Scoring of B.E.S.T. Writing in 2023–2024 and Beyond

Florida Organization of Instructional Leaders

May 17, 2023
Definitions:

• **Human scoring**: Traditionally used in Florida, highly-trained and qualified human scorers independently review and score student responses, with extensive quality controls in place before, during, and after scoring.

• **Automated scoring (AS)**: Previously used on a trial basis in Florida, AS is the use of human-scored responses to train an engine that models scoring of student responses.

• **Hybrid scoring**: Hybrid scoring uses automated scoring as the primary scorer, while routing a subset of responses for human scoring.
The Human Component

• Florida educators
  • Rubric Development
  • Passage and Prompt Review
  • Field Test Rangefinder
  • Operational Rangefinder

• Human scorers
  • All human scorer training and qualifying materials are approved by Florida educators.
  • All field test responses are minimally double-human scored.
  • Exact agreement
  • 5,000 student responses per field test prompt

• Resulting materials are used in AS training.
AS Responses Routed to Humans

- Responses the AS engine has not been trained to score
- Creative/unusual responses
- Condition codes
- Responses with low confidence scores
- A percentage of all responses
# Planned Approach to Scoring B.E.S.T. Writing

<table>
<thead>
<tr>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autoscore</td>
</tr>
<tr>
<td>Training Methods</td>
</tr>
<tr>
<td>Hybrid Scoring</td>
</tr>
<tr>
<td>Retraining</td>
</tr>
</tbody>
</table>
Autoscore
Automated Scoring (AS)

• AS produces scores more quickly, ensures consistent score application within and across test administrations, reduces cost to taxpayers, and produces high-quality scores.

• AS engines will be used in Florida in conjunction with additional human scoring for certain types of student responses.

• AS is used across the country in several statewide, summative assessments, as well as in several interim assessment programs.

• Outside of Florida, Cambium’s AS system currently scores more than 3 million responses in a typical school year.
High-Level AS Flow
Two Models Used to Score Each response

Classical Model

• Writing quality features include syntax; grammar; spelling; sentence and paragraph quality.
• Semantic features via *Latent Semantic Analysis*, which analyzes the distribution and relationships among terms and concepts found in the stimulus, prompt, and response.

Language Model

• Representation of language, based on modeling on prompt and a large number of responses, which is then fine-tuned based on each prompt.
• More sensitive to words not appearing in the responses used to train the AS engine due to use of natural language processing of root words and *word-pieces*.
• Considers word order in modeling.
Ensembling

• The purpose of the ensemble is to use outputs from both models to produce an accurate score.
  • Language model typically outperforms the classical model.
  • Ensemble performs slightly better than each individual model.
## Condition Codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Response</td>
<td>Response was empty or consisted only of white space (space characters, tab characters, return characters).</td>
</tr>
<tr>
<td>Not Enough Data</td>
<td>Response has too few words to be considered a valid attempt.</td>
</tr>
<tr>
<td>Duplicate Text</td>
<td>Response contains a significant amount of duplicate or repeated text.</td>
</tr>
<tr>
<td>Prompt Copy Match</td>
<td>Response consists primarily of text from the passage.</td>
</tr>
<tr>
<td>Common Refusals</td>
<td>Response is a refusal to respond, in a form such as &quot;idk&quot; or &quot;I don't know.&quot;</td>
</tr>
<tr>
<td>Non-Scorable Language</td>
<td>Response is written mostly in another language</td>
</tr>
<tr>
<td>Unusual vocabulary</td>
<td>Most words in the response do not appear in typical responses.</td>
</tr>
<tr>
<td>Non Specific</td>
<td>Response displays characteristics of condition codes assigned by humans that do not fall under the above condition code categories.</td>
</tr>
</tbody>
</table>
Confidence

• Autoscore produces a confidence index for each response.
• This index reflects the degree to which Autoscore ‘thinks’ it is producing an accurate score, or the score an expert scorer would have assigned.
• Based upon a statistical approach
• Lower-confidence responses will be routed for human verification.
Automated Troubled Child Alert Identification

• Automated detection system for ‘crisis’ papers or alerts uses the Hotline system, which is separate from AS.
• Scans student-written text, including notes, for phrasing indicating harm to self or others.
• Combined with human review, ensures systematic and timely review of every piece of text written by students.
• Typically provide alerts within 24 hours of identification.
Student Alert Response Flow

1. Response
2. Hotline Alerting System
   - Human Review
     - Alert? (Yes -> Client via Program Management, No -> Alert Status: Alert)
Autoscore Training Methods
Overall Process

1. Data
2. Hand-score
3. Train engine/Validate scores
4. Deploy and QC models
5. Hybrid scoring
6. Monitor results
Data

• Models built for each prompt
• Identify pool of available responses
  • Administration conditions match anticipated conditions
  • 2,500-4,000 recommended
  • Stratify to ensure sufficient score point representation, if possible
  • Typically part of embedded or stand-alone field test, but could be drawn from operational samples
    • Will use Spring 2023 Writing Field Test responses
    • May draw from future operational samples as needed
Handscoring

• Obtain the highest-quality score on which to train the engine

• Training Materials
  • Scores and condition codes
  • Rater training, qualification, and monitoring materials

• Scoring Responses
  • Rater training, qualification, and monitoring
  • Two independent reads
  • Non-exact adjudication
Training

• Divide the sample into three sets: model training, ensembling, and validation.

• Train classical and language models separately using the model training sample.

• Use the score outputs from each of the classical and language models and train the engine using the ensembled data.

• Once the ensemble is built, use the ensemble to predict scores on the validation sample.
Criteria for Evaluation

- Consider human scoring to be the ‘gold standard’
  - Engine-final resolved scores compared to the two human scores
- Florida will use multiple measures to monitor AS and to adjust as needed:
  - Does the engine give exactly the same score in the same proportions as humans do?
  - Does the engine agree with a human beyond what would be expected simply by chance the same way the two humans do?
  - Does the engine produce similar average scores compared to the humans?
Quality Control: Engine Changes and Model Deployment

• Standardized scripts for engine training and validation
• Test cases and models used to examine the impact of any software change
• Scripts to re-score validation data on deployed models
  • Must return same scores, condition codes, and confidence values
• Checks to assess adherence to scoring specifications
Monitoring Performance

• First N sample
  • Helps to identify any early issues
  • Examines performance early in the window but not throughout
  • Not representative
  • We should expect that the engine agreements with the human raters to be similar to those observed in the held-out validation sample.

• Lower-confidence sample
  • We should see generally lower agreements with the human raters.
AS Validation Best Practices

- Engine design (deep-learning based)
- Engine performance evaluation (including bias)
- Unusual paper identification
- Lower-confidence identification
- Operational monitoring
  - Adherence to routing conditions
  - Agreements and mean differences
- Technical reporting and transparency
- Educator comparability workshops
- Help desk ticket reviews/responses

Continuous efforts to improve based upon findings
Hybrid Scoring
More about Hybrid Scoring

• In 2023–2024 and beyond, Florida will use a hybrid of AS and human scoring.

• What kinds of responses will be routed for human scoring?
  • Unusual responses
    • Certain condition codes
    • Lower-confidence responses
  • Monitoring responses
    • Can be set number of first responses or random sample

• Routing decisions are configurable, specifications-based, and will be annually approved by Florida Department of Education staff.
Routing Florida

Responses routed for human scoring are scored by trained professional scorers.
Retraining
Rationale

• Causes of low AS-HS agreement can include:
  • Changes in how students respond to test items
  • Insufficient data in engine training
  • Changes in handscoring

• It can be difficult to unpack the source of the issue.
  • Changes in handscoring should be examined with validity papers, potentially including the data used to train the engine.

• Possible recalibration with training data that includes both the original data and new operational data; this is most appropriate when we suspect training data are insufficient or responses have changed.
  • Need representative operational data
  • Need to ensure adherence to original rubric interpretation
Methods

• Train with original training sample and subset of operational data.

• Two validation sets:
  • Original held-out validation
  • Operational held-out validation

• Ensures adequate performance on both validation samples:
  • First, ensures adherence to the original interpretation of the rubric.
  • Second, examines performance in live scoring.

• If both datasets meet criteria, then use retrained model.