Divide and Correct: Using Clusters to Grade Short Answers at Scale

Michael Brooks$^{1,2}$, Sumit Basu$^1$, Chuck Jacobs$^1$, Lucy Vanderwende$^1$

Learning At Scale 2014: Assessment
March 5$^{th}$, 2014, Atlanta
Not All Tests Are Created Equal

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

• Impact beyond grades
  • Opportunity for rich feedback to students
  • Getting a snapshot of student understanding
  • Adapting teaching to student needs

What Can We Do with Short Answers at Scale?

• **Automatic Grading**
  - Hand-constructed answer templates, regular expressions
  - Paraphrase recognition (C-Rater)
  - Similarity metric between response and answer key

  Mitchell et al. 2002; Nielsen et al. 2009; Coursera

  Leacock & Chodorow 2003

  Mohler & Mihalcea 2009; Mohler et al. 2011; Hahn & Meurers 2012

• **Peer Grading**
  - Debiasing and aggregation
  - Crowdsourced grading
  - Possible learning benefits

  Piech et al. 2013; Reily, Finnerty & Terveen 2009

  Weld, Adar & Chilton 2012

  Sadler & Good 2006
Our Approach: Divide and Correct

Automatically group student answers into a hierarchy of clusters*

Teachers give grades, feedback on clusters, subclusters, and answers

Amplify teachers’ capabilities

Grading Clustered Short Answers

What about real teachers?
Do clusters make sense?
Can they find outliers?
Is it confusing?
Design
Clustered Grading Interface
Clustered Grading Interface Design

• Exploratory navigation
• Grade and feedback is inherited and overridden
Clustered Grading Interface Design

• Exploratory navigation
• Grade and feedback is inherited and overridden
• Visual summary of cluster contents
see http://bit.ly/powergrading
Evaluation
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 Design

• Online study
• Within-subjects, 20 minute tasks, 1 hour total
• Tutorial video for each interface
• Gathered log data and questionnaire answers
• Examined speed, accuracy, feedback, and comments
Participants

- Recruited 25 teachers
- Screened for teaching experience
- Teaching experience was vetted individually

<table>
<thead>
<tr>
<th>Subject Area</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government, Politics, Civics</td>
<td>9</td>
</tr>
<tr>
<td>English</td>
<td>23</td>
</tr>
<tr>
<td>Literature</td>
<td>12</td>
</tr>
<tr>
<td>STEM areas</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middle/high school</td>
<td>22</td>
</tr>
<tr>
<td>College/graduate</td>
<td>12</td>
</tr>
</tbody>
</table>
Answer Data for Evaluation

• Powergrading Short Answer Grading Corpus
  • Selected 2 questions from the US Citizenship Exam
  • Total of 698 answers per question, from Mechanical Turk

<table>
<thead>
<tr>
<th>Question</th>
<th>Distinct answers</th>
<th>Avg. words/ans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q4</td>
<td>196</td>
<td>3.9</td>
</tr>
<tr>
<td>Q6</td>
<td>205</td>
<td>6.4</td>
</tr>
</tbody>
</table>
Results
“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>Preference</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 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>
Efficiency

Flat Interface, answers graded vs. time for Q4

Clustered Interface, answers graded vs. time for Q4
Efficiency

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

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

Grading a cluster

Refining and searching
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
Grading Quality

Accuracy vs. Speed for Both Questions

- Q6 Flat
- Q6 Cluster
- Q4 Flat
- Q4 Cluster
Grading Quality

Accuracy vs. Speed for Both Questions

Participant 4
### Feedback for Students

**Question:** “Who or what makes federal (national) laws in the US?”

<table>
<thead>
<tr>
<th>Answer from students</th>
<th>Feedback from teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td>“state legislation (legislators)”</td>
<td>“State legislators do not make national laws. The congressional members from the States do, but not the states themselves.”</td>
</tr>
<tr>
<td>“congress along with the president and the supreme court”</td>
<td>“Be clear with your response. what role if any does the Supreme Court play?”</td>
</tr>
</tbody>
</table>
Amplifying Feedback

• No difference in median amount of feedback written
• Clustering distributed feedback to 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

• Insights elicited after grading appeared equally rich
• 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)
Findings

**Preferences:** Faster, easier to use, more enjoyable and effective

**Efficiency:** Makes rapid grading possible

**Quality:** Effectively allows the grader to control quality

**Feedback:** Impact of feedback increased

**Reflection:** Clustering aids reflection
Conclusions

• Clustered grading of short answers is promising
  • More scalable grading without loss of accuracy
  • Better support for feedback and instructor reflection

• Future work
  • Evaluation at larger scale, in a real course
  • Refinement to clustering, visual cluster summaries
  • Integrate auto-grading features
Divide and Correct: Using Clusters to Grade Short Answers at Scale

Michael Brooks¹,², Sumit Basu¹, Chuck Jacobs¹, Lucy Vanderwende¹

Learning At Scale 2014: Assessment
March 5th, 2014, Atlanta

bit.ly/powergrading
Backup Slides
Question: Who or what makes federal (national) laws in the US?

Congress

Senate and House

Senate and House of Representatives

Congress makes federal laws.

Congress is in charge of making federal laws.

The legislative branch is responsible for making federal laws.

Who or what makes federal (national) laws in the US is Congress.
Q: Who or what makes federal (national) laws in the US?

Congress
Senate and House
Senate and House of Representatives
Who or what makes federal (national) laws in the US?

- Congress
  - Senate
  - House of Representatives
- Legislative branch
- Senate and House of Representatives
  - Congress makes laws
- President
- Supreme Court
- Legislative branch (House of Representatives and the Senate)
- Congress, House of Representatives, Senate, House of Representatives, Senate
Question: Who or what makes federal (national) laws in the US?

- Congress
- Senate
- House of Representatives

Subclusters:
- Congress (senate and house of representatives)
- Senate
- House and Senate
- The legislative branch (house of representatives and the senate)

Clusters:
- Legislative branch
- Senate and house of representatives
- President, congress

Answer Key:
- Congress
- Senate and House
- Senate and House of Representatives