KC3 ($n = 17$, k = 0.971, $p < 0.01$), KC4 ($n = 7$, k = 0.933, $p < 0.01$), KC5 ($n = 174$, k = 0.954, $p < 0.01$), KC6 ($n = 430$, k = 0.768, $p < 0.01$), and KC7 ($n = 153$, k = 0.757, $p < 0.01$). These results indicate that the functions (or extraction 'rules') that we generated in SPSSTA are broadly applicable to different evolutionary scenarios (i.e., bacterial resistance, cheetah running speed, and salamander vision loss). This result is encouraging, as it suggests that the development of new prompts may *not* necessitate the development of new Lexical Analysis (LA) rules.

Software Training Analysis 3

The third analysis explored the inter-rater reliability (human vs. computer) for Key Concept Diversity (KCD), which is a measure of the number of *different* Key Concepts of natural selection that were employed by students *among* their three essay prompts. An Inter-Class-Correlation (ICC) analysis for pre- and post-course responses ($n = 122$) revealed high and significant correspondence between human and computer scoring (ICC [2,1], 0.914, $F_{121,121} = 22.77$, $p < 0.01$). Given the promising results in the previous two analyses, it is perhaps unsurprising that this coarse-grained analysis produced the highest magnitude of human–computer agreement. Overall, these results indicate that SPSSTA is able to grade large numbers of evolutionary responses and produce measures of the complexity of accurate evolutionary elements in undergraduate biology students' explanations that are comparable to those derived by an evolutionary biologist.

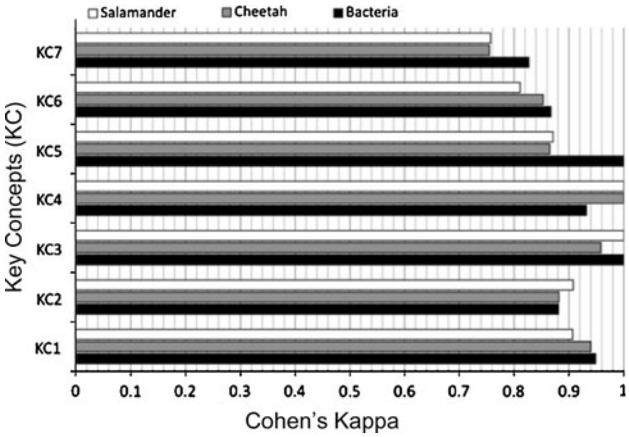


Fig. 1 Inter-rater agreement between Human Rater 1 scores and SPSSTA scores for the seven key concepts of natural selection using Cohen's Kappa (Bacteria $n = 364$, Cheetah $n = 224$, Salamander $n = 249$). All correlations are significant ($p < .001$) and approach or exceed the Kappa benchmark of 0.80. KCs are: (1) The presence and causes of variation (mutation, recombination, sex); (2) The heritability of variation; (3) competition, (4) Hyper-fecundity or 'overproduction' of offspring; (5) Limited resources; (6) The differential reproduction and survival of individuals; and (7) A change in the distribution of produced phenotypic/genotypic variation in the next generation

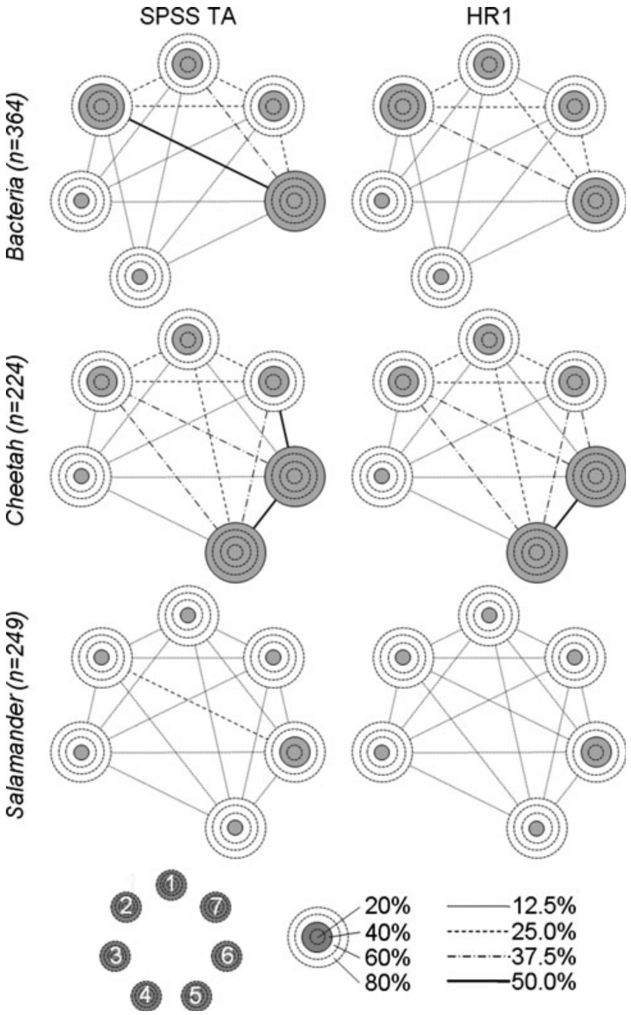


Fig. 2 Key Concept co-occurrence patterns between the expert human rater and SPSSTA. See Fig. 1 caption for KC descriptions

Software Training Analysis 4

The fourth analyses explored whether a different construct conceptualization (that is, using Core Concept Diversity instead of Key Concept Diversity) had any impact upon human–computer agreement. An Inter-Class-Correlation (ICC) analysis of core concept diversity for pre- and post-course responses ($n = 122$) revealed high and significant

www.manaraa.com

correspondence between human and computer scoring (ICC [2,1], 0.936, $F_{121,121} = 30.26$, $p < 0.01$). Thus, regardless of construct definitions, human–computer agreement was high and significant. Further, regardless of whether the scorer was a computer (SPSSTA) or a human (HR1), CCD was highly and significantly correlated with KCD in all comparison cases (Pearson r > 0.80, $p < 0.01$). Thus, both construct measures are significantly and highly correlated, and human–computer scoring produced comparable measures of both CCD and KCD.

Software Training Analysis 5

In addition to our statistical analyses of correspondence patterns, we developed a new graphical organization technique for visually displaying how the key concepts extracted from responses using different methods (i.e., human and computer) are structured. This method provides a visual snapshot of both the abundance and co-occurrence of different key concepts in a dataset for each assessment item. Figure 2 illustrates score patterns derived from the human raters and SPSSTA. It is apparent that while the overall structures of knowledge are remarkably similar, differences may also be noted. For example: As the human expert and SPSSTA differ in the abundance of KC 6 and KC7 detected, the co-occurence between these two KCs differs for the Cheetah and the Bacteria Item. Thus, while highly concordant, these new graphical displays provide an additional way to evaluate the correspondence of our CAS system to expert human raters.

Testing The Efficacy of The Software with Expert Human Scorers and a New Dataset

All of the software training analyses discussed above produced strong and significant associations between the scores produced by an expert human rater and the SPSSTA software. Nevertheless, the SPSSTA extraction rules were built using text manually extracted by the expert human rather. A more important series of analyses are consequently needed to explore whether: (1) the software performs well on a dataset upon which the SPSSTA software was *not* trained and (2) the trained software performs effectively relative to an expert human rater who was not involved in the rule development. Thus, the second set of analyses that we report are tests of the efficacy of the software on a new data set and compared to another expert human rater and show whether the results are expert—and dataset—independent.

As above, the inter-rater agreements were calculated using Inter-Class Correlations, Cohens kappa or Yules Y (see "Methods"), but unlike above, three comparisons were made instead of two: (1) Human Rater 1 vs. Human Rater 2; (2) HR1 vs. SPSSTA; and (3) HR2 vs. SPSSTA. For these analyses, each expert rater blindly scored 110 randomly chosen open-response answers for each of the three items (bacteria, cheetah, and salamander). Specifically, 55 responses were randomly drawn from the pre-test responses and 55 responses were drawn from the post-test responses. Thus, a total of 330 responses were used in tests of the efficacy of the software with expert human raters.

Software Testing Analysis 1

Our first analysis explores the efficacy of the extraction rules established in the training of SPSSTA on the 2009 dataset (see "Methods"). Specifically, we compare KC scores for HR1, HR2 and SPSSTA separately for each of the three assessment items (bacteria, cheetah, and salamander) (See Fig. 3). The vast majority of agreements are "substantial," with kappa values >0.60; only KC7 does not meet this inter-rater agreement level. Additionally, for the Salamander item and KC5, HR1 and HR2 agreement is below 0.60. Key Concept 2 is a good example of the interdependence of the SPSSTA results. As explained above, in some cases it was not possible to create a single set of rules that lead to sufficient agreement for all items.

Software Testing Analysis 2

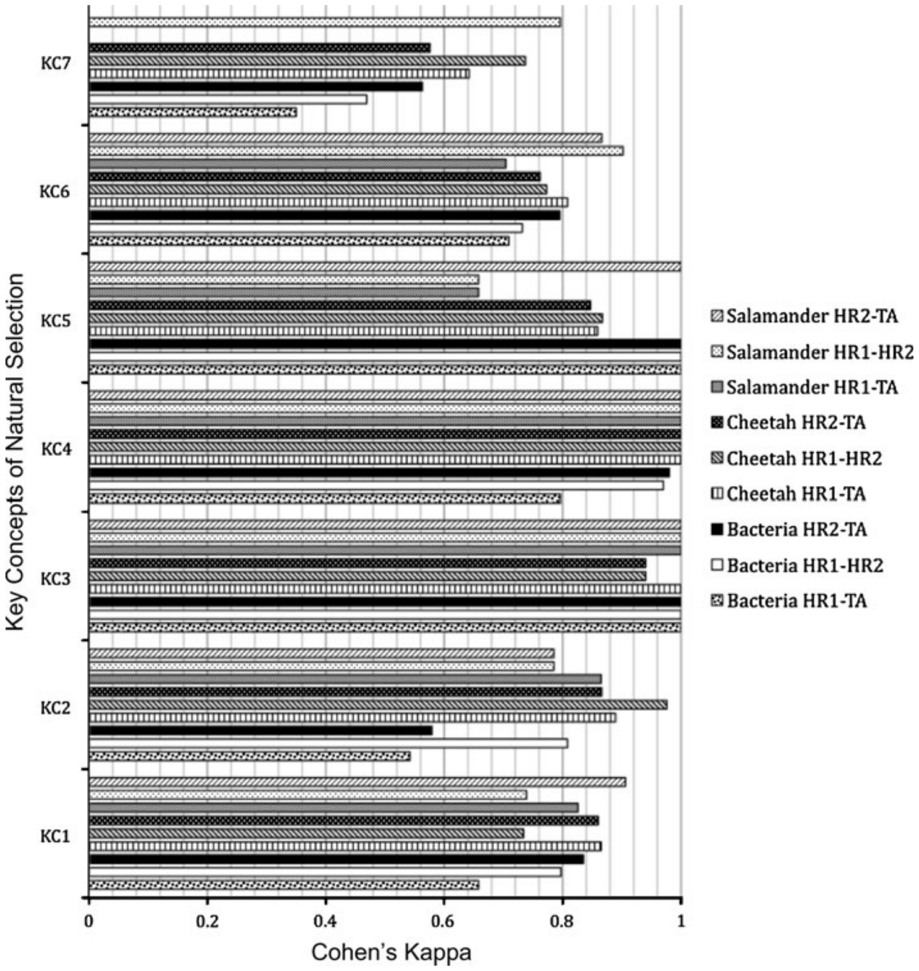
It is also possible to examine KC correspondence across all three items. As explained above, the categories and rules used in SPSSTA are exactly the same across all items. Therefore, the software does not differentiate between the differing item features (e.g., salamander, bacteria, etc.). Figure 4 provides an overview of the findings, which are "near perfect" in very many cases. All KCs reach a "sufficient" agreement except KC seven. Notably, poor agreement between the human raters is also apparent in this instance.

Software Testing Analysis 3

The third analysis explored inter-rater agreements among HR1, HR2, and SPSSTA for Key Concept Diversity (KCD). As there are seven possible Key Concepts (KC) in an essay response, KC Diversity ranges from 0 to 7. Therefore, inter-rater reliability was measured with Inter-Class-Correlations (ICCs). An Inter-Class-Correlation (ICC) for all 110 responses revealed high and significant correspondence between human and computer scoring and exceeded a value of 0.80 (ICC [2,1], 0.857, $F_{109,218} = 19.142$, $p < 0.01$). Furthermore, there were no significant differences in KCD among HR1, HR2, and SPSSTA ($p > 0.05$ in all comparisons). Thus, these results indicate that Key Concept

المنارة للاستشارات
www.manaraa.com

Fig. 3 Inter-rater agreement for all three items (bacteria, cheetah, salamader) for all key concepts (KC). Agreement statistics include Cohen's kappa and Yules Y. *HR1* human rater 1, *HR2* human rater 2, *TA* SPSSTA. All reported values are significant at $p < 0.001$. See Fig. 1 caption for KC descriptions



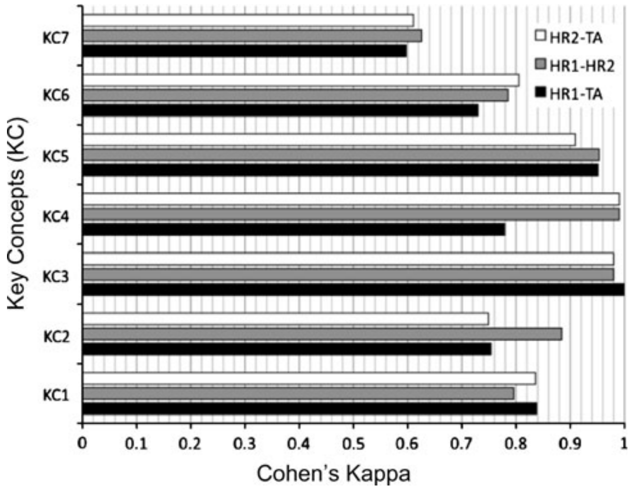


Fig. 4 Inter-rater agreement among Human Rater 1, Human Rater 2, and SPSSTA for the seven key concepts of natural selection (KC) using Cohen's Kappa ($n = 330$). All correlations are significant ($p < .001$). Notably, KC7 scores do not meet benchmark targets. See Fig. 1 caption for KC descriptions

Diversity scores generated using SPSSTA are comparable to scores independently generated by two biologists.

Software Testing Analysis 4

Similar to the third analysis above, the fourth analysis explored the inter-rater agreements among HR1, HR2, and SPSSTA for Core Concept Diversity (CCD). Recall that these three core concepts of natural selection are considered to be the most important of the seven elements of an evolutionary explanation in the professional literature on evolution (Nehm and Schonfeld 2010). As such, they warrant focused and independent analysis. As above, inter-rater reliability for CCD was measured with Inter-Class-Correlations (ICCs). This analysis, using all 110 responses, revealed high and significant correspondence between human and computer scoring and again exceeded 0.81 (ICC [2,1], 0. 858, $F_{109,218} = 19.230$, $p < 0.01$). Thus, similar to KCD results, CCD results also indicated that the scores generated using SPSSTA are comparable to scores

www.manaraa.com

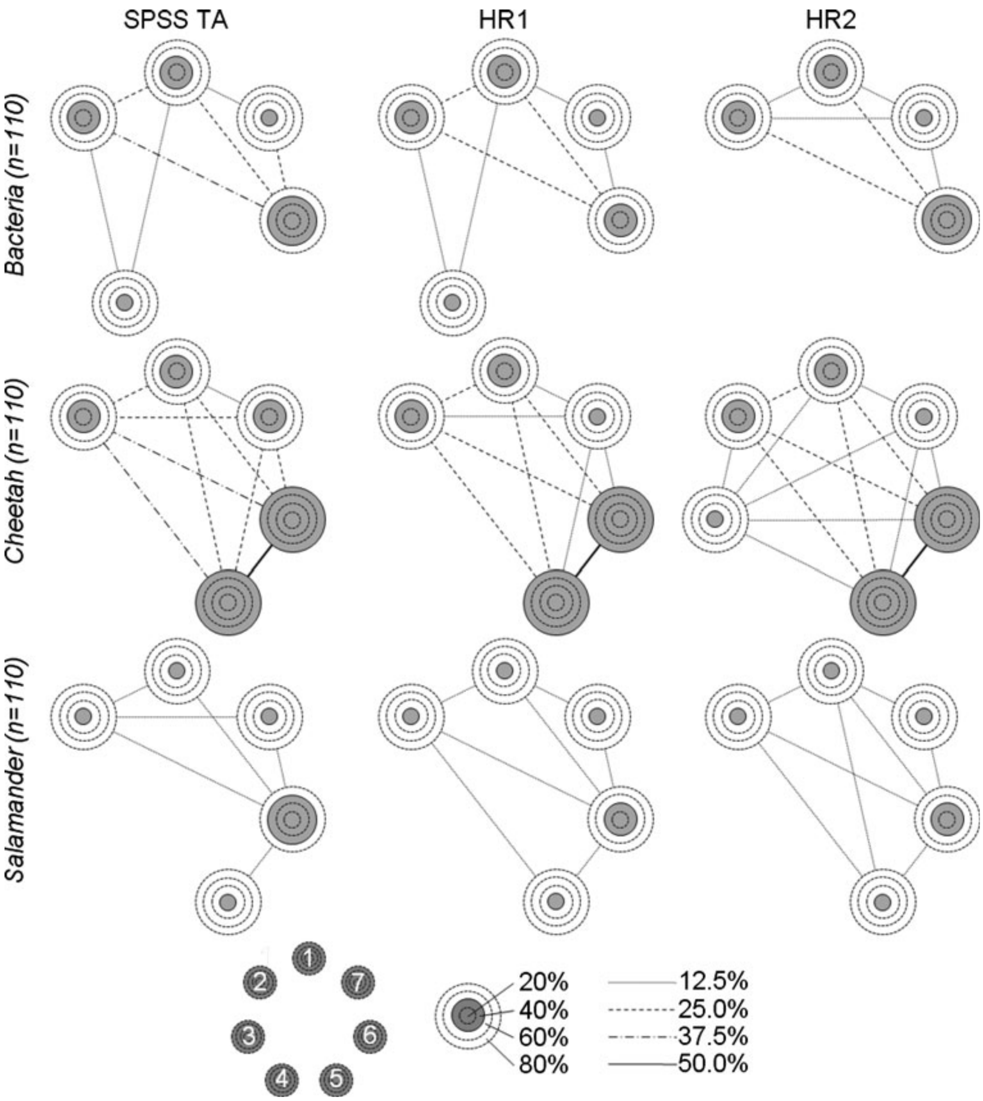


Fig. 5 Key Concept co-occurrence patterns among SPSSTA (TA), Human Rater 1 (HR1) and Human Rater 2 (HR2). See Fig. 1 caption for KC descriptions

independently generated by two biologists. Finally, given that KCD and CCD are different construct conceptualizations, Pearson correlation coefficients were used to compare the two measurements among raters. For HR1, CCD and KCD were highly and significantly correlated ($n = 110$, r = 0.898, $p < 0.01$), as were those for HR2 ($n = 110$, r = 0.894, $p < 0.01$) and SPSSTA ($n = 110$, r = 0.896, $p < 0.01$). Thus, regardless of human or computer scoring, CCD and KCD are strongly and significantly correlated.

Software Testing Analysis 5

In addition to our statistical analyses of correspondence patterns, we visually displayed how the key concepts extracted from responses using different scoring methods (i.e., human and computer) are composed and structured. This method provides a visual snapshot of both the abundance and co-occurrence of different key concepts extracted from the dataset. Figure 5 illustrates score patterns derived from the two expert human scorers and SPSSTA. It is apparent that while the overall structures of knowledge are very similar, differences may also be noted. It is interesting to note, for example, that co-occurrence patterns for the salamander and bacteria items differ among all three raters. The first human rater found more Key Concepts in students' answers to the bacteria item than the second human rater. The second human rater also found more co-occurrences among KCs for the Cheetah item than did SPSSTA. Overall, however, the structures of student explanations are concordant (Fig. 5).

Discussion

Over the past 40 years, numerous studies have demonstrated the utility and efficacy of Computer Assisted

المنارة للاستشارات
www.manaraa.com

Scoring tools relative to expert human raters, notably PEG (Page, 2003), C-rater (Sukkarieh and Bolge 2008), E-rater (Burstein 2003), and Intelligent Essay Assessor (Landauer et al. 2001). Our study expands upon this growing body of work by exploring the efficacy of a new CAS tool (SPSS Text Analysis 3.0) and a never before explored content domain—evolutionary biology. Our study employed a large corpus of open-response data derived from a previously validated natural selection instrument (the Open-Response Instrument of Nehm and Schonfeld 2008) to test the efficacy of SPSSTA scoring relative to expert human raters (that is, biologists with graduate training in evolution). At the outset, it is important to reiterate that our work is anchored in the pursuit of formative diagnosis of student thinking about natural selection, and this purpose must frame our data, evidence, and interpretations.

Software efficacy relative to expert human scoring

Our five analyses of the correspondence between human-derived and computer-derived measures of students' natural selection knowledge produced consistent findings; regardless of granularity (i.e., individual key concepts or composite measures of key concept diversity) or construct definition (core concepts vs. key concepts) SPSSTA generated measures of student knowledge of natural selection comparable to those generated by two trained biologists, one of whom is an expert in evolution. In the vast majority of cases, SPSSTA and human agreement—measured using Kappa or Yule's Y—met or exceeded the benchmark of 0.8, which is considered very good or excellent agreement (Landis and Koch 1977). Specifically, out of 282 human–computer comparisons, 42.9% ($N = 121$) achieved outstanding agreement values (Kappa > 0.9); 64.9% ($N = 183$) achieved "near perfect" agreement values >0.8; 82.3% ($N = 232$) achieved agreement values >0.7; and 89.4% ($N = 252$) achieved agreement values >0.6. Given that the medical research community considers Kappa values >0.6 to be acceptable (Altman 1991), 89.4% of human–computer comparisons met this benchmark.

Certain key concepts of natural selection were detected with greater correspondence than others, however. Key Concept 7, for example, displayed lower levels of agreement not only between SPSSTA and the expert human rater, but also between the two human raters. This suggests that this particular concept may not be clearly conceptualized or scored, and consequently the text extraction rules may likewise be in need of refinement. This finding highlights the fact that extraction success is, perhaps unsurprisingly, dependent upon clear scoring criteria. It is difficult, if not impossible, to build effective text extraction rules using ambiguously circumscribed constructs.

In some instances, the frequency of student use of particular core or key concepts of natural selection in the sample was very low (e.g.,<five). These very low case numbers make it remarkably difficult to reach high agreement levels, or draw any robust conclusions about the efficacy of the software. Specifically, in some cases the reported agreement is 100%, but this is simply a product of both raters failing to find a key concept in the response set. Therefore, such "almost perfect" agreement tells us only that the rules used in SPSSTA do not overestimate the number of key concepts. But it is not possible to conclude in such cases that SPSSTA would in fact find the key concept if it was present. However, the result was still encouraging, as more than 73% of the agreements are above a value of 0.70, or "substantial" (Cohen's Kappa and Yules' Y, $p < 0.05$). Surprisingly, the comparison of the second human rater to the software resulted in higher agreement than between the software and the first human rater (on whose rating the rules were built in part I of the study). The comparison between the two human raters shows the difficulty of reaching a high agreement, as they reach sufficient values for 33 of 42 comparisons.

Our findings are in line with prior work using other CAS systems. In a study of the automated scoring of architectural mental models, Williamson, Bejar and Hone (1999) reported moderate to high Kappas (between 0.32 and 0.92; mean of 0.53). For the E-Rater software, Powers et al. (2001) reported Cohen's Kappas of 0.85 between human raters and 0.49 and 0.27 between the E-rater and two human expert raters. As the authors note, these human–computer agreement scores are significantly lower than the usually reported agreement indices hovering above 0.80. Chodorow and Burstein (2004) also used the E-rater system but used it to judge TOEFL essays. They found a slightly higher agreement between the human experts (0.56) than between the E-rater and human rater comparison (0.53). Wang, Chang and Li (2005) used Pearson product-moment correlations (r) to measure the agreement between human experts and a CAS system for open-ended problem solving. They also reported higher agreement between the human raters (.89) than between the human raters and CAS tools (0.69–0.82; all correlations were highly significant). Overall, much like our findings using SPSSTA to score open-response evolution items, human–human agreement is generally higher than or equal to CAS-human agreement.

From a formative and summative assessment perspective, the measurement of Key Concept or Core Concept Diversity (KCD, CCD respectively) provides the broadest measure of evolutionary competency (Nehm and Reilly 2007). KCD and CCD measure students' abilities across three items that differ in surface features (e.g., taxa, traits, and change types) and in so doing provide students with multiple opportunities to demonstrate their evolutionary

المنارة للاستشارات
www.manaraa.com

knowledge. The KCD/CCD measures are calculated as the sum of *different* KCs used across all items by one student; they do not award more credit for the repeated use of the same concept. Therefore, this measure indicates whether a student is able to apply conceptual knowledge in a variety of situations. The particularly high inter-rater agreement between human experts and SPSSTA for KCD and CCD demonstrates the broad applicability of the tool. Thus, our results indicate that the best use of SPSSTA is a substitute for the time consuming process of human scoring for this particular assessment measure.

Importantly, our findings only apply to the *scientific* elements of evolutionary explanations; that is, we did not attempt to build term and type libraries or extraction rules for naïve ideas or evolutionary misconceptions (Nehm and Reilly 2007). Given that many students build evolutionary explanations comprised of assemblages of both accurate (key concept) and misconception elements, it is important to expand our current work to test whether comparable success may apply to other explanatory elements. Overall, however, our findings indicate that computer assisted scoring of constructed-response evolutionary explanations are comparable to human-generated assessment scores in the vast majority of cases. Collectively, these findings affirm our view that CAS may be a transformative method for STEM assessment in general and natural selection measurement in particular. This is of particular importance given that several documented limitations characterize extant multiple-choice natural selection concept inventories (Nehm and Schonfeld 2008, 2010).

Assessing the composition and structure of scientific explanations

Students' ability to identify and assemble scientific information into explanatory models is a core skill that is receiving increasing emphasis in science education. The newly proposed *Framework for Science Education*, for example, has identified 'generating and evaluating scientific explanations' as a central—but neglected—feature of scientific literacy (Alberts 2010: 491). Constructed response assessments, such as the ORI, were developed in part to evaluate students' abilities in this regard (Bishop and Anderson 1990; Nehm et al. 2010a, b). CAS scoring (performed using SPSSTA), coupled with new representational approaches for displaying students' explanatory models (e.g. Fig. 5), provides progress in this neglected area of science assessment. No other instruments in the domain of evolution assess the composition and structure of student-built explanations.

Our comparisons of the composition and structure of students' explanatory models of evolutionary change generated by human and computer scoring revealed in a majority of cases strong correspondence in terms of Key Concept (KC) presence, abundance, and association (see Figs. 3 and 5). They also revealed the KCs that were (or were not) employed by students, as well as their relative co-occurrences. For example, among all items, KC3 (hyperfecundity) and KC4 (resource limitation) were rarely used by students; in the Cheetah item, however, KC5 (competition) and KC6 (differential survival) were employed very commonly (>80%). Furthermore, these diagrams visually document how parallel items differing in surface features (i.e., bacteria, cheetah, salamander) are associated with different explanatory structures among items (e.g., compare Fig. 5 top and middle rows). Given that CAS using SPSSTA produced explanatory structures generally concordant with human scoring, our work demonstrates that the complex task of representing students' explanations of evolutionary change may be successfully facilitated by SPSSTA. Further work identifying and extracting causal language and linkages among concepts, and visually integrating this information with our visual diagram methods, would enhance our understanding of students' scientific explanations in the context of evolution.

It is important to note that an explanation composed of exclusively correct (or scientific) explanatory elements need not imply a correct explanatory *structure*. A response including the Key Concepts (1) differential survival, (2) mutation, and (3) heritable variation, for example, would be scored by humans (and in most cases by SPSSTA) as containing three Key Concepts (see the Key Concept rubrics of Nehm et al. 2010a, b). Nevertheless, these three accurate elements could be arranged in an inaccurate explanatory structure, such as: "The differential survival of individuals caused mutations to happen, and these mutations caused variation to be heritable." Thus, analysis of explanatory structure will typically require tests of structural accuracy as well as compositional accuracy.

Importantly, cases of compositional accuracy but structural inaccuracy were not found to occur in our sample or in our analyses of Key Concepts. However, as we expand our work to encompass computer-based diagnoses of both key concepts and misconceptions, we suspect that this issue will become much more relevant and common. Indeed, the co-existence of both Key Concepts and misconceptions in student explanations is quite common; such "mixed models" are known to constrain the validity and utility of existing concept inventories in science (Nehm and Schonfeld 2010). It remains to be determined whether SPSSTA equations may be built that are capable of reliably distinguishing differences in more complex causal structures, such as mixed models. For whatever reason, in our sample the co-occurrence of accurate Key Concepts (albeit in different magnitudes, see Fig. 2) was also indicative of structural accuracy.

المنارة للاستشارات
www.manaraa.com

Advantages of CAS of open-response evolution assessments

CAS of open-response items has many conceptual advantages over closed-response formats such as multiple-choice; three that have been noted in the context of natural selection assessment include more precise documentation of: (1) which knowledge elements *students* consider to be important to explain a phenomenon (as opposed to the selection of a prescribed "either-or" menu of options); (2) how knowledge recruitment is affected by the context or situation represented in the item (such as bacterial resistance to antibiotics, cheetah running speed, or salamander vision loss, see Kirch 2008; Nehm and Ha 2011); and (3) how students assemble and structure chosen knowledge elements into an explanation (e.g., Figs. 3 and 5) (Nehm and Ha 2011). Currently, assessing these three aspects of students' evolutionary reasoning is limited in scope because of the prohibitive time, money, and expertise required to hand-score open-response evolution assessments (such as the ORI). Thus, CAS tools such as SPSSTA may be used to not only address the practical constraints of grading open-response data (time, cost, etc.), but leverage improvements in the quality of assessments that attempt to assess student thinking about evolution and natural selection (cf. NRC 2001) (However, see "Discussions" of cost effectiveness below).

Several additional advantages characterize the SPSSTA scoring system. The development of new assessment items, for example, often necessitates significant changes in scoring procedures, rubrics, coding manuals, and human training. The type and category libraries built in SPSSTA, however, are often of a general nature and need not be fundamentally changed if item surface features are changed. For example, changing the ORI item features from "bacteria" to "roses," and "resistance" to "thorns," would only entail a few changes to the SPSSTA term and type libraries. These changes would be made within the SPSSTA library, and the program would automatically alter all rules and categories with only a few mouse 'clicks' (Galt 2008). As noted above, only two 'types' in our SPSSTA library were item specific ($n = 2/20$, 10%). Our finding that inter-rater reliabilities were "substantial" across different items (e.g., bacteria, salamander) using similar terms and types, provides empirical support for the utility of general rule-building approaches. Nevertheless, further empirical work should be completed in order to test this finding more rigorously. Overall, however, SPSSTA appears to offer significant advantages in terms of the practical necessity of updating the features of open-response assessment items.

Formative assessment of student knowledge is an important activity in school, university, and online environments and is known to positively impact student-learning gains (Wood 2004). Consequently, several technological tools have been developed for use in large enrollment classes to perform rapid and meaningful formative assessments, such as the well-known 'clicker' response systems (Caldwell 2007). While these clicker-type response systems are often not limited to closed-response items (e.g., multiple-choice), SPSSTA scoring tools could be used to leverage rapid, automated scoring of longer open-response answers about evolution in large enrollment classes. Currently, open-ended answers are used infrequently in lecture contexts because of the practical problem of rapidly evaluating large response sets (which is not a constraint for multiple-choice). Indeed, while open-ended questioning is a common feature of classroom instruction, such actions typically only permit the evaluation of a few orally delivered student responses. In contrast, SPSSTA, in concert with electronic student response systems, could be used to rapidly evaluate hundreds of open-ended responses in a brief time, improving the validity of formative assessment inferences (as a consequence of having results from a larger sample). Finally, formative assessment performed during online cognitive tutoring sessions (cf. Koedinger et al. 1997) could leverage open-response answers scored using SPSSTA to more efficiently diagnose learning deficiencies and direct students along appropriate branches of a learning trajectory. Thus, overall, CAS tools such as SPSSTA may open many new avenues for formative assessment in both classroom and online learning environments.

A final advantage of using SPSSTA is that it may potentially mitigate the myriad limitations of human scoring, most notably inconsistent grading; conscious or unconscious bias; fatigue; working memory overload; discouragement or mood changes while scoring; and many others (Shermis and Burstein 2003). These advantages are not unique to SPSSTA, however, and may characterize many types of CAS systems (e.g., IEA, PEG, C-rater, etc.).

Disadvantages of our approach to CAS

Several disadvantages characterize our particular approach to automated text analysis. First, the domain that we investigated (evolution and natural selection) was not well suited to the off-the-shelf capabilities of SPSSTA 3.0. Specifically, the software did not include the language (scientific vocabulary or common verbs) necessary for analyzing student responses about evolutionary scenarios. Consequently, considerable expertise was needed to define and build the term and type libraries necessary for text extraction in the domain of evolutionary biology. In our study, we needed to create a very large library composed of approximately 450 terms and types (e.g., meiosis, genetic

المنارة للاستشارات
www.manaraa.com

recombination, point mutation, adaptation, selection, etc.). Term library construction required a considerable amount of time (hundreds of hours) and expertise (Ph.D. training in evolutionary biology). Nevertheless, the term and type library may now be used, expanded, and refined by other biologists and biology educators; the initial setup of such libraries need only be completed once per subject or domain. For domains that do not require specialized language or verbs—unlike evolutionary biology—such investments will not serve as a disadvantage to using SPSSTA.

Our study utilized an a posteriori approach to the construction of term libraries; that is, we compiled and subsequently mined a large corpus of existing responses to evolution prompts to determine which terms were likely to be informative in our assessment tasks and useful as text extraction rules. This strategy was costly, as it required gathering a large corpus of text focusing on intuitive and scientifically accurate explanations of evolutionary change. A further limitation of our approach was that annotating this corpus for text elements considered representative of the construct being measured could not be completed by a novice; hence, considerable expert time was needed to identify and categorize the terms central to the diagnosis of student thinking about evolution.

Given the aforementioned limitations, the question may be raised as to whether using SPSSTA to automatically score student explanations of evolution (or for that matter other content areas) is cost effective given the time, money, and expertise needed to use the program (e.g., purchase the program, build term libraries, construct analysis rules, and test the efficacy of the program, as we have done). Indeed, would it be more cost effective to hire humans to exclusively hand-score responses? While providing a generalizable answer to this question is difficult given the idiosyncrasies of our particular case, as well as uncertainty concerning the final number of responses that will eventually be scored using our SPSSTA work, we have nevertheless attempted to do so. (Note that cost estimates are in US dollars).

The cost of the SPSSTA program was approximately $1,000.00. The time cost of expert work (e.g., learning the program, building the term and type libraries, setting up and performing the analyses, modifying the system, and scoring the responses) was approximately 500 h. At a rate of $50 per hour, for our study, the financial cost of SPSSTA development is conservatively estimated to have been $26,000.00. In our experience, a trained rater may score 30 responses in one hour. The cost of a trained expert rater (including training time and preparation) may be estimated at $20 per hour. Given these estimates, human scoring would have cost significantly less than SPSSTA scoring ($541.00 vs. $26,000.00). While it is clear that SPSSTA scoring is not cost effective at present, the institution at which the present study took place enrolls nearly 9,000 students in its introductory biology program each year, and if each student were to complete the three item instrument, 27,000 responses would need to be scored. This would amount to a cost of $18,800.00 per year (assuming the instrument is used only once). Thus, while initially not cost effective, it is likely that the long-term benefits may far exceed these start-up costs, especially if the rules and libraries are used at more than one institution. Nevertheless, our individual case can only serve as a rough estimate. Overall, now that much of the start-up work has been completed, other researchers may readily expand upon and refine our accomplishments in order to leverage more cost effective scoring than characterized our present study.

The disadvantages that we note with the CAS methods used in our study primarily pertain to the SPSS Text Analysis software. Other methodological approaches, such as machine learning, may also be used to leverage automated scoring and may not entail the costs that we outlined. That is, rather than having a human develop and test the efficacy of particular analysis rules, machine learning approaches may be used to analyze a set of human-scored essays and 'discover' and save computational rules that are predictive of human scoring (see, for example, Witten and Frank 2005). Such machine learning approaches do not require: (1) human identification of specialized terms and language; (2) the development of term libraries; or (3) the construction and testing of extraction rules. Thus, the costs that we outline should not be generalized to all text analytic approaches associated with computer assisted scoring systems.

Finally, despite many advantages, open-response assessments—regardless of whether they are implemented using paper and pencil or electronically—have intrinsic limitations that must not be ignored. Constructed response instruments, such as the ORI, may be biased by students' aversion to writing and consequent errors of omission. That is, the format itself may constrain the valid measurement of student knowledge. Further, poor writing skills may hamper clear communication, preventing both human and computer scorers from recognizing the true extent of the student's knowledge; both situations will lead to inaccurate knowledge measures. Thus, although we demonstrated outstanding correspondence between human and computer-generated scores, we did not demonstrate that these scores were valid measures of student knowledge. Nevertheless, prior work has indicated that the instrument we employed in our study (the ORI) produced valid measures of student knowledge of natural selection compared to oral interviews (Nehm and Schonfeld 2008, 2010).

المنارة للاستشارات
www.manaraa.com

Conclusions

Numerous psychometric constraints characterize extant multiple-choice Concept Inventories of natural selection and evolution (Nehm and Schonfeld 2008, 2010). Our study of computer assisted scoring of constructed response evolutionary explanations by biology undergraduates demonstrated that: (1) text analysis tools (i.e., SPSS Text Analysis 3.0) may be used to successfully diagnose fine-grained explanatory elements comprising students' mental models of natural selection as represented in open-response text and (2) text analysis assessment scores are comparable to expert human-generated assessment scores in the vast majority of cases. Collectively, these findings affirm our view that text analysis may be a transformative method for STEM assessment in general and natural selection measurement in particular. Numerous disadvantages also characterize our approach, however, and should be weighed carefully relative to other text analytic strategies such as machine learning. Our future research will leverage the advances made in our term library expansion and rule generation to tackle other knowledge elements common to student evolutionary explanations—notably naïve ideas or misconceptions—and attempt to build more sophisticated and holistic representations and measures of students' evolutionary thinking using computational tools.

Acknowledgments We thank Judy Ridgway for hand scoring some of the responses; Minsu Ha for help with the figures; the editors and anonymous reviewers for very helpful feedback on the manuscript, and the National Science Foundation REESE program (0909999) for funding parts of this study.

References

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

Altman DG (1991) Practical statistics for medical research. Chapman and Hall, London

American Educational Research Association, American Psychological Association, National Council of Measurement in Education (1999) Standards for educational and psychological testing. AERA, Washington, D.C

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

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

Braun HI, Bennett RE, Frye D, Soloway E (1990) Scoring constructed responses using expert system. J Educ Meas 27:93–108

Bridgeman B (1992) Conscious vs. unconscious processes. Theor Psychol 2(1):73–88

Brumby MN (1984) Misconceptions about the concept of natural selection by medical biology students. Sci Educ 68(4):493–503

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, NJ, pp 113–122

Caldwell JE (2007) Clickers in the large classroom: current research and best-practice tips. Life Sci Educ 6(1):9–20

Chi MTH, Feltovich PJ, Glaser R (1981) Categorization and representation of physics problems by experts and novices. Cognit Sci 5:121–152

Chodorow M, Burstein J (2004) Beyond essay length: evaluating e-rater's performance on TOEFL essays (TOEFL Research Rep. No. RR-73). ETS, Princeton, NJ

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, NJ, pp 23–40

Clauser BE, Harik P, Clyman SG (2000) The generalizability of scores for a performance assessment scored with a computer automated scoring system. J Educ Meas 37(3):245–261

Clough EE, Wood-Robinson C (1985) How secondary students interpret instances of biological adaptation. J Biol Educ 19(2):125–130

D'Avanzo, C., Morris, D., Anderson, A., Griffith, A. Williams, K., & Stamp, N. (2008). Diagnostic question clusters to improve student reasoning and understanding in general biology courses: Faculty Development Component. Proceedings of the CABS II conference. Available online at: http://bioliteracy.net/manuscripts08.pdf

Dagher ZR, BouJaoude S (1997) Scientific views and religious beliefs of college students: the case of biological evolution. J Res Sci Teach 34(5):429–445

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

Donnelly LA, Boone WJ (2007) Biology teachers' attitudes toward and use of Indiana's evolution standards. J Res Sci Teach 44(2):236–257

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

Field AP (2009) Discovering statistics using SPSS. SAGE Publications, London

Fleiss JL, Levin B, Paik MC (2003) Statistical methods for rates and proportions, 3rd edn. John Wiley & Sons, Inc., Hoboken

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

Gould SJ (2002) The structure of evolutionary theory. Harvard University Press, Cambridge

Grose EC, Simpson RD (1982) Attitudes of introductory college biology students toward evolution. J Res Sci Teach 19(1):15–23

Ha M, Cha H (2009) Pre-service teachers' synthetic view on Darwinism and Lamarckism. Paper presented at the National Association for Research in Science Teaching conference, Anaheim, CA

Keith TZ (2003) Validity and automated essay scoring systems. In: Shermis MD, Burstein J (eds) Automated essay scoring: A cross-disciplinary perspective. Lawrence Erlbaum Associates, Inc, Mahwah, NJ, pp 147–168

Kingston NM (2009) Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: a synthesis. Appl Meas Educ 22(1):22–37

Kirsh D (2009) Problem solving and situated cognition. In: Philip Robbins, Aydede M (eds) The Cambridge handbook of situated cognition. Cambridge University Press, Cambridge, MA, pp 264–306

المنارة للاستشارات
www.manaraa.com

Koedinger KR, Anderson JR, Hadley WH, Mark MA (1997) Intelligent tutoring goes to school in the big city. Int J Artif Intell Educ 8:30–43

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

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

Kuechler WL, Simkin MG (2004) How well do multiple choice tests evaluate student understanding in computer programming classes. J Infor Sys Educ 14:389–400

Landauer TK, Laham D, Foltz PW (2000) The intelligent essay assessor. IEEE Intell Syst 15(5):27–31

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 testing: emerging technologies and opportunities for diverse applications conference, Tucson, AZ

Landauer TK, Laham D, Foltz PW (2003) Automated scoring and annotation of essays with the Intelligent Essay Assessor. In: Shermis MD, Burstein J (eds) Automated essay scoring: a cross-disciplinary perspective. Lawrence Erlbaum Associates, Inc, Mahwah, NJ, pp 87–112

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

Lewontin R (2010) Not so natural selection. New York Review of Books, New York

Liu X (2010) Using and developing measurement instruments in science education: A Rasch modeling approach. Information Age Publishing, Charlotte, N.C

Mislevy RJ, Steinberg LS, Almond RG (2002) Design and analysis in task-based language assessment. Language Test 19(4):477–496

Morgan R, Maneckshana B (1996) The psychometric perspective: meeting four decades of challenge. In Lessons learned from 40 years of constructed response testing in the advanced placement program. Symposium conducted at the NCME Conference

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

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

Nehm RH (2006) Faith-based evolution education? BioScience 56(8):638–639

Nehm RH, Ha M (2011) Item feature effects in evolution assessment. J Res Sci Teach. doi:10.1002/tea.20400

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

Nehm RH, Schonfeld I (2007) Does increasing biology teacher knowledge about evolution and the nature of science lead to greater advocacy for teaching evolution in schools? J Sci Teac Educ 18(5):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 Teac 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, Kim SY, Sheppard K (2009) Academic preparation in biology and advocacy for teaching evolution: biology versus non-biology teachers. Sci Educ 93(6):1122–1146

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

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 January 10, 2011 at: http://evolutionassessment.org

Newport F (2004) Third of Americans say evidence has supported Darwin's evolution theory. The Gallup Organization, Princeton, NJ

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

Page EB (2003) Project essay grade: PEG. In: Shermis MD, Burstein J (eds) Automated essay scoring: a cross-disciplinary perspective. Lawrence Erlbaum Associates, Mahwah, NJ, pp 43–54

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

Powers DE, Burstein JC, Chodorow MS, Fowles ME, Kukich K (2002a) Comparing the validity of automated and human scoring of essays. J Educ Computing Res 26(4):407–425

Powers DE, Burstein JC, Chodorow M, Fowles ME, Kukich K (2002b) Stumping e-rater: challenging the validity of automated essay scoring. Comput Hum Behav 18(2):103–134

Resnick LB, Resnick DP (1992) Assessing the thinking curriculum: new tools for educational reform. In: Gilford BR, O'Conner MC (eds) Changing assessments: alternative views of aptitude achievement and instruction. Kluwer, Boston, pp 37–75

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

Sinatra GM, Southerland SA, McConaughy F, Demastes JW (2003) Intentions and beliefs in students' understanding and acceptance of biological evolution. J Res Sci Teach 40(5):510–528

Spitznagel EL, Helzer JE (1985) A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry 42:725–728

SPSS Inc (2006) SPSS text analysis for surveys^{TM} 2.0 user's guide. SPSS inc, Chicago, IL

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. Proceedings of the 9th international conference on intelligent tutoring systems, ITS 2008, Montreal, Canada, June 23–27, 2008, vol 5091. Springer-Verlag, New York, pp 779–783

Traub RE, MacRury K (1990) Multiple-choice vs. free response in the testing of scholastic achievement. Test und Tends 8:128–159

Wang HC, Chang CY, Li TY (2005) Automated scoring for creative problem solving ability with ideation-explanation modeling. Paper presented at the 13th International conference on computers in education, Singapore

Williamson DM, Bejar II, Hone AS (1999) 'Mental model' comparison of automated and human scoring. J Educ Meas 36:158–184

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

Wood WB (2004) Clickers: a teaching gimmick that works. Dev Cell 7(6):796–798

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

Zimmerman M (1987) The evolution-creation controversy: opinions of Ohio high school biology teachers. Ohio J Sci 87(4):115–125

المنارة للاستشارات

www.manaraa.com

Copyright of Journal of Science Education & Technology is the property of Springer Science & Business Media B.V. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

المنارة للاستشارات

www.manaraa.com