What is Artificial Intelligence (AI) Scoring?

DRC uses specialized proprietary software to automate the process of scoring open ended item responses (i.e., "AI Scoring"). This technology is used to score student responses to the Short Write Tasks in the ELA test. DRC's AI Scoring Engine has been used to score hundreds of thousands of student responses with high reliability and accuracy.

How accurate is AI scoring?

DRC's AI Scoring Engine is trained using human-adjudicated data, which helps ensure high levels of accuracy. In recent independent studies, AI scores matched or exceeded human reliability. DRC's AI Scoring Engine results are commensurate with agreements studies comparing expert human scorers to one another. Since the engine is taught to "think" like a human scorer, its accuracy is driven by the high-quality data used in training. Additionally, AI Scoring provides perfect reliability in terms of intra-rater reliability, in that it is never fatigued and always scores a given writing sample the same way, every time. In short, DRC's AI Scoring Engine is very accurate when compared to traditional human scoring.

How are the AI models built?

DRC's AI Scoring Engine utilizes natural language processing to analyze the semantic content, grammar, syntax, vocabulary, and dozens of other detailed descriptions of students' responses. These analyses produce quantified measures of response quality which are processed by a statistical model to produce predicted scores. Similar analyses are conducted on the item rubrics, human scoring training materials, range-finding data, and exemplar item responses. All materials are combined for thousands of examples of student responses that have been independently scored twice by expert raters. These data are used to train the models to distinguish distinct levels of response quality with high accuracy. Detailed measurements are conducted to ensure congruence between the human- and AI-scored distributions, including separate calculations of recall and precision at every possible score point.

Is AI still able to recognize "Alert Papers"?

Built into DRC's AI Scoring Engine are a variety of algorithms used to evaluate the "scoreability" of student responses and to analyze any potentially disturbing content they may contain (often called "alert papers"). DRC's AI Scoring Engine flags responses that lack proper development, lack enough content to be scored, or are written in an unsupported language. Alert papers that contain inappropriate language or represent a bad faith effort to complete the test are also identified. Responses that cannot be validly scored are marked for further review by expert raters, routed to DRC's performance assessment handscoring system.

Quality Assurance

As part of the quality control process, DRC works closely with expert human raters to verify that AI scoring rules are commensurate with expectations. They are also consulted throughout the model training process to ensure that AI Scoring practice is commensurate with the human approach. This enhances the validity of scores and the extent to which AI model scores can be explained and understood. Further, one of DRC's standard practices during handscoring is to have readers score validity papers, student responses that have been scored and reviewed by content experts. To monitor the accuracy of the AI and human scoring, DRC routes the same validity papers through both processes. DRC professional staff scores a set percentage of student responses (20% for the first year) after AI scoring has been completed. This double check of scoring provides further information as to the accuracy of the AI scoring efforts.