Autoscore Benchmark Writing Module Scoring
Frequently Asked Questions
American Institutes for Research
2019-2020 Academic Year

Why is automated essay scoring being used?

Automated essay scoring provides many benefits to teachers, students, districts, and states. It saves on teacher grading time and hastens 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 students, and potentially ensures that writing continues to be evaluated in large-scale assessment. Automated scoring, backed by human review, improves the quality of overall scores, providing the consistency of the latest technology supported by highly trained human judgement.

How does automated essay scoring work?

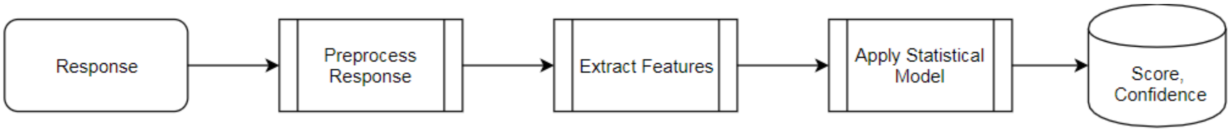
Automated essay scoring uses specialized software to model how human raters would assign scores to essays. Essentially, the automated scoring analyzes essay characteristics and human-provided scores and predicts what a human scorer would do.

The automated scoring engine is trained on specific questions. It is taught how to predict human responses on a specific prompt by exposing the engine to scores provided by experienced and trained human scorers. After initial training is completed, 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 that of two humans. In the comparison and in the training, humans are considered to be the “gold standard.”

The scoring engine scores each response in stages: preprocessing, feature extraction, and score modelling. These are outlined at a high level in Figure 1.

- 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 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 the words and phrases used.
- During score modelling, the values from the feature extraction phase are combined with scoring model weights to produce a score and a confidence level.

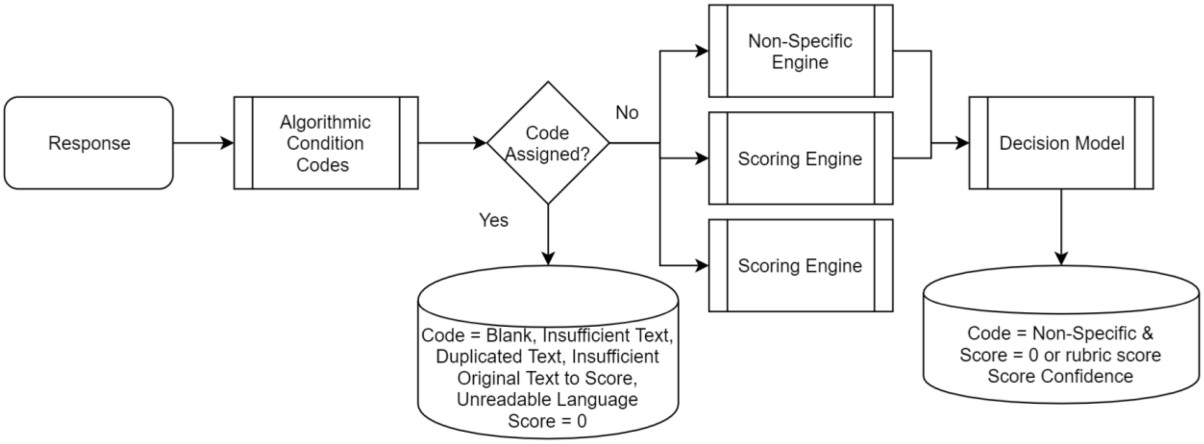
Figure 1. Automated Essay Scoring Process Flow



What is the overall scoring process?

When a test is submitted, responses are routed to the scoring engine. Once in the scoring engine, the response follows a multi-stage process. The steps of this process are conducted separately for each rubric dimension (e.g., Purpose, Focus, and Organization) and are illustrated in Figure 2 below.

Figure 2. Autoscore Process


Response
Algorithmic Condition Codes
Code Assigned?
No
Yes
Non-Specific Engine
Scoring Engine
Scoring Engine
Decision Model
Code = Blank, Insufficient Text, Duplicated Text, Insufficient Original Text to Score, Unreadable Language Score = 0
Code = Non-Specific & Score = 0 or rubric score Score Confidence

The first stage of the process evaluates the response to determine whether it meets the criteria for a "Blank," "Insufficient Text," "Duplicated Text," "Insufficient Original Text to Score," or "Unreadable Language" condition codes. If it meets any of these criteria, then the appropriate code is stored in a database and a score of zero is assigned.

If the response is not assigned a condition code via the first process, then it is routed to the following stages: the engine for assigning "Non-Specific" condition codes, the essay scoring engine, and an outlier engine. The results of each of these stages are then submitted to a decision model, which uses a statistical process to determine whether the response should receive a "Non-Specific" condition code and score of 0 or a valid score based on the item's rubric and confidence level, which is the measure of how sure the machine is that the score is assigned is correct. The confidence level is based on two factors: how close a score is predicted to be to the line between two adjacent scores; and, whether the essay seems dissimilar to the essays seen in the training set.

How are condition codes assigned?

If your student's response received a condition code, this means that the engine determined that the response did not successfully pass one of six filters that examine the response for length, extent of copying of the passage, duplicate text, or relationship to the prompt. The table below provides a brief description of each condition code. In the AIRWays platform, we alert you to responses that received an 'Unreadable Language' or 'Non-Specific' code so that you can review those scores if you choose.

Table 1. Condition Codes, Descriptions

Condition Code	Description
Blank	The response was empty or consisted only of white space (space characters, tab characters, return characters)
Insufficient Text	The response has too few words to be considered a valid attempt.
Duplicated Text	The response contains a significant amount of duplicate or repeated text.
Insufficient Original Text to Score	The response consists primarily of text from the passage or prompt.
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 condition code categories.

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 to ensure that the score is accurate. 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 USBE and based upon a review of the data. We believe that the use of confidence levels with thresholds to route summative responses to hand-scoring allows Utah to obtain the most accurate scoring performance across the set of responses *and* for each individual response.

How does the scoring process differ between scoring the benchmark and summative writing prompts?

In summative scoring, the engine is tuned specifically to model the human scores on the writing prompts appearing in that assessment. At the start of the testing window, the first set of responses (often 500) are scored by Autoscore and then routed for professional hand-scoring. Once these responses are scored by the human scorers, the performance of the engine is evaluated relative to the human scores. If the agreement of the engine with humans is appropriate, then we use Autoscore as the primary scorer. Responses that the engine deems it cannot score with confidence are routed for expert human scoring. The responses receiving certain condition codes, as decided by USBE but always including the "Non-Specific" and the "Unreadable Language" codes, are routed for human scoring as well. In practice, the proportion of responses which received hand-scores scoring ranges between 20 and 40%, and this proportion is determined in partnership with USBE.

In benchmark scoring, the engine is also tuned to model human scores on writing prompts appearing in the assessment. However, scores are not automatically routed for evaluation by trained human scorers. This approach allows USBE to offer automated writing evaluation to teachers and students without the expense of professional human scoring and to offer scoring from an engine modeled upon professionally hand-scored data. Educators are alerted to low-confidence responses (and ones receiving the two condition codes) so they can review the writing completed by their students.

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. In Table 2, we present the exact agreement rates of two human raters against those of Autoscore with a highly-vetted human-based score for the benchmark essay prompts (four items per grade). As Table 2 demonstrates, we often see higher agreement between Autoscore and a highly vetted human rater score compared to that of two human raters for each of the rubric dimensions and for each grade.

Table 2. Percentage of Exact Agreement for Responses Across Grades and Dimensions

	Editing & Conventions		**Evidence & Elaboration**		**Purpose, Focus, & Organization**	
Grade	**Human**	**Autoscore**	**Human**	**Autoscore**	**Human**	**Autoscore**
3	67%	76%	59%	64%	64%	68%
4	64%	72%	58%	68%	56%	66%
5	75%	81%	57%	68%	59%	73%
6	70%	75%	60%	66%	61%	64%
7	72%	80%	68%	71%	58%	68%
8	81%	84%	66%	71%	63%	71%
9	70%	78%	61%	74%	62%	74%
10	72%	80%	63%	73%	62%	72%
11	78%	84%	67%	74%	66%	73%
All Grades	**72%**	**79%**	**62%**	**70%**	**61%**	**70%**

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 for the benchmarks, please be sure to compare the condition code and description available in this FAQ against the response. If there seems to be a serious problem, please follow the recommendations of USBE for reporting concerns.

Why did this very brief response receive a high score?

If the response was not given a condition code, then the response was routed to the essay scoring engine to produce a score. The essay scoring engine processes the response, extracts feature variables (such as number of grammar errors) and combines the feature variables using a statistical process to produce a score.

The feature extraction process includes measures of ideas, grammar, spelling, word choice, organization, and voice. While there is generally a correlation between response length and scores, the engine usually does not explicitly look 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. Please note that the scoring rubric does not explicitly define length in the scoring process.

One student’s essay 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 it agrees with a trained scorer as often as another scorer would agree. When evaluating the response in the benchmarks, consider whether another teacher might give a slightly higher or lower score. Also, please make sure that each response did not receive a low-confidence code.

Will the use of automated scoring disadvantage my students?

In general, the use of automated scoring has not been shown to favor any group of students. Many studies have been published examining score agreement at the dimension level, the prompt level (i.e., across dimensions) and at the test level. There have been a few studies on the performance of automated essay scoring engines for particular student groups such as English learners (EL), students with disabilities (SWD), or differences between genders. The results of these studies indicate that automated essay scoring engines have shown similar agreement with human scores for most student subgroups.