Comparative Efficacy of Two Computer-assisted Scoring Tools for Evolution Assessment

Minsu Ha, The Ohio State University, USA
Ross H. Nehm, The Ohio State University, USA

Abstract:
The growing use of computer assisted scoring (CAS) in many academic disciplines is driven by the numerous disadvantages that characterize human scoring of constructed response items. These include high costs (in terms of scoring time and expert training) and delayed feedback to test takers. Furthermore, human scoring is problematic because of grading fatigue, inconsistent training and/or background knowledge of graders, and the intrinsic subjectivity associated with interpretation. The central question asked of CAS systems is whether they can score written responses as accurately as trained human scorers. In line with this question, our study compares the efficacy of two CAS tools (SIDE and SPSSTA) for measuring student knowledge of a core idea in the biological sciences: natural selection. Methodologically, we use Cohen's Kappa to quantify levels of agreement between CAS scores and human expert raters and to compensate for chance inter-rater agreements. We employ a corpus of 1572 responses from the ORI instrument of Nehm and Reilly (2007). Our results indicate that both SIDE and SPSSTA approximate human-human agreement levels (i.e., Kappas > 0.7) and are effective tools for scoring evolution essays. Nevertheless, the machine learning approach employed by SIDE requires significantly less programming, set up time, and content expertise than SPSSTA.

Citation: Ha, M., Nehm, R.H. (2011). *Comparative Efficacy of Two Computer-assisted Scoring Tools for Evolution Assessment.* Paper presented at the National Association of Research in Science Teaching, Orlando, FL, April 3-6.

Comparative Efficacy of Two Computer-Assisted Scoring Tools for Evolution Assessment

Introduction

Announcements of the impending revolution in computer assisted scoring (CAS), begun in the 1960s, are finally coming to fruition (Page, 1966; Yang et al. 2002; Shermis and Burstein, 2003). Many commercial CAS tools have emerged in recent years, notably C-rater (Sukkarieh and Bolge, 2008), E-rater (Burstein, 2001), Intelligent Essay Assessor (Landauer et al., 2001), and SPSS Text Analysis (SPSS, 2006). These automated scoring systems are being employed with increasing frequency in educational settings. Moreover, CAS systems (and related tutoring systems) are beginning used to capture and measure more advanced performance skills in large populations, particularly in medical fields, higher education, and secondary school classrooms (Clauser et al., 2000; Mislevy et al., 2002; Braun et al., 1990).

As reviewed by Nehm and Haertig (2011), the growing use of CAS tools in many academic disciplines is driven in part by the many 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. Human scoring is problematic for many well-known reasons, notably grading fatigue, inconsistent training and/or the background expertise of graders, and the intrinsic subjectivity associated with interpreting text (Yang et al. 2002). The development of CAS has consequently been justified by its purported ability to produce greater reproducibility, objectivity, reliability, and efficiency in comparison to human scored responses (Williamson et al. 1999). Finally, the comparability of computer-administered and paper-and-pencil administered test scores provides additional justification for making use of such readily available electronically formatted responses (Keith, 2003; Kingston, 2009; Nehm and Haertig, 2011).

An important question is whether CAS systems can in fact score written responses as accurately as trained expert 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 agreements between CAS systems and human raters. These measures have their disadvantages, however; they are influenced by the number of cases that are analyzed (Yang et al., 2002). As noted by Bejar (1991), 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. Other measures have also been used, such as "average judgment scores" (as estimates of "true scores") or consensus scores (among human experts). Correlations among rating scores from many experts (or methods) have also been employed in the literature (Yang et al., 2002). In summary, many different approaches have been used to test how well CAS systems work in comparison to human scoring.

Regardless of the validation method, CAS system-generated scores have been repeatedly found to display very similar agreement patterns with expert raters. Beginning with the PEG system (Page, 2003), researchers have found correspondences between human and computer scoring > 0.80. Studies of the Intelligent Essay Assessor (IEA) have also demonstrated outstanding agreement with human raters: "…IEA score[s] agreed with single readers as well as single readers agreed with each other" (Landauer et al., 2003). Such promising findings have also been found with Educational Testing Service's C-rater (Sukkarieh and Bolge, 2008).

Despite such remarkable 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 trivial aspects of text would be of little use to educators (Landauer et al., 2000). Indeed, agreement metrics are very sensitive to the "grain size" of the analysis. Very fine-grained scoring is much less likely to produce very high levels of agreement in comparison with course gain size comparisons, such as whole-essay scores. Consequently, analysis scale is an important factor to keep in mind when interpreting the efficacy of CAS systems relative to human graders.

Our study explores the efficacy of two different CAS systems: SIDE (The Summarization Integrated Development Environment) and SPSSTA (SPSS Text Analysis for Surveys 3.0) at a very fine grain size: ability to detect key concepts of natural selection (Nehm and Reilly, 2007). We discuss each program in detail below.

SIDE: The Summarization Integrated Development Environment

SIDE (The Summarization Integrated Development Environment) is a computer program designed to manage and analyze text data (Mayfield et al., 2009). This program utilizes machine learning to summarize text data, organize information, or analyze conversational data. In the present study, we focus on the usefulness of this program as a computer assisted scoring (CAS) tool for analyzing short-answer responses to a natural selection test known as the ORI (Open Response Instrument; Nehm and Reilly, 2007). SIDE works by building a scoring model using several different machine-learning approaches. Specifically, the SIDE program analyzes the individual features of text elements in a text corpus and builds a 'feature table' for the text corpus. Then, the program builds a model (using machine learning) based upon the feature table and applies this model to the raw text that has been provided to the program. SIDE subsequently compares a human rater's scoring with the program's own scoring model and displays agreement statistics (i.e., a Kappa value). This so-called "training" step is used to build a scoring model that can be saved and subsequently applied to new data sets containing similar text responses.

Overall, SIDE "learns" how to analyze text by (1) examining human scoring patterns, (2) building a model using machine learning based on a feature table, and (3) testing the model by comparing it to the human-derived scores. Importantly, SIDE permits users to revise the final model manually in situations in which the Kappa values are not sufficient. This revised model may also be applied to new text corpora. It is important to note that *training* is a different step than *testing* of the model. That is, it is possible that a SIDE scoring model derived from training—and displaying excellent Kappa values—may be a poor match using a new data set. Thus, both training *and* testing must be used to establish the efficacy of the scoring model derived from SIDE.

SPSS Text Analysis 3.0

We have previously employed the SPSS Text Analysis 3.0 program as a CAS tool for evolution assessment (Nehm and Haertig, 2011). Because a complete user's guide for SPSSTA is available, we only provide a very brief summary of how the program works. Overall, SPSSTA uses linguistic-based techniques to identify, extract, and classify text. These classifications are based upon semantic networks and term co-occurrences that must be built, largely manually, using knowledge of the content area that will be analyzed (in our case, evolution). For scoring a particular concept such as evolution, it is necessary to develop a lexical library (list of specialized words and phrases) and text "rules" (directions for what text should be identified by the program) because SPSSTA does not include verbs or most specialized biology terms. Hence, a very large amount of work in SPSSTA—which is not

needed in SIDE—is required for the designation of term libraries and the development of semantic rules prior to system training and testing. Previous work (Nehm and Haertig, 2011) built the term libraries and lexical rules for SPSSTA and tested them on a response corpus from the ORI instrument.

Study purpose and methodology

The primary aim of this study is to compare the efficacy of two CAS tools (SIDE and SPSSTA) for measuring student knowledge of a core idea in the biological sciences: natural selection. Methodologically, we use Cohen's Kappa to quantify and compare score agreement magnitudes among CAS-derived scores and human expert derived scores (note that Kappa statistics compensate for chance inter-rater agreements). We employ a dataset of undergraduate biology majors' written responses to examine the comparative efficacy of the two computer assisted scoring tools (SIDE and SPSS). Student essays were also scored by evolution experts using published scoring rubrics (Nehm et al. 2010). Our text corpora are comprised of student responses to the ORI instrument (Bishop and Anderson, 1990; Nehm and Reilly, 2007) and the EGALT instrument (Nehm et al., 2010). The ORI data sets were collected in 2008 and 2009 from biology major students at a major public research university. The EGALT data set was collected in 2009 from a different group of biology majors at the same university. The expert-derived scores of ORI and EGALT are used as a benchmark to compare the efficacy of SIDE and SPSSTA.

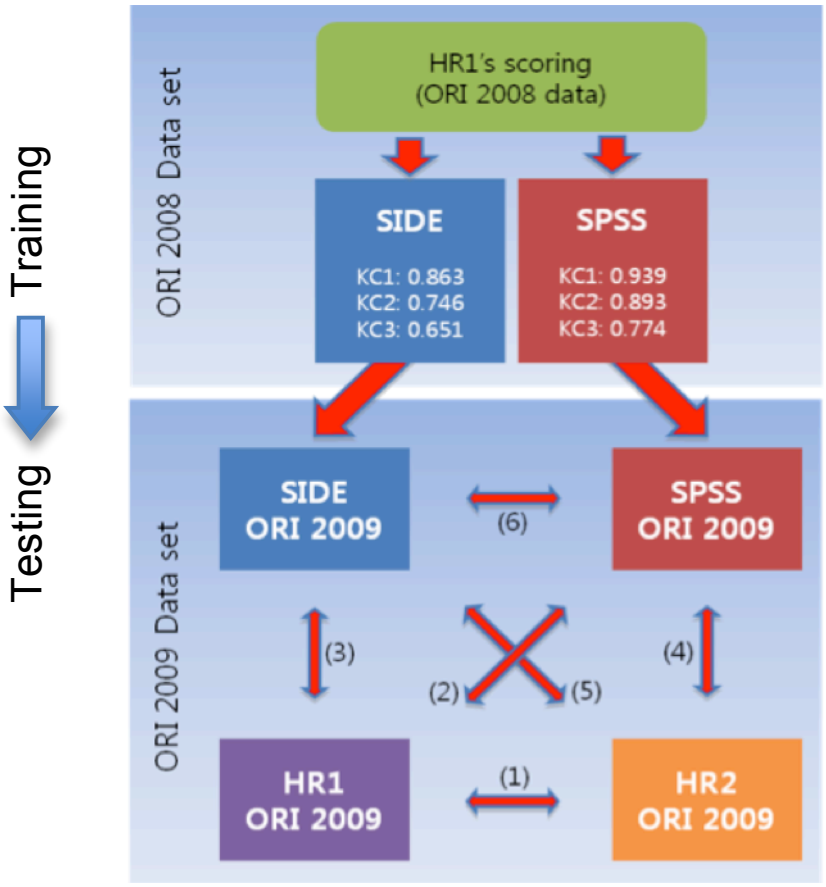
This study focuses on the ability of SIDE and SPSS to measure three core elements of natural selection. Although there is disagreement in the literature about the number of "essential" elements of natural selection, at a minimum, three key or core concepts (KCs) are universally recognized as necessary and sufficient to explain evolutionary patterns using the natural selection model: (1) The presence and causes of variation (mutation, recombination, sex) (KC1); (2) The heritability of variation (KC2); (3) The differential reproduction and/or survival of individuals (KC3) (Lewontin, 1970; Pigliucci & Kaplan, 2006). All students' responses were scored for these three key concepts by (1) the human expert rates (2) SIDE and (3) SPSSTA. Kappa agreement statistics were calculated among the scoring methods.

Results

Testing the training models

The first step of our analysis involved testing the efficacy of the CAS methods by examining the correspondence among the training results for SIDE, SPSSTA, and the expert-derived scores. Kappa values are indicated in the top box of Figure 1 [SIDE: KC1 (0.863), KC2 (0.746), KC3 (O.651); SPSSTA: KC1 (0.939), KC2 (0.893), KC3 (0.774)]. Overall, it is apparent that while both methods performed very well (all displayed Kappa values > 0.6 and many reached 0.9), SPSSTA performed slightly better on the training data set. However, the overall superiority of SPSSTA must take into account the fact that hundreds of hours were spent developing the lexical libraries and rules for SPSSTA, whereas almost no time was spent performing these activities in SIDE. Nevertheless, both CAS methods are effective at scoring evolution essays.

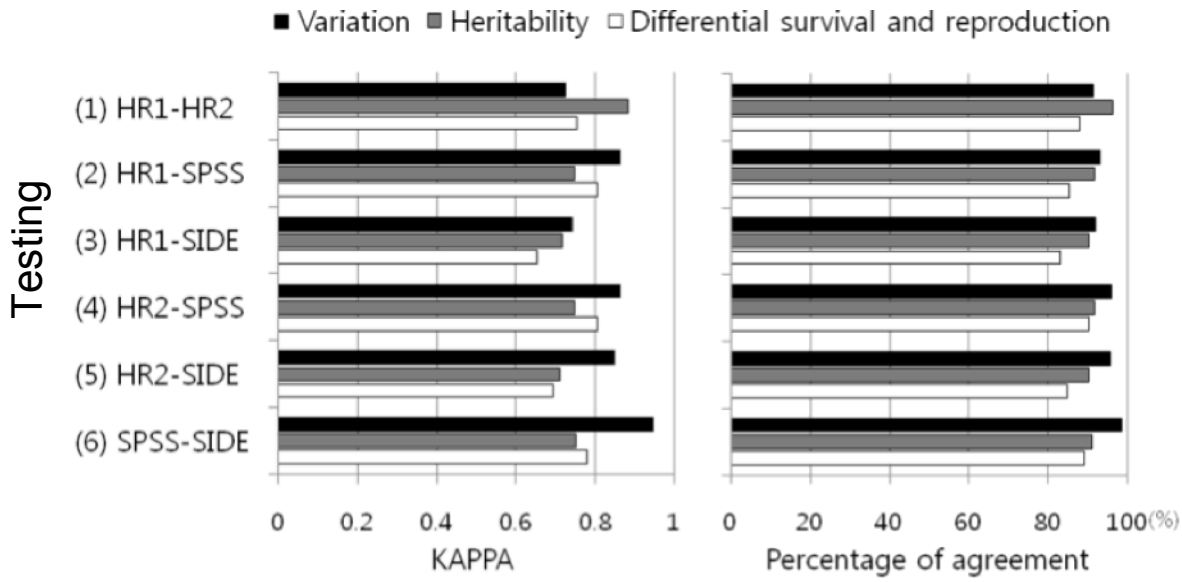
Figure 1.*Diagram illustrating the two key steps of the present study (upper blue box and lower blue box). In the first step, an expert rater's scores are used to train both SIDE and SPSS. The resulting Kappa values are shown to be acceptable for the three core concepts of natural selection (see below). In the lower box, which is the focus of the present study, the SIDE and SPSS program are tested on a new dataset, and Kappa comparisons are made among CAS and human expert-derived scores.*



Testing the training models on a new data set

The aim of the second analysis was to determine if the scoring models built during the training step generalize successfully to a *new* data set (that is, responses from *different* students but the *same* items/instrument). We performed six comparisons (See Figure 1, lower panel, numbers 1-6). The results for these six comparisons are shown in Figure 2, below.

Figure 2. *Testing the training models on a new data set (ORI responses). Left panel: Kappa values. Right panel: Percentage agreements.*

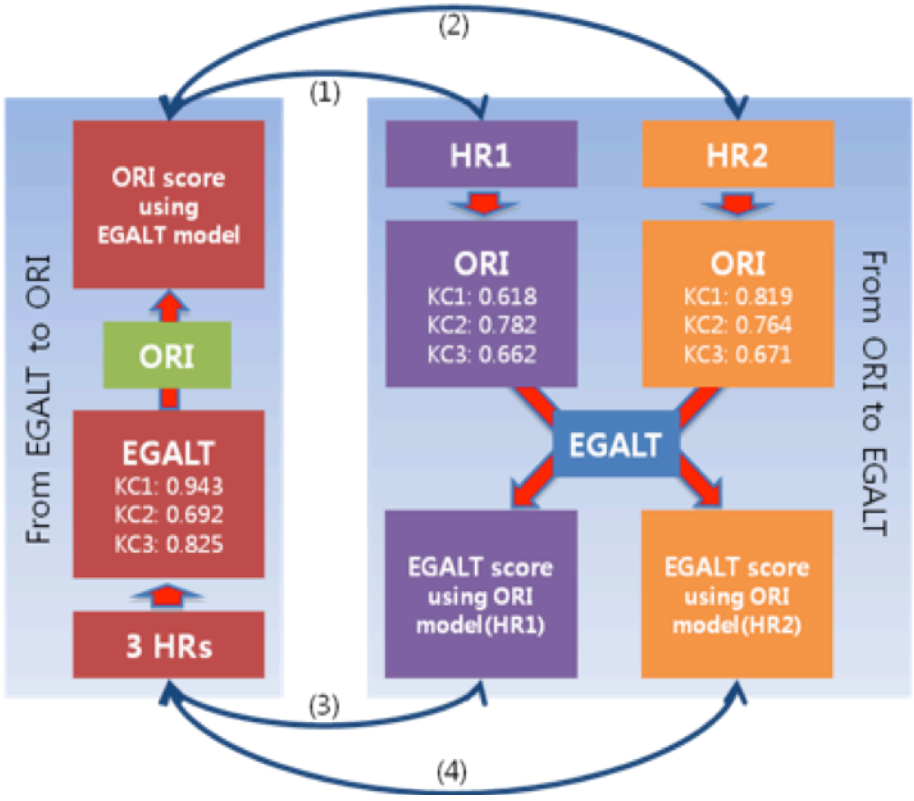


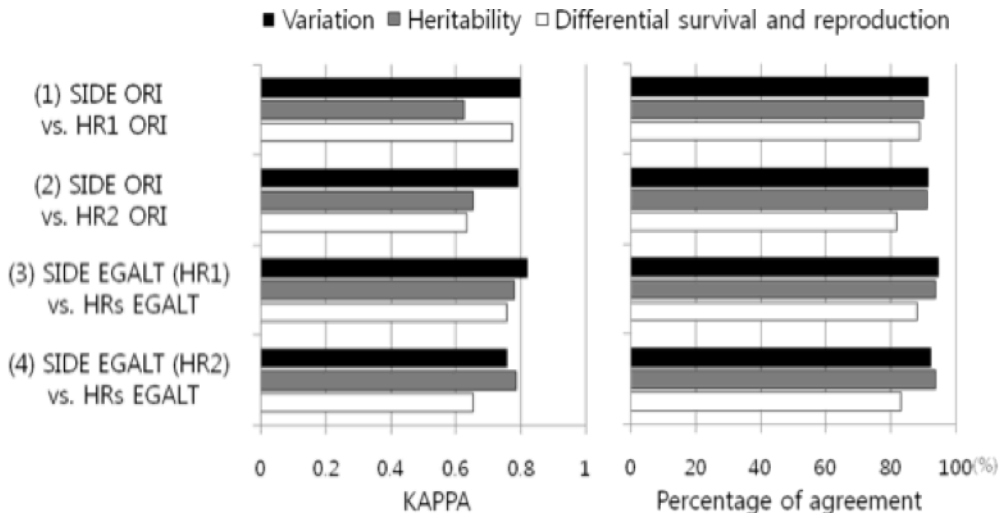
As illustrated in Figure 2, the training models generated in SIDE and SPSSTA successfully generalized to new items. In all comparisons, Kappas were > 0.6, and in several cases they exceeded 0.9), and in no instances did agreement percentages drop below 80%. Surprisingly, the two CAS tools agreed with one another almost as much as the two expert human raters did with each other. Overall, it is clear that both SIDE and SPSSTA produce very similar results, although SPSSTA produced slightly higher agreements with the human raters. Recall that similar results were found for the training dataset.

Testing the training models on a new evolution instrument

Our final analysis explores whether the training models that were shown to be successful on a new data set from the same instrument (see Figure 2, above) generalize to responses from a *different* instrument (the EGALT). In other words, while our results (displayed above) show that both SIDE and SPSS work effectively on new sets of responses to the ORI, will they work with completely different items about evolution? In order to answer this question, we performed a similar set of analyses as above (see Figure 3 left panel for a schematic of the comparisons, and Figure 3 right panel for the results). As Figure 3 illustrates, using a different set of items (and associated student responses) decreases the Kappa values. Nevertheless, in all cases Kappas were > 0.6, and many reached 0.8. In all comparisons, agreement percentages were > 80%.

Figure 3. *Comparisons of computer assisted scoring among human raters, SIDE and SPSSTA. In this case, efficacy was tested using an instrument that was different from the one in which the model was developed.*





Conclusions

Our study of computerized text analyses of evolutionary explanations written by biology undergraduates demonstrated that: (1) CAS tools (both SIDE and SPSSTA) successfully identified fine-grained explanatory elements (Key Concepts of natural selection) and (2) in the vast majority of cases the CAS tools generated scores that were comparable to human-generated assessment scores. Our analyses indicate that both SIDE and SPSSTA approximate human-human agreement levels and are effective tools for scoring evolution essays. Nevertheless, the machine learning approach used in SIDE requires significantly less programming, set up time, and content expertise than SPSSTA.

Acknowledgments

We thank Dr. Hendrik Haertig for his collaboration and help with SPSSTA; Judy Ridgway for help with data collection; and the National Science Foundation (REESE 0909999) for funding parts of this study. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF.

References

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. Journal of Applied Psychology, 76(4), 522-532.

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

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. Journal of Educational Measurement, 27(2), 93-108.

Burstein, J. C. (2001). Automated essay evaluation in criterion. Paper presented at the Association of Test Publishers Computer-Based Testing: Emerging Technologies and Opportunities for Diverse Applications conference, Tucson, AZ.

Clauser, B. E., Harik, P., &Clyman, S. G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. Journal of Educational Measurement, 37(3), 245-261.

Keith, T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 147-168). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kingston N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. Applied Measurement in Education, 22(1), 22-37.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000).The Intelligent Essay Assessor. IEEE Intelligent Systems, 15(5), 27–31.

Landauer, T. K., Laham, D., & Foltz, P. W. (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, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of

essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 87-112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Mayfield, E., Kang, M., Rosé, C. (2009). SIDE: The Summarization IDE: User manual. Carnegie Mellon University.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. Applied Measurement in Education, 15, 363–389.

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

Nehm, R. H., and Haertig, H. (2011). Human vs. computer diagnosis of mental models of natural selection: Testing the efficacy of lexical analyses of open response text. Journal of Science Education and Technology.

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

Page, E. B. (1966). The imminence of grading essays by computer. Phi Delta Kappan, 48, 238–243.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.). Automated essay scoring: A cross-disciplinary perspective (pp. 43-54). Mahwah, NJ: Lawrence Erlbaum Associates.

Pigliucci, M. & Kaplan J. (2006) Making Sense of Evolution: The Conceptual Foundations of Evolutionary Biology. Chicago: University of Chicago Press.

Shermis, M. D., & Burstein, J. (2003). Introduction. In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. xiii-xvi). Mahwah, NJ: Lawrence Erlbaum Associates.

SPSS (2006). SPSS Text analysis for surveys 2.0 user's guide. Chicago, IL: SPSS, Inc.

Sukkarieh, J., & Bolge, E. (2008). Leveraging c-rater's automated scoring capability for providing instructional feedback for short constructed responses. In B. P. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (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 (pp. 779-783). New York: Springer-Verlag.

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). “Mental model” comparison of automated and human scoring. Journal of Educational Measurement, 36, 158–184.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. Applied Measurement in Education, 15(4), 391-412.