Powergrading: Helping Teachers Grade Short Answers at Scale

Sumit Basu, Chuck Jacobs, Michael Brooks, Lucy Vanderwende
Challenges of Testing & Grading at Scale

- Modern web technologies are solving the one:many problem, but there are many challenges to the many:one direction...
Not all Tests are Created Equal

- True/False
- Multiple Choice
- Fill-in the Blank
- Short Answer
- Essay/Application
- Project

- Laufer and Goldstein, 2004, on the difficulty of Recall tasks vs. Recognition
- Anderson and Biddle, 1975, “On Asking People Questions About What They are Reading.”
Grading is More than Assigning Scores

- Grading Practices
  - Grading is complex and individualized
  - Maintaining consistency between students is a challenge
- Instructor grading has important benefits
  - Opportunity for rich feedback
  - Getting a snapshot of student understanding
  - Adapting teaching to student needs
- Issues with Auto-Only Grading
  - some answers are always left on the table
  - no feedback for teachers
- Assisted Grading: can we leverage the best of human and machine?

Core Idea:
Perhaps clustering could help deal with a large number of answers, while still allowing for completeness, feedback, and reflection

Hypothesis and Metric:
using this procedure, the number of “grading actions” required to complete the task could drop dramatically.

# Actions Remaining after Auto + 3 User Actions

<table>
<thead>
<tr>
<th>Q#</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>Num Unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>57</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>12</td>
<td>13</td>
<td>132</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>87</td>
<td>145</td>
<td>586</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>26</td>
<td>43</td>
<td>205</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>18</td>
<td>26</td>
<td>138</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>30</td>
<td>50</td>
<td>219</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>12</td>
<td>11</td>
<td>395</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>157</td>
</tr>
<tr>
<td>13</td>
<td>75</td>
<td>54</td>
<td>50</td>
<td>367</td>
</tr>
<tr>
<td>20</td>
<td>11</td>
<td>14</td>
<td>10</td>
<td>276</td>
</tr>
</tbody>
</table>
Designing and Evaluating a User Interface for Powergrading
Clustering Grading Interface
Flat Grading Interface (Baseline)
Questions

Preferences: Did teachers like it?
Efficiency: Was grading speed improved?
Quality: Was accuracy affected?
Feedback: Was giving feedback supported?
Reflection: Were teachers able to reflect on student answers?

Study: 25 MS/HS/College teachers, within-subjects (2 interfaces, 2 questions)
“When initially viewing the video on this interface, I was a little worried that it might be somewhat complicated and time consuming due to the subcategories. However, I was incorrect. This interface was quite efficient and easy to use.” (P15)

“[The clustered interface] worked very well for me, especially given the large number of total responses. I found [the flat interface] quite tedious … The clustered interface] helped me to identify student patterns in thinking quite well.” (P12)

<table>
<thead>
<tr>
<th></th>
<th>Clustered</th>
<th>Flat (Baseline)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster</td>
<td>21</td>
<td>4</td>
</tr>
<tr>
<td>More</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Enjoyable</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Easier to Use</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>More Effective</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>Better Overall</td>
<td>21</td>
<td>4</td>
</tr>
</tbody>
</table>

21 of 25 participants preferred the Clustered interface
Efficiency

Flat Interface, answers graded vs. time for Q4

Clustered Interface, answers graded vs. time for Q4
Speed Calculation

**Flat Interface**, answers graded vs. time for Q4

120 ans. / 20 min.  
= 6 ans/min

**Clustered Interface**, answers graded vs. time for Q4

196 ans. / 9 min.  
= 21 ans/min
Grading Quality

Accuracy vs. Speed for Both Questions

- Q6 Flat
- Q6 Cluster
- Q4 Flat
- Q4 Cluster

Participant 4
Amplifying Feedback

- No difference in median instances of entered feedback
- However, clustering distributed the same number of instances to many more answers: (median 75 vs. 18)
- Clustered interface rated higher for supporting feedback

“Being able to grade categorized responses makes it easier on the grader and allows them to pay closer attention to types of feedback needed.” (P24)

“Because [the clustered interface] was so much faster, more time could be spent giving feedback.” (P14)
Instructor Reflection

• Participants felt the clustered interface was better for this:

“This interface does make answer trends more easily identifiable.” (P6)

“I liked this [clustered] interface better; breaking the answers down into clusters allowed me to spot patterns, to be more consistent in grading, and to devote more time to individual answers where it wasn’t clear whether they were right or wrong. The information seemed less overwhelming when presented this way, so I felt like I was less apt to mis-read or mis-grade any one answer.” (P8)
Powergrading: a “Power Tool” for Grading

• Keep teachers in the loop, but amplify their capabilities so they can handle larger classes

• Current work: making Powergrading broadly available; let us know if you want to try it out with your data

http://bit.ly/powergrading  sumitb@microsoft.com