Faster and Better: The Continuous Flow Approach to Scoring

Presenters:
- Joyce Zurkowski
- Karen Lochbaum
- Sarah Quesen
- Jeffrey Hauger

Moderator:
- Trent Workman

CCSSO NCSA 2018

Summer and Fall of 2010: Assessment Subcommittee and Stakeholder Meetings

Resulting Expectations for the Next Assessment System

- Online assessments
- More writing
  - Continued commitment to open-ended responses (legislation and consequences)
  - Alignment to standards
    - Different types of writing
    - Text-based
- Move test to closer to the end of the year
- Get results back sooner than before
Leverage Technology

Content - new item types

Administration - reduce post-test processing time

Scoring - increase efficiency and reduce some of the challenges with human scoring
  • Practical: time, cost, availability of qualified scorers
  • Technical: drift within year, inconsistency across years which limited use as anchors and pre-equating, influence by construct-irrelevant variables, etc.

Reporting - online reporting to reduce post-scoring processing time
Prior to Initiating RFP:
Information Gathering

Investigated a variety of different scoring engines

- Types
  - Surface features (algorithms)
  - Syntactic (grammar)
  - Semantic (content-relevant)

- How is human scoring involved?

- How does the engine deal with atypical papers?
  - Off topic
  - Languages other than English
  - Alert
  - Plagiarized
  - Unexpected, just plain different
  - Test-taking tricks
RFP Requirements

A minimum of five (5) years of experience with practical application of artificial intelligence/automated scoring

Item writers trained to understand the implications of the intended use of automated scoring in item writing

Commitment to providing assistance in explaining to a variety of (distrusting/uncomfortable) audiences
- Believers in the art of writing
- Technology anxious
To expedite the return of results to districts, CDE would like to explore options for automated scoring using artificial intelligence (AI) for short constructed response, extended constructed response and performance event items.

- Current capacity for specified item types and content areas (quality of evidence)
- Description of how the engine functions, including training in relationship to content
- Projected (realistic) plans for improving its AI scoring capacity
- Procedures for ensuring reliable and valid scoring
  - Training and ongoing monitoring
  - Validity papers? Second reads?
  - Reliable and valid scoring for subgroups
Scoring System Expectations

Need a system that:

• Recognizes the importance of CONTENT; style, organization and development; mechanics; grammar; and vocabulary/word use
• Has a role for humans in the process
• Is reliable across the score point continuum
• Is reliable across years
• Is proven reliable for subgroups
Initial Investigation with CO Content

Distribution between human and AI scored items determined based on the number of items the AI system has demonstrated ability to score reliably.

• Discussions on minimum acceptable values versus targets
• Adjustments in item specific analysis
• Score point specific analysis
  • Uneven distribution across score points became an issue
  • Conversations about how many items are needed by score point
  • Identification of specific score ranges for specific items

The use of AI had to provide for equity across student populations supported by research.
  • We had low n-counts.
So where did we go from there?

Found some like-minded states!

PARCC
Automated Scoring

• Each prompt/trait is trained individually
• Learn to score like human scorers by measuring different aspects of writing
• Measure the content and quality of responses by determining
  • The features that human scorers evaluate when scoring a response
  • How those features are weighed and combined to produce scores
What is Continuous Flow?

- A hybrid of human and automated scoring using the Intelligent Essay Assessor (IEA)
- Optimizes the quality, efficiency, and value of scoring by using automated scoring along side human scoring
- Flows responses to either scoring approach as needed in real time
Why Continuous Flow?

• **Faster**
  • Speeds up scoring and reporting

• **Better**
  • Continuous Flow improves automated scoring which improves human scoring which improves automated scoring which improves ...
Continuous Flow Overview
Responses flow to IEA as students finish

IEA requests human scores on responses
- Likely to produce a good scoring model
- Selected for subgroup representation
As human scores come in, IEA
- Tries to build a scoring model
- Requests human scores on additional responses
- Suggests areas for human scoring improvement
Once the scoring model passes the acceptance criteria, it is deployed.
IEA takes over 1st scoring
Low confidence scores are sent to humans for review

Human scorers second score to monitor quality

*Set by customer

Pearson
How Well Does It Work?
Performance on the PARCC assessment

Starting in 2016, we used Continuous Flow to train and score prompts for the PARCC operational assessment.
PARCC Performance Statistics

% AI First Score | % Exact Agreement
---|---
82 | 72
88 | 72
93 | 73

65% IRR target
# Reading Comprehension/Written Expression Performance 2018

<table>
<thead>
<tr>
<th>Grade</th>
<th>Exact</th>
<th>SP0</th>
<th>SP1</th>
<th>SP2</th>
<th>SP3</th>
<th>SP4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Blue** means IEA exceeded human performance.
- **Green** within 5 of human.
- **Orange** lower by more than 5.
Conventions Performance 2018

<table>
<thead>
<tr>
<th>Grade</th>
<th>Exact</th>
<th>SP0</th>
<th>SP1</th>
<th>SP2</th>
<th>SP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
</tr>
<tr>
<td>3</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
</tr>
<tr>
<td>4</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Orange</td>
<td>Blue</td>
</tr>
<tr>
<td>5</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Orange</td>
<td>Blue</td>
</tr>
<tr>
<td>6</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Orange</td>
<td>Blue</td>
</tr>
<tr>
<td>6</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Orange</td>
<td>Blue</td>
</tr>
<tr>
<td>7</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Orange</td>
<td>Blue</td>
</tr>
<tr>
<td>7</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Orange</td>
<td>Blue</td>
</tr>
<tr>
<td>8</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Orange</td>
<td>Blue</td>
</tr>
<tr>
<td>8</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Orange</td>
<td>Blue</td>
</tr>
<tr>
<td>9</td>
<td>Blue</td>
<td>Blue</td>
<td>Orange</td>
<td>Blue</td>
<td>Blue</td>
</tr>
<tr>
<td>9</td>
<td>Blue</td>
<td>Blue</td>
<td>Orange</td>
<td>Blue</td>
<td>Blue</td>
</tr>
<tr>
<td>10</td>
<td>Blue</td>
<td>Orange</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
</tr>
<tr>
<td>10</td>
<td>Blue</td>
<td>Orange</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
</tr>
<tr>
<td>11</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
<td>Blue</td>
</tr>
</tbody>
</table>

*Blue* means IEA exceeded human performance

*Green* within 5 of human

*Orange* lower by more than 5
Summary

• Continuous Flow combines human and automated scoring in a symbiotic system resulting in performance superior to either alone
  • It’s efficient
    – Ask humans to score a good sample of responses up front rather than wading through lots of 0’s and 1’s first
  • It’s real time
    – Trains on operational responses
    – Informs human scoring improvements as they’re scoring
  • It yields better performance
    – Performance on the PARCC assessment exceeded IRR requirements for 3 years running
  • And it doesn’t disadvantage subgroups!
Overview: IEA fairness and validity for subgroups

• Predictive validity methods
  • Prediction of second score
  • Prediction of external score

• Summary

“Fairness is a fundamental validity issue and requires attention throughout all stages of test development and use.”
- 2014 Standards for Educational and Psychological Testing, p. 49
Williamson et. al (2012) offers suggestions for assessing fairness: “whether it is fair to subgroups of interest to substitute a human grader with an automated score” (p. 10).

Examination of differences in the predictive ability of automated scoring by subgroup:

1. **Prediction of Second Score**: Compare an initial human score and the automated score in their ability to predict the score for the second human rater by subgroup.

2. **Prediction of External Score**: Compare the automated and human score ability to predict an external variable of interest by subgroup.
## Summary of sample sizes (averaged across items)

<table>
<thead>
<tr>
<th>Group</th>
<th>Human-Human</th>
<th>IEA - Human</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Female</td>
<td>557</td>
<td>119</td>
</tr>
<tr>
<td>English Language Learner</td>
<td>135</td>
<td>90</td>
</tr>
<tr>
<td>Student with Disabilities</td>
<td>203</td>
<td>83</td>
</tr>
<tr>
<td>Asian</td>
<td>120</td>
<td>17</td>
</tr>
<tr>
<td>Black/AA</td>
<td>230</td>
<td>54</td>
</tr>
<tr>
<td>Hispanic</td>
<td>344</td>
<td>114</td>
</tr>
<tr>
<td>White</td>
<td>349</td>
<td>105</td>
</tr>
</tbody>
</table>
Multinomial logit model

Scores treated as nominal (0-3 or 0-4). A logistic regression with generalized logit link function was fit in order to explore predicted probabilities of the second score \( y \) across levels of the first score \( x \).

\[
g(\hat{y}) = \text{logit} = \log \left( \frac{\hat{π}}{1-\hat{π}} \right) = X\hat{β},
\]

where \( \hat{π} = \frac{e^{X\hat{β}}}{1-e^{X\hat{β}}} \)
Models showed quasi-separation (meaning that the DV separated the IV almost perfectly across some levels). For example, for an expressions trait model, we likely will find:

\[ \text{probability } (Y=0 \mid X>3) = 0 \text{ and probability } (Y=4 \mid X<1) = 0 \]

Given the goal of this analysis, quasi-separation was tolerated in order to get predicted probabilities that were not cumulative and not strictly adjacent.

Some subgroups have insufficient data to estimate predicted probabilities at all score points.
Sampling for IEA Subgroup Analysis

Interpretation of polybar charts

E_H2 is the 2nd human score for expressions trait (DV)

E_H1 is the 1st human score for expressions trait (IV)

Colors of bars represent the score point for second score (blue=0, red=1, green=2, beige=3, purple=4)

Heights of bars represent the predicted probability of the second score given the first score

Predicted prob. of $E_{H2} = 0$ approx. 0.85

Predicted prob. of $E_{H2} = 1$ approx. 0.15

If $E_{H1} = 0$

Predicted prob. of $E_{H2} = 0$ approx. 0.85
Predicted prob. of $E_{H2} = 1$ approx. 0.15
**Research question:** Are the patterns of predicted probabilities among human-human and IEA-human similar for subgroups of interest?

**Caution:** charts should not be over-interpreted at each score point.
Predicted probabilities human-human IEA-human Grade 5 Female

Human – Human (n=482)  IEA – Human (n=6,175)

Written Expressions
Predicted probabilities human-human IEA-human Grade 5 Female

Human – Human (n=482)

IEA – Human (n=6,175)

Written Conventions
Predicted probabilities human-human IEA-human
Grade 7 Black or African American

Human – Human (n=291)

IEA – Human (n=2,944)

Written Conventions
Predicted probabilities human-human IEA-human
Grade 11 Students with Disabilities

Human – Human (n=151)

IEA – Human (n=810)

Predicted Probabilities for E_H2

Written Expressions
Predicted probabilities human-human IEA-human Grade 11 Students with Disabilities

Human – Human (n=151)

Predicted Probabilities for C_H2

IEA – Human (n=810)

Predicted Probabilities for C_H2

Written Conventions
• This analysis is primarily descriptive.

• For the subgroups with sufficient sample sizes across score points, the patterns of predicted probabilities appear similar between human-human and IEA-human.
Subgroup analyses for fairness and validity

Williamson et. al (2012) offers suggestions for assessing fairness: “whether it is fair to subgroups of interest to substitute a human grader with an automated score” (p. 10).

Examination of differences in the predictive ability of automated scoring by subgroup:

1. **Prediction of Second Score**: Compare an initial human score and the automated score in their ability to predict the score for the second human rater by subgroup.

2. **Prediction of External Score**: Compare the automated and human score ability to predict an external variable of interest by subgroup.
External variable: PARCC ELA/L assessments provide separate claim scale scores for both Reading and Writing. Reading raw scores typically range from 0 to 60-65 points.

Ordinary least squares model for reading score (y) predicted by the score (x), where score is treated as continuous.

$$ y = X\beta + \varepsilon $$

Model 1: Reading predicted by human score
Model 2: Reading predicted by IEA score
Research Question: Does IEA score predict reading score similarly to human scorers for subgroups of interest?

\[ SD(y) = \text{std. dev. of reading score} \]
\[ b = \text{estimated slope} \]
\[ \text{RMSE} = \text{root mean square error} \]
\[ R^2 = \text{R-squared} \]
Predicting reading by human or IEA score
RMSE boxplots for subgroups

English Language Learners

- Conventions - Human
- Conventions - IEA
- Expressions - Human
- Expressions - IEA
Predicting reading by human or IEA score
RMSE boxplots for low sample size subgroups

Black / African American

12
10
8
6
4
2
0

Conventions - Human  Conventions - IEA  Expressions - Human  Expressions - IEA
Predicting reading by human or IEA score
RMSE boxplots for low sample size subgroups

Asian

Conventions - Human  Conventions - IEA  Expressions - Human  Expressions - IEA
Predicting reading by human or IEA score
RMSE boxplots for low sample size subgroups

Hispanic

Convention - Human
Convention - IEA
Expressions - Human
Expressions - IEA
• Comparing RMSE as a measure of model fit suggests that scores from IEA predict reading scores similarly to human scorers.
Overall summary

• IEA-human follows similar trends to human-human agreement when looking at predicted probabilities.

• IEA scoring appears fair to subgroups.

• Results indicate students are not disadvantaged when scored by IEA.
Limitations

• Subgroups tend to have lower score scales, and oftentimes have sparse (or no) observed scores at the top score points.
  • This restriction of range inflates agreement rates.
  • Sparse data may cause model assumption violations for regression.

• Data for agreement analyses is limited by the second human scores.
  • Other than the 10% of IEA scores that receive a second score, a second scorer is requested through smart routing when the engine has low-confidence in its first score.
Better and Faster

- Demonstrated Success through PARCC Operational scoring with continuous flow
  - Population of 900k students results in 2-3M responses to score each administration; able to do this with a much shorter scoring window
  - Significant cost savings
  - Performance data supports the use when comparing to human scoring
- Continuing forward – Quick turn around of scoring and reporting is a key priority for New Jersey stakeholders
# Charting The Path Forward

<table>
<thead>
<tr>
<th>Phase</th>
<th>2018</th>
<th>2019</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>May</td>
<td>June</td>
</tr>
<tr>
<td><strong>Phase 1 (short-term planning)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School and Community Listening Tour</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statewide Assessment Collaboratives</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summary Findings Shared</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Phase 2 (long-term planning)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Next steps including additional outreach as determined by Phase 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Collaboratives and Community Meetings