

**April
2011**

[CLA]

The Council for Aid to Education

[COMPUTER-ASSISTED SCORING OF PERFORMANCE TASKS FOR THE CLA AND CWRA]

As with most major assessment programs, the Collegiate Learning Assessment (CLA) is constantly looking for ways to improve the quality of its assessments. An important part of this effort is improvements in scoring of student written responses. To increase scoring accuracy and reduce the amount of time between administration and report, CLA now employs a computer-assisted scoring process. This process uses Latent Semantic Analysis (LSA)—a mathematical/statistical technique—to create computer models that can be used to assist in the scoring of written responses. The computer-assisted process requires extensive scoring of written responses by trained human scorers using a well-articulated scoring rubric. Research shows that the computer-assisted scoring used by CLA is as accurate as or more accurate than human expert scoring.

COMPUTER-ASSISTED SCORING OF PERFORMANCE TASKS FOR THE CLA AND CWRA

INTRODUCTION

The computer has become pervasive in nearly all aspects of our lives. From communications to transportation, computers have changed how we accomplish virtually all that we do. We rely on computers to assist with a wide range of academic and daily activities. Education is no exception.

The Collegiate Learning Assessment (CLA), like most large-scale assessment programs, now makes extensive use of computers to carry out program activities. Institutions and students are enrolled via computer, the assessments are delivered via computer, and scores are created and delivered by computer. Computer-assistance is central to CLA's ability to deliver assessments and associated results accurately, rapidly, and cost effectively.

One innovative application of computer technology within the CLA program is in the scoring of student responses to assessment tasks. To increase scoring accuracy, reduce the amount of time between administration and report delivery, and reduce costs, CLA employs a computer-assisted scoring process. While the grades assigned by human scorers remain the basis for scoring CLA 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.

Computer-assisted scoring is not unique to CLA. Computer-assisted scoring is used by many large-scale assessment programs. For example, both the Graduate Management Aptitude Test (GMAT) and the Graduate Record Exam (GRE) employ computer-assisted scoring.

HOW COMPUTER-ASSISTED SCORING WORKS

Computer-assisted scoring employs computer models to score open-ended assessment responses. These models are created from the scores assigned by trained and calibrated graders. The computer uses these grades to operationally infer the rubric and scoring scale. The computer-assisted scoring process is therefore dependent on human expert scoring. Several hundred expert-scored

student responses are used to train the computer-assisted scoring engine. The computer-assisted scoring engine learns the features and characteristics of the scoring rubric and each score point from the expert-scored responses, which it uses to evaluate student responses. The engine relies on the collective wisdom of the expert scorers as reflected by the scores they assigned to a representative set of actual student responses. Much like the training of human scorers, the engine learns how to score student responses through repeated exposure to expert-scored examples at each score point.

Computer-Assisted Training Process. Each CLA task is first developed by a team composed of psychometricians and experts in the skills measured by the assessment. The tasks are then field tested at several colleges and universities. After a task is developed and field tested, it is provisionally introduced for operational use. Following an initial administration, each student response is scored by two expert scorers, trained using the CLA scoring rubric. The results of the operational scoring by expert scorers is reviewed to ensure that the task is performing as expected and is technically sound.

Once approximately 500 student responses have been double scored by experts and the quality of the task has been verified, the results of the human expert scoring are used to generate the computer scoring model. The computer-assisted scoring engine is presented with the complete text of the approximately 500 student responses along with the expert scores assigned. The engine examines the content and structure of each response and associates the information with the score assigned to create a model of what each score point looks like. Simply put, the computer-assisted scoring engine learns the characteristics of each score point based on the 500 or so responses used to train the engine.

Latent Semantic Analysis. The CLA uses a computer-assisted technology known as Latent Semantic Analysis (LSA) (Landauer, Laham and Foltz, 2003). This technique extracts the underlying meaning in written text. LSA is a mathematical/statistical technique that captures the essential relationships between text documents and word meaning, or semantics, the knowledge base which must be accessed to evaluate the quality of content. 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 and Dumais, 1997). The underlying idea is that the totality of information 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 and 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, Laham and Foltz, 2003).

In short, the computer-assisted scoring process used by the CLA relies on the collective expertise of expert scorers derived from a mathematical analysis of about 500 student responses for each essay prompt (i.e., each answer has its own grading algorithm that was inferred by the computer from the pattern of scores assigned by the readers who graded the responses by hand). The computer-assisted scoring applies what it has learned from the expert scorers to score previously unseen student responses.

WHY THE CLA USES COMPUTER-ASSISTED SCORING

There are several reasons why the CLA uses computer-assisted scoring to evaluate student responses.

Accuracy. Our research and the results obtained by researchers throughout the academy suggest that the computer-assisted scoring is as accurate if not more accurate than expert scorers (see research results below).

Speed. Expert scoring is a very time consuming process. Computer-assisted scoring allows for us to complete scoring thousands of times faster and allows us to get reports back to colleges more quickly. Moreover, this permits instructors to use open ended assessment in teacher-driven, on-demand testing programs, which require tests to be scored and returned to the student and instructor quickly.

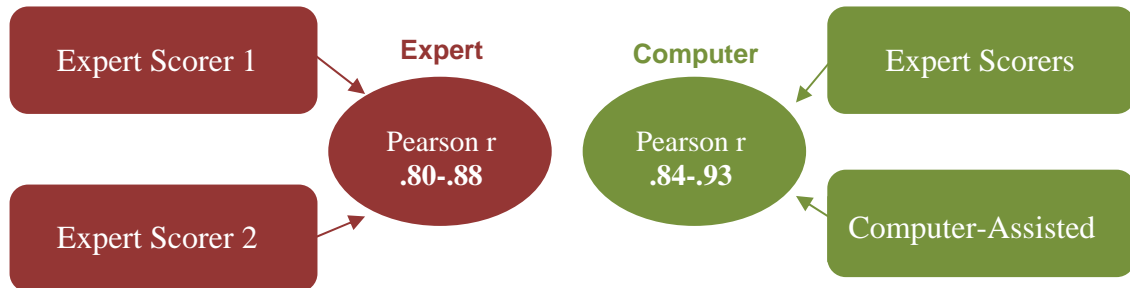
Cost. The use of computer-assisted scoring is more cost effective when there are more than several hundred answers to be scored. Computer-assisted scoring allows CAE to offer the CLA at a reasonable cost and enables institutions of higher education to affordably obtain critical value-added information.

HOW ACCURATE IS CLA'S COMPUTER-ASSISTED SCORING?

CLA computer-assisted scoring is as—and, in some cases, more than—accurate as two human scorers.

CAE conducted a study of using five CLA Performance Tasks, four CLA Make-an-Argument tasks, and four CLA Critique-an-Argument tasks. Each task was independently scored by two trained human expert scorers and then scored by the Intelligent Essay Assessor (IEA), computer-assisted scoring engine used by CAE. To evaluate the scoring model for each task, 300-500 student responses were scored by two human experts and by IEA. The percentage of times the humans agreed with each other was compared to the agreement percentage between the IEA and the human scorers. The correlation (Pearson r) between the two scorers and between IEA and the scorers was also calculated.

Scoring Accuracy: Expert v. Computer-Assisted Scoring



With the exception of one of the thirteen tasks examined, computer-assisted scoring (IEA) agreed more often with the scores of a single expert than the two expert scorers agreed with each other. For the remaining task, the percentage of agreement was the same. The correlation between the two experts ranged from .80 to .88. The correlation between the computer-assigned score and the score assigned by a single expert average ranged from .84 to .93.

While no single study can conclusively establish the accuracy of computer-assisted scoring, this study strongly suggests that computer-assisted scoring is accurate for CLA tasks. This conclusion is further supported by the more than 100 computer-assisted scoring studies conducted in education over the past decade (c.f., Shermis and Burstein, 2003).

CONCLUSION

CAE makes extensive use of computer technology to deliver the CLA program. Among the more innovative applications of computer technology is its use in the scoring of student responses. The computer-assisted scoring approach used by the CLA has been shown to be as—and, in some cases, more than—accurate as expert scorers. The use of computer-assisted scoring allows CLA to offer accurate, fast, and cost-effective value-added assessment services to institutions of higher education.

We are committed to an active program of research and development and will work to continue to enhance the quality of the CLA experience for students, faculty, and institutions of higher education.

REFERENCES

Elliot, S. (2003) IntelliMetric: From here to validity. In M. D. Shermis & J. Burstein. *Automated essay scoring: A cross disciplinary perspective* (pp. 71-86). Mahwah, New Jersey: 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), 211-240.

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

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. *Automated essay scoring: A cross disciplinary perspective* (pp. 87-112). Mahwah, New Jersey: Lawrence Erlbaum Associates.

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

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