WYOMING
DEPARTMENT OF EDUCATION
edu.wyoming.gov

WY-TOPP

FAQ AutoScore

American Institutes for Research

Why is automated essay scoring being used?

Automated essay scoring provides many benefits to teachers, students, districts, and states. It saves teacher's 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 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 assign scores to essays. Essentially, the automated scoring analyzes essay characteristics and human-provided scores, and predicts what a human scorer would do.

The AI 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 (QC) process by professional psychometricians. One criterion for approval is 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 prediction weights to produce a score and a confidence level.

FAQ AutoScore (Continued)

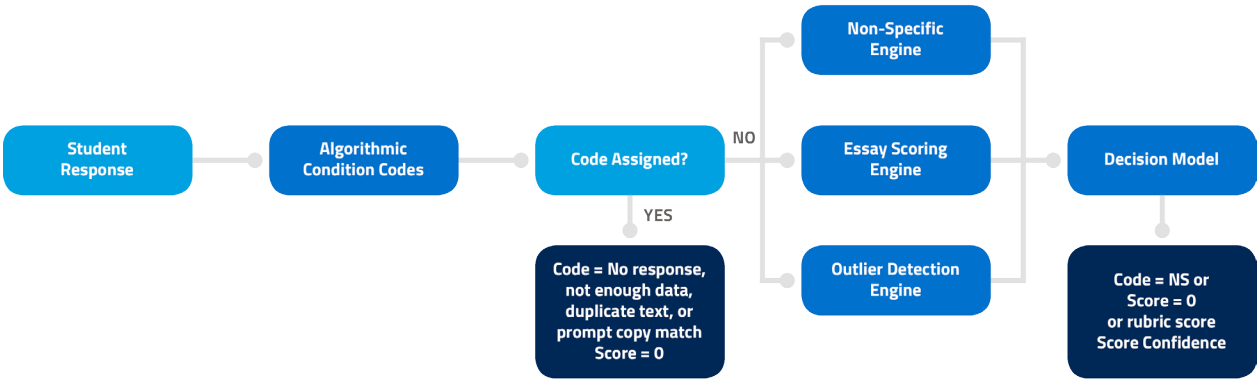
WY-TOPP

Figure 1. Automated Essay Scoring Process Flow

What is the overall scoring process?

When a test is submitted, responses requiring hand scoring are routed to the scoring engine. Once in the scoring engine, it follows a multi-stage process. The steps of this process are conducted separately for each rubric dimension and are illustrated in Figure 2 below.

Figure 2. AutoScore Process



The first stage of the process evaluates the response to determine whether it meets the criteria for a 'No Response,' 'Not Enough Data', 'Duplicate Text', or 'Prompt Copy Match' condition code. 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 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 zero (0) or a valid score based on the item's rubric and confidence level, 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.

FAQ AutoScore (Continued)

WY-TOPP

How are condition codes assigned?

A condition code is a description given to a student response which did not successfully pass one of several filters. The filters examine each student response for length, extent of copy and language from the passage, duplicate text in the response, or how the student response relates to the prompt. The table below provides a description of each condition code.

Table 1. Condition Codes, Descriptions

Condition Code	Description
No Response	The response was empty or consisted only of white space (space characters, tab characters, return characters).
Not Enough Data	The response has too few words to be considered a valid attempt at the prompt.
Duplicate Text	The response contains a significant amount of duplicate or repeated text.
Prompt Copy Match	The response consists primarily of text from the passage or prompt.
Non Specific	The response displays characteristics of condition codes assigned by humans that are do not fall under the other condition code categories.

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

The scoring process for the Interim and Summative Assessments is not perfect; the engine models human judgement, which can have errors and be influenced by multiple factors. Humans tend to agree with each other 60-70 percent of the time on scores and 80-95 percent of the time on condition codes. As part of the engine training process, the human to engine match must be similar.

Here are some steps you can take when you observe a condition code with which you disagree on the Interim Assessment:

1. Review the condition code and description available in this FAQ.
2. Review the response relative to the code description to see whether the condition code assignment is reasonable.

FAQ AutoScore (Continued)

3. If you feel the condition code is not reasonable, then please let us know by contacting the Help Desk. We will review your comment, the response, and the score.

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

The same engine is used in interim and summative scoring of the writing prompts. However, there are other differences in the scoring process.

In summative scoring, the engine is tuned specifically to model the human scores on the writing prompts appearing in that assessment. The responses that the engine deems it cannot score with confidence are routed for expert human scoring. Responses receiving the "Non Specific" code, are routed for human scoring as well. Finally, approximately 15% of responses in Wyoming will also receive a human score, often referred to as a read-behind. In practice, the proportion of responses routed for human scoring ranges between 20 and 40%.

In interim 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 raters. This approach allows the state to offer automated writing evaluation to teachers and students without the expense of professional human scoring.

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

The essay scoring is modeled after human-assigned scores, and humans often do not agree with one another on the same score. This is because the evaluation of writing involves nuance, the relative prioritization of some aspects of writing over others, and the ways in which student's writing can be quite variable. Thus, two experienced and trained scorers may assign similar but not the same scores to a response. In many scoring situations, two experienced and trained human raters will exactly agree 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 engine is as well.

Here are some steps you can take when you observe results with which you disagree on the Interim Assessment:

1. Review the response relative to the rubric to see whether the assignment is reasonable. Consider whether another teacher might give a slightly higher and lower score using another way of viewing the essay.
2. If you feel the score is not reasonable, then please let us know by contacting the Help Desk. We will review your comment, the response, and the score.

FAQ AutoScore (Continued)

WY-TOPP

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.

One of 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 QC process to ensure it agrees with a trained scorer as often as another scorer would agree. When evaluating the response consider whether another teacher might give a slightly higher or lower score.

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 few studies on the performance of automated essay scoring engines for particular student groups such as English language learners (ELL), students with disabilities (SWD), or differences between genders. The results of these studies indicate that automated essay scoring engines have not favored any particular student group.