

Pearson’s Automated Scoring Knowledge Technologies

Pearson’s automated scoring technology, the Intelligent Essay Assessor (IEA), delivers fast, accurate, and valid assessment scores. IEA combines background knowledge about English, along with prompt specific algorithms, to learn how to match student responses to human scores. In short, IEA is adapted especially for each prompt to the types of answers that hundreds of students write in response to that prompt and to the way in which human readers score those answers.

Machine learning methods analyze relevant collections of tens of thousands of background documents to see how words are used and combined into meaningful passages in the content domain of the prompts. IEA analyzes the text through a combination of Latent Semantic Analysis (LSA)—a powerful matrix algebra-based approach pioneered by Pearson principals—and methods widely used in automatic speech recognition, computational linguistics and other forms of statistical artificial intelligence. This combination of measures is designed and empirically demonstrated to accurately characterize both an answer’s substantive content and its written exposition.

With the background knowledge in place, we train the system individually on each prompt to mimic human readers in assigning scores to student responses. Using a representative sample of two hundred papers double-scored by humans, the computer compares the content and relevant qualities of writing of each student response to every other response, along with the scores given to the responses by the human readers. From these comparisons, a prompt-specific algorithm is derived to predict the scores that the same readers would assign to new responses.

“... we train the system individually on each prompt to mimic human readers in assigning scores to student responses.”

In short, IEA is trained to score like human scorers based on a training set of scored student responses. IEA measures the quality of essays by determining the language features that human scorers evaluate when scoring a response and how those features are weighed and combined to produce a score.

Score Reliability

In tests over thousands of essays, the Intelligent Essay Assessor has been shown to be as reliable as professional human scorers and more predictive of the average of two human scorers than what is predicted by inter-rater reliability. This automated scoring technology can also be used to judge traits, such as organization and conventions, with reliabilities equivalent to human graders.

The technology underlying IEA is based on the Knowledge Analysis Technologies™ or KAT engine, including Pearson’s unique implementation of LSA, an approach that is trained to measure the semantic similarity of words and passages by analyzing large bodies of relevant text. LSA can then closely approximate the degree of similarity of meaning of two texts as judged by human readers. This ability has been documented in several [articles](#) published in top-ranked refereed professional journals. The KAT engine also includes additional customized development and proprietary mathematical techniques to optimize the prediction of human scores for automated scoring applications.

Empirical Evidence

For essay scores, the same measure is customarily used to evaluate reliability and validity, how well humans—and in this case humans and an automated scoring system—agree, expert human opinions about writing being in some respects the final arbiter of writing quality. Our various studies have included reliability and validity measures of several different kinds.

Our most extensive reliability evaluation study with operational data was conducted with results from 33,205 essays written to 81 different prompts, 10 to 15 different prompts per grade level, over seven grade levels, 6th through 12th. The table below provides the average inter-rater reliability and agreement indices for holistic grades.

Reliabilities Over Grade Levels 6-12		
	Human/Human	IEA/Human
Correlation	0.86	0.90
Exact agreement	61.7	61.1
Exact + adjacent agreement	97.7	98.1

The reliability of the artificial intelligence and human scoring was, overall, as accurate as the human scorers. IEA-Human correlations were higher than Human-Human for all seven grade levels, the difference in the correlations ranging from .03 to .06. Overall, IEA

scores were correlated with average pairs of human scores on essays significantly better than the human scores were correlated with each other. Exact agreement was almost identical for human to human and IEA to human scores.

The practical conclusions from the body of reliability and validity evaluation data are that human and machine scoring methods are very nearly equivalent. Reviews of IEA's performance scoring over thousands of student responses show that it is as reliable as professional human scorers and agrees with independent humans as often as they agree with each other. In addition, the number of essays that require expert resolution with IEA is comparable to that required by an all-human scoring process. IEA is also more predictive of the average of two human scorers than the inter-rater reliability, providing consistent and accurate results.

A detailed description and explanation of validity testing using Pearson's automated scoring technology is available on the web resource:

<http://pearsonkt.com/research.shtml>.