The Efficacy of Automated Essay Scoring for Evaluating Student Responses to Complex Critical Thinking Performance Tasks

Jeffrey T. Steedle
Council for Aid to Education

Scott Elliot
SEG Measurement

Please send correspondence to:

Jeffrey T. Steedle
Council for Aid to Education
215 Lexington Ave., 21st Floor
New York, NY 10016
212-217-0728
jsteedle@cae.org

Abstract

Automated essay scoring (AES) is commonly employed to score content-based and persuasive essays, but there exists pervasive skepticism about the use of AES to score essay tasks requiring substantial critical thinking and problem solving abilities. In this investigation, AES engines were developed to score five Performance Tasks administered as part of the Collegiate Learning Assessment. Such tasks give examinees 90 minutes to analyze a set of documents and compose an essay explaining their solutions to a problem. AES performed favorably in comparisons of AES to human and human to human agreement, suggesting that AES is an effective tool for scoring CLA Performance Tasks.

The Efficacy of Automated Essay Scoring for Evaluating Student Responses to Complex Critical Thinking Performance Tasks

Automated Essay Scoring (AES) is widely used in scoring student written responses to assessment tasks and has been demonstrated as reliable and valid in many settings (cf. Shermis & Burstein, 2003). From K-12 instructional tools (e.g., Criterion™) to higher education admissions (e.g., GMAT™), AES is routinely used in the operational scoring of written responses.

Written performance assessments are likely to play a major role in the new generation of assessments as the Common Core State Standards movement progresses. Performance assessment is seen as a critical vehicle for helping ensure that students *of all ages become successful members of our global society* (the theme for the 2012 Annual Meeting of NCME). As the use of performance assessment grows, we are likely to see a corresponding growth in the use of AES technology.

While AES is increasingly accepted for use in evaluating content essays (i.e., those with correct responses grounded in content knowledge) and for student responses to narrative, informative, and persuasive essay prompts, AES technology has not been widely applied for tasks requiring critical thinking, analysis, and problem solving. The reasons for this are not well documented. However, most arguments against using AES for this purpose center on a belief that sophisticated levels of thinking are the purview of the human brain and are not well-suited to AES.

In early 2009, CAE investigated the use of AES to evaluate written responses to Performance Tasks included in the Collegiate Learning Assessment (CLA). These tasks require students to analyze multiple sources of information and provide recommendations, decisions, or problem solutions based on that analysis. The perceived benefits of using AES included increased scoring accuracy, reduced costs, and reduced time between the test administration and delivery of score reports. The results presented here demonstrate that AES can be employed successfully for scoring complex, critical thinking tasks.

Methods

The CLA

The CLA is a computer-delivered assessment administered at hundreds of colleges and universities to evaluate student growth in critical thinking, analytic reasoning, problem solving, and written communication skills. Students participating in the CLA complete a Performance Task requiring them to analyze a set of documents containing a mixture of credible and unreliable information and propose a course of action to solve a problem. In one Performance Task, for example, students must evaluate the relative efficacy of plans for reducing crime in a small city.

Each CLA task is developed by a team of subject matter experts and psychometricians. The tasks are field tested at several colleges and universities. Following any necessary revisions, it is provisionally introduced for operational use. After the initial administration, each student response is scored by two expert scorers trained to apply the CLA scoring rubric.[1] The results of the operational scoring by expert scorers are reviewed to ensure that the task is performing as expected and is technically sound.

Latent Semantic Analysis

In this study, a technique known as Latent Semantic Analysis (LSA) (Landauer, Laham, & Foltz, 2003) was employed for AES. LSA is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (Landauer & Dumais, 1997). The underlying idea is that the totality of information

[1] http://www.collegiatelearningassessment.org/files/CLAScoringCriteria.pdf

about all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meanings of words and sets of words to each other. It uses singular value decomposition, a general form of factor analysis, to condense a very large matrix of word-by-context data into a much smaller representation (Landauer & Dumais, 1997). The similarity between resulting vectors for words and contexts, as measured by the cosine of their contained angle, has been shown to closely mimic human judgments of meaning similarity and human performance based on such similarity in a variety of ways (Landauer, et al., 2003).

AES Training Process

To develop AES models, a broad sample of approximately 500 responses for each of 5 Performance Tasks was drawn from previous test administrations. Each response was then scored by two trained scorers using the established CLA rubric. The rubric includes 4 trait scores (2 emphasizing critical thinking, 2 emphasizing writing) and an overall score equal to the sum of the trait scores. The average score across the two scorers was calculated for each response and used to train the Intelligent Essay Assessor (IEA) AES engine developed by Pearson Knowledge Technologies. The engine “learns” the characteristics of each score point based on the responses used to train it. Thus, the grades assigned by human scorers remain the basis for scoring CLA assessment tasks; computers are used to codify, organize, and synthesize the collective knowledge of the human scorers to assist in the scoring process and improve its speed and accuracy.

The scores produced by the IEA engine were compared to those assigned by the two human scorers on each of the traits and for the overall score for each of the five tasks. Specifically, the Pearson correlation, exact agreement, and adjacent (±1) agreement were calculated for the following scorer combinations: Human 1 to Human 2, IEA to the average Human scores, and the average of IEA to Human 1 and IEA to Human 2.

Results

In all, 25 AES models were developed [5 tasks × (4 trait scores+1 overall)]. With only one exception (Problem Solving on PT2), Intelligent Essay Assessor (IEA) had better agreement with the average of two expert scorers than the two experts had with each other (Table 1). As illustrated in Figure 1, the correlation between two experts ranged from .67 to .83 (average .77), and the correlation between IEA and the average of the experts ranged from .79 to .93 (average .87) considering only the 1-6 scale "trait" scores ("Overall" scores tend to have higher correlations). These correlations were computed using a "leave-one-out" methodology that eliminates potential bias caused by the inclusion of training papers in the evaluation.

The same pattern of results is apparent in the columns showing exact and adjacent agreement rates. Although the correlations were high for the "Overall" scores, one would expect fairly low agreement rates because the scale ranges from 4 to 24 (not just 1 to 6).

A fairer comparison between IEA and human scoring can be obtained by examining agreement between IEA and individual humans (rather than their average, which is what IEA uses to develop the AES engine). A comparison of the first and third sets of columns in Table 1

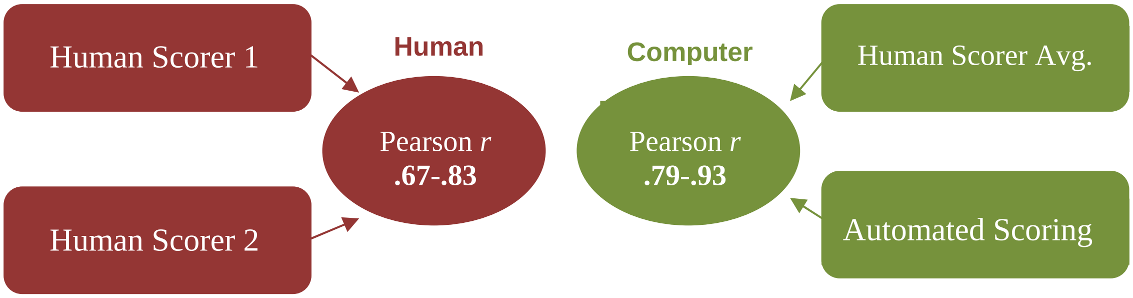

Figure 1. Scoring accuracy: human vs. automated scoring (trait scores only).

Table 1

Comparison of human and automated essay scoring

		Human1-Human2			IEA-Average Human			Average of IEA-Human1 and IEA-Human2		
Prompt	Score	*r*	Exact	±1	*r*	Exact	±1	*r*	Exact	±1
PT1	ARE	.81	.49	.94	.91	.62	.98	.84	.55	.98
	PS	.79	.42	.93	.88	.64	.97	.86	.52	.97
	WE	.79	.38	.90	.91	.65	1.00	.85	.52	.98
	WM	.70	.44	.92	.86	.62	.98	.77	.54	.97
	Overall	.86	.14	.40	.93	.20	.65	.90	.20	.52
PT2	ARE	.83	.48	.93	.85	.55	.94	.80	.49	.92
	PS	.82	.51	.94	.79	.50	.91	.74	.43	.90
	WE	.81	.46	.88	.87	.58	.97	.71	.20	.56
	WM	.82	.43	.94	.85	.57	.97	.79	.51	.95
	Overall	.88	.17	.44	.88	.17	.40	.85	.16	.45
PT3	ARE	.80	.46	.94	.92	.65	.99	.85	.57	.97
	PS	.82	.47	.94	.90	.56	.99	.84	.52	.97
	WE	.81	.44	.90	.93	.64	.99	.88	.48	.98
	WM	.69	.36	.87	.84	.50	.97	.76	.47	.93
	Overall	.84	.14	.40	.93	.28	.59	.89	.17	.51
PT4	ARE	.79	.48	.95	.89	.61	.98	.83	.58	.97
	PS	.68	.45	.93	.83	.60	.97	.73	.54	.95
	WE	.67	.45	.91	.91	.72	1.00	.82	.57	.97
	WM	.70	.51	.96	.82	.63	.98	.72	.58	.97
	Overall	.85	.20	.54	.93	.26	.72	.89	.22	.64
PT5	ARE	.81	.43	.94	.87	.52	.97	.82	.50	.95
	PS	.80	.44	.89	.89	.56	.97	.83	.46	.96
	WE	.77	.45	.91	.90	.53	.97	.82	.51	.95
	WM	.73	.44	.90	.83	.54	.98	.76	.45	.95
	Overall	.86	.14	.41	.92	.21	.52	.88	.16	.46

Note: ARE = Analytic Reasoning and Evaluation (identifying and interpreting relevant information, evaluating the credibility of information), PS = Problem Solving (synthesizing information, making a decision, recognizing where matters are left uncertain), WE = Writing Effectiveness (constructing an organized and cohesive essay with support for positions), WM = Writing Mechanics (demonstrating command of Standard Written English), and Overall = sum of trait scores.

reveals that IEA agrees better with human experts than humans agree with each other in all cases except PT2. This aberrant result may reflect unique feature of this prompt or some deficiency in the training data (e.g., responses that were not fully representative of the range of common responses).

As a summary of results, Table 2 presents average correlations in the five score scales across the five Performance Tasks. On average, IEA agreed better with expert scorers then the expert scorers agreed with each other on all four trait scores and the overall score. Comparing the first and third columns in Table 2, the largest difference in the average correlations (.05) occurred on the Writing Effectiveness scale. Differences were typically about .02 favoring IEA.

Table 2
Efficacy index (average correlations) across five Performance Tasks

Score	Human1-Human2	IEA-Average Human	Average of IEA-Human1 and IEA-Human2
ARE	.81	.89	.83
PS	.78	.86	.80
WE	.77	.90	.82
WM	.73	.84	.76
Overall	.86	.92	.88

Discussion and Conclusion

This study suggests that AES is an effective tool for scoring the CLA Performance Tasks. The agreement rates produced by the IEA AES engine are, for the most part, better than those produced by human scorers. This was true on scales reflecting writing quality as well as those indicating examinees' demonstration of critical thinking and problem solving skills. Explanations for deviations in this trend may be obtained as AES engines are developed for additional CLA Performance Tasks.

No single study can conclusively establish the accuracy of AES. Moreover, as Bennett (2006) cautions, agreement and correlation represent only one component of the validity of automated essay scoring. While this study strongly suggests that AES is effective for scoring CLA Performance Tasks, additional studies need to be conducted. Most notably, the scores produced by AES will need to be compared to other measures of the construct as part of future validation efforts.

As with most studies of this nature, the results raise an important question: Why is AES at least as effective as human scorers? This question is particularly salient because AES was employed here to score critical thinking tasks, a feat that many assume AES is incapable of. The future answer to this question must extend beyond the general descriptions of latent semantic analysis offered by the engine vendor. Despite the lack of an answer, AES will be employed by the CLA to provide accurate, fast, and cost-effective value-added assessment services to institutions of higher education.

References

Bennett, R. E. (2006). Moving the field forward: Some thoughts on validity and automated scoring. In D. M. Williamson, R. J. Mislevy & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 221-240.

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. C. Burstein (Eds.), *Automated essay scoring: A cross disciplinary perspective* (pp. 87-112). Mahwah, NJ: Lawrence Erlbaum Associates.

Shermis, M. D., & Burstein, J. C. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Additional Readings

Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross disciplinary perspective* (pp. 71-86). Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*(2 & 3), 259-284.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kaplan, 76*(6), 561-566.