Using machine learning to interpret NGSS tasks at scale

Jay Thomas
Jay.Thomas@act.org
ACT Next, ACT Inc.

Challenges with assessing NGSS with ML

- Unlike essay scoring which can use general algorithms and engines across multiple prompts (Altattal et al, 2010; Shermis, 2015; Shaw, et al, 2019), NGSS based items need a separate scoring rubric for each task because the integration of science content with argumentation is critical.
- Scoring multidimensional constructs that involve SEP, DCI, and CCC.
- Assessment tasks should include multiple components to fully assess a given concept (NRC, 2014) using authentic data.
- Student errors in spelling, typing, etc. should not negatively impact scoring if not relevant to the construct
- Must be able to maintain acceptable reliability (QWK>.7) across multiple testing cycles
- Need to be able to use composite items with forced choice and constructed response to assess multiple facets of constructs and concepts in time efficient way
- Provide feedback for future instruction (formative)

Items, rubrics, and scoring

![Image]

<table>
<thead>
<tr>
<th>Levels</th>
<th>Indicators</th>
<th>Sample Student Responses for Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 6: Students recognize the periodicity of the figure and identify plant processes as the primary cause.</td>
<td>1. Explains the seasonal pattern of the figure with an incorrect mechanism (e.g., people's fossil fuel use).</td>
<td>2. Not a cause / A minor cause / The main cause: A minor cause / Does not explain the seasonal CO2 variation in plant growth is more important than fossil fuel use. For instance, the CO2 levels in the atmosphere grow because there is less photosynthesis.</td>
</tr>
<tr>
<td>Level 3: Students recognize the periodicity of the figure but make mistakes explaining the mechanisms to cause or they create plant processes as the primary cause but don't explicitly state those processes to the seasonal pattern.</td>
<td>1. Accounts for the seasonal pattern in the figure with an incorrect mechanism (e.g., people's fossil fuel use).</td>
<td>2. Not a cause / A minor cause / The main cause: A minor cause / Does not explain the seasonal CO2 variation in plant growth is more important than fossil fuel use. For instance, the CO2 levels in the atmosphere grow because there is less photosynthesis.</td>
</tr>
<tr>
<td>Level 2: Students identify fossil fuels as a carbon source.</td>
<td>1. Explains that fossil fuels generate CO2 (carbon) and also identify other causes, too.</td>
<td>2. Not a cause / The main cause: A minor cause / Does not explain the seasonal CO2 variation in plant growth is more important than fossil fuel use. For instance, the CO2 levels in the atmosphere grow because there is less photosynthesis.</td>
</tr>
</tbody>
</table>

Examples of larger group analysis

Feedback loops in assessment system using ML

![Diagram]

Recursive Feedback Loops for Item Development

- Increase in the size of the usable data set to increase power of statistics.
- Increased confidence in reliability of scoring through back-checking samples and revising models.
- Reduced costs by needing fewer human coders.
- Model to show that the kinds of assessments envisioned by Pellegrino et al (2014) for NGSS can be reached at scale with low cost.
- Allows for comparison of learning gains because of scope of data.
- Models that fail to meet reliability guidelines can be replaced and all responses rescored quickly.
- Every student response from the entire year can be used for statistical analyses.
- Unit test (pre and post).
- Full year (pre and post).

ML engines CANNOT score items that humans score poorly. This does not mask problems in assessment but it will help to identify problematic issues: poor item design, incomplete rubrics, inconsistent human scoring.

This allows for iterative development of items that are able to assess the desired constructs consistently. Many items in assessment fail either during review or pre-test stages. These feedback loops allow for some of these items to be used through improvements in the rubric or human scoring while driving some items to be replaced so that they better measure the desired constructs.

<table>
<thead>
<tr>
<th>School year</th>
<th>Responses scored</th>
<th>Unique items scored</th>
<th>Assessments scored</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-16</td>
<td>175,265</td>
<td>33</td>
<td>27,981</td>
</tr>
<tr>
<td>16-17</td>
<td>532,825</td>
<td>39</td>
<td>61,475</td>
</tr>
<tr>
<td>17-18</td>
<td>693,086</td>
<td>41</td>
<td>66,335</td>
</tr>
<tr>
<td>18-19</td>
<td>409,266</td>
<td>39</td>
<td>42,117</td>
</tr>
<tr>
<td>Total</td>
<td>1,810,442</td>
<td>57</td>
<td>197,908</td>
</tr>
</tbody>
</table>

Citations available. Please email the author.

ACKNOWLEDGEMENT. This poster is adapted from an earlier version of a paper co-authored with Andy Anderson, Qinyun Lin, and Kenneth Frank (MSU) as well as Karen Draney and Shruti Bathia (BEAR). When submitted for publication rather than a presentation all co-authors will receive the credit that they richly deserve. Karen and Shrutis work is primarily contained in another poster in this symposium. Qinyun, Andy, and Kenneth are presenting in another session at NARST.