Frequently Asked Questions
Smarter Balanced Interim Automated Scoring
2022-2023 Academic Year

Why is automated scoring being used?

Automated scoring provides many benefits to educators, students, districts, and states. It reduces educator 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 educators, 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 raters would assign scores to student responses. Essentially, automated scoring analyzes response characteristics and human-provided scores and predicts what a human rater would do.

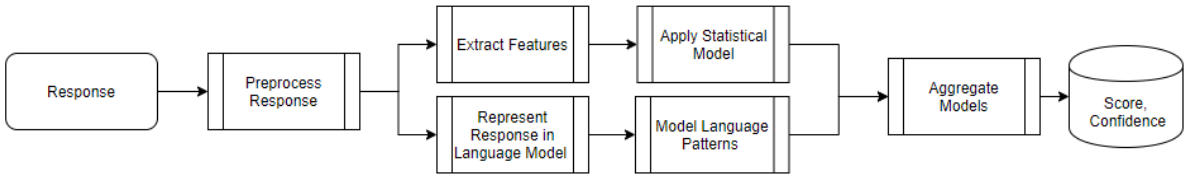
The engine is trained on specific questions. It is taught how to predict human responses to a specific prompt through exposure to scores provided by experienced and trained human scorers. When the initial training is complete, the engine is run through an extensive quality-control process by professional 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 scoring engine first preprocesses the responses and the submits the response to two separate data flows, as outlined in Figure 1. In the top flow of the diagram, features measuring semantics and writing quality are extracted and then these features are submitted to a statistical model. In the bottom flow of the diagram, the response is routed to a process that maps the response to a language model, and then this language model representation is used to model patterns that are associated with score. Finally, the outputs of the two processes are combined to predict the score and the confidence associated with the score. A few key elements of this process are described further below.

- 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.
- The top flow in the diagram uses a classical machine learning framework whereby features are derived by experts and then used to predict scores.
    - During feature extraction, the processed response is analyzed both for writing quality and meaning. Writing features include grammar and spelling errors, elements of sentence variety and complexity, elements of voice and word choice, and discourse or organizational elements. Semantic features model patterns of word usage by students in

their response. The writing quality features are used only for full writes and not for the short answer items. Only the semantic features are used both for full writes and short answer items.
    - During score modeling, the values from the feature extraction phase are combined with prediction weights to produce a score.
- The bottom flow in the diagram represents a more modern framework whereby features are learned during the training process based upon a model of language developed over a large set of written text. This approach is intended to measure the semantics or meaning of the response and how the meaning predicts scores that humans would assign.
    - The response is mapped to a set of locations in the language model that are intended to represent the meaning of each word in the response. Words that are used in similar ways appear in similar locations on the model. This allows the engine to know that "supermarket" is very similar to "grocery store." Additionally, the same word used in different contexts (e.g., "bank of the river" vs. "bank account") appears in different locations to account for the different meanings of the word in this context.
    - Then, the system models the language patterns to optimally mimic how raters would score. In this sense, the system is using multiple patterns of word usage to predict score.
- During model aggregation, the engine combines the results of two processes 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.

Figure 1. Automated Essay 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, educators 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
Unreadable Language	The response consists primarily of words that are unusual (e.g., gibberish, unusual words).
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 reporting system does not report on the condition code status for ELA Research performance tasks. This is due to a change in how these items are scored. We expect to be able to offer such reporting at a later date.

What is a confidence level?

The confidence level reflects the confidence the engine has in the accuracy of the score that it has predicted. The reporting system flags responses that have a low-confidence value, specifically a value that is in the bottom 15% of all confidence values in a validation sample. This flag means that the response and score should be reviewed to ensure that the score is accurate and should be changed if it is not. When scored during operational summative testing, these responses are routed to human scorers.

The intent of the confidence level is to give our clients the ability to identify responses that are unusual relative to the training sample and to route those responses for human review. The thresholds for the confidence value are set in consultation with the state and based upon a review of the data. We believe that the use of confidence levels with thresholds to route responses to hand-scoring allows your state to

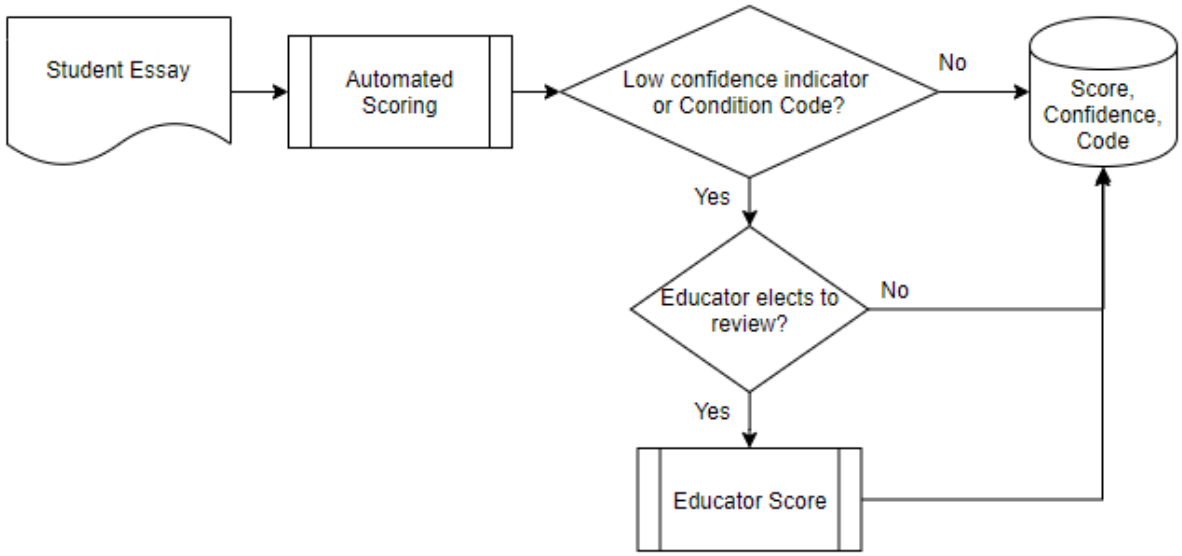
obtain the most accurate scoring performance across the set of responses *and* for each individual response.

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

For those states that use automated scoring 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 your state to offer automated scoring to educators and students without incurring the added expense of professional human scoring. Figure 2 illustrates this flow.

Figure 2. Scoring Process for Interim Items



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, automated scoring often outperforms the human scorers across the Smarter Interim items

Table 3. Percentage of Exact Agreement for Responses

Items	Engine	Human
Math Short Answer (8 items)	81%	79%
ELA Brief Write (50 items)	82%	73%
ELA Reading (24 items)	83%	76%
ELA Full-Writes (7 items)		
Purpose/Organization	76%	65%
Evidence/Elaboration	81%	71%
Conventions	82%	72%

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 60–75% of the time on scores and 80–95% of the time on condition codes. As part of the engine training process, we require that the human-to-engine match be similar across a set of responses and scores; however, we expect to see differences in scoring for any individual response. This is because humans may disagree with one another on scores for any individual response.

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 code/score assigned in the reporting 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 60–75% of the time and disagree the remaining 25–40% of the time. Two human scorers are almost always within one point of each other, and the same is true of the engine.

For short answer items, we often 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 assignment is reasonable. Consider whether another educator 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 instead using the one assigned in the reporting 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 response receive a high score?

If an essay was not given a condition code, the response was routed to the scoring engine to produce a score. Although there is generally a correlation between response lengths and scores, the engines do 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 for full writes. 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 educator might give a slightly higher or lower score.