What is Artificial Intelligence (AI) Scoring?

DRC is using automated essay scoring to score student responses to the short write task in the ELA test. DRC's partner for artificial intelligence (AI) scoring, Measurement Incorporated (MI), is a recognized leader in the field of automated essay scoring. MI's essay scoring engine (PEG) has been used to provide over ten million scores to students in formative and summative writing assessments over the past six years.

How accurate is AI scoring?

In recent independent studies, MI's PEG scores matched or exceeded human reliability. PEG's results are comparable to expert human scorers in terms of reliability and validity. The results have been validated in independent third party studies and in research conducted on behalf of clients. In other words, AI scoring of essays is very reliable when compared to traditional human scoring.

How are the AI models built?

Successful models rely heavily on accurately scored student responses from which the AI training sets are derived. To develop the AI training sets, student responses are scored twice, independently, by DRC's professional hand scoring staff. Once a representative sample is scored, responses and corresponding scores are delivered to the AI team at MI for model development.

Using previously scored student responses, MI's linguists, software developers, psychometricians, and human-computer interactions specialists create task specific algorithms that are then used to accurately predict how humans would score these student responses.

Is AI still able to recognize "Alert Papers"?

Built into MI's automated scoring engine are a variety of triggers for identifying alert papers (responses written by students that require some type of an intervention by the school). MI's AI scoring software 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. For the responses that it cannot score, PEG returns a condition code to DRC indicating why the response could not be scored. The flagged responses will then be routed to DRC's performance assessment hand scoring system. DRC will complete human hand scoring for the limited number of responses that cannot be scored by AI.

Quality Assurance

One of DRC's standard practices during hand scoring 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 will route the same validity papers through both processes. DRC professional scoring staff will score a set percentage of student responses (20% for the first year) after AI scoring has been completed. This double check of scoring will provide further information as to the accuracy of the AI scoring efforts.