R&D Connections
April 2005

E-rater as a Quality Control on Human Scores

William Monaghan and Brent Bridgeman

Can natural language processing evaluate the quality of writing?

Should computers replace humans in analyzing student essays?

The answers depend on whom you ask.

Opponents of automated essay evaluation systems claim that computers lack the intrinsic human capacity to determine good writing from bad. However, testing organizations see such capabilities as being a necessity to efficiently score essay tests (Flam, 2004). A suitable compromise would be to have human readers score essays in tandem with an automated essay evaluation system, such as the ETS-developed e-rater®. The approach benefits those in the testing industry by creating less reliance on expensive readings and lessens the concerns of critics, as human readers are an integral element in the system.

The debate over the efficacy of using an essay format in tests has a long history (Cooper, 1984). Testing programs, such as the Graduate Record Examinations® (GRE®) program, have come to recognize that essays can play an important role as indicators of student ability and have added essay sections (Powers, Fowles, & Welsh, 1999). The Advanced Placement Program® has always utilized essays, and the College Board® has added an essay section to the SAT®.

For those in the testing industry, however, essays present a practical problem—how to efficiently develop, administer, and score tests with essay sections. This paper focuses on the scoring of essays and the role automated essay evaluation systems can play in the process.

Why Automated Essay Scoring

ETS made its mark by standardizing and then automating much of the testing process. This was done out of necessity as much as for creating systems in which all test takers can demonstrate their proficiency in a common, fair way. Few reasonable options are available to administer tests to millions of students and complete the reporting of scores in a timely manner. While ETS has focused primarily on multiple-choice tests in these efforts, the organization has been a pioneer in ways to use essays in such testing.

When using essays for assessment purposes, ETS has found that having a single essay question or prompt and a single reader per essay does not produce reliable scores (Breland, Bridgeman, & Fowles, 1999). The remedy is to have test takers write two essays if not more and to have at least two people read and rate each essay. Scoring costs for such a test are substantial. ETS has held annual massive readings for some of its test administrations involving essays. This has meant bringing together a small army of educators to a single location and having them read through and score essay after essay. Even moving such a system online requires hours of training and logistical support for each rater. Recruitment for each of these systems alone can be a daunting task as qualified individuals are relatively few, and they usually have pressing schedules. Compensation for the raters' time and possible travel is a huge expense that is passed along to test takers as additions to their registration fees.

That is why the organization has invested in and developed automated essay evaluation

capabilities such as e-rater. In the e-rater system, the computer is fed thousands of essays that human raters have scored. The essays range from those deemed to be high-quality responses to ones seen to be less than adequate. To score an essay, the system is set up to look for patterns that are evident in better essays. The system accomplishes this task in seconds. Studies show a high level of agreement between the scores human raters assign to an essay and what e-rater awards (Attali & Burstein, 2005).

Text vs. *Context*

Even with this high-level of agreement and e-rater's apparent efficiency, a number of people still object to the idea of automated essay evaluation. They argue, and rightly so, that such systems can be fooled by clever nonsense or the inclusion of well-constructed sentences that together make no sense at all. This assumes that a human reader, who would detect such cases, is not in the scoring model at all. The opposite fear is to have brilliant writing constructed in such a nonconformist manner that the machine assigns a poor score. Again, a reader should be an effective guard against such a situation. Of course, students seeking instruction would have little to gain in using e-rater outside of its intended function.

Another worry is that automated essay systems might be less valid for use in the scoring of essays written by English language learners. Will a machine that is trained on the writing of native English speakers work in a situation where the majority of the testing population doesn't speak English as a native language? Will systems like e-rater have the same kind of validity in such instances?

Bridgeman (2004) says that a possible solution is to use e-rater to check the scores assigned by human raters. By having e-rater run in the background, the score e-rater provides can be compared to the one assigned by a single human rater. If there is no discrepancy, the score stands. If the scores are discrepant, a second human reader receives the essay to see if a factor such as fatigue affected the score the first rater assigned or if the essay has elements that are unduly influencing the automated system. In this system, the essay score would always be based solely on human raters.

The approach allows testing organizations to streamline the essay evaluation process while still providing valid score reporting.

Testing e-rater as Quality Control

To test his model, Bridgeman (2004) turned to the GRE analytical writing section, which has each test taker write two essays—one on an issue prompt and the other on an argument prompt. Jill Burstein, the lead scientist on the e-rater system and a computational linguist, developed e-rater scoring models for more than 100 prompts of each type (issue and argument). For the issue prompts, the e-rater scores agreed with the scores assigned by a human rater at the same rate that one human agreed with another. For the argument prompts, agreement of e-rater and human raters was slightly lower, but still quite high. The correlation between the scores assigned by two humans was .81, and the correlation of a human score and e-rater score was .76.

To evaluate the effectiveness of using e-rater as an additional score or as a check on the score from one human rater per prompt, Bridgeman studied 5,950 examinees who had taken the GRE analytical writing section twice. He used the final score based on at least four human ratings (two for each prompt) from one administration as the criterion.[1] This criterion provides an estimate of writing ability that is totally independent of the estimate made from using e-rater either as an additional rater or as a check. The criterion was predicted from scores on a different administration that were based on two humans

[1] A single score is reported for the analytical writing section. Each essay receives a score from two trained readers using a 6-point holistic scale. In holistic scoring, readers are trained to assign scores on the basis of the overall quality of an essay in response to the assigned task. If the two assigned scores differ by more than 1 point on the scale, the discrepancy is adjudicated by a third GRE reader. Otherwise, the scores from the two readings of an essay are averaged. The final scores are based on two essays (one a response to an issue prompt and the other a response to an argument prompt) that are then averaged and rounded up to the nearest half-point interval (e.g., 3.0, 3.5).

Table 1

Agreement When Criterion Is Analytical Writing Total From a Different Administration

Readers per prompt	Within ½ point	Within 1 point
2 humans	76.6%	94.0%
1 human	72.9%	92.5%
Checked human	75.5%	93.9%
1 human + e-rater	77.7%	94.2%

per prompt, one human per prompt, one human with the e-rater check procedure resulting in a second human rating about 15% of the time, or one human plus e-rater. Results are summarized in Table 1.

The highest agreement, even higher than two human readers per prompt, was found when the score assigned from one human reader was combined with the e-rater score. But if test users are uncomfortable with having a score assigned by a machine being part of a person's score, the checked human approach results in agreement rates that are nearly as high.

Summary

Automated essay evaluation systems, such as e-rater, have a very high threshold to meet to gain people's full confidence as a valid scoring approach. This skepticism is healthy, and until these systems reach a level of sophistication to make such concerns unwarranted, educational measurement organizations should be judicious in the use of these systems, especially in assessments that help in making high-stakes decisions, such as those used in admissions.

However, automated essay evaluation systems do have value if properly used. One such valid application, as this paper establishes, is as a quality control check on humans rating essay prompts. To produce reliable scores when using essays in assessment, multiple topics and multiple readers are necessary. Arranging for human readers is a time-consuming and costly task and one for which educational measurement organizations can considerably lessen the burden with e-rater. The results described here show that the highest reliability was obtained by combining a human reader's score with that generated by e-rater.

ETS has and continues to explore other uses for e-rater as it works to perfect the system. Even this seemingly limited usage of this capability can reap awards by making essay scoring more efficient and less costly. Of course, test takers are the ultimate beneficiaries, as they will have another avenue besides multiple-choice testing to demonstrate their true abilities.

References

Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater v.2.0* (ETS RR-04-45). Princeton, NJ: ETS.

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework* (College Board Research Rep. No. 99-03, GRE Board Research Rep. No. 96-12R, ETS RR-99-03). New York: College Entrance Examination Board.

Bridgeman, B. (2004, December). *E-rater as a quality control on human scorers.* Presentation in the ETS Research Colloquium Series, Princeton, NJ.

Cooper, P. L. (1984) *The assessment of writing ability: A review of research* (GRE Board Research Report No. 82-15R, ETS RR-84-12). Princeton, NJ: ETS.

Flam, F. (2004, August 30). An apple for the computer. *The Philadelphia Inquirer,* p. D-01.

Powers, D. E., Fowles, M. E., & Welsh, C. K. (1999). *Further validation of a writing assessment for graduate admissions* (GRE Board Research Rep. No. 96-13R, ETS RR-99-18). Princeton, NJ: ETS.

How e-rater Works

Earlier versions of e-rater had some 50 features, and a subset of these features would be selected to score the particular set of essays. The newer version of e-rater uses a fixed set of about 10 features in seven categories from which it derives the final score.

Explanation of the seven score categories

- *Grammar score* – based on errors such as those in subject-verb agreement among others
- *Mechanics score* – derived from errors in spelling and other like errors
- *Usage score* – based on such errors as article errors and confused words (an example would be an instance in which the essay writer uses a word that although phonetically similar has a different meaning from the intended word; using "to" where it would have been proper to use "too")
- *Style score* – based on instances of overly repeated words and the number of very long or very short sentences as well as other such features
- *Lexical complexity score* – drawn from information such as the level of vocabulary the essay writer uses in the essay
- *Organization/development score* – based on the identification of sentences that correspond to the background, thesis, main idea, supporting idea, and conclusion
- *Prompt-specific vocabulary usage score* – derived from e-rater's evaluation of the word choice in an essay and the similarity to the word choice in samples of low- to high-quality essays written on the same topic

In addition to these seven score categories, essay length also may be considered and weighted in a controlled way.

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001

Send comments about this publication to the above address or via the Web at:

http://www.ets.org/research/contact.html

Listening.
Learning.
Leading.