Frequently Asked Questions
Smarter Balanced Interim Automated Scoring
Fall 2018
American Institutes for Research

Why is automated scoring being used?

Automated scoring provides many benefits to teachers, students, districts, and states. It reduces teacher grading time and speeds up the return of scores and feedback to students. At the state and district levels, it lowers the costs of scoring, ensures consistency in scoring within and across test administrations, decreases turnaround time to return scores to teachers, and potentially ensures that writing can continue to be evaluated in large-scale assessments. Automated scoring, backed by human review, improves the quality of overall scores by providing the consistency of the latest technology, supported by highly trained human judgment.

How does automated scoring work?

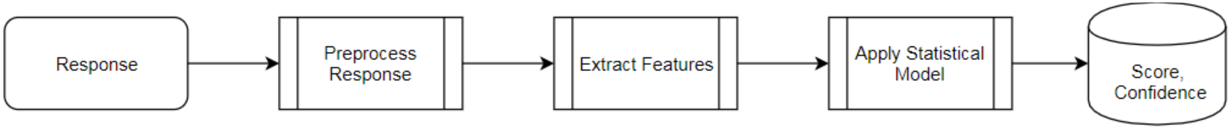
Automated scoring uses specialized software to model how human scorers assign scores to student responses. Essentially, automated scoring analyzes response characteristics and human-provided scores and predicts what a human scorer would do.

The engine is trained on specific questions. It is taught how to predict human responses to a specific prompt through exposure to scores provided by experienced and trained human scorers. When the initial training is complete, the engine is run through an extensive quality-control process by professional psychometricians. Criteria for approval include ensuring that the agreement of the engine with humans is similar to the agreement of two humans. In comparison and training, humans are considered the “gold standard.”

The engine performs differently depending upon the item. For full-writes, the scoring engine scores each response in stages: preprocessing, feature extraction, and score modeling. These are outlined in Figure 1.

- During preprocessing, the response text is prepared for the scoring engine. Blank responses are flagged, as well as responses that have too little original text to be scored by humans or the engine.
- During feature extraction, the processed response is analyzed using functions built to reflect common evaluations of writing quality. Features include grammar and spelling errors, elements of sentence variety and complexity, elements of voice and word choice, and discourse or organizational elements, in addition to words and phrases used.
- During score modeling, the values from the feature extraction phase are combined with prediction weights to produce a score and a confidence level.

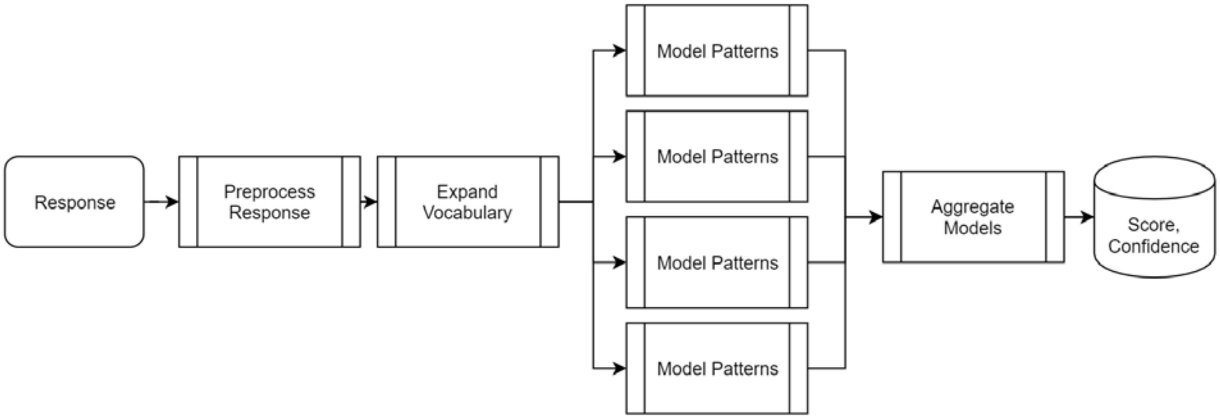
Figure 1. Automated Essay Scoring Process Flow



For other open-ended items, the process is slightly different. Here, unlike with full-write responses, the focus is on the meaning of the student response rather than spelling, grammar, sentence structure, and other measures of writing quality. This process occurs in four layers, as outlined in Figure 2.

- During preprocessing, the response text is prepared for the scoring engine. During this phase, blank responses are flagged, as are responses that have too little original text to be scored by humans or the engine.
- During vocabulary expansion, the engine converts the response into a semantic representation of each word in the response. This allows the engine to know that a misspelled word such as "oxigen" is very similar to "oxygen," or that "supermarket" is very similar to "grocery store."
- During pattern modeling, the engine uses multi-layer neural nets to examine patterns in word usage so that these patterns can be associated with the score. Thus, the engine should detect the critical difference between the statements, "The conclusion suggests that Suzy was not happy with her decision" and "The conclusion suggests that Suzy was happy with her decision."
- During model aggregation, the engine combines the results of multiple pattern models and provides a score that can be thought of as the aggregate of these models. This is akin to having multiple humans score a response and helps to ensure that the score is stable and reliable. Like the automated essay scoring engine, this engine outputs a score and a confidence level.

Figure 2. Automated Short Answer Scoring Process Flow



What do the condition codes mean?

If your student's response received a condition code, this means the engine determined that the response did not successfully pass one of the seven filters. Table 1 provides a brief description of each condition code.

Table 1. Condition Code Descriptions

Condition Code	Description
No Response	The response is empty or consists of only white space (space characters, tab characters, return characters).
Not Enough Data	The response has too few words to be considered a valid attempt.
Common Refusals	The response is a refusal to respond, in a form such as "idk" or "I don't know."
Non-Scorable Language	The response is written mostly in Spanish. In mathematics, teachers should be able to review these responses and score them as appropriate.
Duplicate Text	The response contains a significant amount of duplicate or repeated text.
Insufficient Original Text	The response consists primarily of text from the passage or prompt for essays or consists of only text from the passages for brief writes.

Condition Code	Description
Non-Specific	The response displays characteristics of condition codes assigned by humans that do not fall under the other artificial intelligence (AI) condition code categories. Unlike the other condition code functions that use algorithmic functions that are independent of the training sample, the nonspecific condition code is assigned using statistical features modeled on the features of the training sample.

The use of condition codes varies by the type of item and content area, as outlined in Table 2. These choices were made based upon the characteristics of the response as elicited by the item. For instance, "Not Enough Data" is not appropriate for items other than full writes, because it is possible to earn a non-zero score with a one-word response for many of the items.

Table 2. Condition Code Usage by Item

Condition Code	English Language Arts				Mathematics
	Performance Task Full-Writes	Performance Task Research	Brief-Writes	Short-Answer Items	
No Response	x	x	x	x	x
Not Enough Data	x				
Common Refusals		x	x	x	x
Non-Scorable Language		x	x	x	x
Duplicate Text	x		x		
Insufficient Original Text	x		x		
Nonspecific	x		x	x	x

Note: The AIRWays system reports condition code status for only full write items on ELA Performance Tasks.

What is a confidence level?

The confidence level reflects the confidence the engine has in the accuracy of the score that it has predicted. AIRWays flags responses that have a low-confidence value, specifically a value that is in the bottom 15% of all confidence values in the validation sample. This flag means that the response and score should be reviewed and/or altered to ensure that the score is accurate. When scored during operational summative testing, these responses are routed to human scorers.

How does the scoring process differ between the interim and summative items?

For summative scoring, low-confidence responses and responses receiving certain condition codes are routed for professional human scoring. Additionally, as a monitoring step, a portion of all other responses also receives a human score. In practice, the portion of responses routed for human scoring ranges between 20% and 40%, and this portion is determined in partnership with the individual state.

For interim scoring, low-confidence responses are flagged for educators to review and score. This approach allows the state to offer automated writing evaluations to teachers and students without incurring the added expense of professional human scoring.

How does the engine perform relative to human scorers?

Overall, the engine agreement with professional human scorers is similar to the agreement of human scorers with one another. As Table 3 demonstrates, when we restrict our evaluation to responses in which the engine is confident, the engine often outperforms the human scorers. When the engine is not confident, it often underperforms relative to the human scores. This indicates that these types of responses are difficult to score. You may see that as you evaluate such responses.

Table 3. Percentage of Exact Agreement for Responses in Which the Engine is Confident and Not Confident

	Percentage of Exact Agreement			
	Confident		Not Confident	
Item Set	Engine	Human	Engine	Human
All (73 items)	86	78	55	59
Math Short Answer (8 items)	86	83	54	64
ELA Short Answer (65 items)	86	77	56	58
ELA Full-Writes (7 items)	77	75	55	66
Purpose/Organization	80	77	57	66
Evidence/Elaboration	80	76	57	66
Conventions	71	71	50	66

I disagree with the condition code assigned to the response. What should I do?

Like the results of human scorers, automated scoring is not perfect. The engine models human judgment, which can have errors and be influenced by multiple factors. Humans tend to agree with one another 60–70% of the time on scores and 80–95% of the time on condition codes. As part of the engine training process, the human-to-engine match must be similar.

If you disagree with the condition code assigned to the response, please be sure to compare the condition code and description available in this FAQ against the response. If you still disagree, you can provide your own code or score instead using the AIRWays system. If there seems to be a serious problem, please follow the recommendations of your state assessment agency for reporting concerns.

I disagree with the score assigned to the response. What should I do?

The engine is modeled after human-assigned scores, and humans sometimes do not agree with one another on the same score. Therefore, we cannot always expect the engine to agree with your score.

We most often see disagreements in full writes because the evaluation of writing involves nuance and the relative prioritization of some aspects of writing over others. Furthermore, the ways in which students write can vary. Thus, two experienced and trained scorers may assign similar scores, but not the same scores, to a response. In many scoring situations, two experienced and trained human raters agree exactly on a score about 60–70% of the time and disagree the remaining 30–40% of the time. Two human scorers are almost always within one point of each other, and the same is true of the engine.

For other items, we see higher agreement rates. However, even these items involve some level of interpretation that can result in disagreements.

If you observe results with which you disagree, please first review the response relative to the rubric to see if the computer-assigned score is reasonable. Consider if another teacher might give a slightly higher or lower score using another way of viewing the essay. If you still disagree, you can provide your own code or score using the AIRWays system. If there seems to be a serious problem, please follow the recommendations of your state assessment agency in reporting concerns.

Why did this very brief full-write response receive a high score?

If an essay was not given a condition code, the response was routed to the essay scoring engine to produce a score. The essay scoring engine processes the response and extracts syntactic feature variables (such as number of grammatical errors) and semantic features (such as clusters of words used). These features are combined using a statistical process to produce a score.

Although there is generally a correlation between response lengths and scores, the engine usually does not look explicitly at length. A short response can be a good response, and often human scorers will assign a high score, as well. Similarly, long responses may receive a low score.

One of my students' essays received a higher score than another student's essay, but the first student's essay is better. Why?

The essay scoring engine predicts how a human would score the test based on many factors, including measures of ideas, grammar, spelling, word choice, organization, and voice. The engine's agreement with humans is reviewed during the quality-control process to ensure that it agrees with a trained scorer as often as another scorer would agree. When evaluating the response, consider if another teacher might give a slightly higher or lower score.