Scaling Short-answer Grading by Combining Peer Assessment with Algorithmic Scoring

Chinmay Kulkarni, Richard Socher
Michael S. Bernstein
Stanford University
Stanford, CA 94305-9035
{chinmay,socherr,msb}@cs.stanford.edu

Scott R. Klemmer
University of California, San Diego
La Jolla, CA 92093-0440
srk@ucsd.edu

ABSTRACT

Peer assessment helps students reflect and exposes them to different ideas. It scales assessment and allows large online classes to use open-ended assignments. However, it requires students to spend significant time grading. How can we lower this grading burden while maintaining quality? This paper integrates peer and machine grading to preserve the robustness of peer assessment and lower grading burden. In the identify-verify pattern, a grading algorithm first predicts a student grade and estimates confidence, which is used to estimate the number of peer raters required. Peers then identify key features of the answer using a rubric. Finally, other peers verify whether these feature labels were accurately applied. This pattern adjusts the number of peers that evaluate an answer based on algorithmic confidence and peer agreement. We evaluated this pattern with 1370 students in a large, online design class. With only 54% of the student grading time, the identify-verify pattern yields 80-90% of the accuracy obtained by taking the median of three peer scores, and provides more detailed feedback. A second experiment found that verification dramatically improves accuracy with more raters, with a 20% gain over the peer-median with four raters. However, verification also leads to lower initial trust in the grading system. The identify-verify pattern provides an example of how peer work and machine learning can combine to improve the learning experience.

Author Keywords

assessment; online learning; automated assessment; peer learning

INTRODUCTION

Short answer questions are a powerful assessment mechanism. Many real-world problems are open-ended and require students to generate and communicate their response. Consequently, short-answer questions can target learning goals more effectively than multiple choice; instructors find them easier to construct; and short answers are relatively immune to test-taking shortcuts like eliminating improbable answers [13].

Many online classes could adopt short-answer questions, especially when their in-person counterparts already use them. However, staff grading of textual answers simply doesn't scale to massive classes. In our experience, grading each answer takes approximately a minute. Grading a hundred students is feasible, taking two hours per question. For an online class of 5,000 students this involves two person-weeks of grading per question. Automated grading and peer assessment both offer ways to scale assessment [17, 29], but in isolation, both introduce an unsatisfactory tradeoff.

While algorithmic grading consistently applies criteria to all student work [29], it has many shortcomings. It frequently relies on textual features [28], rather than semantic understanding. For instance, automated essay scoring software uses counts of bigrams and trigrams (sequences of two or three words) [8]; NLP techniques like syntactic parsing [5]; dimension reduction techniques such as PCA [10]; or a combination of these features [7]. This reliance on textual features reflects algorithms' limited ability to capture the semantic meaning of student work. This limited understanding can cause grading errors because answers using unconventional phrasing may be penalized. Furthermore, students may game algorithms with answers that match patterns, but are otherwise incorrect [26]. This has, in turn, led to public skepticism about algorithmic grading [1].

Algorithmic grading for short answers is especially challenging, because the limited text provides fewer lexical features. Algorithms can still use features like word overlap, but accuracy suffers [14].

In contrast, peers can more robustly handle ambiguity and differences in phrasing, and students learn by assessing others' work. However, peer assessment requires students to spend time grading several (e.g., five) peers. Student raters need training, and still may differ in how they apply grading criteria, and ratings may drift over time [29]. Raters also suffer from systematic cognitive biases including the Halo Effect (wrongly generalizing opinions on one characteristic to the entire answer), stereotyping (e.g. gendered/nationalistic cues affect grading [17]), or perception differences (grading of prior answers affects grading of the current answer) [29].

L@S 2014, Mar 04-05 2014, Atlanta, GA, USA
ACM 978-1-4503-2669-8/14/03.
http://dx.doi.org/10.1145/2556325.2566238

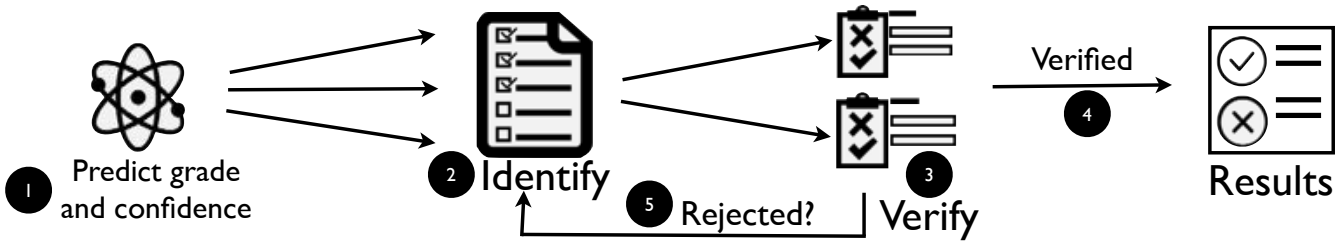

Figure 1. Overview of the assessment process. (1) Machine learning algorithm predicts grades and confidence. Number of independent identifications decided based on confidence (2) Peers identify attributes in answer using rubric (3) Two other peers verify existence of attributes. Final score is sum of verified attributes (5) if attributes are rejected, one more rater is asked to Identify. If two independent identifications are identical amongst raters, one is considered a verification (4).

Could machine-learning algorithms mitigate grader biases and minimize human effort? Crowdsourcing algorithms can correct inter-rater differences [22], and recruit more raters when they encounter unreliable raters [16, 21]. Inspired by these successes, this paper introduces a workflow that intelligently combines algorithmic and peer assessment to provide the benefits of both, while mitigating their individual drawbacks.

The *identify-verify* workflow uses algorithmic grading to estimate how many independent peer assessments are needed. The algorithm estimates "ambiguity" of the answer using its prediction confidence. More raters are assigned to highly ambiguous answers and fewer to less ambiguous ones. In this paper, the range was 1 to 3 raters. Peers then identify key features of the answer using a staff-provided rubric. Other peers verify whether these feature labels were accurate. Few peers are needed when initial human ratings agree with a high-confidence machine rating. The algorithm seeks more assessments when raters disagree. The algorithm automatically seeks higher quality assessment if more raters are available.

An experiment compared hybrid grading with peer grading; 1370 students from an online human-computer interaction class participated. Compared to a baseline of aggregating independent peer ratings using a median, integrating machine grading yields comparable accuracy with lower effort. For binary questions, using the machine grading with identify (and no verify step) yields 83% of the peer-median accuracy, and only needs 54% of human effort. For an enumerative short-answer question, 70% of the effort yields 80% accuracy. For both types, adding verification yields higher accuracy and more reliable information about the answers' attributes, but increases human effort. A follow-up experiment investigated how identify-verify works with a varying number of graders, compared to the baseline of median of peer grades. Adding the verify step yielded a 20% gain in accuracy over the peer-median method with four raters.

In addition to saving time, this hybrid also provides students richer, structured feedback about their answers in addition to their scores. Students see both a list of features of the answer they got right, and common errors they made.

This paper makes two contributions. First, it introduces the identify-verify pattern for combining peer and machine grading. Second, it presents experimental results demonstrating the accuracy benefits and the tradeoffs in human effort of the identify-verify pattern in various configurations.

CLASS SETUP

We evaluated the identify-verify approach in a large, online class introducing human-computer interaction. This class is based on an in-person class that uses short-answer questions to assess if students students' knowledge. For instance, short answers assess if students can construct well-formed interview questions, if they understand prototyping strategies, and can explain differences between experimental designs. The system described in this paper introduced these short-answer questions to the online class. Students answer short answer questions on two quizzes, one in Week 3 of the class, and once on the final (Week 9).

PILOT: LENIENT PEERS, STRICT MACHINES

We piloted short-answer questions in the May 2013 offering of the class. The pilot explored whether simply combining peer and machine scores using a median yielded accurate results. In addition, it aimed to understand the relative merits of machine and peer grading.

Three independent peer raters scored each student answer. The site provided raters with a grading rubric and staff-graded examples to calibrate themselves (similar to Calibrated Peer Review [6]). After grading a staff-provided example, students assessed peer answers. A machine classifier reliant on textual features scored all answers as well. The system combined human and machine scores by taking the median of all four scores. Other methods of combining grades, such as linear regression, were sensitive to outliers.

To assess accuracy, we compared the median grade to the staff grade for 200 submissions. We found that accuracy increased with increasing number of peer raters, consistent with prior work [12, 17]. In addition, we made the following observations:

- **Peers were more lenient than staff, and writing fluency swayed judgments on correctness**: Peers sometimes awarded points to plausible-sounding but incorrect answers. For instance: "Rewrite the interview question 'Do you like the WordArt feature from Microsoft Word?' to address problems with it". The problems with the interview question are that it is leading and it assumes users

have an opinion on the feature. One incorrect student answer was "With respect to your experience, how much do you like the WordArt feature, on a scale of 1-5?" Three peer raters marked this as correct, even though it has the same problems as the original question. We also found that cues such as how confidently the answer was written, or whether it used fluent language seemed to affect the peer's rating. Prior work has shown similar Halo effects influence human grading more generally [29].

- **Peers understand ambiguous answers better**: For example, for the same WordArt question, machine grading marked the correct answer "How do you add images or text in different styles into your documents in Microsoft Office?" as incorrect (possibly because training examples had few correct answers without the word WordArt). However, two of three peer raters marked it to be correct.

  Together, these two factors meant algorithmic grading was stricter, since it only awarded credit when the answer matched example answers closely (the average machine grade was 16% lower than staff). Peer grading was more lenient than staff: the average peer grade was 14% higher than staff.

- **High-confidence predictions from machine grading were generally accurate, and agreed with peer assessment.** For binary questions, when the algorithm reported confidence larger than 80%, staff and machine grades matched 85% of the time (staff and a single peer agreed 78% of the time). In addition, for low-confidence predictions, staff/machine disagreement was larger than staff/median-peer disagreement. (When confidence was 50-60%, staff and machine grades agreed 53% of the time. For these same submissions, a single peer agreed with staff grade 52% of the time, but the median of three raters agreed with staff 68% of the time.) Therefore, low-confidence predictions are somewhat informative, but cannot be trusted reliably.

This pilot suggests that few peers are needed for answers graded with high algorithmic confidence, but more peers may be necessary for assessing questions with low confidence. However, a simple median for combining human grades and machine grades cannot handle machine grades are not uniformly reliable. This suggests that a grade-combination scheme should tune the number of raters based on algorithmic confidence. Essay scoring on standardized tests uses one such scheme: the GMAT compares a human essay score with the machine score, and recruits more human raters if the scores differ [4].

Combination schemes could also leverage peers' ability to understand ambiguous answers, but should account for them being biased and lenient. Prior work suggests it is possible to create processes that mitigate cognitive biases [19, 15], but simply alerting students to their biases does not help mitigate them [25]. Therefore, this paper seeks to create a workflow and interface to mitigate biases and improve accuracy.

THE IDENTIFY/VERIFY ARCHITECTURE

Based on these pilot insights, we designed a grading system to combine the strengths of human and machine grading. This system seeks to minimize human effort while still retaining current accuracy. We choose to reduce human effort, rather than improve accuracy, because many large, online classes (including our evaluation class) are pass-fail, and we found accuracy from the pilot (between 67% and 82%) reasonable. At this accuracy, we estimate the number of students who should have passed but didn't due to grading errors to be less than 3%. This paper leverages the insight that partitioning tasks so people can audit each other improves quality and efficiency [2, 18].

Identify-verify comprises three steps (Figure 1). First, a machine-learning algorithm predicts a grade and confidence score for each submission. The system assigns a number of peers is assigned to grade the answer based on the confidence score. Second, peers use a grading rubric to *identify* which features the answer contains (Figure 2). Third, they *verify* other peers' feature identification for other answers (Figure 3). Identify-verify assigns a final grade by combining the grade for verified features in the answer; our prototype uses the sum of feature grades. For instance, if a student submission is identified to have two features each worth one point, the submission is awarded two points, the sum of feature scores. Below, we describe each step in the assessment process.

Step 1: Algorithm estimates grade and number of raters

Before peer assessment begins, a machine-learning algorithm predicts the grade for each answer. We built a generic text classifier using etcml.com with the predicted grade as the output. This classifier uses textual features such as word, bigram and trigram counts, length of answers, and letter n-grams (to capture use of word fragments like "creati-", which match "creativity", "creative", "creation" etc.).

Teaching assistants provided numeric scores and correct/incorrect attributes for about 500 student responses per question. The numeric grades were used as labels to train the classifier. Instructors provided teaching assistants an initial rubric for grading. TAs then expanded this rubric with correct/incorrect attributes they identified, and added example student answers with those attributes. Future work could bootstrap attributes and examples using prominent features from the trained classifier.

The system then uses the classifier trained on staff-graded answers to grade all answers. The classifier outputs the most likely grade (the prediction), as well as the probabilities of all possible grades (e.g., an answer may have a grade of 1 with probability of 0.2, and a grade of 0 with probability 0.8). For the rest of the grading process, we use the probability of the most likely grade (in our example 0.8) as the algorithm's confidence in the grade. (Future work could consider using other statistics).

The algorithm's confidence determines the initial number of peer raters assigned to each answer. The intuition behind this is that confidence represents a measure of ambiguity—

answers with high confidence are usually those that are clearly right or wrong. Conversely, ambiguous answers often have low confidence, and therefore should have more independent human assessments. We require answers with high confidence ($> 90\%$) to have a single rater, those with medium confidence (75%-90%) required two, and all other answers required three raters. Overall, 34% of student submissions had grades predicted with $> 80\%$ confidence, and 16% of submissions had grades predicted with $> 90\%$ confidence.

This paper seeks to demonstrate the feasibility of combining human and machine grading. It does not determine the most suited machine-grading algorithm. Therefore, while our classifier represents the state-of-the-art in text classification, it does not use any special logic for answer grading. We hope that demonstrating feasibility with a generic classifier will also inspire other researchers to create better ones.

Step 2: Peers identify answer attributes

In this step, randomly-chosen peers independently identify correct/incorrect attributes in student answers. Raters select these attributes from the expanded grading rubric from Step 1 (Figure 2). Staff associated a score with the presence of each attribute, which could be negative.

To minimize the impact of too-few ratings, the system solicits ratings in order of greatest need. Specifically, the system finds the student answer that has the largest number of required assessments, with the fewest completed. Ties are broken randomly.

The grading page displays this answer along with the grading rubric. Peer raters mark each attribute present by clicking a checkbox next to it. To encourage students to be critical (and reduce the leniency we saw in our pilot), the grading rubric is initially shown with incorrect attributes displayed, and correct attributes collapsed (Figure 2). Raters expand the correct attribute section by clicking the drop-down arrow.

Raters are asked to identify attributes in four student submissions. After a rater completes identification, the answer and its attributes are queued for verification. If two identifiers independently select the same attribute, that also constitutes verification. Such answers skip the separate verify step.

Even with high-confidence machine predictions, it is important that student grades do not suffer due to an over-optimistic algorithm. The current system requests one additional identification for high-confidence answers where the peer and algorithm grades differ by one or more points. (In this paper, answers are worth up to 3 points, and only whole point values are awarded.)

Step 3: Other peers verify attributes correctly identified

Now, independent raters verify attributes identified in the previous step by other peers. This interface groups answers according to the identified attribute, e.g. grouping all answers marked as "More sharing of features between designs" (Figure 3). Peers then verify whether answers contain the marked attribute. We hypothesize that grouping submission marked with the same attribute increases accuracy because verifiers are presented with a group of nominally-similar responses for comparison.

When two raters independently verify an identified attribute, the system marks the attribute as verified and removes it from the verification pool. If two raters reject an identified attribute, the system returns the submission to the identify pool for one additional identifier, since the initial identification was inaccurate.

Similar to the identification step, the system presents submissions to verifiers in decreasing order of the number completed, and breaks ties randomly. This again provides every submission with some data quickly. This algorithm also needs at most three verifications: after three, each attribute will either have been verified, or rejected.

Optimizing the number of raters

Identify-verify reduces the grading workload by recruiting fewer raters when the grading algorithm reports high confidence. This scheme is also cautious. First, we increment the number of identifications required for high-confidence predictions if peers disagree with the predicted grade. Second, identified attributes for an answer that are rejected may indicate the answer was difficult to grade, so we request additional assessments.

Display results and feedback

A student's final score is the sum of scores of all verified attributes, clamped to the minimum and maximum score for the question. Students see their score along with the features

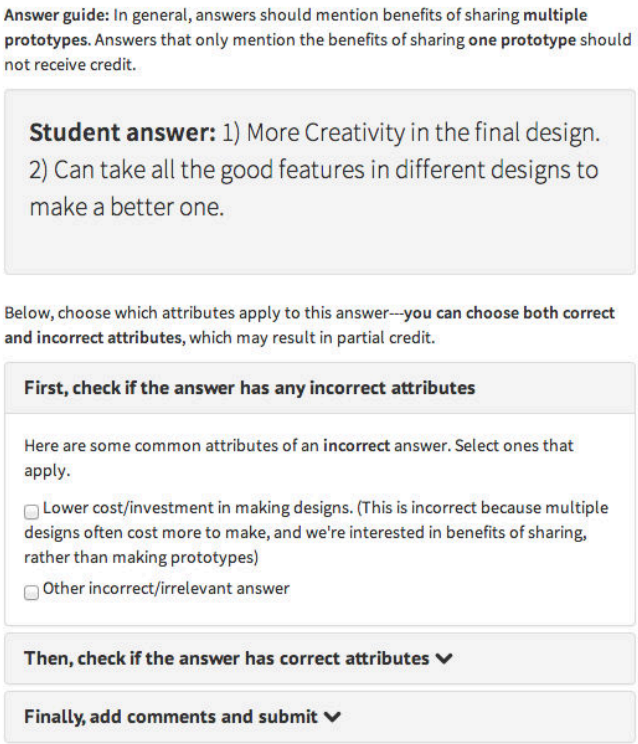

Figure 2. Identify UI: Students identified whether student answers had staff-provided features (which indicated right/wrong answers)

that peers identified, and correct attributes that their answer missed (Figure 4). Thus, students receive more than a grade: they receive detailed information about what they did well and poorly.

EVALUATION

Identify-verify seeks comparable accuracy to using the median grade of independent peers, but with less human effort. Our comparison baseline asks three peers to grade a student answer.

Experiment 1: Does identify-verify yield accurate grades?

This controlled experiment explored two questions: First, does identify-verify grade accurately and lower effort? Second, does identify-verify reduce leniency from our pilot? (We hypothesize that leniency is due to the Halo effect, and using a structured process and interface would reduce this bias [15].)

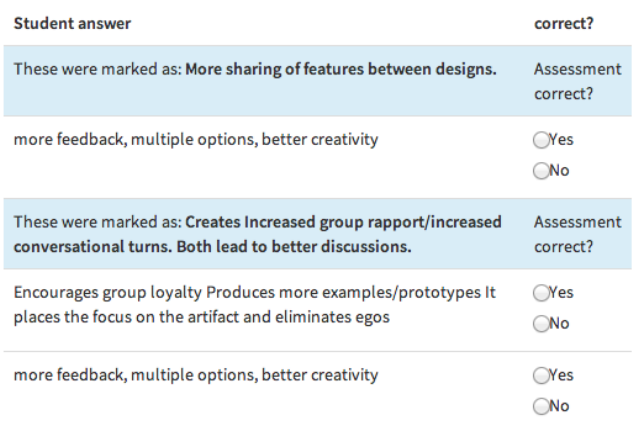

Figure 3. Verify UI: Students verified if other peers had assessed answers correctly.

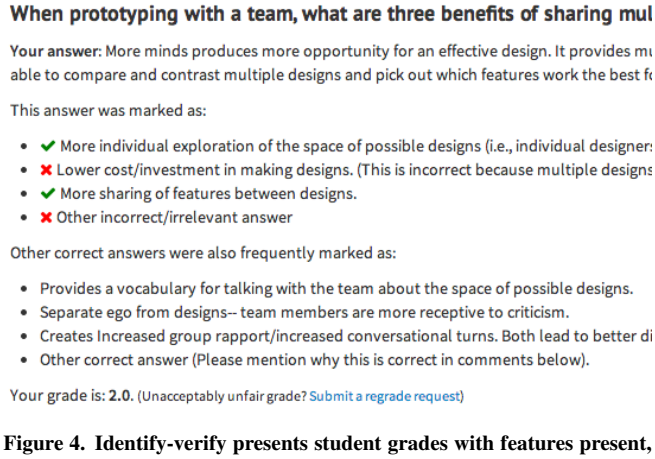

Figure 4. Identify-verify presents student grades with features present, and those missing in answers

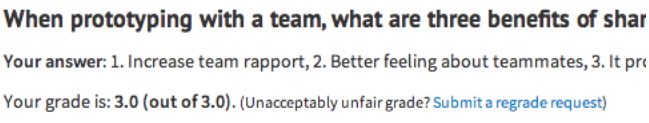

Figure 5. Student grade display in baseline condition (Grades are computed using Identify-verify, but detailed feedback is hidden.)

Conditions

This between-subjects experiment had three conditions. In the *peer-median* condition, students assess four peers using a grading rubric, and enter their grade into a text field (Figure 6). In the *identify-only* condition, students assess four peers using the same grading rubric, but would instead use the Identify interface to select which aspects of the rubric were present in the student answer (Figure 2). In the *identify-verify* condition, students assessed four peers using the Identify interface. Then, they would verify assessments of eight answers that other students had created in the Identify step (Figure 3).

We wanted to reduce grading burden in the class, and since we hypothesize that Identify-verify would save student effort, the experiment used an unbalanced assignment; 20% of students randomly assigned to the *peer-median* condition, and the rest split evenly between *identify* and *identify-verify*.

Questions

Students assessed answers to two short-answer questions. Question 1 asked students to rewrite an interview question: "Rewrite the following interview question to address its problems: 'Do you like the Word Art feature of Microsoft Office?"', and had a binary grade (credit or no-credit). Question 2 asked students to enumerate "three benefits of sharing multiple designs with your team members, instead of sharing only one design?". Students could earn 0-3 points on this question, one per enumerated benefit. Students assessed four submissions per question, so there were a total of eight assessments per participant.

After they had completed grading, the system invited students to participate in a short survey. The survey measured trust in the system, and time taken for grading vis-a-vis their initial expectations.

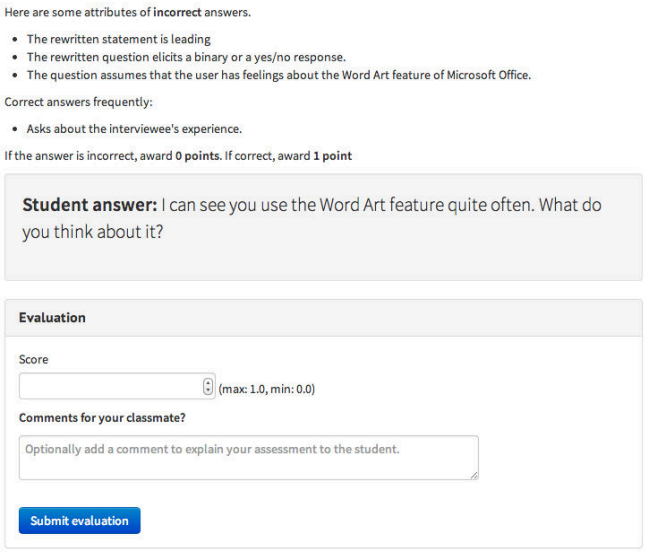

Figure 6. Peer-median UI: Students entered grades in a text box.

Table 1. Peer-median was faster for each rating, but employed more raters, so took more time overall. Overall, the time each condition took and its quality correlated.

Type	Method	#assessments: Median (mean)	Accuracy	κ	Human effort (seconds)
Binary	Peer-median	3	0.85	0.57	109
Binary	Peer-median	2	0.68	0.21	73
Binary	Identify only	1 (1.15)	0.71	0.41	59
Binary	Identify-verify	1 (1.15) + 2 (2.08) verifications	0.72	0.41	91
Binary	Machine prediction	–	0.60	0.19	–
Enum.	Peer-median	3	0.49	0.32	103
Enum.	Peer-median	2	0.33	0.19	68
Enum.	Identify only	1 (1.42)	0.39	0.15	71
Enum.	Identify-verify	1 (1.42) + 3 (3.1) verifications	0.45	0.22	104
Enum.	Machine prediction	–	0.28	0.09	–

The system showed students their final grades a day after the peer assessment period ended. All students saw grades computed using Identify-verify. To measure the effects of detailed feedback, the system showed those in the peer-median condition only the final score (Figure 5), students in other conditions saw both the score and identified attributes (Figure 4). After they saw results, we invited students to a second survey, which gauged how accurate they perceived grading to be and how satisfied they were with feedback.

Participants 2,556 students submitted answers; 1,370 performed assessment (the others dropped the class). 620 students participated in the pre-results survey, and 102 participated in the post-results survey. In all, students created 11006 assessments and 12264 verifications.

Measures

For both the *peer-median* and the *identify-verify* strategies, course staff looked at 100 student answers for each question with three peer-median assessments, and 100 more answers with two peer-median assessments. (We did not select based on the number of identify assessments, because the system dynamically determined this number for each answer). For each student answer, we compared the staff grade to the computed grade.

Results

In terms of both effort and accuracy, the ranking of conditions was the same: *Peer-median* was highest, *identify-verify* was the middle, and *identify-only* least (See Table 1.) *Peer-median* had three raters. *Identify-only* had median one rater. *Identify-verify* had median one rater, with two verifiers for the binary question and three verifiers for the enumeration.

How accurate is identify-verify assessment?

Peer-median required disproportionately more effort than *identify-only* to achieve its results. *Identify-only* consumed 54% of the effort to achieve 83% of the accuracy in the binary question, and 71% of effort for 80% of accuracy in the enumeration question. Identify-verify consumed 84% of effort for 85% of accuracy in the binary question, and identical effort for 92% of accuracy for the enumeration question. This study only examined one effort level. The second study simulates multiple effort levels.

Verification provided a large benefit for the enumeration question, but minimal benefit for the 1-level question. Labels were rejected at similar rates (19.8% for 1-level and 18.6% for enumeration). For a binary question, not all attributes need to be identified to accurately grade it (for example, if the answer is wrong for two reasons, identifying just one is sufficient). Therefore, we hypothesize that the benefits of verification are larger for questions that are non-binary, and investigate this in Experiment 2.

Identify assessments take longer, more accurate

Students took significantly longer to select an attribute label than to select a score (see Figure 7), log-transformed $t(6789) = 28, p < 0.01$). Labeling also yielded more accurate work (see Table 1). Identify-verify reduced leniency, while retaining peers ability to assess unusual answers better than machines (see Table 2 and Table 3).

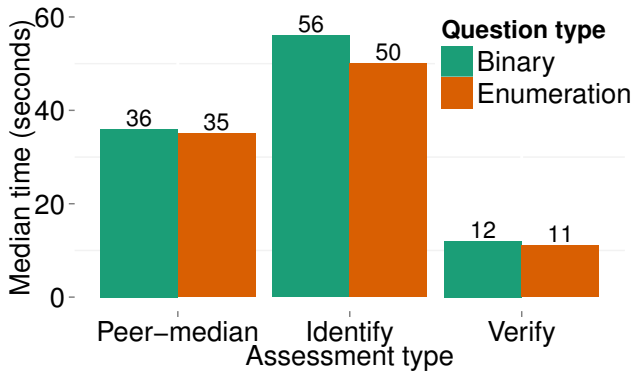

Figure 7. Assessment took longer using the Identify interface, but yields more accurate results.

Identify-verify reduces voluntary acceptance

Fewer students in the *identify-verify* condition reported wanting to continue using the grading interface for other quizzes

Question	Peer-median 3 raters	Identify-verify	Staff	Machine
Yes/no (1 point)	0.57	0.33	0.31	0.17
Enumeration (3 points)	2.17	1.65	1.74	1.35

Table 2. Peer grade averages in points. Identify-verify reduces leniency compared with peer-median.

Table 3. Sampling of errors in assessment. Peer ratings help when machines are less confident of the grade.

Student answer	Remarks
"How do you use the Word Art feature and how does it help you to meet your goals?"	Machine marked as incorrect, possibly because of leading bigrams "does it", "help you". Peers marked as correct. Staff graded as correct.
"What do you think of the Word Art feature of Microsoft Office?"	Construction marked as incorrect in the grading rubric (because it assumes opinion); yet, two of three peers in the peer-median condition marked as correct (possibly because it's less leading than "do you like…"). Both machine, and identify peers marked as incorrect.
"What would you like to see changed in the 'Word Art' feature on Microsoft Office?"	Possibly useful interview question asks how to change, instead of understanding current use (and so, is wrong): 3 peers in the peer-median condition marked correct; one rater identified it as 'Other correct answer', but verification rejected it. Staff graded as incorrect.
"Inspiration. Innovation. Social" (for benefits of sharing prototypes)	Uses keywords without context. Machine awarded one point (possibly due to 'Inspiration'), but Identify peers did not (this answer had no peer-median assessments), nor did staff
"Because the best way to have a good idea is to have lots of ideas." (for benefits of sharing prototypes)	Pithy and plausible, but irrelevant. Awarded 1 point (out of 3) in peer-median evaluation, none in Identify. Staff graded at 0.

(64% said yes, $t(732) = 2.9, p < 0.01$); no significant differences existed between *peer-median* and *identify-only* (78% and 75% respectively). Usability challenges with the verify interface may have reduced interest. Some students reported that the "the layout was very confusing" others were initially unsure if they were verifying the student answer or the label. 15.8% of students in the *peer-median* condition completed more assessments than required, while 8% of students in the *identify-only* condition completed more than required.

Fewer students in the identify-verify condition believed the process would give them a fair grade (Asked as Yes/No: $\beta = 0.12, t(734) = 2.7, p < 0.05$). This may be because verify explicitly revealed individual peers work; reducing trust. One student said that based "on the verification step of the peer assessment I'm not confident that people's quizzes are being assessed correctly." Furthermore, *identify-only* students reported more accurate grades ($\mu = 1.9, t(93) = 2.04, p < 0.05$) than those in the *peer-median* or *identify-verify* conditions ($\mu = 2.5$, 4-point Likert scale with 1: 'very accurate').

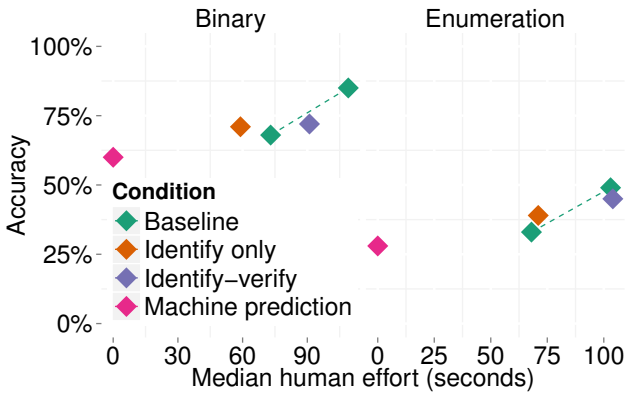

Figure 8. With median one rater, the peer-median method takes disproportionately more human effort to get high accuracy compared to *Identify-only* for the yes/no question. For the enumeration question, both methods need nearly the same effort for comparable accuracy.

Experiment 2: How number of raters affects accuracy

A second experiment investigated how the number of raters affects accuracy. As before, students were assigned to either the *identify-verify*, *identify-only* and the *peer-median* condition. All raters graded one of fifty randomly-selected submissions. 634 students participated.

The final had three enumeration questions asking students to a) mention one disadvantage of a between-subjects experimental design, b) list three ways of visually grouping related information, c) list two situations where heuristic evaluation is preferable to user testing. The experimental setup was identical to Experiment 1.

Measures

We performed a bootstrapped simulation of the peer assessment. This simulation chooses a random sample of raters for each question. We then calculate the final grade using ratings only from this sample of raters, and compare it with the staff-assigned grade. Repeating this process multiple times estimates peer agreement with staff [17]. Figure 9 shows median results from 20-repetition sampling, with one to eight raters.

We benchmark each condition against its peak accuracy: the highest accuracy seen in that condition in our simulation. More raters did not always improve accuracy, so peak accuracy was achieved with fewer than eight raters in the *identify-only* and *peer-median* conditions.

Results

A few raters identify most features

A small number of raters can identify most attributes present. Figure 9 shows that accuracy quickly plateaus, and four raters yield 92% of the peak accuracy with the *identify-only* method. Overall, the peak *identify-only* accuracy was 55% with six

raters, the *peer-median* had a peak accuracy of 66% with seven raters. This early saturation is similar to heuristic evaluation of interfaces [20], suggesting similar processes may be involved.

Identify raters satisfice, Identify-only errors accumulate

Identify-Only accuracy was lower than *peer-median*, and much lower than *Identify-Verify* (see Figure 9). First, most raters select only one attribute, even though the answer may match multiple attributes. Of the 1488 assessments collected, only 173 had more than one selected attribute. In contrast, staff assessments averaged 1.4 selected attributes. Second, because identifiers sometimes mislabel answers and there is no mechanism (i.e. verification) that catches this, asymptotically optimal performance is with relatively few raters and relatively low quality. In contrast, the peer-median approach uses the median of peer grades in the peer-median approach, so grades become more accurate with more raters as outlier ratings are discarded.

Many identifiers appear to have selected the first relevant label (Figure 10). Randomizing order across raters should mitigate ordering effects. Future work could investigate interfaces that incent raters to select all relevant labels.

Verification improves accuracy, especially with more raters

Identify-verify yielded the highest accuracy: the peak accuracy was 82% with six raters. The simulation required labels to have one peer verification and no peer rejections. (Actual student grading requires two verifications. Because the system solicits verifications in decreasing order of need, the median staff-graded submission had only one verification, or was rejected.)

Even single-peer verification dramatically increases accuracy. With three raters, accuracy is 28% higher than *identify-only*, and 18% higher than *peer-median*. Peer-median assessments took a median time of 19 second, identifications took 40s. Verification took 12s, similar to Experiment 1. Therefore, this 18% boost in accuracy comes with approximately two extra minutes of human effort per answer.

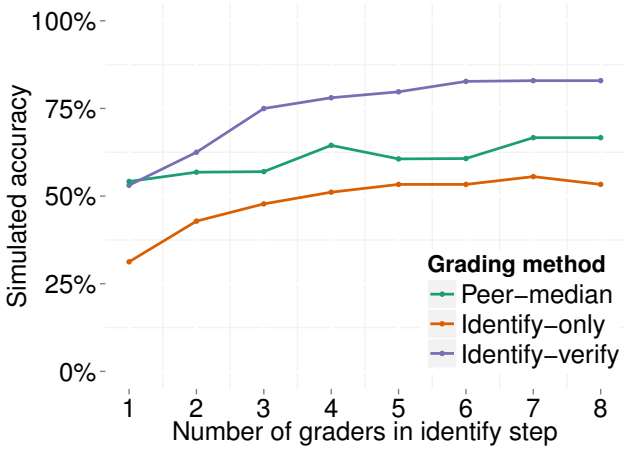

Figure 9. For enumeration questions, identify accuracy is lower than the peer-median method. Identify-verify obtains better accuracy than peer-median, especially with three or more raters.

Because verification filters out erroneous identifications, its benefit is larger with more raters: verification with one rater yields a 22% benefit in accuracy, with four raters, it yields a 27% benefit. In our simulation, three identifiers identified most attributes, and inaccuracies with three or more raters are due to wrongly identified attributes.

DISCUSSION

Identify-verify represents one choice in the trade-offs between human effort and grading accuracy. This choice was optimized for a large, pass-fail class.

Is verification necessary?

Our results demonstrate how erroneous identification can be detected with an easier operation (verification), similar to Soylent [2]. This is especially useful for questions where all attributes need to be correctly identified. While verification increases grading time, it yields more yields more descriptive, actionable, and accurate student feedback, which helps students learn.

Opportunities for early feedback

To explore the possibility of automatic, early feedback, we trained a classifier using etcml.com to detect the most common errors for each question (Table 4). Because students unlikely to revise work without external feedback [24], even somewhat unreliable feedback (e.g., "Check to see that…") may have benefits.

Identify-verify uses its auto-graders confidence to indicate ambiguity. Might students benefit from knowing that peers may have trouble understanding them? Evidence from automated essay scoring suggests that well-designed early feedback may help students write clearer answers [11, 23].

Coping with fewer graders than submitters

In Experiment 1, almost twice as many students submitted work as performed assessment; the rest dropped the class in the meanwhile. Experiment 2 was conducted later in the course, and a much larger fraction of the 850 students who submitted answers also assessed. Intelligently rationing raters is important in large online systems with voluntary participation. Identify-verify system handles this problem by rationing fewer graders for unambiguous answers. Because of the smaller number of raters, the system asked a median of

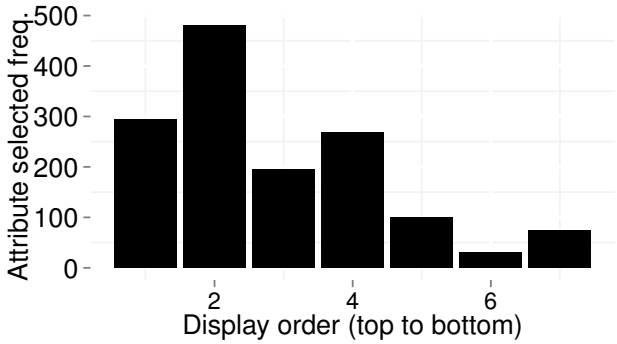

Figure 10. Raters were more likely to choose attributes displayed earlier on the page.

Table 4. Algorithmically predicting errors could automate early feedback.

Attribute	Accuracy	Precision	Recall
Incorrect attribute: "The question assumes that the user has feelings about the feature" (Q1)	0.79	0.58	0.41
Missed attribute: "More individual exploration in the space of designs" (Q2)	0.59	0.64	0.79
Incorrect attribute: "Other incorrect/irrelevant answer" (Q2)	0.90	0.27	0.73

only one identification per question, saving more identifications for the most ambiguous answers. For this experimental system, students were not penalized for not participating in assessment. Future work could explore penalties for non-participation, or incent assessment in other ways.

When should instructors use hybrid grading?

Peer assessment works best when staff spot-grade some student submissions because it helps staff refine assessment materials and baseline peer grades [3, 6]. However, courses may not have the resources for staff to grade several hundred examples that can train a machine-learning algorithm. (Even if it enables richer questions.) Furthermore, requiring large amounts of training data may dissuade instructors from revising questions. We see two opportunities exist for future work. First, an online-learning algorithm may improve prediction accuracy as students assess each other. However, because the system would demand fewer assessments as its prediction accuracy increases, this may encourage free-riding. Future work could leverage such algorithms, while balancing for fairness. More immediately, assessment data from peers may be used to train algorithms. For example, an advanced cohort takes the class a week ahead of the general class. There are many exciting opportunities for integrating peer and algorithmic assessment to increase student learning and leverage the rater's time better.

FUTURE WORK AND CONCLUSION

This paper demonstrated the feasibility of combining machine and peer grading through the identify-verify workflow. It showed how this workflow results in more detailed student feedback, and can be leveraged to provide early feedback. further instructor experimentation and research, our open-source code is available at `https://github.com/StanfordHCI/peerstudio`. In addition, a hosted version of the platform is available at `http://www.peerstudio.org`.

Future work falls in three categories: First, this paper assumes the final grade for a short-answer response can be expressed as a summed combination. Deploying this workflow in other classes may suggest other ways to structure assessment and verification, for e.g., as a decision tree. Second, many techniques in this paper may be extended with algorithmic improvements. For instance, our system currently implements a fixed-control method for dynamically controlling the number of peer raters for a submission. A decision-theoretic model may result in even lower grading burden [21]. Similarly, an online learning algorithm [27] could dynamically update estimates of the predicted grade to guide which ratings are collected. Third, in this paper, the system decided which answers a rater should assess and which assessments to verify based on what information was most valuable to determine the final grade. Because performing peer assessment is a valuable learning activity [6], future work may select submissions for raters that optimize both score/feedback quality and student learning (e.g. by choosing submissions for peer raters that they can learn most from).

We propose that the combination of machine and human grading can offer strengths that neither has in isolation. The large scale of online classes enables machines to effectively improve the educational experience [9]. By lessening grading burden, machines can focus peers on providing more detailed feedback. Automatic feedback may also focus students on topics they have not fully mastered. Likewise, peers can help machines identify "unknown unknowns" that are blind spots in their models, and help bootstrap that model quickly. Hybrid peer-machine approaches may also help in-person classes and many social computing areas, including crowdsourcing.

ACKNOWLEDGMENTS

We thank Zhenghao Chen and Brennan Saeta at Coursera for building systems that enabled our experimental grading; Kathryn Papadopoulos, Lalida Sritanyaratana and community TAs for grading student submissions; Kanit (Ham) Wongsuphasawat for helping run the pilot experiment, and students that enrolled in our class and participated in this experiment. Chinmay's research was supported by a Siebel Scholarship.

REFERENCES

1. Professionals against machine scoring of student essays in high-stakes assessment (www.humanreaders.com).
2. Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, ACM (2010), 313–322.
3. Boud, D., and Brew, A. *Enhancing learning through self assessment*, vol. 1. Kogan Page London, 1995.
4. Burstein, J. The e-rater scoring engine: Automated essay scoring with natural language processing. *Automated essay scoring: A cross-disciplinary perspective* (2003), 113–121.
5. Burstein, J., Chodorow, M., and Leacock, C. Automated essay evaluation: the criterion online writing service. *AI Magazine 25*, 3 (2004), 27.

6. Carlson, P., and Berry, F. Calibrated peer review and assessing learning outcomes. In *Frontiers in Education Conference*, vol. 2, STIPES (2003).
7. Chen, H., and He, B. Automated essay scoring by maximizing human-machine agreement.
8. Chodorow, M., and Leacock, C. An unsupervised method for detecting grammatical errors. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, Association for Computational Linguistics (2000), 140–147.
9. Fast, E., Lee, C., Aiken, A., Bernstein, M., Koller, D., Smith, E., and Institute, K. Crowd-scale interactive formal reasoning and analytics.
10. Foltz, P. W., Laham, D., and Landauer, T. K. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning 1*, 2 (1999).
11. Grimes, D., and Warschauer, M. Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment 8*, 6 (2010).
12. Heimerl, K., Gawalt, B., Chen, K., Parikh, T., and Hartmann, B. Communitysourcing: engaging local crowds to perform expert work via physical kiosks. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (2012), 1539–1548.
13. Hirschman, L., Breck, E., Light, M., Burger, J. D., and Ferro, L. Automated grading of short-answer tests. *Intelligent Systems and their Applications, IEEE* (2000), 31–37.
14. Hirschman, L., Light, M., Breck, E., and Burger, J. D. Deep read: A reading comprehension system. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (1999), 325–332.
15. Kahneman, D., Lovallo, D., and Sibony, O. Before you make that big decision. *Harvard Business Review 89*, 6 (2011), 50–60.
16. Karger, D. R., Oh, S., and Shah, D. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems* (2011), 1953–1961.
17. Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. R. Peer and self assessment in massive online classes. *ACM Trans. on Computer-Human Interaction 20* (2013), Preprint.
18. Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., et al. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society 389*, 3 (2008), 1179–1189.
19. Lovallo, D., and Sibony, O. The case for behavioral strategy. *McKinsey Quarterly* (2010), 30–43.
20. Nielsen, J. Usability inspection methods. In *Conference companion on Human factors in computing systems*, ACM (1994), 413–414.
21. Peng Dai, M. D., and Weld, S. Decision-theoretic control of crowd-sourced workflows. In *In the 24th AAAI Conference on Artificial Intelligence (AAAI10* (2010).
22. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. Tuning peer grading. In *Proceedings of the 6th International Conference on Educational Data Mining* (2013).
23. Shermis, M. D., Garvan, C. W., and Diao, Y. The impact of automated essay scoring on writing outcomes. *Online Submission* (2008).
24. Sommers, N. Revision strategies of student writers and experienced adult writers. *College composition and communication 31*, 4 (1980), 378–388.
25. Wetzel, C. G., Wilson, T. D., and Kort, J. The halo effect revisited: Forewarned is not forearmed. *Journal of Experimental Social Psychology* (1981).
26. Winerip, M. Facing a robo-grader? just keep obfuscating mellifluously. *New York Times* (2013).
27. Yang, B., Sun, J.-T., Wang, T., and Chen, Z. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM (2009), 917–926.
28. Yannakoudakis, H., Briscoe, T., and Medlock, B. A new dataset and method for automatically grading esol texts. In *ACL* (2011), 180–189.
29. Zhang, M. Contrasting automated and human scoring of essays. *R & D Connections* (2013).