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

Michael Brooks[1,2]
[1]University of Washington
Seattle, WA
mjbrooks@uw.edu

Sumit Basu[2], **Charles Jacobs**[2], **Lucy Vanderwende**[2]
[2]Microsoft Research
Redmond, WA
{sumitb, cjacobs, lucyv}@microsoft.com

ABSTRACT

In comparison to multiple choice or other recognition-oriented forms of assessment, short answer questions have been shown to offer greater value for both students and teachers; for students they can improve retention of knowledge, while for teachers they provide more insight into student understanding. Unfortunately, the same open-ended nature which makes them so valuable also makes them more difficult to grade at scale. To address this, we propose a cluster-based interface that allows teachers to read, grade, and provide feedback on large groups of answers at once. We evaluated this interface against an unclustered baseline in a within-subjects study with 25 teachers, and found that the clustered interface allows teachers to grade substantially faster, to give more feedback to students, and to develop a high-level view of students' understanding and misconceptions.

Author Keywords

Grading; grading interfaces; assessment; MOOCs; clustering; user interfaces; clustering interfaces.

ACM Classification Keywords

H.5.2 Information interfaces and presentation (e.g., HCI): User interfaces – Evaluation/methodology; K.3.1 Computers and Education: Computer uses in education.

INTRODUCTION

While the impressive scale of modern online courses allows a teacher to easily deliver lectures to massive numbers of students, interactions in the other direction are still a challenge. Dealing with hundreds or thousands of student exams can be overwhelming, particularly if they contain responses to open-ended questions. At the same time, open-ended assessments have substantial value to both students and teachers, as we discuss in detail in the background section. Automated approaches to grading open-ended questions reduce the workload for teachers, but some benefits of open-ended questions depend on the teacher actively reviewing and assessing their students' answers.

In this work, we propose a means of maintaining the teacher's involvement while still allowing them to work with short answer responses at scale. Our approach uses clustering to group student responses into clusters and subclusters. We develop an interface giving teachers access to these groupings, allowing them to read, grade, and provide feedback to large numbers of responses at once.

Previous research has demonstrated that automatic clustering of answers could reduce the number of grader actions required, potentially improving scalability of instructor grading for an algorithmically "optimal" grader [2]. In order for a grading interface to be effective at scale, though, speed alone is not enough. While efficiency is important, teachers must also be able to get a sense for trends in students' understanding, as well as give helpful feedback to students with misconceptions. In this work, we investigate whether a user interface for grading clustered answers would improve efficiency for real teachers, while promoting high quality grading, feedback for students, and instructor reflection.

We have created a web application that allows a teacher to grade and give feedback on hundreds of responses to short-answer questions; in our work this refers to answers that range from a few words to a sentence in length. In a within-subjects experiment with 25 teachers, we compared our proposed *clustered* version of the system, where the grader works with automatically clustered responses, to an unclustered (*flat*) baseline system. We found that teachers were able to grade far more quickly using the clustered version, and that the resulting grades were of equivalent accuracy when compared to a gold standard. Feedback was given to roughly three times as many answers, and teachers reported being better able to reflect on trends in student understanding. Furthermore, teachers found the new interface to be superior in terms of ease of use and overall effectiveness.

BACKGROUND AND RELATED WORK

Below, we provide background on the role that assessment plays in the education process, and discuss related work on peer and automated grading in the context of education research on grading practices, feedback, and reflection.

L@S 2014, March 4–5, 2014, Atlanta, Georgia, USA.

ACM 978-1-4503-2669-8/14/03…$15.00.
http://dx.doi.org/10.1145/2556325.2566243

Assessment

Educational assessment is concerned with measuring student ability and aptitude, but the goals and impacts of assessment are broader than individual students. Assessment is used for quality assurance for educational institutions and programs, and influences ways of teaching [7]. Testing also assists knowledge retention and can guide learning [1,20]; for example, "it is not until students start to work on their assignment that they know whether or what they have learned from their studies" [20]. Assessment indirectly influences student learning by shaping curriculum design and learning goals [7], and teachers often intentionally adapt their teaching based on formative assessments [3,19]. Assessment is deeply intertwined with most aspects of education; additional research is needed to understand assessment at MOOC scale.

There are many different forms of assessment available, and their use varies widely. For example, in a survey of secondary teachers, McMillan found that social studies and science teachers reported using objective assessments and quizzes significantly more often than English teachers did, while English, science and social studies teachers tended to use constructed- or open-response assessments more than math teachers, and English teachers more than science teachers [11]. With MOOCs and other approaches to large-scale education becoming increasingly important, multiple-choice questions and other highly-constrained assessment instruments offer scalability since submissions can be automatically scored against an answer key. However, Anderson and Biddle established that open-ended, constructed response questions such as short answers and essays are preferred forms of assessment [1]. Open-ended questions are more valuable for measuring understanding, application, and reasoning [11], and play a critical role in consolidating learning [9]. Our work is concerned with making it practical for teachers to use short answer questions in MOOCs, where there may be hundreds or thousands of responses to grade.

Grading

As with assessment strategies, approaches to grading differ across grade levels, subject areas, and from teacher to teacher. Grading has been a controversial topic in education research for most of the past century [5], with much discussion focused on the factors that teachers take into account when assigning grades. Studies of grading practice find that teachers use a "hodgepodge" of factors in grading, including not only achievement but also effort and ability [4,5]. Although this contradicts the recommendations of measurement specialists [5], more complex and subjective judgments may support teachers' practical needs of managing classrooms and motivating students [4]. Whether or not a more objective and achievement-oriented approach to grading is desirable, we must be mindful that, in practice, there is far more to grading than just marking an answer "right" or "wrong."

For grading open-ended assessments in MOOCs, one approach has been automated grading against a carefully authored answer key that attempts to anticipate all possible student answers [6,8]. However, designing the answer keys is time-consuming, and tuning may require linguistics expertise, making this an unreasonable method for most teachers. Alternatively, automated grading can be formulated as a similarity task in which a score is assigned based on the similarity between student answer and teacher answer [12]. While these automatic methods promise improved scalability of constructed-response assessment, it must be mentioned that the accuracy of both methods of automatic grading is reported in the range of 84% to 92%, which is less than the 100% which a teacher strives for, and so understandably, there is a lively debate regarding the trade-offs [10,14].

Other approaches to grading open-ended assessments include peer-grading and self-grading. In peer-grading, students use a scoring rubric to grade each other's answers [15,17,18]. Students are typically from the same class or a parallel class, but some researchers have explored whether peers could also be drawn from a *crowd* of experts [21]. There have been claims that peer-grading provides learning benefits to graders through the grading exercise itself [17]. However, Sadler and Good found no evidence that peer-grading improves peer performance on subsequent tests on the same material [18]. Moreover, while this approach shows promise, student biases, self-interest, and lack of expertise remain as limitations.

Feedback

Feedback, a major area of study in education, provides "the comparison of actual performance with some set standard of performance" [13] and is known to facilitate learning. A great deal of education research has focused on how, when, and what feedback should be provided to students in order to maximize various learning benefits; recently, focus has shifted to student perspectives and perceptions about feedback [16]. Research on scalable assessment in MOOCs should take into account how effectively feedback can be provided to students. Feedback can be pre-authored for multiple-choice answers, as seen in the tutorial-like feedback used by the Khan Academy quizzes (kahnacademy.org), but automatically providing feedback for open-response assessments would require pre-authoring feedback for all anticipated answers as well as the ability to match answers to feedback automatically; this would have many of the same drawbacks as automated grading. As with grading, there is the possibility of peer feedback: in a peer-grading exercise for a Coursera MOOC, students gave each other feedback alongside grades [15]. Analysis showed that negative comments were typically longer, but that overall, the comments ranged "from neutral to quite positive, suggesting that rather than being highly negative to some submissions, many students make an effort to be balanced in their comments to peers."

Reflection

Mastery Learning, which has been shown to improve student performance in a traditional classrooms, relies on frequent use of *formative assessment*, where teachers use assessment to learn about and improve their teaching methods [3,19]. For example, teachers may "determine when [students] comprehend the explanations and illustrations" so that they can supply additional clarification as needed [3]. Grading is an important avenue for teachers to gain comprehension of student understanding and misunderstanding, but automated grading and peer grading cannot offer teachers the same insights they get from grading students themselves.

CLUSTERED GRADING APPLICATION

We designed and implemented a web application enabling teachers to *efficiently* grade and give feedback on a large volume of short answers, while still allowing teachers to control the quality of their work and to learn about the general state of students' understandings. This efficiency is created through the use of clustering to group and organize similar answers. We iteratively designed and developed the application with continual testing and evaluation by the authors, consultation of the literature on grading practices reviewed above, and informal discussions with teachers, educators, and graders. The interface is shown in Figure 1.

A video showing the interface in action can be seen at http://research.microsoft.com/~sumitb/grading .

Target Users and Context

In our design process, we envisioned an instructor or grader in a large online course using open-ended questions on an exam or other assignment. The instructor collects the students' answers to a question and provides them to our software, which runs a *hierarchical clustering algorithm* on the answers and displays an interface for grading the clustered answers. In this work, we focus only on grading, the final stage in this process. Below, we briefly describe the clustering algorithm that our system uses, and then discuss our most significant design challenges and decisions.

Clustering Algorithm

We clustered short answers using the metric clustering approach developed and evaluated in [2], which builds a hierarchy with two levels: the hierarchy may have up to 10 top-level clusters with up to 5 subclusters in each cluster. There may also be *miscellaneous* clusters or subclusters, containing answers that did not fit well into any of the other clusters.

The clustering algorithm computes a learned distance metric over pairs of answers, based on difference in answer length, words with matching base forms, *tf-idf* vector similarity, lowercase string match, and Wikipedia-based LSA similarity. The clustering is computed from these distances using a version of the k-medoids algorithm. Further details on rationale, implementation, and evaluation are in [2].

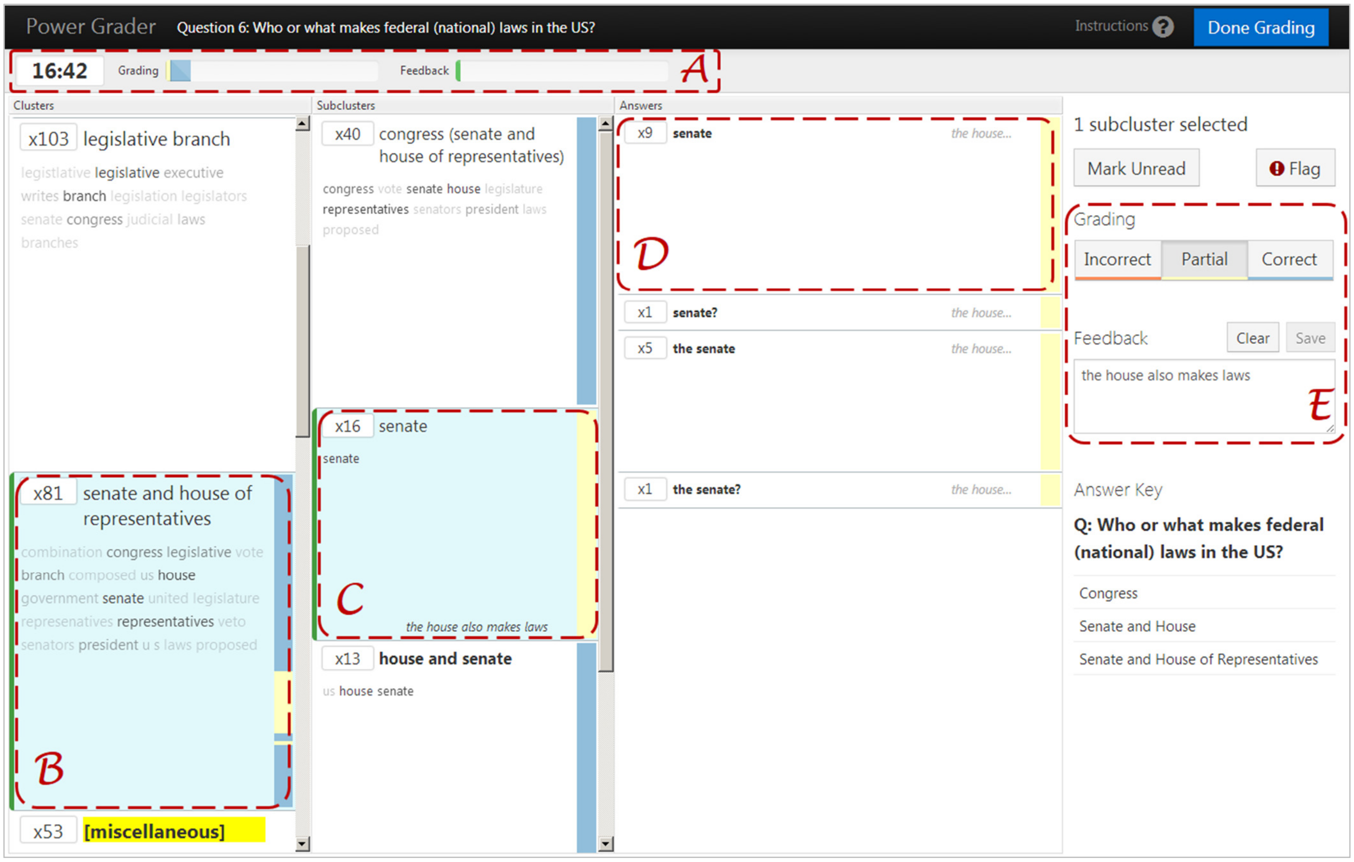

Power Grader
Question 6: Who or what makes federal (national) laws in the US?
Instructions ?
Done Grading
16:42
Grading
Feedback
A
Clusters
x103 legislative branch
legistlative legislative executive writes branch legislation legislators senate congress judicial laws branches
x81 senate and house of representatives
combination congress legislative vote branch composed us house government senate united legislature represenatives representatives veto senators president u s laws proposed
B
x53 [miscellaneous]
Subclusters
x40 congress (senate and house of representatives)
congress vote senate house legislature representatives senators president laws proposed
x16 senate
senate
C
the house also makes laws
x13 house and senate
us house senate
Answers
x9 senate
the house...
D
x1 senate?
the house...
x5 the senate
the house...
x1 the senate?
the house...
1 subcluster selected
Mark Unread
Flag
Grading
Incorrect
Partial
Correct
Feedback
Clear
Save
the house also makes laws
E
Answer Key
Q: Who or what makes federal (national) laws in the US?
Congress
Senate and House
Senate and House of Representatives

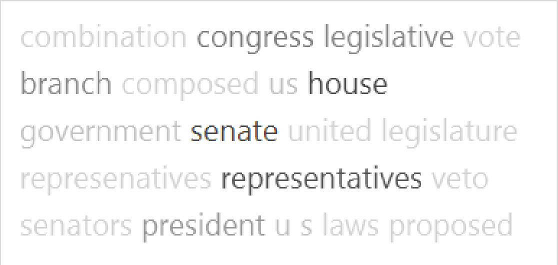
Figure 1: Grading a cluster in Question 6. (A) progress bars; (B) a cluster of 81 students, graded mostly correct (blue); (C) the selected subcluster, graded partially correct (yellow); (D) the answer "senate" by 9 students; (E) grading and feedback controls.

Cluster Exploration

Our first challenge was designing a layout and organization for the clusters that would be easily understandable and support an efficient process for grading. We considered several approaches: for example, we sketched a focused, one-at-a-time presentation of clusters and subclusters for grading, like checking phone messages on an answering machine. However, we decided this kind of guided design would be too constrained; instead, we designed an open-ended hierarchy-browsing layout that we supposed would be familiar from email or file system navigation programs. Users view and interact with lists of clusters, subclusters, and answers in a three-column layout (Figure 1). Selecting a cluster causes its nested subclusters and answers to appear, while subsequent selection of a subcluster filters the answers displayed. This design allows users to quickly and easily explore or drill down into the answers wherever the contents of a challenging cluster may demand it.

Cluster Summaries

This suggests another design challenge: how to present summaries of clusters and subclusters so as to enable users to make *informed* decisions about where to explore. We prototyped several visualization-based techniques with the goal of representing the "cohesiveness" or "compactness" of clusters and subclusters. For example, since each answer within a cluster has an estimated distance from the cluster centroid, we created plots showing the distribution of distances within the cluster. It became apparent that the distances were often not intuitively distributed, leading to confusing or misleading visualizations, and that these visualizations wasted space in clusters with few answers.



combination congress legislative vote
branch composed us house
government senate united legislature
represenatives representatives veto
senators president u s laws proposed

Figure 2: A summary of cluster contents. The list of words is ordered by average position in the answer text.

We developed a type of *word cloud*, a representation of word frequency in unstructured text (Figure 2). Unlike typical word clouds, we decided to keep the font size of words constant for readability, and to instead vary lightness so that the least used words appear nearly transparent. The elegance of this solution is that clusters with few answers, or little diversity of answers, have relatively small word clouds. Meanwhile, clusters with a great diversity of answers, which likely warrant closer inspection, have large, eye-catching word clouds. A weakness of traditional word clouds is that information about the order of words in the original text is utterly lost. In our word clouds, we order the words according to their *average position* in the answers (normalized by answer length). This technique proved a useful improvement on the word clouds in many clusters we inspected, though not all. In addition, we scaled the height taken up by clusters and subclusters based on how many *students*, not distinct answers, were contained in each cluster (this number is also shown in the top left of each cluster, subcluster, or answer item). This encourages the grader to spend time first on those areas that will benefit the most people. Given the rich literature in text summarization and visualization in recent years, there is a substantial research opportunity in experimenting with alternative visualizations for clusters of answers; we hope that we as well as others will explore this space further.

Grading and Feedback Interactions

The third challenge was to design the operations that the interface should support. It was clear that there would need to be a *grade* action, but it was less clear what the user would expect to happen if, for example, they mark a particular cluster *Correct*, one of its child subclusters *Incorrect*, and one of its child answers *Partial*. We settled on the following solution: when the user marks a grade on a cluster as a whole, its contained subclusters "inherit" that grade. The user can "override" that inheritance on one of the subclusters by marking it with a grade. Similarly, individual answers inherit the grade of their parent subcluster or cluster unless they have been specifically marked by the user. The effect is that the user can use grade inheritance to mark a possible grade to every answer very rapidly, simply by grading the clusters. Following up with closer inspection of the subclusters or answers allows discovery of exceptions which may need to be overridden. We developed an identical model for feedback—writing a feedback message for an entire cluster effectively gives feedback to all of the students contained in that cluster, but this can also be overridden. To make this behavior more apparent to users, we added color-coded grading indicator bars to the right side of every cluster, subcluster, and answer (Figure 1 B and C). These show the current grade (whether inherited or marked directly) in a kind of spatial map of every answer in the given cluster or subcluster.

Grading Controls and Progress Bars

Finally, we added several features to the interface that make it more convenient and more satisfying to use. In the right-panel of the grading tool (Figure 1 E), we provide a group of three buttons for marking grades on the selected cluster, subcluster, or answers, with colors associated consistently wherever grades appear throughout the interface. Below the grading controls, feedback can be typed for the current selection. It is also possible to *reuse* feedback that was given previously, a time-saving feature. At the bottom of the right panel is the question currently being graded, and the answer key.

Assuming that graders, facing an overwhelming number of answers to grade, would want to see that they were making progress, we added progress bars showing the proportion of answers with grades and feedback (Figure 1 A). This is more complex than it might first seem, since there is actually no clearly-defined end point to grading. Anecdotally, it is not uncommon in traditional grading to make multiple passes through student answers until a satisfactory level of consistency and quality is achieved. Thus, the grading progress bar could rapidly fill at the beginning of the grading task, when in fact more exploration and checking is needed to improve grading quality. Likewise, since many answers need no feedback (particularly *Correct* answers), the feedback progress bar usually never fills. To the left of the progress bars is a timer, which counted down from the time-limit of 20 minutes for the purpose of our study.

METHOD

We used a within-subjects experiment design to compare *clustered* vs. *flat* grading of hundreds of responses to short answer questions in terms of efficiency, grade quality, feedback to students, and grader insight. The study was conducted online with 25 individuals with teaching experience; participants spent up to 20 minutes grading about 200 distinct answers from each of two questions. Below, we provide details on the study participants, question and answer data, and experiment design.

Participants

Grading is a complex practice balancing many competing priorities and concerns, and graders develop expertise and tacit knowledge that could have a significant impact on how they interact with the software. Therefore, we sought participants with teaching or grading experience—at least one year of teaching or grading experience within the past five years. We also required that they occasionally use email, chat or forums, and spreadsheets; and that they used some of these tools in their teaching or grading work. They also had to be native speakers of English.

To more easily reach this population, we conducted the study online. For recruitment, we worked with a company specializing in linguistics crowd-work, which sent our initial screening survey to a population of 110 teachers or former teachers from its worker pool. Of the 80 who responded, we invited 40 qualified individuals. Of these, 25 completed the one-hour study successfully, and were paid $15.

In the group of 25 participants, eight are aged 22 to 34, eleven are 35 to 44, and six are 45 or over. All participants reported at least 2 years of teaching experience, and about half had more than 5 years; the company that assisted with recruitment individually vetted these teaching backgrounds. Participants' teaching experience was in a wide variety of subject areas, and most people had taught multiple subjects; due to the topic of the grading tasks (discussed below) we prioritized recruitment for participants who had taught Government, Politics or Civics (9 people), but others were also included: 23 participants had taught English, and 12 had taught literature; 12 had teaching experience in science, technology, engineering, or math. Most (22) had experience at the middle or high school level; 12 had college- or graduate-level teaching experience. All reported using communication technology such as email, forums, or chat multiple times a day, and all but one had used email for communicating with students. All participants reported that they use or have recently used the Internet at least once a week for their teaching roles, and use spreadsheets at least monthly. All but 4 said that they have used computers for grading work, 13 out of 25 had done this at least daily. Most (19) had taught an online course at least once, and 14 had done so many times. None had experience with MOOCs; the largest class size reported was 120 students.

Questions to Grade

Participants in our study graded answers to each of two different short answer questions. We selected a pair of questions from the Powergrading Short Answer Grading Corpus [2], which includes short answers given by 698 Mechanical Turk workers, for 20 questions from the United States Citizenship Exam. The questions vary in terms both of scope and difficulty. The corpus also contains grades independently assigned by three judges, which we used as benchmark grades for comparison to grades from our study participants.

Because the kind and structure of answers received depends on the question, the efficiency of clustered grading may be sensitive to the question being graded. We decided to select a pair of "average" questions: for each of the 20 questions, we calculated the number of *distinct* answers (some answers are identical), the average answer length, the percent of correct answers *vs.* the researcher grades, and agreement among the researcher grades. We selected Question 4 and Question 6 (Table 1) because, in terms of these four metrics, they were in the middle of the distribution for the corpus, and were similar to one another. Question 4 had 196 distinct answers (1-25 words in length, mean=3.9), and Question 6 had 205 (1-97 words, mean=6.4). From this point on, we refer to these distinct answers as simply answers, for brevity.

Q4	*What is the economic system in the United States?*
Q6	*Who or what makes federal (national) laws in the US?*

Table 1: The two questions that were selected for grading in the experiment. Participants graded both questions.

Clustered and Flat Grading Interfaces

Participants graded the answers to one of the two questions using the *clustered* grading application, discussed above. To determine how interacting with clustered answers altered the process and results of grading, we created a second, *flat* version of the application which does not use clustering. The flat version was designed to provide a realistic baseline grading experience, while also preserving as much from the clustered version as possible.

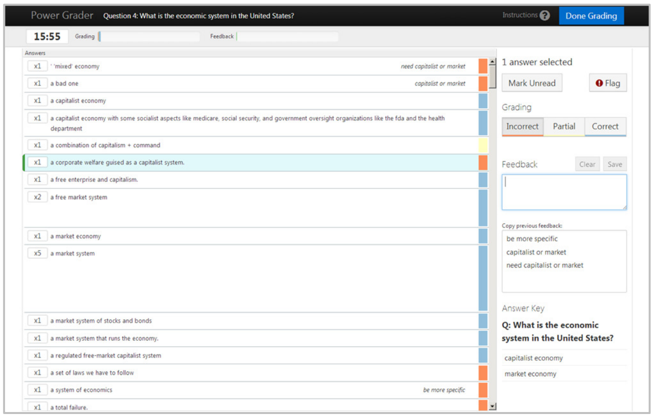
Our assumption in designing the flat version was that in the absence of a more specialized tool, most teachers and graders would probably attempt to grade short answers by going through them one by one. For example, they could grade using a spreadsheet program, since such software is commonly available and familiar. The flat grading interface displays a flat list of all of the answers, allowing a workflow similar to grading in a spreadsheet. We sort the answers alphabetically, a basic step that provides some time-savings because answers that start with the same words sometimes receive the same grade or feedback. As in the clustered interface, we collapse duplicated answers. The flat interface (Figure 3) is implemented as a minor variation of the clustered interface. We simply remove the two left panels which display the clusters and subclusters, allow the Answers panel to expand horizontally, and show all of the distinct answers in alphabetical order. All other aspects of the tool are identical between the two versions, facilitating comparison in our evaluation.

Measures

Because the study was conducted over the Internet, most data was collected through logging of user activities. Selection of clusters, subclusters, or answers was logged, as was the application of grades or feedback. These events were timestamped to show how users progressed through the task.

In addition to actions over time, we also evaluated the final grades given by participants. We created a set of "gold standard" grades based on the three independent sets of grades from the Powergrading Corpus. We selected the subset of answers where the three independent graders had unanimous agreement (about 82% of answers, for each question). These grades are binary (Correct or Incorrect) without partial credit as we had in the current study, so we converted our participants' grades to match by re-coding partial credit as Correct, as this best reflected the rubric used by the graders from the corpus.

We spaced several questionnaires at various points throughout the study. A *Pre-Study Questionnaire* confirmed the participant's teaching experience and background information. A *Post-Task Questionnaire*, completed immediately after each grading task, asked for impressions and reflections on the task, including Likert-type questions about how the interface supported consistent and fair grading, giving feedback, and overall difficulty of use. Finally, a *Post-Study Questionnaire* asked for general comments and direct comparison of the two interfaces in terms of speed, ease-of-use, enjoyment, and effectiveness. We analyzed the Post-Task questions using non-parametric Wilcoxon Signed Ranks tests. For the Post-Study comparison, we compared how often participants chose one interface or the other using chi-square tests.



Question 4: What is the economic system in the United States?

15:55

x1 ' 'mixed' economy — *need capitalist or market*
x1 a bad one — *capitalist or market*
x1 a capitalist economy
x1 a capitalist economy with some socialist aspects like medicare, social security, and government oversight organizations like the fda and the health department
x1 a combination of capitalism + command
x1 a corporate welfare guised as a capitalist system.
x1 a free enterprise and capitalism.
x2 a free market system
x1 a market economy
x5 a market system
x1 a market system of stocks and bonds
x1 a market system that runs the economy.
x1 a regulated free-market capitalist system
x1 a set of laws we have to follow
x1 a system of economics — *be more specific*
x1 a total failure.

be more specific
capitalist or market
need capitalist or market

Q: What is the economic system in the United States?

capitalist economy
market economy

Figure 3: The flat grading interface. Identical to the clustered interface but without the cluster and subcluster controls.

One of the items on the Post-Task Questionnaire also invited participants to reflect on the students' answers: "summarize how the students did on the question you just graded." This was intended to elicit insights that the teachers may have gained while grading which they could have used to improve future lessons.

Procedure

The within-subjects study consisted of two grading tasks; participants graded both questions, and used both interfaces. We began by sending the 40 qualified participants a unique link to the study website, which randomly assigned participants to one of four study groups, comprised of both combinations of assignments of conditions (clustered and flat) to questions (4 and 6) as well both orders between conditions to control for effects from having been exposed to one interface or question before the other. Due to attrition, the final set of 25 had 15 participants who started with the *flat* interface, while 10 started using the *clustered* interface.

After assigning participants to an ordering of Question and Interface, the study website guided them through the pre-study questionnaire and a text description of the study, followed by the two interface conditions. Each condition consisted of a narrated video tutorial, the grading task itself, and a post-task questionnaire. The video tutorial contained instructions for the task and an explanation of the given interface. The study then ended with the post-study questionnaire. Each task took 20 minutes; the entire study took 1 hour.

RESULTS

We begin with an analysis of general participant comments about the interfaces, then examine grading speed, quality (accuracy), the amount/quality of feedback given to students, and reflections on students' understanding.

	Clustered	Flat
Faster	21	4
More Enjoyable	20	5
Easier to Use	20	5
More Effective	19	6
Better Overall	21	4

Table 2: Number of participants who preferred each interface across various attributes. All differences significant (p < 0.01).

Interface Preferences

After using both interfaces, the Post-Study Questionnaire asked participants to rate which was faster, more enjoyable, easier to use, more effective, and better overall (Table 2). We analyzed these responses using chi-square tests; the clustered interface was preferred significantly more often in all categories (p < 0.01 in all cases).

Additional chi-square tests indicated that neither the question being graded nor the order of using the interfaces had significant effects on these choices. The comments were positive:

> *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)*

A less skeptical participant described how the clustering was helpful for understanding how the students were doing:

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

Of the few who preferred the flat interface, the main reason was the complexity of working with clustered answers:

> *[The clustered interface] was just a little too complicated, even though some elements of it were easier. There was just too much to keep track of [...] I found myself having to backtrack and recheck a lot of answers. (P1)*

Grading Speed

We next investigate whether the proposed interface led to an increase in grading speed. We have provided the raw trace of answers graded over time for all participants grading Question 4 with the clustered interface (Figure 4) and the flat interface (Figure 5). These charts illustrate the marked difference in the general *shape* of the curves (the differences were similar for Question 6). Even the fastest participants using the flat interface had essentially linear progress; a few created sudden jumps by selecting multiple answers and grading the bunch. In contrast, the fastest users of the clustered interface show a different and decidedly nonlinear progression.

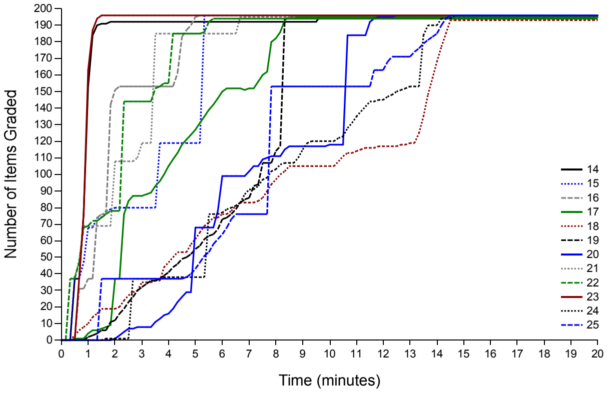


Figure 4: Answers graded vs. time for Q4, clustered interface

With the clustered interface, we observe early use of the high-level action of grading clusters and subclusters (vertical jumps), followed by no actions (horizontal segments) as the timer continues. This time was used to check subclusters and answers to see if refinements were needed after the first pass. With the clustered interface, all participants grading Question 4 had assigned grades for all answers after 15 minutes. If pressed for time, they could have spent less time checking the individual answers, but still would have assigned at least a first-approximation grade for every answer. With the flat interface, participants could only have progressed more quickly by increasing the slope, i.e., by examining more items per unit time. If forced to end early, answers would have been ungraded; indeed, several participants did not finish with the flat interface (Figure 5).

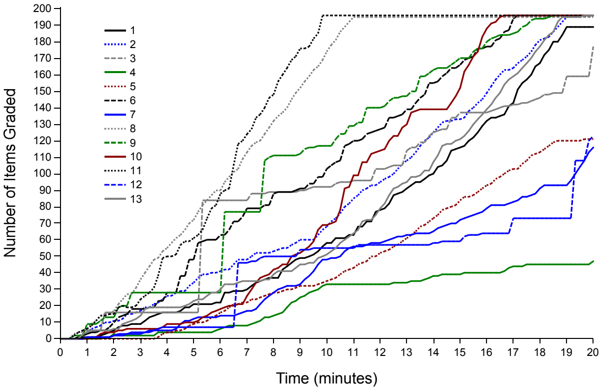


Figure 5: Answers graded vs. time for Q4, flat interface

In order to quantitatively compare the efficiency of the two interfaces, we sought an intuitive measure of speed that reflects the differences we observed. However, the choice of a metric is not straightforward: with the clustered interface, users both assigned grades rapidly using clusters and drilled down more slowly to obtain more accurate grades where needed. These two activities were unpredictably interleaved, and participants allocated their 20 minutes in different ways.

We chose to focus on comparing the speed at which participants could assign an initial grade to all the answers they would eventually grade, even if they later corrected some of those grades. While such corrections complicate the interpretation, this measure does represent a maximal rate of how quickly users could process answers, helping us extrapolate to larger datasets. To compute our speed metric, we determine how many answers the participant ultimately graded, and then divide this by the earliest time at which the participant reached that level of completion. Looking at the curves, this is the maximum value divided by the first time at which this value is reached; the resulting quantity has units of answers graded per minute. Using this measure of speed, the participants had an average speed of 11±3.9 answers/min using the flat interface and 33±40 answers/min using the clustered interface. This was a statistically significant difference (paired t-test, df=24, t=2.92, p<.007). No significant accuracy differences were found between questions (4 vs. 6) or order (first trial vs. second).

While this gives an overall sense of the average improvement in speed, the relative gain from the clustered interface changes over time as it is used. We calculated the amount of gain provided by the clustered interface *at each point in time*, averaged across participants. To do this, we combined the individual curves (such as in Figure 4) to create average progress curves for each interface; we then normalized by the total number of answers for each question, and plotted the difference between the average *clustered* and *flat* curves. The result is shown in Figure 6: at each time point, the middle curve reflects the fraction of all answers that users of the clustered interface had graded *beyond* what users of the flat interface had graded, at that point in time. For example, at about 7 minutes, an additional 45% of the answers (about 90 answers) had been graded in the clustered condition on average. The greatest gains occurred early in the task; by the end, users in the flat condition caught up, as the clustered participants had already run out of new answers to grade.

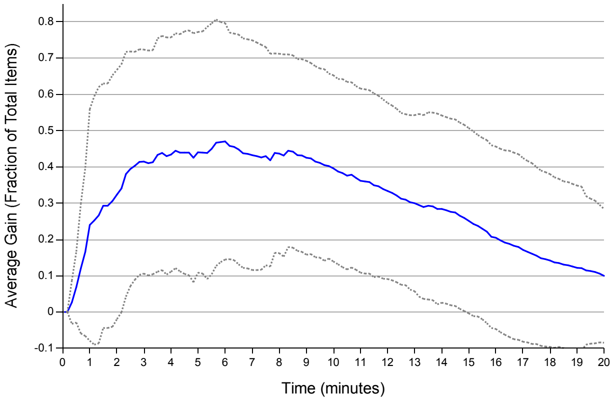


Figure 6: Average gain (± one std. dev.) over time, in fraction of answers graded, with the clustered vs. the flat interface.

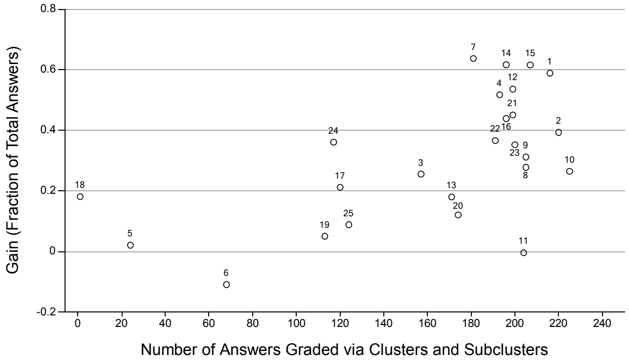


Figure 7: Gain for each participant in the clustered condition (over flat) vs. the number of items marked using a cluster or subcluster-level action.

Finally, we were curious to understand where these gains were coming from – were participants grading so many more answers because they were using the clusters and subclusters, or was it due to some other aspect of the interface? To study this, we computed, for each participant, the *gain* between their *clustered* trial and their *flat* trial, averaged over time (i.e., the time-average of Figure 6, but specific to a participant). We compared this to the number of answers graded *using actions at the cluster or subcluster level* (Figure 7), revealing a trend that increased use of high-level actions was associated with greater gains.

Grading Quality

Though the gains in speed using the clustered interface are clear, we were concerned that these speeds might be at the expense of grading accuracy. We measured grading quality by comparing grades against a gold standard, and also asked participants about their perceived consistency and fairness.

In order to test accuracy, we first needed an independent standard that we could measure both conditions against. Although grading is an individualized process and there is often no absolute correct grade for any answer, we chose to use as a gold standard the subset of items in which the three graders from the Powergrading Corpus [2] had perfect agreement. This corresponded to 167 of the 196 items for Q4 and 160 out of the 205 for Q6; the percent of items judged *Correct* (vs. *Incorrect*) were 53% and 67% respectively.

We then measured the accuracy of our participants' grades with respect to this gold standard, only counting those items that were marked (i.e., accuracy was not penalized for not completing the task). There was not a statistically significant difference in accuracy between clustered vs. flat conditions either for both questions together (92% vs. 90%, respectively) or either question individually (95% vs. 91% for Question 4, 88% vs. 90% for Question 6). In Figure 8, we show the accuracy for each participant/condition against speed. Contrary to our fears, it appears that the greater speed of the clustered interface does not hurt accuracy. This is remarkable given that the fastest *clustered* participant (8) graded at the rate of 176 answers/minute, compared to the fastest *flat* participant (11), at only 20 answers/minute.

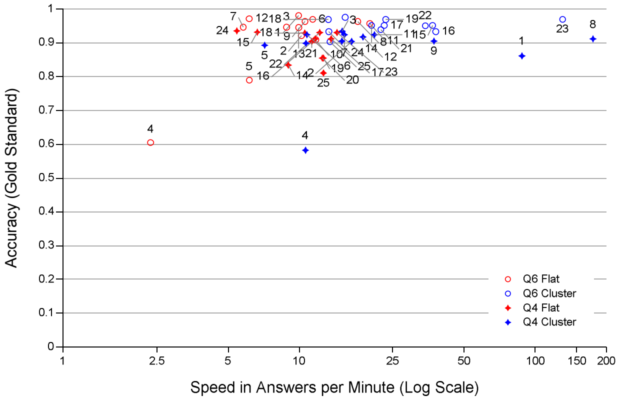


Figure 8: Accuracy vs. speed (items per minute) for all participants in all conditions. Note that speed is shown on a log scale.

Beyond the quantitative measures, we also wanted to know how participants felt about the quality of grading they were able to achieve. We analyzed the Post-Task questions about how well the interface supported grading *consistency* and *fairness* using a Wilcoxon Signed Ranks test, finding that significantly more participants rated the *clustered* interface higher than *flat* for support of *consistent* grading (Z=-2.3, p<0.022). There was no significant difference for *fairness*.

Feedback

In addition to grading efficiency and quality, we also assessed how well each interface supported graders giving students feedback about their answers. We analyzed the amount of feedback that participants gave, the feedback itself, and participants' comments about giving feedback.

Qualitative analysis of the feedback messages showed that most participants took their task seriously and wrote helpful, clear messages for students, but there were differing levels of effort, ranging from copy-pasting from the answer key to writing longer messages incorporating outside knowledge. Strategies differed as well: some asked students questions in their feedback messages, while others offered specific suggestions for improvement. Below we present a few selected examples of feedback messages, for illustration:

> *State legislators do not make national laws. The congressional members from the States do, but not the states themselves. (P7, Question 6)*

> *Gave you partial credit, as "market economy" is correct. "Free" market is too specific. (P14, Question 4)*

> *Be clear with your response. what role if any does the Supreme Court play? (P24, Question 6)*

Participants typed a median of 3 different feedback messages (ranging from 0 to 17) over the course of each grading task, which did not differ significantly between the two interfaces. Each of these distinct messages could be applied more than once, to clusters, subclusters, or individual answers, but this also did not differ significantly between the two interfaces. Participants attached feedback to a median of 11 objects with the clustered interface, vs. 18 objects with the flat interface.

However, using the clustered interface, more answers actually *received* feedback (median of 75 answers) than with the flat interface (median 18), because this feedback was often applied at the cluster or subcluster level and inherited by many answers. Because of non-normal distributions, we used a Wilcoxon Signed Ranks test to confirm that this difference is significant (Z=-3.23, p<0.001). In summary, while exerting the same or slightly less effort to create and attach feedback messages with the clustered interface, more than three times as many distinct answers received feedback.

Comments in the Post-Study Questionnaire also indicate that the clustered interface made giving feedback easier. Participants said that the efficiency of grading with the clustered interface let them think more carefully about 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)*

The Post-Task Questionnaire asked about how satisfied participants were with the amount and usefulness of feedback they were able to give. Wilcoxon Signed Ranks tests showed that significantly more people were more satisfied with the amount of feedback they gave using the *clustered* interface than with the *flat* interface (Z=-2.4, p<0.018). There was no significant difference in reported usefulness of feedback.

Although participants spent no more effort providing feedback, clustering effectively amplified their efforts to impact more students, giving them more time to give good feedback.

Reflection

One benefit teachers can derive from assessment and grading is the opportunity to reflect and learn about student knowledge and misunderstandings; we asked participants to note patterns and trends in the students' answers and to comment on how the students did on the question. Across both interfaces, most of these reflections were substantive, noting several patterns or observations, many of which would have been informative to a teacher. For example, P12, after using the clustered interface, noted several misunderstandings:

> *The vast majority had the response being sought. A fairly large subgroup of students correctly identified that the legislative branch and executive branch act in concert with the President either vetoing a bill or signing the bill into law. A smaller, but still significant subgroup was clearly confused regarding the branches of government, thinking that the judiciary actually creates the laws instead of interpreting the laws. (P12)*

A content analysis of these reflections, where we coded for attributes such as type and number of patterns detected, tone, and length, showed some noticeable differences between the two questions that were graded, but did not reveal any significant difference between the two interfaces. It appears that the participants were able to produce equally detailed and insightful reflections using either interface. Although we did not detect a difference in the reflections, other comments indicate that the clustered interface made finding patterns and misconceptions easier:

> *This format [clustered interface] made it easy to see patterns in student thinking, whether correct patterns or errors. (P12)*

> *This interface does make answer trends more easily identifiable. (P6)*

Note that we did not prompt participants with any questions that directly asked about this aspect of the system.

> *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)*

As this comment summarizes, the clustered interface allowed participants to get more answers graded quickly, to give feedback to more students, and to extract insights which could be used to inform teaching.

DISCUSSION AND FUTURE WORK

Our goal in introducing a new approach and interface for grading short answer questions was to improve scalability and efficiency while allowing teachers to achieve high quality grades, give feedback to many students, and reflect on student understanding to inform their teaching. From our quantitative and qualitative results, it seems our clustered interface was successful in all three areas, and substantially outperformed the flat baseline. The comments from teachers in our study confirmed these benefits and demonstrate the usability and efficiency of the clustered interface.

There are still a variety of questions to consider about this approach. First, while we have shown that the interface is fast, is it fast enough? In our study, participants had 20 minutes to grade 698 student response (collapsed to around 200 distinct answers). For a MOOC with 10,000 students (14 times larger), while the number of distinct answers would probably grow sublinearly, we might still expect around 2,000 distinct answers. If graders took the entire 20 minutes to grade 200 answers, it could take over three hours to grade a class of 10,000. However, the fastest users of the clustered interface in our study assigned their first-pass grades at a rate of 100–200 answers per minute, leaving time to go in at the more detailed levels to check and improve their work. Assuming this strategy, the first pass over 10,000 students could take only 10 to 20 minutes (teachers could still identify trends in answers and give quality feedback during this time). The teacher could then check and improve grades as carefully as time permitted, reflecting the flexibility of this approach. However, further studies with larger data sets are needed to see if these extrapolations hold. Different types of questions could lead to changes in the relative gains of the method – while previous results show similar performance over 10 questions of varying scope [2], we expect to encounter a far wider range of response distributions in practice; we hope to explore this range in future work via a wider deployment to real courses and classrooms.

Opportunities also exist to further refine and improve the clustered interface. In previous work, using the answer key and the learned distance metric to "autograde" answers improved efficiency [2]. However, this might increase the risk of teachers missing out on insights and ignoring autograded clusters that require closer inspection. In fact, one participant mentioned that the interface "should emphasize a little more that the individual responses must be viewed, as there are variations within each category" (P14). This points to another avenue for expansion, that of allowing the teachers to refine the clustering with their own judgments and letting the algorithm improve its results based on their changes. On a related note, the work of one teacher grading a set of answers could be leveraged for future classrooms or graders. Finally, there is the opportunity to explore more sophisticated text visualization and cluster summarization methods which could allow teachers to better allocate their attention. We hope our further developments, as well as contributions from others, can address many of these questions in future work.

REFERENCES

1. Anderson, R. and Biddle, W. On asking people questions about what they are reading. *Psychology of learning and motivation 9*, (1975).
2. Basu, S., Jacobs, C., and Vanderwende, L. Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading. *TACL 1*, (2013), 391–402.
3. Bloom, B. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Edu. Researcher 13*, 6 (1984), 4–16.
4. Brookhart, S. Teachers' grading: Practice and theory. *Applied Measurement in Edu.*, (1994).
5. Cross, L. and Frary, R. Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Edu.*, October 2013 (1999), 37–41.
6. Hearst, M. The debate on automated essay grading. *Intelligent Systems and their Applications*, (2000).
7. Heywood, J. *Assessment in higher education: Student learning, teaching, programmes and institutions*. 2000.
8. Jordan, S. and Mitchell, T. e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback. *British Journal of Edu. Tech. 40*, 2 (2009), 371–385.
9. Karpicke, J.D. and Roediger, H.L. The critical importance of retrieval for learning. *Science 319*, 5865 (2008), 966–8.
10. Markoff, J. Essay-Grading Software Offers Professors a Break. *The New York Times*, 2013.
11. McMillan, J. Secondary teachers' classroom assessment and grading practices. *Edu. Measurement: Issues and Practice*, (2001).
12. Mohler, M.A.G., Bunescu, R., and Mihalcea, R. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. *Proc. ACL*, (2011).
13. Mory, E. Feedback research revisited. In D.J. Mahwah, ed., *Handbook of Research on Educational Communications and Technology*. 2004, 745–784.
14. Perelman, L. *Critique (Ver. 3.4) of Mark D. Shermis & Ben Hammer, "Contrasting State-of-the-Art Automated Scoring of Essays: Analysis."*2013.
15. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. Tuned Models of Peer Assessment in MOOCs. *Proc. EDM*, (2013).
16. Poulos, A. and Mahony, M.J. Effectiveness of feedback: the students' perspective. *Assessment & Evaluation in Higher Edu. 33*, 2 (2008), 143–154.
17. Reily, K., Finnerty, P.L., and Terveen, L. Two peers are better than one: aggregating peer reviews for computing assignments is surprisingly accurate. *Proc. GROUP*, ACM Press (2009), 115.
18. Sadler, P. and Good, E. The Impact of Self- and Peer-Grading on Student Learning. *Edu. Assessment 11*, 1 (2006), 1–31.
19. Scriven, M. The methodology of evaluation. In R.E. Stake, ed., *AERSA Monograph Series on Curriculum Evaluation*. Rand McNally, Chicago, 1967.
20. Thorpe, M. Assessment and 'third generation' distance education. *Distance Edu. 19*, 2 (1998), 265–286.
21. Weld, D., Adar, E., and Chilton, L. Personalized Online Education—A Crowdsourcing Challenge. *Proc. AAAI, workshop on Human Computation*, (2012), 159–163.