Where's the Learning in Education Crowdsourcing?

Ha Nguyen
University of California-Irvine
Irvine, CA
thicn@uci.edu

June Ahn
University of California-Irvine
Irvine, CA
junea@uci.edu

William Young
Competitive Intelligence Solutions
wyoung@autodealerdata.com

Fabio Campos
New York University
New York, NY
fabioc@nyu.edu

ABSTRACT
Crowdsourcing has shown promise in education domains. For example, researchers have leveraged the wisdom of the crowd to process grading in MOOCs and develop learning resources. An untapped domain is harnessing the crowd to systematically process educational data in classrooms – data that provide key instructional insights but take time to process, such as paper-based assessments. In this paper, we describe an experiment of a crowdsourcing task to effectively process classroom-based data and explore the potential of crowdsourcing as a learning opportunity for the crowdworkers. We discuss implications for designing crowdsourced assessment tasks to yield both high quality output and enriching learning experiences for those involved in the crowdsourcing task.

Author Keywords
crowdsourcing; K-12 education; data processing

CCS Concepts
•**Human-centered computing** → **Empirical studies in collaborative and social computing;**

INTRODUCTION
There is resurgent interest in processing data for improvement in educational contexts. For example, educators frequently collect data in their classrooms through various informal assessments, such as ad-hoc written responses. Processing these non-standardized forms of data at scale is a persistent challenge. For educators, self-processing data takes time. Meanwhile, partnering with researchers to process assessments or to build computational approaches requires fine-tuning large amounts of data, often at substantial costs. To find robust systems to process data in everyday school settings, we turn to potential applications of crowdsourcing.

L@S '20, August 12–14, 2020, Virtual Event, USA.

ACM ISBN 978-1-4503-7951-9/20/08.
http://dx.doi.org/10.1145/3386527.3406734

Crowdsourcing describes a class of human-computation systems that leverage a large number of people (crowdworkers) to solve tasks and process data, rather than relying on a few domain experts [6]. Little is known about the feasibility of crowdsourcing to efficiently process non-standardized educational data that educators already collect. In addition, researchers have only recently attended to how much workers learn from the experience [6], despite findings that enriching tasks help workers stay engaged and produce high-quality data [1]. These observations motivate our project to design crowdsourcing systems to process assessment data from a local middle school. We recruited the crowd from Education undergraduates to score a set of science assessments. The assessments involved open-ended responses that required critical evaluation against a rubric. We ask the following questions:

RQ1. To what extent can crowdsourcing effectively process open-ended assessment data?

RQ2. To what extent can crowdsourcing influence crowdworkers' understanding of task-related concepts, compared to a control group who did not participate in crowdsourcing?

This study has two main contributions. First, our experiment illuminates the potential to leverage crowdsourcing to efficiently process non-standardized educational data. Second, findings suggest learning benefits from the crowdsourcing experience for those individuals who helped with the collective task, compared to a control group. We discuss implications for designing crowdsourcing tasks to process data for educators and create positive learning experiences for workers.

THEORETICAL BACKGROUND
Crowdsourcing for Education
The literature on crowdsourcing for education is nascent but growing. A recent review [9] identified the main domains in which researchers have leveraged crowdsourcing: to generate content, provide training experiences, and gather feedback [2, 3, 6]. Of those areas, using the crowd to grade assessment pertains to our study. Researchers have developed systems to crowdsource grades, provide qualitative feedback on student work, and increase student knowledge in self or peer assessment [3, 7]. These tasks help to process assessment ***for*** learning [12] and provide instructors immediate insights.

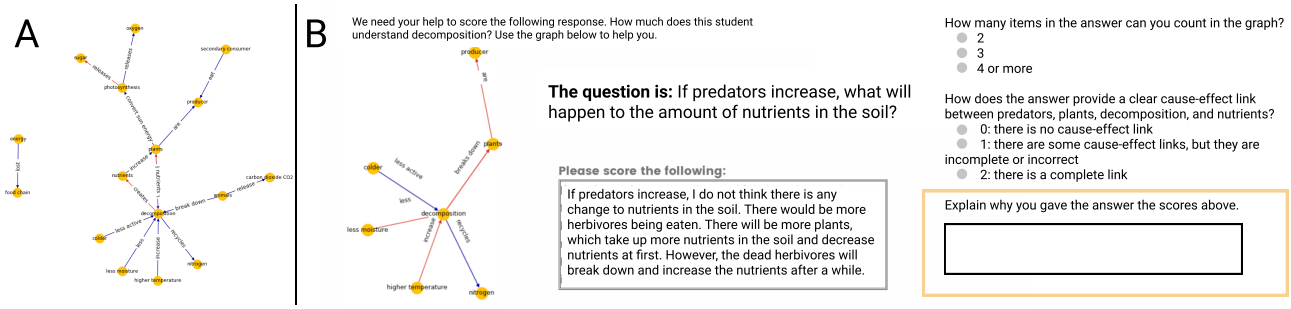

Figure 1. Example Prompt for the Grading Task

Creating crowdsourced assessments requires familiarity with the assessment rubric on the part of workers who are grading student artifacts. Many studies overcome this barrier by recruiting workers from specific pools, such as past learners of the same course [10], submitters of the same homework assignment to grade others [3], or the end-users of a system [7]. In education tasks, the requirement to have domain knowledge appears to be a key design constraint. We overcame this constraint in grading K-12 assessments by recruiting our workers from Education undergraduate courses (K-12 Education, Education Evaluation, Education Technology). We verbally confirmed with participants prior to the task that they had experiences with scoring classroom-based assessment rubrics.

Assessment for Learning

Researchers find that crowdworkers are motivated to perform tasks when they find them interesting, such as fostering new learning experiences [9]. One learning mechanism is to build one's knowledge from assessment tasks. We turn to the idea of assessment ***for*** learning to understand how we may help crowdworkers learn [12]. Assessment for learning is employed in way that goes beyond mere evaluation, one that also provides learning values to the graders of the task. For example, in self-grading assessment, participants have the agency to use feedback to realize gaps in their knowledge and adjust understanding. There also exist learning benefits when grading others' work [11], although the crowdsourcing literature has rarely examined the workers' learning experiences [9].

Design Features for Crowdsourcing Tasks

To improve the quality of the crowdsourced assessments, researchers have explored different forms of training workers. Most assessment studies utilize rubrics to direct the crowds' attention to specific scoring dimensions. Providing worked-examples for training workers and calibrating work generates higher quality performance, while minimizing costs such as payments, attrition, and time spent [5]. Researchers have also leveraged different task designs. For example, workers who are asked to provide reasoning to their answers (rather than just giving a score) produce higher quality results [8].

METHOD

Study Context

This study is part of a project in middle school science. The project's goal is to help learners develop an understanding of systems as complex, cyclical causal links (e.g., carbon cycle) instead of linear links (e.g., plants need water). We collected pre and post-test assessments that can be scored using a rubric.

As a pilot to develop a processing pipeline for the assessment, we recruited participants from three Education courses (n = 168) at a Hispanic-serving university in the U.S. We prepared 5 assessment sets, each containing 3 student answers that represent a range of complexity in causal thinking. All participants were told that their participation would help to build a data pipeline to process educational data. The experiment took place on graphlab.net, a website we created.

Informed by work on design for crowdsourcing, we designed our crowdsourced task to include a clear scoring rubric and a training phase. We further tested the extent to which explaining why a response would receive certain scores enhanced the accuracy of the crowdsourced assessments. Participants received extra credit for completing all tasks, outlined below.

Baseline. Participants answered 20 multiple choice questions to establish baseline knowledge about ecosystems causal links.

Training. All participants were prompted to explore an individualized concept map, which highlighted their incorrect answers from the baseline test (panel A, Figure 1). The concept map illuminates the types of elements and causal links we want the middle school students to show in their responses.

Experiment. Participants were randomly divided into one of the three conditions:

- Control group: watched a three-minute video about the decomposition process and did not grade (n = 71).
- Grade Only group (n = 45). Each participant was randomly assigned a set of student responses (panel B, Figure 1).
- Grade & Explain group (n = 45). To test the benefit of self-explanation to accuracy, we included an additional grading condition where participants had to explain the rationale for their scores (highlighted in yellow; Figure 1).

Post-test. Participants in all 3 experimental groups answered a post-test about science knowledge and science attitude [4]. The science post-test prompted participants to explain the population balance in a hypothetical ecosystem and provide examples of how they may teach causal links to a fifth grader. The post-test's elements and links (e.g., decomposer-producer-

Variable	Test Conditions					
	Control		Grade Only		Grade Explain	
	M	SD	M	SD	M	SD
Post-test	6.25	2.35	7.27	2.44	6.87	2.29
Pre-test science	14.77	3.11	15.07	2.54	15.24	2.85
Science attitude	16.00	2.94	15.62	3.90	16.16	3.42
Time on survey	10.76	4.95	12.09	5.06	13.78	5.38
Observations	71		45		45	

Table 1. Descriptive Statistics

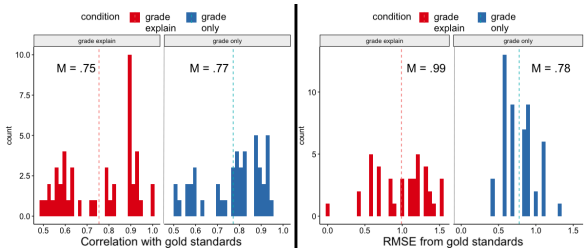

Figure 2. Correlation and RMSE between Crowdworkers and Expert Scores

predators) parallel those in the pre-test to assess the same domain understanding of causal links. The post-test examined potential learning from the crowdsourcing experience.

Grading Rubric

Participants scored student responses along two dimensions: Elements and Coherence. For Elements, participants counted the target vocabulary (the training concept map; panel B) that appeared in students' responses. For Coherence, participants determined whether the responses reflected coherent links. For example, a lowest score of 0 would show only linear links (e.g., rabbits eat grass). A highest score of 2 would indicate coherent feedback loops (e.g., foxes eat rabbits so rabbits number declines, then foxes die; rabbits eventually increase). A score of 1 would indicate some links, but incomplete evidence.

Analytical Strategies

We report results from 161 scores after removing blank answers (4% of the sample). Table 1 presents the descriptive statistics. There was no difference among conditions for science understanding and science attitude at pre-test.

RQ1. Assessment Quality. We reported two reliability metrics for the crowdsourced grading from the overall sample and each treatment group. The first is inter-rater reliability (Krippendorff's alpha, i.e., ratio of observed disagreement to expected disagreement). Alpha ranges between 0 and 1, with 1 suggesting perfect agreement and a threshold of .70 indicating acceptable agreement among crowdworkers. The range for crowdworkers in prior research is .40 to .60 [1]. The second is internal consistency (Cronbach's alpha; threshold of .70).

To examine the quality of the crowdsourced grades, we calculated the correlation between the crowdsourced grades and grades given by three researchers in the project (i.e., expert scores). The domain expert's scores has often been treated as the gold standard in crowdsourced assessment work. We also calculated the deviation of the crowdsourced scores from the expert scores using Root Mean Square Error (RMSE).

RQ2. Learning Benefits. To parallel the grading rubric for the middle school students, we scored participants' post-test science knowledge along similar dimensions: Elements and Coherence. We applied linear regression to predict the post-test scores by experimental conditions. Covariates included baseline science understanding and science attitudes. We controlled for multiple testing with the Benjamini-Hochberg procedure at the false discovery rate of .10.

FINDINGS

Crowdsourced Scores Showed Substantial Agreement

We observed that the inter-rater reliability and internal consistency in our sample were higher than prior crowdsourcing work [1]. The scores were substantial in terms of agreement among raters (average Krippendorff's alpha across the five test sets: Overall: M = .61, SD = 02; Grade only: M = .66, SD = .06; Grade & explain: M = .59, SD = .07).

The ratings had moderate internal consistency, with variation among the question sets. Cronbach's alpha for the overall sample: M = .68, SD = .14; Grade only, M = .53, SD = .22; Grade & explain: M = .70, SD = .12.

Scores were Strongly Correlated with Expert Scores

The crowdsourced scores were strongly correlated with the expert scores (M of Pearson's r = .76, SD = .15). Figure 2 presents the distribution of correlations between individual learner's scores and the expert scores. Wilcoxon signed rank test suggested no difference in the correlation distribution between the two grading conditions, W = 1020.5, p = .91.

Scores in Grade Only Group were More Accurate

We compared the accuracy for the two grading conditions by calculating the RMSE between the crowdsourced scores and the expert scores. Smaller RMSE values suggests closer alignment with the expert scores. Interestingly, we found that the gap between the participants' scores and the expert scores in the Grade & Explain condition was significantly larger, M = .99, SD = .36, than the Grade Only condition, M = .78, SD = .21, W = 1411.5, $p < .001$. Overall, the scores in Grade Only condition were closer to the expert scores.

There were Learning Benefits to Crowdsourcing

We found that grading facilitated learning, compared to the control group who did not grade. On average, those in the two grading conditions scored higher in their post-test science total score (Grade Only: M = 7.27, SD = 2.44; Grade & Explain: M = 6.87, SD = 2.29, respectively), compared to the control group (M = 6.25, SD = 2.35).

For the Grade Only group, we found a significant, positive learning outcome after accounting for pre-test science understanding and attitudes towards science, β = .42, SE = .18, adjusted p = .04. Participants who graded the essays scored .42 standard deviation higher than those who watched a video and did not grade. We did not find a significant difference between the Control and the Grade & Explain group (Table 2).

Variable	β	*SE*	*t*	*p*	*Adj. p*
Grade only	.42	.18	2.32	.02*	.04*
Grade & explain	.21	.18	1.16	.24	.25
Pre-test science	.28	.08	3.64	***	***
Science attitude	.15	.08	1.98	.05*	.07
Observations	161				

*p < .05. ** p < .01. *** p < .001

Table 2. Predictors of post-test science scores

DISCUSSION

Findings suggest the promise of crowdsourcing in processing formative school data. The crowdsourced scores were internally reliable and highly correlated with the experts'. Importantly, the workers showed enhanced understanding of task-related concepts, compared to the group who did not grade.

At the same time, results illuminate the need for new crowdsourcing design guidelines in education domains. In our study, self-explanation did not show substantial benefits in grading accuracy and learning, compared to the Grade Only condition. An explanation is that the self-explanation task does not provide enough guidance for the assessment dimensions on which crowdworkers can justify their answers. This explanation suggests tasks should concretely map onto the assessment rubrics. Another explanation is workers may have suffered from survey fatigue, since they spent more time answering the tasks (Table 1). This explanation points to the need to balance reward, task features, and cognitive load. Design implications from the learning sciences for future iterations include suggestions for providing timely, task-specific feedback on the crowdsourced assessments to improve workers' quality [5, 7].

CONCLUSIONS & FUTURE WORK

Our experiment illustrates that crowdsourced experience can build towards the vision of assessment to improve learning rather than merely evaluating learning [12]. A future direction is to test other task designs to improve assessment accuracy and learning benefits in education domains. Our experience suggests that the crowdsourced output will result in variation, regardless of the task designs. Representing this variance to our end users (educators who want to act upon this data), illuminates another challenge that we plan to pursue.

ACKNOWLEDGMENTS

We thank our partner educators who inspired this work, the volunteer participants, and the conference reviewers for their helpful feedback.

REFERENCES

[1] Omar Alonso. 2019. The Practice of Crowdsourcing. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 11, 1 (2019), 1–149.

[2] Dmytro Babik, Lakshmi S. Iyer, and Eric W. Ford. 2012. Towards a Comprehensive Online Peer Assessment System. Springer, Berlin, Heidelberg, 1–8. DOI: http://dx.doi.org/10.1007/978-3-642-29863-9{_}1

[3] Julia Cambre, Scott Klemmer, and Chinmay Kulkarni. 2018. Juxtapeer: Comparative Peer Review Yields Higher Quality Feedback and Promotes Deeper Reflection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, New York, New York, USA, 1–13. DOI: http://dx.doi.org/10.1145/3173574.3173868

[4] J. Chung, M. A. Cannady, C. Schunn, R. Dorph, and P. Vincent-Ruz. 2016. *Measures Technical Brief: Competency Beliefs in Science*. Technical Report. Activation Lab.

[5] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a Learning Science for Complex Crowdsourcing Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, New York, New York, USA, 2623–2634. DOI: http://dx.doi.org/10.1145/2858036.2858268

[6] Shayan Doroudi, Joseph Jay Williams, Juho Kim, Thanaporn Patikorn, Korinn S. Ostrow, Douglas Selent, Neil T. Heffernan, Thomas Hills, and Carolyn P. Rosé. 2018. Crowdsourcing and education: Towards a theory and praxis of learnersourcing. In *Proceedings of International Conference of the Learning Sciences, ICLS*.

[7] Steven Dow, Elizabeth Gerber, and Audris Wong. 2013. A pilot study of using crowds in the classroom. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. ACM Press, New York, New York, USA, 227. DOI: http://dx.doi.org/10.1145/2470654.2470686

[8] Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

[9] Yuchao Jiang, Daniel Schlagwein, and Boualem Benatallah. 2018. A Review on Crowdsourcing for Education: State of the Art of Literature and Practice. In *Pacific Asia Conference on Information Systems*. 180. https://aisel.aisnet.org/pacis2018/180

[10] Hassan Khosravi, George Gyamfi, Barbara E Hanna, and Jason Lodge. 2020. Fostering and supporting empirical research on evaluative judgement via a crowdsourced adaptive learning system. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 83–88.

[11] Carmen E Sanchez, Kayla M Atkinson, Alison C Koenka, Hannah Moshontz, and Harris Cooper. 2017. Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology* 109, 8 (2017), 1049.

[12] Richard J Stiggins. 2002. Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan* 83, 10 (2002), 758–765.