Why is automated essay scoring being used?

Automated essay scoring provides many benefits to teachers, students, districts, and states. It hastens the return of scores and feedback to teachers, parents, and students. It ensures consistency in scoring within and across test administrations, decreases turnaround time to return scores to teachers, 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 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 prediction weights to produce a score and a confidence level.

Figure 1. Automated Essay Scoring Process Flow

How does Autoscore work?
When a test is submitted, writing responses 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 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 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 a response receives a condition code, this means that the engine determined that it did not successfully pass one of five filters that examine the response for length, extent of copying of the passage, duplicate text, relationship to the prompt, or nonspecific text. The table below provides a brief description of each condition code.

<table>
<thead>
<tr>
<th>Condition Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Response</td>
<td>The response was empty or consisted only of white space (space characters, tab characters, return characters)</td>
</tr>
<tr>
<td>Not Enough Data</td>
<td>The response has too few words to be considered a valid attempt at the prompt.</td>
</tr>
<tr>
<td>Duplicate Text</td>
<td>The response contains a significant amount of duplicate or repeated text.</td>
</tr>
<tr>
<td>Prompt Copy Match</td>
<td>The response consists primarily of text from the passage or prompt.</td>
</tr>
<tr>
<td>Non-Specific</td>
<td>The response displays characteristics of condition codes assigned by humans that do not fall under the other condition code categories. Unlike the other condition code functions which use algorithmic functions that are independent of the training sample, the non-specific condition code is assigned using statistical features modeled on the features of the training sample.</td>
</tr>
</tbody>
</table>

When a response receives a condition code, one of two things happen. If the response receives a “No Response,” “Not Enough Data,” “Duplicate Text,” or “Prompt Copy Match” condition code from the engine, then the response receives score of 0 in each dimension for a total score of 0. When a response receives a “Non-Specific” condition code, then the response is routed for expert human review. This approach is taken because the “Non-Specific” code indicates that the response is not similar to typical responses (e.g. many mis-spellings, mostly written in a language other than English, off topic or purpose) but still may merit a score from the scoring rubric.
What is the confidence value?
The confidence value is the level of confidence that AutoScore has when predicting a score or a condition code. The value ranges between 0 and 1, with a low value indicating lower confidence and a high value indicating higher confidence. The value is generated during engine calibration using a combination of modelling condition codes and scores and examining multivariate outliers relative to the calibration sample.

The intent of the confidence value is to give our clients the ability to identify responses that are unusual relative to the training sample and to route those responses for hand-scoring. 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 is Autoscore used when scoring student responses?
In your state’s summative testing program, Autoscore is used in conjunction with trained human raters to assign scores to student responses. Responses receive a score and/or condition code by Autoscore or by human raters.

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 raters, the performance of the engine is evaluated relative to the human rater scores. If the agreement of the engine with humans is appropriate, then the responses are routed through a process described in Figure 3.

When a test is submitted, writing responses are routed to the scoring engine. If Autoscore assigns a targeted condition code (e.g., Non-Specific, and as defined by your state) then that response is routed for hand-scoring, which provides a score and/or condition code.

If AutoScore assigns a condition code that is not targeted for human review (e.g., Not Enough Data), then the response earns a score of 0 in each dimension and an overall score of 0. If Autoscore assigns a valid score, then the confidence level associated with that score is evaluated relative to a threshold. If the confidence value is below the threshold, then the response is routed for hand-scoring. If the confidence value is above the threshold, then the score and/or condition code are returned.
Figure 3. Scoring Process for Responses

Is Automated Scoring valid and reliable?

The validity and reliability of automated scoring engines is typically examined by comparing the automated scoring results to human scoring results. The assumption underlying these comparisons is that the engine should predict scores that are similar to those that would have been assigned by humans.

The statistics used to make comparisons focus on score distributions and rater agreements. The distribution of scores assigned by the engine should be practically and statistically similar to those assigned by humans. In addition, the agreement of the engine to a human rater should be practically and statistically similar to the agreement of two human raters with one another. One measure of agreement is the percent of response in which two human raters, who are blind to one another’s rating, assign the same score to a response. This measure is called “Exact Agreement.” We can compare the agreement of two humans with the agreement of the Autoscore with a human rater.

Figure 4 presents exact agreement rates for humans and for Autoscore for essays in a multi-state sample of five grades and a rubric with three dimensions. The exact agreement rates for human raters range between 62 and 83%, and the exact agreement rates for Autoscore range between 65 and 82%. This means that professionally-trained human raters often disagree on 17-38% of scores. Visual inspection of the figure shows the similarity of the rates between humans and Autoscore for the five grades and three dimensions.
I disagree with the score assigned to the response. 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 fact is made clear by Figure 4, above. This is because the evaluation of writing involves nuance and the relative prioritization of some aspects of writing over others as well as the ways in which students write 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 60-80% of the time, and disagree the remaining 20-40% 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:

1. Review the Writing Performance section of the student’s Individual Student Report for details on each of the writing dimensions that comprise the student’s writing performance.
2. Review the Writing and Language reporting category under the Performance on the NDSA ELA/Literacy test section of the student’s Individual Student Report for information on the student’s performance on this specific reporting category, what the results mean, and next steps.
Does the length of the response impact the score?

If the response was not given a condition code related to length (see Table 1), 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 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 QC process to ensure it agrees with a trained scorer as often as another scorer would agree.

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.