

Transforming Everyday Information into Practical Analytics with Crowdsourced Assessment Tasks

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

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

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

William Young
wyoung@autodealerdata.com
Competitive Intelligence Solutions LLC.
Irvine, CA, USA

ABSTRACT

Educators use a wide variety of data to inform their practices. Examples of these data include forms of information that are commonplace in schools, such as student work and paper-based artifacts. One limitation in these situations is that there are less efficient ways to process such everyday varieties of information into analytics that are more usable and practical for educators. To explore how to address this constraint, we describe two sets of design experiments that utilize crowdsourced tasks for scoring open-ended assessments. Developing crowdsourced systems and their resulting analytics introduced a variety of challenges, such as attending to the expertise and learning of the crowd. In this paper, we describe the potential efficacy of design decisions such as screening the crowd, providing multimedia instruction, and asking the crowd to explain their answers. We also explore the potential of crowdsourcing as a learning opportunity for those participating in the collective tasks. Our work offers key design implications for leveraging crowdsourcing to process educational data in ways that are relevant to educators, while offering learning experiences for the crowd.

CCS CONCEPTS

• Human-centered computing → User interface design.

KEYWORDS

crowdsourcing; worker training; assessment; education

ACM Reference Format:

June Ahn, Ha Nguyen, Fabio Campos, and William Young. 2021. Transforming Everyday Information into Practical Analytics with Crowdsourced Assessment Tasks. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21)*, April 12–16, 2021, Irvine, CA, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3448139.3448146>



This work is licensed under a Creative Commons Attribution-ShareAlike International 4.0 License.

LAK21, April 12–16, 2021, Irvine, CA, USA

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8935-8/21/04.

<https://doi.org/10.1145/3448139.3448146>

1 INTRODUCTION

In the United States, the use of data to inform instruction is a substantial expectation for K-12 teachers [5, 16]. The data that educators utilize is quite varied, but the most common forms are standardized assessments from state-mandated tests [19]. In contrast, the field of learning analytics (LA) forwards the idea that data collected unobtrusively from digital platforms could also be useful to inform teaching practices. In this paper, we begin from a different starting point with the observation that *K12 educators are already awash in rich, everyday, and locally relevant data*. These data come in forms such as students' work on paper, oral presentations, or "exit slips" that are short surveys teachers might give to check in with students [42].

Predating the rise of data science and LA, these forms of information were used for formative assessment, to provide timely feedback on students' educational progress and improve learning and teaching approaches [39]. For example, a teacher might administer formative assessments iteratively to identify content areas to focus on, experiment with different teaching strategies, analyze the results from the next assessment, and make additional pedagogical changes. However, a challenge in utilizing formative data is that they are often expensive and time-consuming to process at scale or in an efficient enough manner to be usable for educators. Teachers either have to self-process (e.g., educators or students collect, process, and evaluate their own data) or partner with other institutions (e.g., researchers) to develop validated scoring systems and data pipeline processes. Each of those approaches has unique limitations, and educators' choices tend to be limited by time and resources. Meanwhile, leveraging other stakeholders such as researchers can be time-consuming and costly [43]. While computation can reduce the time and cost to process data, there are particular forms of data that are commonplace in school settings but too complex and ambiguous to rely solely on automated scoring [3]. This observation motivates our research question:

How might we develop systems to improve the processing and usability of diverse forms of education data that are abundant in school settings and already collected by teachers?

Exploring this question is vital for many reasons. First, it is important to value the diverse forms of data that educators utilize to make sense of instructional events and inform future pedagogical

moves. Researchers and practitioners should not assume that standardized assessments or learning analytics collected from digital platforms are the only valid or useful forms of data. Second, designing analytics that are directly relevant to teachers, can better connect such analytics to instructional work. A common finding is that teachers are more likely to connect with and use data to make instructional improvements when that data is locally-generated, relevant to questions they have, and directly suggest changes in practice, versus abstract constructs or metrics that leave practice implications vague [16, 39].

In thinking about how to efficiently process everyday education data, we were motivated by the rich body of research in LA and Human-Computer Interaction (HCI) on *crowdsourcing*. The term crowdsourcing describes a class of human-computation systems where members of the public perform tasks that would typically be done by a designated agent or expert [28, 35]. Crowdsourcing systems leverage elements of collective intelligence, where a large number of people solve discrete and simpler tasks, that in the aggregate can accurately approximate what an expert might do. Crowdsourcing can be seen as outsourcing tasks to humans that artificial intelligence cannot yet accomplish [28]. Of particular interest to our work are studies that employ crowdsourced assessment to reduce the workload for educators and provide diverse perspectives to students about their learning progress [3, 6]. However, additional design studies are needed to explore how to create crowdsourcing systems that can process the diverse forms of data found in K-12 contexts.

This paper details two sets of design experiments for developing crowdsourcing tasks to process diverse forms of educational data, to make three major contributions. First, there has been extensive work in the human computation field to improve the technical and algorithmic aspects of crowdsourcing systems [9], but only a small body of work focuses on **practical applications of crowdsourcing in education** [12, 21]. We contribute to this emerging research area by documenting how such systems can be applied to a different genre of problem: developing more robust practical analytics that align with the contextual conditions and needs of complex, K-12 school environments.

Second, we illuminate **key design implications** for employing crowdsourcing in education assessment, such as screening workers, providing multimedia instruction and worked examples, and adding explanation prompts that may help improve the quality of the assessments. Researchers and practitioners can consider these locally grounded principles in developing workflows for other crowdsourcing tasks for education data.

Third, our experiments suggest **learning benefits for the crowd** as they engage with the collective tasks. In processing assessment data that require domain knowledge, the crowd may acquire knowledge that helps to improve their performance in subsequent tasks. From an ethical standpoint, our findings highlight how the crowd are not merely data points, but could be positioned as partners with educators to personally benefit from the crowdsourcing workflow.

2 THEORETICAL FRAMEWORK

Our project integrates an interdisciplinary set of research areas. First, we articulate a new problem-space in education settings that

aligns with the affordances of crowdsourcing systems: the need to develop practical analytics to inform pedagogical improvement. Second, we review relevant work in LA and HCI on the design and application of crowdsourcing systems in education, and highlight design lessons learned from broader crowdsourcing studies. Third, we discuss the importance of providing enriching learning experiences for those involved in the crowdsourcing tasks.

2.1 A Design Problem: Practical Analytics in Education

There has been growing interest in using data to improve teaching and learning in K-12 settings [5, 16]. A key element of improvement-oriented approaches for promoting positive changes in K-12 schools is to engage educators - such as policymakers, school leaders and teachers - in collecting practical data that are (1) efficient and low-cost to collect, (2) directly inform educators' improvement goals, and (3) provide useful information to change school practices [43].

While using practical data seems logical, collecting and processing such information into usable metrics - what we call *practical analytics* - is difficult to do, and often not utilized in systematic ways in K-12 schools [5]. Instead, the most prevalent forms of data in K-12 schools are "accountability data" such as standardized test scores; "research data" such as observations, surveys, validated assessments, interviews, and other forms of data collected by outside researchers; and a growing prevalence of digital platforms that provide automatically generated analytics of student behavior.

Education researchers have highlighted how accountability and research data are often not practical nor useful for school improvement. For example, standardized test scores are useful to track performance and enforce government accountability, but teachers often cannot make substantial classroom decisions from test scores alone [16]. Similarly, researchers spend many years carefully developing data collection protocols, validating survey measures, and developing data analysis plans. By the time findings from such studies are disseminated, schools and educators have moved on as they must solve problems on a daily basis. Thus, there has been renewed interest in finding efficient and effective ways to leverage more practical data that can more directly inform improvements in school settings [5, 27, 29, 43].

Educators often rely on everyday sources of information that are present in classrooms to inform their decisions on a daily basis [38, 42]. Examples of such information include local assessments in formats that are difficult to organize at scale (e.g., pen and paper), group discussion, or responses in class. These data types allow educators to quickly gauge student understanding and engagement, and make more immediate instructional decisions. We argue that these rich information sources are better suited to help drive local, school improvement. However, such information is often difficult to collect, organize, code or score, validate, and represent back to educators. Processing data places excessive demands on educators' already limited time and resources [40].

Finally, while a rich body of work in intelligent tutoring systems and computational linguistics has explored automated scoring of student work [17, 18], more complex data and local, idiosyncratic assessment needs may not easily lend themselves to general computational approaches [3]. For example, computational approaches

can provide proxies for writing quality and complexity – such as essay length, counts of parts of speech – but may not capture the rich semantic meaning in short, open-ended responses [25]. In addition, automated assessment systems often rely on a large corpus of training data that may not readily adapt to local assessment contexts [1]. Pre-trained natural language processing models have shown promise in automated scoring of student essays, but fine-tuning these models comes at substantial computational costs [31]. The design challenge to convert fast-changing, everyday data into usable analytics led our research to explore human computation approaches, specifically crowdsourcing applications.

2.2 Crowdsourcing Systems in Education

The literature on crowdsourcing for education is nascent but growing. A review of 51 studies [21] identified the main domains that have leveraged crowdsourcing: to generate content, provide training experiences, exchange knowledge, and gather feedback from other experts and novices [10, 12, 20]. Researchers have developed systems to crowdsource grades, provide feedback on student work, and increase student knowledge attainment as they assess their own or peers' work [6, 10, 13, 33]. However, most of these cases occur in self-contained courses where students assess each other, often in higher education settings, and where students can be expected to have some domain knowledge.

Most educational crowdsourcing studies also take place in digital environments such as online courses or communities, where the collection of data and crowdsourced output happens on a larger scale [23]. Contexts that can lead to practical analytics in K-12 classrooms present a potential use case that the crowdsourcing literature has largely underexplored.

A key issue in educational crowdsourcing is how to increase the overall quality of the work [6, 10, 14, 26]. Educational crowdsourcing tasks often require specialized knowledge (e.g., teaching and learning knowledge, or skills to assess a given assignment) that cannot be solved effectively by unskilled crowdworkers [24]. Researchers have sought to overcome this knowledge barrier by recruiting workers from specific pools with relevant knowledge or experience, such as submitters of the same homework assignment, who can then grade others [6, 10] or the potential end-users of a system [13]. Another strategy is to screen workers [9]. Here, workers are asked to demonstrate certain knowledge and skills by passing a threshold of correct answers before they can access the main task. Qualification tasks can identify prepared workers, thereby enhancing the output quality. We explored how these known issues may play out in educational domains by screening workers (Study 1) and recruiting those with prior knowledge (Study 2).

2.3 Designing Crowdsourcing Systems

To increase performance in crowdsourcing tasks, researchers also attend to the task design features, such as developing clear scoring rubrics or providing feedback to workers. Most crowdsourced assessment studies utilize rubrics to direct the crowds' attention to specific scoring dimensions. Providing worked-examples for training crowdworkers and calibrating work with other workers or experts generates higher quality performance, while minimizing costs such as payments, attrition, and time spent [11, 32]. For

example, Doroudi and colleagues [11] find that reviewing expert examples was associated with higher task accuracy than other conditions, such as simple training (i.e., learning by solving additional problems), seeing solutions after trying tasks, or no training.

Researchers have explored the utility of giving feedback to crowdworkers to improve their performance iteratively [4]. Task-specific feedback that is provided internally (i.e., self-assessment) or externally (i.e., experts) can yield better overall work than receiving no feedback [14]. In other studies, researchers found that asking crowdworkers to provide reasoning for their assessments (rather than just giving a score) produced higher quality results [15].

Although the crowdsourcing literature acknowledges the importance of training and providing examples, little research has elaborated on the modalities in which workers received task instruction, and how instructional modality (i.e. video-based, text-based, etc.) may affect task performance [11]. In our study, we were interested in developing crowdsourcing tasks for our design problem of processing practical analytics from education information, with a keen eye to ensuring that crowdworkers would have adequate preparation for assessment tasks. We turned to relevant learning sciences research to design instructional modalities to efficiently and effectively train workers on our assessment tasks. For example, a series of experiments employing cognitive theory has found that students scored significantly higher on transfer tests after receiving a multimedia explanation (i.e., words and pictures), rather than words alone [30].

Crowdsourcing task designers are also attuned to theories of how workers learn as they participate, to provide value that go beyond financial incentives [4, 21, 32]. Emerging research has begun to explore how tasks can be designed to provide learning opportunities for crowdworkers as they engage in the work [12]. For example, the idea of "*learnersourcing*" describes systems where learners crowdsource their activities to create materials for future learners [12, 23]. *Learnersourcing* assumes that workers can mutually benefit from each others' curated resources [23].

We argue that crowdsourcing tasks in education, which often require deep knowledge of the subject matter and school contexts, can provide learning opportunities for participants. One mechanism is to build one's knowledge from assessment tasks. We turn to the idea of assessment *for* learning [41] to understand how we may help the crowd learn. While all assessment can be used to develop instructional and learning strategies, assessment for learning is employed in a way that goes beyond mere evaluation, *one that also provides learning value 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 [36].

2.4 Crowdsourcing Systems for Practical Analytics

In sum, we situated our study in a unique problem space of needing to process practical data in school settings. We built on the small but growing number of studies that seek to apply crowdsourcing approaches to education domains, and drew from the LA and HCI literature on how to design tasks to support learning and worker performance. We structure the following paper into two studies in different domains: (1) argumentation in writing, and (2) complex

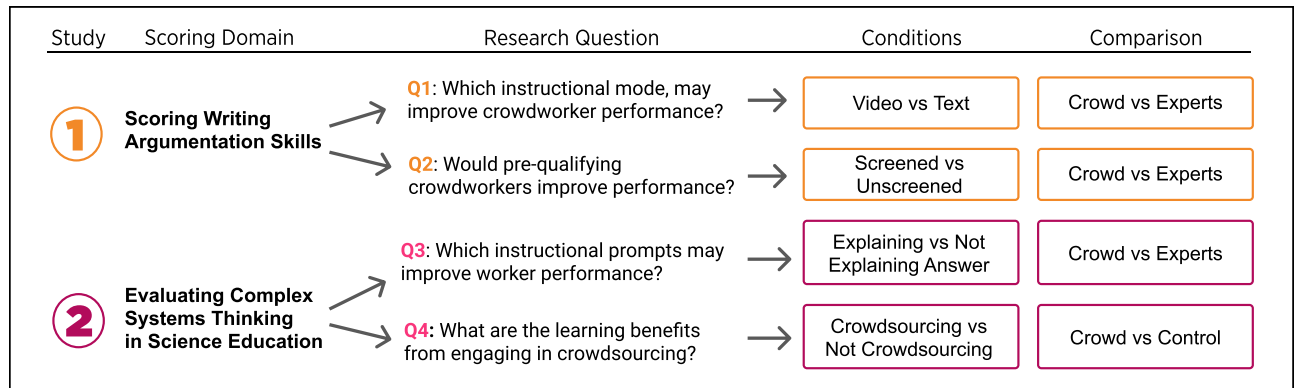


Figure 1: Overview of the two studies

systems thinking in science. To understand the importance of education domain knowledge, we involved two different crowds. The first study involved crowdworkers on Amazon Mechanical Turk (AMT), a general marketplace to crowdsource tasks. The second study recruited a crowd of Education undergraduates, who presumably had prior knowledge in grading student work. Together, the two studies (Figure 1) help us examine the potential of crowdsourcing applications in processing a variety of school-based data, using different task designs.

3 STUDY 1: ASSESSING WRITING SKILLS

3.1 Context

This study was part of a larger participatory design research initiative conducted in partnership with a middle school in the United States. The school was part of a national arts-based education initiative [44] that utilized a curriculum where students developed their writing and argumentation skills by analyzing visual artifacts (e.g., artwork, graphs, etc.). The curriculum and assessments were called Visual Thinking Strategies (VTS).

In our partner school, students took a short assessment at the beginning and end of the school year, where they were asked to write arguments based on an image (Figure 2) and justify their claims with evidence. These critical writing and literacy skills were important for teachers and principals; particularly as they served many students who were English-language learners. Teachers used the data to identify gaps in students' reasoning skills, to create a school-wide curriculum across grades and subjects. A key goal for our partner teachers was to efficiently process these writing artifacts to inform pedagogical decisions within a short time period.

In fact, a practical motivation for this use-case arose when the school principal showed us the pile of hundreds of student essays, tucked in the corner of her office, as the staff had little time or capacity to do any in-depth scoring during the school year. The school was relying on more basic metrics such as essay length to cut down on processing time, even though they acknowledged those metrics did not properly capture aspects of learning that the teachers cared about. In our initial attempt

to score the essays, we explored uses of off-the-shelf algorithmic approaches that had been tested in similar educational contexts. For instance, we experimented with Coh-Metrix, a computational linguistics system that has been trained and validated on a large, representative written text corpus across grade levels [17]. However, we could not find an approach that perfectly aligned with the assessment task.

This use-case represented an instance where crowdsourcing could be utilized. We developed a rubric to guide the crowd, that came from a general population in Amazon Mechanical Turk, to be able to accurately score the students' writing artifacts even though they might not have background knowledge about teaching and learning of writing or visual thinking strategies.

3.2 Assessment Rubric

Our team developed an assessment rubric based on students' work, the school's feedback, and research on writing assessments [37]. The rubric was broken down into two dimensions: Making **Claims** and Providing **Evidence** (Figure 2). A Claim referred to the total number of times students made an inference or assertion, as they wrote about a visual prompt. Evidence referred to the number of times students supported their claims with justification. It is entirely possible for students to make many claims, but not offer supportive evidence; and vice-versa to offer many factual statements but never synthesize their thoughts into an argument. The teachers of our partner school were interested in developing both parts of this writing. Our research team validated the rubric through four rounds of reliability checks, establishing substantial agreement after the fourth round (intraclass correlation - ICC - of .90 and .93 for identifying Claims and Evidence in writing samples).

3.3 Defining Metrics for Worker Performance

Building on prior crowdsourcing work [2], we assessed whether general crowdworkers were able to evaluate student writing, just as experts might do (who had knowledge to accurately assess student writing). First, we explored whether general crowdworkers were consistent in their evaluations of student writing by looking at


CLAIM (a conclusion with ideas not explicitly present in the picture)			EVIDENCE (an explicit description of the picture)		
Score	Definition	Example	Definition	Example	
4	More than three statements of inference.	"Maybe the man in the other room is her dad. I think that they got in an argument / and know the girl is mad at him/ What I could also found is that the girl looks sad/ and the dad looks very mad."	More than three statements of evidence from the picture to support claims.	"because it looks like a stove near his arm/ They look sad because they are not smiling/ and the position that the lady is in/. I can see a fan."	
3	Three statements of inference.	"I think that the dad and daughter got in an argument with his dad/ and know the girl is mad at him/. What I could also found is that the girl looks sad/ and the dad looks very mad."	Three statements of evidence.	"... because he is standing alone. And he is not smiling/. And she is crying."	
2	Two statements of inference.	"I see a family about to move. Or maybe they have just moved in."	Two statements of evidence.	"They are in different rooms. The house looks empty" OR "Because they are in separate rooms. They have straight faces."	
1	One statement of inference.	"In this picture there are two people in a fight"	One statement of evidence.	"It is an empty house" OR "There are boxes"	
0	No response OR Not enough information to show inference.	"I see a room" OR "I see a man standing"	No response OR Not enough information to show evidence.	"I don't know" OR "I think they are arguing."	

Figure 2: Instructions for crowdworkers to score students' essays about visual prompts

their inter-rater reliability. We calculated Krippendorff's α , which is the ratio of observed to expected disagreement among workers. Krippendorff's α ranges between 0 and 1, with 1 suggesting perfect agreement and a threshold of .70 indicating acceptable agreement.

Second, we evaluated how closely worker evaluations resembled the "ground truth" (i.e., correlation with and deviation from a set of scores agreed upon by the research teams and educators). We determined correlation between crowdworkers scores and those from experts as Pearson's r , and calculated Root Mean Square Error (RMSE) to determine worker scores' variation from expert scores. RMSE was computed as the square root of the sum square differences in scores between workers and our research team, divided by the number of all items.

Finally, we explored the assumption of *collective intelligence*: that crowdsourcing tasks could achieve higher quality when the number of workers increased [28]. To further compare the crowdsourced output to the human experts, we ran simulations using the aggregated input from multiple individuals, instead of scores from single workers. For example, for iteration 2, the simulations randomly drew a combination of k scores from n workers from the screened and unscreened data without replacement, $k = (1, 7)$, $n = 7$ in screened and $k = (1, 7)$, $n = 30$ in unscreened condition. We rounded up the average for each worker combination and visualized how much these averages deviated from the expert scores. One expectation is that as more workers score a given piece of writing, their average score would get closer to experts' scores.

3.4 Iteration 1: Multimedia vs. Text Instruction

In this crowdsourcing situation, one has to give instructions to general crowdworkers, to tell them how to do this assessment task. A key question in this first iteration was to explore whether crowdworkers might produce accurate evaluations of the claims and evidence students used, if presented instructions in different modalities. We drew on the Learning Sciences and crowdworker training literatures to design our tasks on Amazon Mechanical Turk

(AMT) [9, 11, 30]. We developed detailed instructions for crowdworkers to score an essay for the quality of claims and evidence present in the writing (see Figure 2).

In the first iteration of our design experiment, we tested multimedia versus text-based instructions. Workers ($n = 195$) who accepted the AMT Human Intelligence Task (HIT) were randomly assigned to either multimedia ($n = 89$) or text conditions ($n = 106$). In the multimedia condition, workers viewed a 45-second video about scoring the essay. The video showed an image of the rubric similar to the text condition, but included narration of what counted and did not count toward the scoring criteria. The text condition gave text-based examples in a rubric (Figure 2). Following the instructions, workers attempted to score ten test writing samples, with a time limit of fifteen minutes. Workers were asked to give a score of 0 to 4 for Claim and Evidence statements they observed in each essay. The sample size difference between the multimedia and text-based conditions is likely due to attrition for workers who watched the instruction video in another browser. We pooled the test tasks from a set of essays that had already been scored by experts and checked for consistency. In this iteration, we tested the following hypotheses:

H1. Crowdsourcing potential. The crowdsourced scores would show substantial agreement with the expert scores.

H2. Instruction. The video condition would be associated with more accurate crowdworker scores. [30].

3.5 Findings from Iteration One

H1. Crowdsourcing showed moderate agreement with experts.

We observed that the crowdsourced scores in both training conditions were moderately correlated with the expert scores. The correlation values (Pearson's r) were: $M = .51$, $SD = .37$ overall; $M = .55$, $SD = .32$ for text, and $M = .46$, $SD = .42$ for video. Scores from individual workers were in low agreement with one another; Krippendorff's alpha was .47 overall, .55 for text, and .37 for video condition.

H2. No difference between instructional conditions. We examined the impact of text versus multimedia instruction on worker

Conditions	Per worker		
	Correlation	RMSE	Completion Time
Iteration 1			
Text (n=106)	.55 (.32)	1.20 (.30)	2.94 (1.15)
Video (n=89)	.46 (.42)	1.22 (.40)	3.00 (.96)
Iteration 2			
Unscreened (n=30)	.45 (.25)	1.42 (.30)	6.86 (2.64)
Screened (n=7)	.74 (.17)	.97 (.28)	8.50 (3.05)

Note: Standard deviations in parentheses

Table 1: Correlation, deviation from experts, and time

performance. Table 1 reports the correlation, deviation from expert scoring, and completion time for workers in the two conditions. Contrary to our expectations, non-parametric Mann-Whitney-U tests indicate no significant difference between conditions in the crowdsourced scores' correlations with and variation from expert scores ($U = 4429$, $p = .46$; $U = 4773$, $p = .89$). An explanation for not finding differences between the two conditions, based on the informal feedback we received from some crowdworkers from the video condition, is that the allocated time for watching the video and then scoring essays was too short. However, this explanation is less likely as we did not find any difference in the completion time between conditions. Workers spent, on average, the same time scoring the writing regardless of how they received instructions,

3.6 Iteration 2: Screening Workers

Although results from the first iteration were promising, the accuracy of the worker evaluations could be further improved, for example, to increase agreement with expert scores to be higher than .70. In the second iteration, we explored whether only allowing crowdworkers who demonstrated ability to score the essays, resulted in more accurate evaluations. Prior studies of crowdsourcing systems suggest screening workers - by giving them a sample task and only allowing those who perform well to complete subsequent work - as one strategy to improve the quality of respective crowd tasks [9]. Thus, we developed a screening procedure, comprised of three different essays and recruited AMT crowdworkers. We restricted the screening task to workers with at least 95% approval rates in previous HITs. Those crowdworkers whose accuracy was at least 0.67 in the screening task were then invited to complete the main task (which was the same as in iteration one). Our conjecture, inspired by prior literature, was that crowdworkers who already demonstrated some ability to conduct this educational evaluation, would perform more accurately than a general population of AMT crowdworkers. We examined the following hypotheses:

H3. Screening. Screening workers would yield higher quality work than unscreened workers [9].

3.7 Findings from Iteration Two

H3. Screened Workers Showed Higher Quality Work. A set of findings support our hypotheses that screened workers showed higher quality work. First, although there was no difference in the

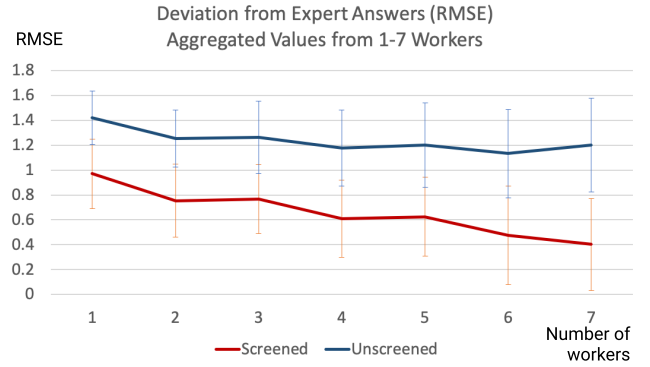


Figure 3: Adding more workers improved performance, as shown by the reduced RMSE, in the screened condition

task completion time, screened workers showed higher agreement with one another. The inter-rater agreement (Krippendorff's alpha) among screened workers was higher ($\alpha = .57$, $n = 7$) than the unscreened condition ($\alpha = .28$, $n = 30$).

Second, workers screened with a pre-qualification task showed substantial correlation with expert scores. On average, the correlation values for screened workers was significantly higher than those in the unscreened condition, $M = .74$, $SD = .17$ for screened; $M = .45$, $SD = .25$ for unscreened; $U = 176$, $p = .004$ (Table 1).

Third, screened workers had less deviation in their scores from expert coders (Table 1). Scores from the screened workers deviated less from the expert scores, Screened: $M = .97$, $SD = .28$, Unscreened: $M = 1.42$, $SD = .30$, $U = 28$, $p = .003$. Finally, principles of collective intelligence worked more for screened workers. As more screened workers scored a task, their evaluations grew closer to expert scores. Figure 3 demonstrates a simulation for aggregating the worker scores in screened versus unscreened conditions (screened: red line; unscreened: blue line). We observed that the aggregated output tended to get closer to the expert scores (i.e., smaller RMSE) as we aggregated more workers in the screened condition, but not the unscreened condition.

A limitation to the study design is that fewer workers than we had anticipated reached the accuracy threshold, resulting in unbalanced sample sizes between the screened and unscreened conditions (n screened = 7; n unscreened = 30). We applied non-parametric tests to address this sample imbalance. In future iterations, we plan to replicate the study design with a larger sample of workers.

In sum, findings indicate the potential of crowdsourcing in scoring short, open-ended essays. Although we did not find any differences in performance between text and multimedia instruction, our findings suggest that screening workers may produce higher quality results more consistently. Furthermore, we found that screening workers could be cost-efficient. With our data set of 800 student essays, using two screened workers for every 10 essays costs about \$440. This cost to process data for the whole school is more economical than paying teachers for overtime work like the school normally had done (\$50/teacher/hour, several hours) or developing a computational system.

4 STUDY 2: ASSESSING SYSTEMS THINKING

4.1 Context

The second study was part of a project in middle school science, aimed at helping learners develop an understanding of causal links in local ecosystems through exploring decomposition processes. A learning goal for educators in this situation, was that students could begin to describe ecosystems as cyclical causal links (e.g., carbon cycle) instead of linear (e.g., plants need water). Students were asked to write essays before and after the curriculum about causal links in decomposition, for example, answering how a surge in predators can result in fewer animals and more plants, affecting the nutrition and decomposition rates. Due to time constraints in the science unit, which took place every day for a short, three-week time window, our partner teachers were only able to mark the essays as complete or not, instead of grading the essays for their argumentative content.

As a pilot to develop a processing pipeline for the assessment, we recruited participants from three Education undergraduate courses ($n = 168$) at a Hispanic-serving university in the U.S. We prepared 5 assessment sets, each containing 3 student essays that represented a range of complexity in causal thinking. We verbally confirmed with participants that they had prior experience with using or scoring educational assessments.

Similar to Study 1, designing the crowdsourcing task was a non-trivial process of developing a clear scoring rubric, designing different task design features, and determining metrics for worker performance. Building on the extant literature in crowdsourcing design, we tested another task feature: whether self-explaining why a response would receive certain scores enhanced the accuracy of the crowdsourced assessments [15]. In addition, because the study took place in Education classes for pre-service educators who may teach similar science content in the future, we were curious to explore whether participating in the crowdsourcing tasks deepened participants' conceptual understanding. Our experiment builds toward designing crowdsourcing to provide learning value to the crowd, to position them more than rote labor for a system [4, 21]. The following hypotheses guided the task design and analyses:

H4. Crowdsourcing potential. We hypothesized that crowdsourced scores would show substantial agreement with the expert scores.

H5. Task design: Explanation. Asking participants to explain their answers would yield higher quality work.

H6. Learning. There would be learning benefits from crowdsourcing, compared to not participating in the task.

4.2 Crowdsourcing Task Design

Our crowdsourced task included a clear scoring rubric and a training phase. Participants received extra credit in their college course for completing all tasks, as outlined below.

Baseline. Participants answered 20 questions to establish baseline knowledge about ecosystem 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 4). The concept map illuminated the types of elements and causal links we wanted the middle school students to show in their responses.

Experiment. Participants were randomly divided into:

- **Control group ($n = 71$):** watched a three-minute video about the decomposition process and did not grade.
- **Grade Only group ($n = 45$):** Participants were randomly assigned a set of student responses (panel B, Figure 4).
- **Grade & Explain group ($n = 45$):** To test the benefit of self-explanation for accuracy, we included another grading condition where participants explained the rationale for their scores (highlighted in yellow; Figure 4).

Post-test. Participants in all 3 experimental groups answered a post-test about science knowledge and their science attitudes [8]. The post-test prompted participants to explain the population balance in a hypothetical ecosystem and provide examples of how they might teach causal links to a fifth grader. The post-test's elements and causal links paralleled those in the pre-test to assess the same domain understanding.

4.3 Rubric Development

The crowd participants scored student responses along two dimensions: Elements and Coherence. For **Elements**, participants counted the target vocabulary (the training concept map; panel B, Figure 4) that appeared in students' responses. For **Coherence**, participants determined whether the responses reflected coherent links. For example, a score of 0 (lowest) would show only linear links (e.g., rabbits eat grass). A score of 2 (highest) 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 non-linear links, but incomplete evidence. The rubric was developed based on frameworks of science argumentation [34]. Prior to the crowdsourcing task, the second author and a research assistant coded 25% of the student responses to establish inter-rater reliability for the rubric: Cohen's $\kappa = .73$ for Elements and .88 for Coherence.

4.4 Defining Metrics for Accuracy and Learning

Similar to Study 1, we defined performance as the extent to which the crowd scores agreed with one another (inter-rater reliability) and how they aligned with experts (correlation and RMSE). As we turned to theories of collective intelligence [28], we aggregated the output from 1-5 participants who were randomly drawn without replacement from the crowdsourcing pool, and calculated how much the average scores deviated from expert scores. The analysis aimed to understand the impact on the overall task quality of adding more individuals to the collective tasks.

4.5 Validity Checks with Algorithmic Approaches

When using practical analytics to inform instructional decisions, educators can self-process data, consult external evaluators, or employ computational tools such as automated scoring of student work. In this study, we were also curious to explore how our alternative approach of crowdsourcing compared to automated, algorithmic approaches, in their correlation with expert scores. Researchers have suggested that fine-tuning algorithmic approaches to score non-standardized data is a nontrivial task [3]. Instead of training a

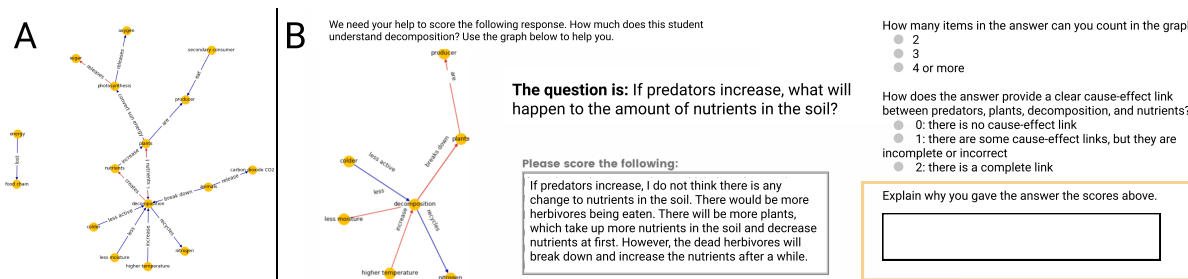


Figure 4: Instruction for the grading task

scoring model from scratch, given the time and resource constraints of our educator partners, we purposefully selected off-the-shelf algorithmic approaches based on two main criteria: (1) the approach did not require time- and cost-intensive fine-tuning; and (2) the approach has been tested with similar educational tasks or domains.

For **Elements**, because the curriculum covered a list of target vocabulary (e.g., plant, nutrients), we wrote a script to count the occurrences of the target words in students' responses. Similar to the human-scoring rubric, essays with fewer than two target elements received a score of 0, 1 for two to four elements, and 2 for more than four elements.

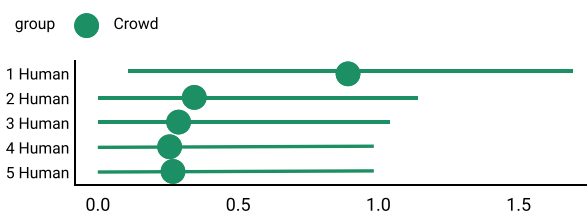
For **Coherence**, we wrote a "model" response for training the algorithms that contained all the causal links in the ecosystem unit (e.g., "dead matter becomes nutrients"). We then calculated the semantic similarity between student answers and the model response based on the overlapping words and sentence contexts. We used two main approaches: word embedding with word2vec and sentence embedding with Universal Sentence Encoder [7].

The word embedding approach took the average of all word embeddings within a sentence, and calculated the Cosine similarity between sentences. Vectors that represent sentences with similar meaning will occupy closer spatial positions, and will thus have similarity closer to 1. However, word embedding does not account for the order in which words appear and potential semantic differences (e.g., "rabbits eat plant" vs "plants eat rabbits"). Encoders that have been trained on word orders, such as Universal Sentence Encoder (USE), can address this constraint [7]. For each approach, we obtained the semantic similarity scores (range of 0 to 1, with 1 suggesting perfect similarity), and calculated their Pearson's correlations with the expert scores. We later compared the correlation values with the crowdsourced results.

Learning Benefits. To parallel the grading rubric, 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 Benjamini-Hochberg procedures (false discovery rate of .10).

4.6 Findings

H4. Crowdsourced scores showed substantial agreement and high correlations with expert scores. The crowdsourced scores

Figure 5: Aggregating Output from Multiple Crowdworkers
Reduced Disagreement with Expert Scores

Notes. Whiskers = 90% Confidence Interval. RMSE is always > 0.

were substantial in terms of agreement among raters (average Krippendorff's α across the five test sets: Overall: $M = .61$, $SD = .02$; Grade only: $M = .66$, $SD = .06$; Grade & explain: $M = .59$, $SD = .07$). The inter-rater reliability values in our sample were higher than the range of .40 to .60 in prior crowdsourcing work with crowdworkers [2]. This finding indicates the potential benefit of recruiting from the crowd with prior domain knowledge.

In addition, the crowdsourced scores were overall strongly correlated with the expert scores (M of Pearson's $r = .76$, $SD = .15$). As a validity check, we compared these values to results from the algorithmic approaches. The correlation values for Coherence were higher than those observed in the algorithmic approaches for word embedding ($r = .14$) and sentence embedding ($r = .66$). Meanwhile, we found that the correlation between crowdsourced scores and expert scores for Element was quite high and did not seem to differ from the computational approach of matching string expressions (Crowdsourced $r = .83$, $SD = .12$; Computation: $r = 1.00$). It is quite likely that the task's instruction to count occurrences of target words was straightforward enough for both human and algorithmic analyses. We discuss the implications of human-computation versus algorithmic approaches in more details in the Discussion.

We also turned to simulating the aggregated output from 2-5 crowdsourcing individuals to explore whether we could improve the crowdsourced performance and reduce the RMSE values (Figure 5). On average, we observed that the aggregated scores were closer to the expert scores than those obtained from one individual alone.

H5. Scores in the grade only condition 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

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

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, $M = .78$, $SD = .21$, $W = 1411.5$, $p < .001$. Contrary to our hypothesis, this finding suggests that overall, the scores in Grade Only condition were closer to the expert scores.

H6. There were learning benefits to crowdsourcing. We found that grading facilitated learning for the workers, 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, $\beta = .42$, $SE = .18$, 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). In short, results from the second study suggest the potential of crowdsourcing in scoring education data that require interpretation. Findings also suggest the learning potential of crowdsourcing tasks, an underexplored area in LA.

5 DISCUSSION AND IMPLICATIONS

Processing practical data, which are abundant in schools but require substantial resources and expertise, is a non-trivial task. Our studies illustrate **the potential to leverage different crowds** – workers on a general crowdsourcing platform and students with prior knowledge in education assessment – in processing practical analytics. We note that the task requirements, worker qualifications, and the computational approaches we selected were different in the two studies, and may have implications for the worker performance that we observed. Across both cases, however, we developed an efficacy-case to show that aggregated scores of a crowd can be accurate, and in line with domain experts scores. These findings illuminate the potential of crowdsourcing to process non-standardized, constantly evolving formative assessments. As learning analytics researchers seek to develop approaches for processing practical data, crowdsourcing may offer a promising venue to validate and improve scoring approaches for local, idiosyncratic, and personally relevant forms of data to teachers.

Our design cases also validate and test prior research findings on designing crowdsourcing tasks, particularly in education domains. **Domain knowledge appeared to be key to worker quality.** Overall, screened workers (who showed prior performance with a

task) and recruited education students (who had some background knowledge about grading student work) had higher correlations (and lower disagreement) with ground-truth scores.

Our experiments with task design also highlight the need to explore design guidelines for crowd tasks in the education domain. In the first study, we found that the practice of **screening workers resulted in higher reliability and accuracy of scores**. Interestingly, the modality of instructions and training (text vs. multimedia) did not seem to matter for workers' ability to reliably and accurately score the assessments. In the second study, contrary to our hypothesis, **we did not find an association between asking workers to self-explain their assessments and better performance**. Our experience suggests that well-designed, and scoped, scoring tasks in education domains may differ from other types of microtasks in the HCI literature, and thus warrant future research to investigate design principles for these domains.

We found that **the crowdsourcing tasks may provide learning value to crowdworkers**. In our review of the crowdsourcing literature, we found limited discussion of how crowdsourcing in education may provide content knowledge for individuals who participated in the crowd tasks. For instance, although not in educational domains, [11] show learning benefits of crowdsourcing tasks. Our findings suggest a potential direction for LA research in human-computation. In rethinking the design of crowdsourcing tasks to be person-oriented, designers may explore the pedagogical value of the tasks to provide meaningful learning opportunities for all individuals involved [22]. From an ethical perspective, integrating learning into crowdsourcing platforms positions workers as more than rote elements of data processing pipelines [4].

In this study, we focused on designing and evaluating the potential for systems to process practical analytics, which can be used to inform pedagogical decisions. The more fine-grained assessment results obtained from the crowdsourcing tasks can help educators to identify concrete areas where students need improvement, and develop more relevant decisions to target these practices. Examining how the analytics from the crowdsourcing tasks can be utilized in instructional practices is a pertinent next step for our research.

Our findings illustrate that **crowdsourced evaluation will result in variation in scores on educational assessments, regardless of task design**. That is, scores are not always accurate. Thus, a compelling area of future research lies in understanding how to represent this variance to educators (who want to act on this data). In preliminary conversations with three educator partners, we presented the crowdsourced output and variation from our various studies. Our partners voiced the desire to view the analysis process – how the crowdworkers were involved, how their scores were used toward the final representation, and the levels of uncertainties associated with the crowdsourced output. This experience points to future research needs around data transparency and uncertainty.

6 LIMITATIONS

We note several limitations of our series of design experiments to contextualize our insights. First, there were unbalanced samples in the screened and unscreened conditions in Study 1. The unbalanced samples arose because the number of workers who

were qualified after the screening task were lower than expected. We employed non-parametric analyses to address this imbalance. Second, our computational approaches served as baselines to compare with crowdsourced performance, and were thus not extensive. Follow-up studies should expand the task designs and explore human-computation strategies to improve the accuracy of assessment scores.

7 CONCLUSION AND NEXT STEPS

Overall, this paper demonstrates how crowdsourcing principles can be employed to process practical analytics in education contexts. Follow-up work can replicate the study with other educational data and contexts. Researchers can also leverage research in crowdsourcing and learning sciences, among other disciplines, to understand other design features that we did not explore here. For example, would providing in-time expert feedback increase workers' expertise and performance? Also, by bringing external parties (e.g., researchers) into the process of creating local analytics, one might introduce misinterpretation of the metrics. Future research may consider how to mitigate these misinterpretations, beyond partnering with educators. Finally, our discussion highlights issues of developing pipelines for crowdsourced assessments and representing the output back to educators in practice. Our next steps involve examining ways to integrate the crowdsourced assessments into teacher routines and how teachers can apply the crowdsourced results for instructional improvement. These research directions may benefit the broader education community, if we were to leverage crowdsourcing applications for processing the wealth of practical analytics in education settings.

ACKNOWLEDGMENTS

We thank the participants who dedicated their time to this study and our educator partners, who inspired this work.

REFERENCES

- [1] Laura K Allen, Matthew E Jacovina, and Danielle S McNamara. 2016. Computer-Based Writing Instruction. *Grantee Submission* (2016).
- [2] Omar Alonso. 2019. The Practice of Crowdsourcing. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 11, 1 (2019), 1–149.
- [3] Dmytro Babik, Lakshmi S. Iyer, and Eric W. Ford. 2012. Towards a Comprehensive Online Peer Assessment System. Springer, Berlin, Heidelberg, 1–8. https://doi.org/10.1007/978-3-642-29863-9_11
- [4] Natã M Barbosa and Monchu Chen. 2019. Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [5] Anthony S. Bryk, Louis M. Gomez, Alicia Grunow, and Paul G. LeMahieu. 2015. *Learning to improve: how America's schools can get better at getting better*. Harvard Education Press, Cambridge, MA.
- [6] Julia Cambre, Scott Klemmer, and Chinmay Kulkarni. 2018. Juxtapaper: 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. <https://doi.org/10.1145/3173574.3173868>
- [7] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, and Chris Tar. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).
- [8] 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.
- [9] Florian Daniel, Pavel Kucherbav, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing. *Comput. Surveys* 51, 1 (1 2018), 1–40. <https://doi.org/10.1145/3148148>
- [10] Luca de Alfaro and Michael Shavlovsky. 2014. CrowdGrader. In *Proceedings of the 45th ACM technical symposium on Computer science education - SIGCSE '14*. ACM Press, New York, New York, USA, 415–420. <https://doi.org/10.1145/2538862.2538900>
- [11] 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. <https://doi.org/10.1145/2858036.2858268>
- [12] 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 learn- ersourcing. In *Proceedings of International Conference of the Learning Sciences, ICLS*.
- [13] 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. <https://doi.org/10.1145/2470654.2470686>
- [14] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. ACM Press, New York, New York, USA, 1013. <https://doi.org/10.1145/2145204.2145355>
- [15] 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*.
- [16] Julie A. Farrell, Caitlin C.; Marsh. 2016. Contributing conditions: A qualitative comparative analysis of teachers' instructional responses to data. *Teaching and Teacher Education* 60 (11 2016), 398–412. <https://doi.org/10.1016/j.TATE.2016.07.010>
- [17] Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher* 40, 5 (2011), 223–234.
- [18] Neil T. Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *International Journal of Artificial Intelligence in Education* 24, 4 (12 2014), 470–497. <https://doi.org/10.1007/s40593-014-0024-x>
- [19] Ilana Seidel Horn, Britnie Delinger Kane, and Jonee Wilson. 2015. Making sense of student performance data: Data use logics and mathematics teachers' learning opportunities. *American Educational Research Journal* 52, 2 (2015), 208–242.
- [20] Michelle Ichinco. 2014. Towards crowdsourced large-scale feedback for novice programmers. In *2014 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 189–190. <https://doi.org/10.1109/VLHCC.2014.6883049>
- [21] 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>
- [22] Eunice Jun, Morelle Arian, and Katharina Reinecke. 2018. The potential for scientific outreach and learning in mechanical turk experiments. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. 1–10.
- [23] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J. Guo, Robert C. Miller, Krzysztof Z. Gajos, Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J. Guo, Robert C. Miller, and Krzysztof Z. Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. ACM Press, New York, New York, USA, 4017–4026. <https://doi.org/10.1145/2556288.2556986>
- [24] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*. ACM Press, New York, New York, USA, 453. <https://doi.org/10.1145/1357054.1357127>
- [25] Richard Klein, Angelo Kyrilov, and Mayya Tokman. 2011. Automated assessment of short free-text responses in computer science using latent semantic analysis. In *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education*. 158–162.
- [26] Pushkar Kolhe, Michael L. Littman, and Charles L. Isbell. 2016. Peer Reviewing Short Answers using Comparative Judgement. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16*. ACM Press, New York, New York, USA, 241–244. <https://doi.org/10.1145/2876034.2893424>
- [27] Andrew E. Krumm, Rachel Beattie, Sola Takahashi, Cynthia D'Angelo, Mingyu Feng, and Britte Cheng. 2016. Practical Measurement and Productive Persistence: Strategies for Using Digital Learning System Data to Drive Improvement. *Journal of Learning Analytics* 3, 2 (9 2016), 116–138. <https://doi.org/10.18608/jla.2016.32.6>
- [28] Edith Law and Luis von Ahn. 2011. Human Computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5, 3 (6 2011), 1–121. <https://doi.org/10.2200/S00371ED1V01Y201107AIM013>
- [29] Catherine Lewis. 2015. What Is Improvement Science? Do We Need It in Education? *Educational Researcher* 44, 1 (1 2015), 54–61. <https://doi.org/10.3102/0013189X15570388>
- [30] Richard E. Mayer. 2002. Cognitive Theory and the Design of Multimedia Instruction: An Example of the Two-Way Street Between Cognition and Instruction. *New Directions for Teaching and Learning* 2002, 89 (2002), 55–71. <https://doi.org/10.1002/tl.47>

- [31] Elijah Mayfield and Alan W Black. 2020. Should You Fine-Tune BERT for Automated Essay Scoring?. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 151–162.
- [32] Tanushree Mitra, C.J. Hutto, and Eric Gilbert. 2015. Comparing Person- and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. ACM Press, New York, New York, USA, 1345–1354. <https://doi.org/10.1145/2702123.2702553>
- [33] Maletsabisa Molapo, Chane Simone Moodley, Ismail Yunus Akhalwaya, Toby Kurien, Jay Kloppenberg, and Richard Young. 2019. Designing Digital Peer Assessment for Second Language Learning in Low Resource Learning Settings. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale - L@S '19*. ACM Press, New York, New York, USA, 1–13. <https://doi.org/10.1145/3330430.3333626>
- [34] Ha Nguyen and Rossella Santagata. 2020. Impact of computer modeling on learning and teaching systems thinking. *Journal of Research in Science Teaching* (10 2020), tea.21674. <https://doi.org/10.1002/tea.21674>
- [35] Alexander J. Quinn and Benjamin B. Bederson. 2011. Human computation. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. ACM Press, New York, New York, USA, 1403. <https://doi.org/10.1145/1978942.1979148>
- [36] 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.
- [37] Emily Saxton, Secret Belanger, and William Becker. 2012. The Critical Thinking Analytic Rubric (CTAR): Investigating intra-rater and inter-rater reliability of a scoring mechanism for critical thinking performance assessments. *Assessing Writing* 17, 4 (10 2012), 251–270. <https://doi.org/10.1016/j.ASW.2012.07.002>
- [38] Kim Schildkamp. 2019. Data-based decision-making for school improvement: Research insights and gaps. *Educational research* 61, 3 (2019), 257–273.
- [39] Kim Schildkamp and Wilmad Kuiper. 2010. Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education* 26, 3 (4 2010), 482–496. <https://doi.org/10.1016/J.TATE.2009.06.007>
- [40] Einar M. Skaalvik and Sidsel Skaalvik. 2011. Teacher job satisfaction and motivation to leave the teaching profession: Relations with school context, feeling of belonging, and emotional exhaustion. *Teaching and Teacher Education* 27, 6 (8 2011), 1029–1038. <https://doi.org/10.1016/J.TATE.2011.04.001>
- [41] Richard J Stiggins. 2002. Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan* 83, 10 (2002), 758–765.
- [42] Peter S. Wardrip and Phillip Herman. 2018. ‘We’re keeping on top of the students’: making sense of test data with more informal data in a grade-level instructional team. *Teacher Development* 22, 1 (2018), 31–50. <https://doi.org/10.1080/13664530.2017.1308428>
- [43] David Yeager, Anthony Bryk, Jane Muhich, Hannah Hausman, and Lawrence Morales. 2013. *Practical Measurement*. Technical Report. Carnegie Foundation for the Advancement of Teaching, 78712, Palo Alto, CA.
- [44] Phillip Yenawine. 2013. *Visual Thinking Strategies: Using Art to Deepen Learning Across School*. Harvard Education Press, Cambridge.