Exploring Trade-offs Between Learning and Productivity in Crowdsourced History

NAI-CHING WANG, Department of Computer Science, Virginia Tech, USA
DAVID HICKS, School of Education, Virginia Tech, USA
KURT LUTHER, Department of Computer Science, Virginia Tech, USA

Crowdsourcing more complex and creative tasks is seen as a desirable goal for both employers and workers, but these tasks traditionally require domain expertise. Employers can recruit only expert workers, but this approach does not scale well. Alternatively, employers can decompose complex tasks into simpler microtasks, but some domains, such as historical analysis, cannot be easily modularized in this way. A third approach is to train workers to learn the domain expertise. This approach offers clear benefits to workers, but is perceived as costly or infeasible for employers. In this paper, we present CrowdSCIM, a workflow that teaches domain expertise (historical thinking skills) to novice crowd workers. We compare CrowdSCIM with two crowd learning techniques from prior work and a baseline to explore the trade-offs between learning and productivity. Our evaluation (n=360) shows that CrowdSCIM allows workers to learn domain expertise while producing work of equal or higher quality versus other conditions, but efficiency is slightly lower.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Collaborative and social computing systems and tools**; *Computer supported cooperative work*; • **Applied computing** → *Interactive learning environments*; Arts and humanities;

Additional Key Words and Phrases: Crowdsourcing; learning; productivity; history; historical thinking; analytical thinking; domain expertise; SCIM-C

ACM Reference format:
Nai-Ching Wang, David Hicks, and Kurt Luther. 2018. Exploring Trade-offs Between Learning and Productivity in Crowdsourced History. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 178 (November 2018), 23 pages.
https://doi.org/10.1145/3274447

1 INTRODUCTION

Crowdsourcing more complex and creative tasks has become a desirable goal for both employers and workers [29]. Employers can crowdsource more tasks with high scalability on demand, while workers have more and greater varieties of tasks from which to choose.

However, complex tasks traditionally require domain expertise. To complete these tasks, employers can recruit expert workers, but this approach does not scale well because relatively few workers may have the required expertise compared to the large number of available tasks. Alternatively, employers can decompose complex tasks into simpler micro-tasks, design workflows, and aggregate results from these tasks. However, in some domains, such as historical analysis, work

2573-0142/2018/11-ART178 $15.00
https://doi.org/10.1145/3274447

cannot be easily modularized in this way (e.g.,[5, 9, 51, 52]). Even when decomposition is feasible, studies show that experts may have to invest significant time and effort to design an effective crowdsourcing workflow (e.g., [6, 7, 30, 35, 36]). A third approach is to train workers to learn new skills or knowledge required to complete the task. This approach offers clear benefits to workers, but may be perceived as costly or infeasible for employers.

In this paper, we built on this third research direction. Prior work has explored techniques for helping crowds learn, but primarily in narrow, goal-oriented contexts, such as learning a specific task [13, 57] or memorizing factual domain knowledge [34]. We considered how crowds might learn domain expertise, i.e. analytical skills and thinking strategies, that may support the task at hand, but also provide more generalized, transferable value to the worker beyond the immediate task. Our work seeks to help bridge the gap between task-oriented learning for paid crowds and the deeper online learning experiences of students participating in massive open online courses (MOOCs).

Our investigation focused on the task domain of analyzing historical documents. Although historians report a variety of challenges in conducting research online [42] and the benefits of crowdsourced support for historical research are increasingly recognized [33], few studies have systemically evaluated the potential of crowdsourced history support. Further, most crowdsourced history projects focus on transcription of primary sources, an important but relatively simple task requiring low domain expertise. We focused on more complex tasks such as writing summaries of historical documents and evaluating their relevance to high-level themes of interest to historians.

To help paid novice crowd workers learn domain expertise, we drew inspiration from a scaffolding technique called SCIM-C [22], originally developed for students in traditional schools to learn historical thinking skills. Through an iterative design process and several pilot studies, we adapted SCIM-C for novice crowd workers to create a microtask workflow called CrowdSCIM.

We compared CrowdSCIM with two other crowd learning techniques from prior work [13, 57], plus a baseline, to understand the trade-offs between learning and productivity in training crowds to analyze historical documents. Our evaluation with 360 crowd workers from Amazon Mechanical Turk (AMT) shows that CrowdSCIM allows workers to learn domain expertise while producing work of equal or higher quality compared to other approaches, though efficiency is slightly lower. Although our empirical results focus on historical document analysis, we propose a generalized workflow for teaching domain expertise within microtasks, towards the goal of more meaningful, beneficial experiences for crowd workers.

2 RELATED WORK

2.1 Historical thinking and history education

A Library of Congress publication colorfully depicts historians as detectives searching for evidence among primary sources [53]. Learning history is more than merely memorizing facts from various sources. Although historians have expertise in different periods or topics of history, they share some common way of thinking history and analyzing historical documents [54]. The analysis of sources includes identifying factual information, evaluating reliability of sources, understanding multiple perspectives, contextualizing sources in time and space, reasoning and inferences, corroborating across multiple sources, and generating possible understandings and interpretations [5, 9, 51, 52].

Several learning approaches have been proposed to help students learn history through historical thinking, such as learning through authorship [20], apprenticeships (or guidance) [10, 43], and confronting questions [55]. These strategies generally require substantive interactions between a human instructor and a student. Some studies have shown that the use of hypertext scaffolding

may support some historical thinking processing [24, 25]. Building on these studies, Hicks et al. developed SCIM-C [22], a strategy that can scaffold the historical thinking process when analyzing historical primary sources. It includes five phases: 1) Summarizing information and evidence from the source, 2) Contextualizing the source in time and space, 3) Inferring from subtexts and hints in the source, 4) Monitoring initial assumptions and overall focus, and 5) Corroborating understanding across multiple sources. Evaluations showed that SCIM-C is an effective strategy to help students learn historical thinking through multimedia embedded scaffolding [21, 39]. While SCIM-C has been shown to be effective in the classroom with collocated students, experienced teachers, and multi-day training sessions, its applicability for novice crowd workers is unknown. This paper explores how SCIM-C can be adapted for a micro-tasking context, providing just-in-time domain expertise for workers completing tasks requiring historical thinking skills.

2.2 Crowd learning and learner-sourcing

Some research has begun exploring the use of crowdsourcing in classroom-related settings. This body of work focuses on improving learning with collective learner activity or receiving feedback from other (paid) crowds, including creating crowdsourced subgoals in how-to videos [27, 28, 49], crowdsourced assessments or exercises [40, 45], personalized hints for problem-solving [16] , receiving design critiques [19, 56], collaborative discussion [8], identifying students' confusions [15], and generating explanations for solving problems [50]. These studies try to address the issue of low ratios of expert teachers to learners, especially in MOOCs. This body of research is also termed learner-sourcing because it focuses on how learners can collectively generate useful learning materials for future learners.

CrowdSCIM differs from these studies in that CrowdSCIM is built on top of a scaffolding technique to be used without the need of other peer learners or crowds. While these learner-sourcing techniques require additional (learner) crowds' participation or content production (e.g., sub-goals in how-to videos, design critiques, and explanations) to facilitate learning, a CrowdSCIM user can learn historical thinking while doing tasks without feedback or participation from others. Further, CrowdSCIM is designed for paid crowd workers, a population with greater limitations of time, interest, and expertise, than students in MOOCs or traditional classrooms.

2.3 Crowd learning on citizen research platforms

While citizen research platforms like Zooniverse have attracted many non-professionals to contribute to major discoveries, these projects are also considered a means of engagement and outreach, such as citizen science and public history. Yet, recent studies show that learning often happens outside the context of the crowdsourced tasks [26]. Most relevant to our work, Crowdclass [34] was among the first efforts to design in-task learning modules for citizen science. Similar to our work, Crowdclass focuses on paid novice crowd workers and uses pre- and post-tests to measure learning. However, unlike our approach, Crowdclass focuses on learning factual knowledge; workers correctly answer multiple-choice questions (and "hybrid questions" synthesizing facts across multiple lessons) to demonstrate mastery and advance in a hierarchy of learning modules. In contrast, CrowdSCIM teaches workers to consider the meaning of a document from multiple perspectives by reflecting on a set of generalized questions. Although direct comparisons are complicated by differences in task and domain (i.e., analyzing historical documents vs. classifying galaxies), CrowdSCIM has the benefit of being content-agnostic within a domain. Crowdclass may require experts to design new questions and answers to teach and test each new type of fact, whereas CrowdSCIM does not require expert intervention when new documents are presented.

CrowdSCIM builds on Incite [3] for summary, tone rating, tag and theme rating in the history domain. Incite is an open-source crowdsourcing system designed to help historians analyze primary sources [3]. We chose Incite as our target crowdsourcing platform for a few reasons. First, it includes a variety of higher-level tasks to support historical research, including summaries, tone ratings, tags, and theme ratings, in contrast to most crowdsourced history platforms that focus on simple transcription (e.g., [1, 4]). Incite groups these tasks into three steps: Transcribe (transcribe, summarize, and rate tone), Tag (tag entities), and Connect (rate theme). This selection of tasks is consistent with what prior work suggested in supporting historical research [14, 42]. Second, Incite has been used by real digital archive projects to support historical research [3]. Third, it is open-source and easy to plug into existing digital archives.

While Incite may also be used to support history education, CrowdSCIM differs in that Incite is optimized for students in classrooms with instructors' intervention over a multi-day time span, while CrowdSCIM is designed as a standalone system with novice, paid crowd workers and micro-tasks. Also, CrowdSCIM excludes the simple transcription task and focuses on higher-level Summary-tone, Tag, and Connect tasks.

2.4 Crowd learning and work quality

While most crowdsourcing studies focus on work quality, some research considers both worker learning and work quality [11–13, 57]. While most of these studies show that learning can help improve quality, others do not. Pandey et al. [41] found that workers who had access to MOOC-style learning materials about microbiomes scored higher on a subject matter test, yet produced similar work quality (i.e., generating creative ideas about microbiome influences) compared to workers without access to the learning materials. Crowdclass [34] shows that a workflow designed for learning may actually lower the work quality. These mixed results motivate our current study.

To understand CrowdSCIM's potential trade-offs between learning, quality and efficiency, we selected two of the most similar approaches from prior work, Reviewing vs. Doing (RvD) [57] and Shepherd [13] as comparison conditions. In RvD, Zhu et al. [57] found that workers who review others' work perform better on subsequent tasks than workers who simply performed more tasks. They theorize that reviewers experience learning benefits seen in offline studies of mentorship. In Shepherd, Dow et al. [13] compare the performance of workers receiving no feedback to workers who either perform a self-assessment using a rubric, or receive an external assessment from an expert. Self-assessment was as effective as expert assessment in improving work quality.

Aiding the comparison to CrowdSCIM, both prior studies reported work quality and learning in detail, and both included some type of writing or summarization tasks, though in different domains. However, it is not clear which technique works better and how they are applicable to other types of tasks and domains. Moreover, both RvD and Shepherd focus on learning the task through provided rubrics, while CrowdSCIM focuses on learning domain expertise through analytical thinking skills. This comparison supports a close examination of which type of learning is more effective for gaining domain expertise (i.e., historical thinking).

Following this literature, we compared how CrowdSCIM, RvD, Shepherd, and a baseline technique affect learning, quality and efficiency on selected crowdsourced tasks (Summary-tone, Tag and Connect) in the history domain. We asked the following research questions:

- RQ1: How do these techniques affect *learning historical thinking* for each of the crowdsourced tasks?
- RQ2: How do these techniques affect *quality of work* for each of the crowdsourced tasks?
- RQ3: How do these techniques affect *efficiency* for each of the crowdsourced tasks?

3 CROWDSCIM

3.1 Pilot studies

Our goal was to design a crowdsourcing technique to help crowd workers learn historical thinking while working on tasks that may contribute to historical research. To achieve this goal, the crowdsourcing technique had to support learning gains without impeding work quality. We began with a workflow resembling Incite and, through the series of pilot studies reported below, made iterative changes to arrive at the current CrowdSCIM workflow.

3.1.1 Incite outside of classrooms (Pilot Study 1). Our first step was to see how well Incite could support in learning historical thinking outside of the classroom and without an instructor's intervention. We first customized Incite for the lab study and tested Incite's workflow on the Amazon Mechanical Turk (AMT) paid crowdsourcing platform. We focused on Summary-tone, Tag and Connect tasks. We first asked each participant to complete a pre-test that involved demonstrating historical thinking skills by writing an interpretation of a historical document. The participant then completed the three crowdsourced tasks in sequence, and finally completed a post-test. The tests required writing a historical interpretation for the given primary source. We measured learning by comparing the pre- and post-test scores using rubrics from prior work [21]. Pilot tests with seven participants showed no learning gains from the scores or from their verbatim feedback. In other words, simply giving Incite to crowd workers did not promote learning.

3.1.2 SCIM intervention (Pilot Study 2). To try to increase learning, we reviewed the social science education literature and identified the SCIM-C framework [22] as a promising candidate to be adapted for crowd workers. We modified our workflow to add reflective questions from the SCIM-C framework that prompted participants to think about the meaning of the historical document from different perspectives. To minimize the gap between the crowdsourced tasks and SCIM questions, we matched the tasks and questions based on similarity in collaboration with a history professor, Historian A. Specifically, we matched Summary-tone with Summarize because both require a good summary of the original text. We matched Tag with Contextualize because both ask users to identify entities such as location and time. And we matched Connect with Infer and Monitor because it requires a solid understanding and inference to see if high-level topics are relevant to a given historical document.

We tested this revised task design with nine participants from AMT using the same procedure as before. The results again showed no learning effects, suggesting the unmodified SCIM framework is not (directly) applicable to the crowdsourcing context.

We observed that although there was no significant learning between tests, the participant's answers often included valuable content that the participant should have included in the post-test. The results seemed to suggest that the participant was able to find required information, but just did not know how to synthesize it in the post-test. Worker feedback also suggested that the task seemed too big for a micro-tasking context. We concluded that it might be too much to ask a crowd worker to complete tasks and learn all four phases of SCIM in one shot and apply all of them in the post-test.

3.1.3 Micro-task design with in-task practice (Pilot Study 3). To reduce the workload, improve focus, and increase flexibility, we revised the crowdsourcing workflow to decompose the process into one task at a time. We began with the Tag task because it showed the lowest learning scores for Contextualizing in Pilot Study 2. Our revised procedure again began with the pre-test. The worker made a first attempt at the crowdsourcing task (Tag, in this case). Then, the worker answered a series of reflection questions from the corresponding SCIM phase (Contextualize) and

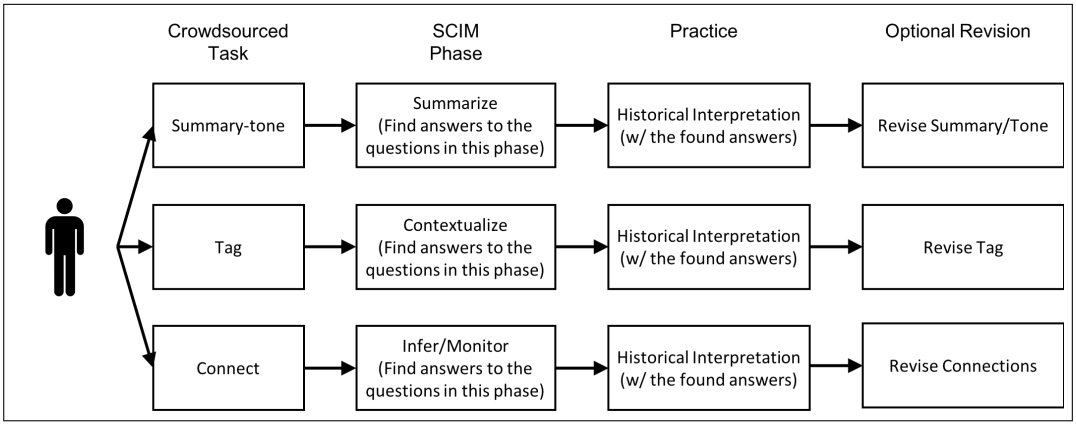

Fig. 1. The CrowdSCIM workflow

practices writing an interpretation. Next, the worker had the option to revise their Tag task, hopefully incorporating the historical thinking skills from the scaffold. Finally, the worker completed the post-test.

We tested this design with six participants from AMT. The results showed a significant learning effect corresponding to Contextualize phase in SCIM with a large effect size (>1.0) for both in-task and post-test interpretations. The learning effect was slightly higher in in-task practice than in the post-test, an expected result when the scaffold is "faded." Based on these promising results, we tested this workflow with the Summary-tone and Connect tasks, each with five or six participants, and observed similar patterns.

3.2 Final workflow

Our iterative process led to the final design of CrowdSCIM, a workflow to help crowd workers learn historical thinking skills while performing micro-tasks supporting historical research. CrowdSCIM consists of three micro-tasks: Summary-tone, Tag and Connect, corresponding to the Summarize, Contextualize, and Infer and Monitor phases in SCIM-C, respectively. With this decomposition, each crowdsourced task can be performed individually and each of the phase of SCIM can be learned separately.

For the Summary-tone task, the user writes a summary of the document and rates the intensity of each tone from a list. The user then answers four questions from the Summarize phase and writes a historical interpretation containing the answers. Finally, the user can choose whether to revise the summary and tone ratings.

For the Tag task, the user tags named entities with categories (e.g., politician, school) for a given primary source. The user then answers the four questions from the Contextualize phase. The user then writes a historical interpretation containing the answers. Finally, the user can revise the tags, if desired.

For the Connect task, the user rates relevance of the given historical primary source to each high-level theme in a list. The user then answers four questions from the Infer and Monitor phases. (To balance the workload with the other tasks, we selected two distinctive questions from each phase.) The user then writes a historical interpretation containing the answers. Finally, the user can decide whether to revise the theme ratings.

Thus, the generalized CrowdSCIM workflow contains four steps (see Figure 1). First, the worker completes an unmodified production microtask (e.g., summarizing, tagging, or connecting). Second, the worker completes a SCIM phase prompting him or her to reflect on the task just completed by answering a set of questions. Third, the worker writes a historical interpretation synthesizing his or her answers to the questions. Finally, the worker has the option to apply his or her newly sharpened historical thinking skills to the initial production microtask by revising his or her work.

Because steps 2–4 in the CrowdSCIM workflow occur after the unmodified production task and do not require content knowledge, CrowdSCIM is relatively straightforward to implement as an enhancement of an existing crowdsourced history workflow. This "add-on" design offers several benefits for requesters. First, it does not require modifying the interface of the initial production task, reducing requester effort and risk. Second, it guarantees at least equal work quality (vs. not using CrowdSCIM) because the intervention is applied after the initial production task. As prior work suggests (e.g., [34]), a primarily learning-oriented workflow may lower work quality, discouraging adoption by requesters. Third, it can be easily turned on and off based on requester needs (see Section 6.4 for a discussion of trade-offs).

4 EVALUATING CROWDSCIM

To evaluate the effectiveness of CrowdSCIM, we conducted an experiment comparing CrowdSCIM to three other conditions in terms of learning, quality, and efficiency.

4.1 Apparatus and procedure

The experiment was conducted entirely online. After completing an online IRB-approved consent form, each unique participant was randomly assigned to one of the four crowdsourcing workflows: CrowdSCIM, RvD [57], Shepherd [13], and a baseline similar to Incite. The participant was also randomly assigned three different historical documents — one for pre-test, one for the task, and one for post-test — from a pool of five documents. The participant then used the web interface we developed to complete a three-stage work process.

First, the participant completed a pre-test on writing a historical interpretation for a historical document.

Second, the participant completed the randomly assigned task (Summary-tone, Tag, or Connect). Three of the conditions involved a three-step process: the initial task, an intervention, and an optional revision to the initial task. The CrowdSCIM intervention involved answering four questions derived from SCIM-C. The RvD intervention involved reviewing existing work from other participants. The Shepherd intervention involved self-assessing the participant's own work. The Baseline condition required completing only the task itself and had no intervention. Unlike the Baseline, other three conditions provided an option for the participant to revise the work after the intervention.

Third, the participant completed a post-test on writing another historical interpretation for a different historical document.

After the work process, the participant completed a post-task survey for demographic information and feedback.

4.2 Participants

We recruited novice crowd workers from Amazon Mechanical Turk (AMT). We restricted to US-only workers to increase the likelihood of English language fluency, with a 95% HIT (human intelligence task) minimum acceptance rate and 50 or more completed HITs. We recruited 360 workers and randomly assigned 30 to each of the three crowdsourced tasks of each of the four conditions

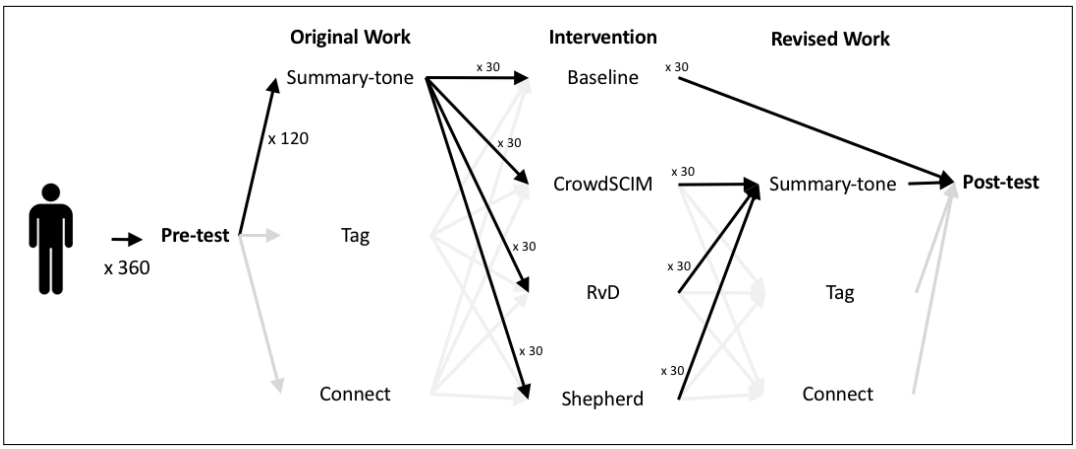

Fig. 2. Experimental design with the process of the Summary-tone task highlighted

(30 participants × 3 crowdsourced tasks × 4 techniques = 360). Each worker was unique and assigned to only one HIT to ensure that the required expertise was learned within that HIT. Thus, there were 30 unique workers per each crowdsourced task per crowdsourcing technique. We paid participants at least minimum wage ($7.25/hour) based on average task times in pilots.

4.3 Materials

To ensure the validity of our test materials, we used the same historical documents and grading rubric used in previous evaluations of SCIM-C [21]. The SCIM-C materials were selected, constructed, and tested by domain experts including a historian, a teacher educator, an educational psychologist, and a high school social studies teacher. The documents also cover a variety of eras and topics in American history, including the American Civil War, the American Revolution, the Great Depression, and Women's Rights. The random assignment and wide coverage of these documents helped eliminate the possibility that the potential effect was caused by some specific topic or document. Two of the sources are shown on the left panel in Figures 3 and 4. Some of the task outputs, such as tone ratings, tags, and connections, can be graded automatically if gold standard data is provided. Other task outputs, including historical interpretations and summaries, requires manual grading (see Appendices A and B for the rubrics).

4.4 Experimental design

We conducted a between-subjects GRCB (Generalized Randomized Complete Block) design with one treatment factor (crowdsourcing technique), one block factor (crowdsourced tasks), and three dependent variables (learning, quality, and efficiency). The overall experimental design is shown in Figure 2 where the process of the Summary-tone task is bolded.

4.4.1 Crowdsourced tasks (block factor). The crowdsourced tasks Summary-tone, Tag, and Connect were adapted from Incite. The Summary-tone task was to write a maximum three-sentence summary about the given historical document, and then, on a four-point Likert scale, rate how intensely each tone was expressed in the document from a given list of tones (e.g., informative, optimistic). The Tag task was to label named entities with correct categorical information (person, location, and organization). The Connect task was to rate how relevant each theme (e.g., Racial Equality) was to the given document on a 4-point Likert scale.

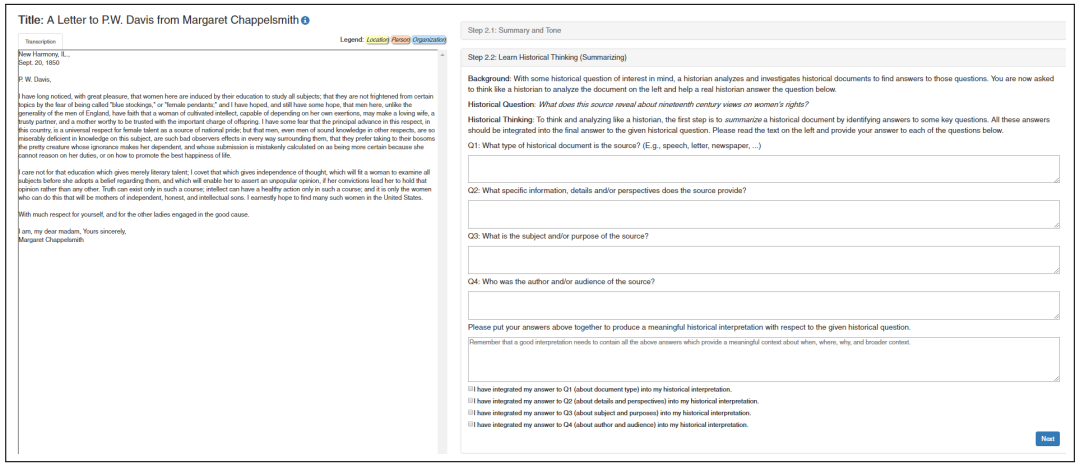

Fig. 3. A screenshot of the CrowdSCIM intervention for the Summary-tone task

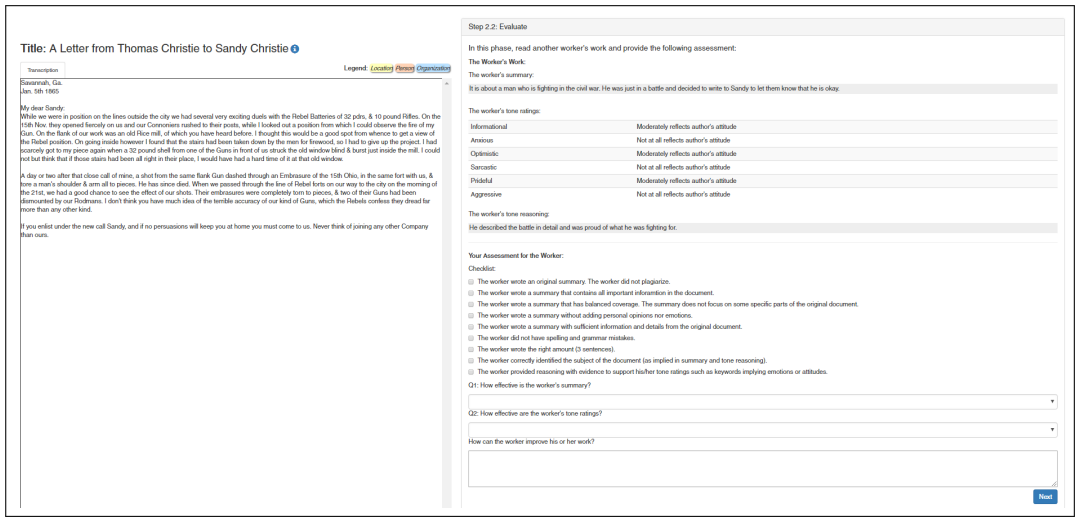

Fig. 4. A screenshot of the RvD intervention for the Summary-tone task

4.4.2 Crowdsourcing techniques (treatment factor). The independent variable, crowdsourcing technique, had four levels: CrowdSCIM, RvD, Shepherd, and Baseline. While Baseline did not contain any intervention nor revision steps, the other three levels contained an intervention and an optional revision step. The revision step is the same across all the three techniques, so the only difference is the intervention.

The Baseline did not contain any intervention nor revision steps. In addition to the pre- and post-tests, the participant was asked to complete the assigned crowdsourced task.

The CrowdSCIM intervention is described above. Participant answered four SCIM questions depending on the assigned task. Figure 3 includes the four questions of the Summarize phase corresponding to the Summary-tone task (see Appendix C for a complete list of these questions).

After the intervention, the participant had a chance to revise their response to the crowdsourced task if so desired.

The RvD and Shepherd mimicked the design from the original studies using the same rubric as the graders. After completing the crowdsourced task, both interventions asked the participant to assess the quality of work based on a given rubric. After the intervention, the participant also had a chance to revise their response to the crowdsourced task if so desired. The major difference was that RvD asked the participant to assess another participant's work, while Shepherd asked the participant to self-assess his or her own work. The RvD intervention is demonstrated in Figure 4. In the Shepherd intervention, we replaced "the worker" and "the worker's" with "I" and "my", as indicated in the original Shepherd study.

4.4.3 Dependent variables. To measure *learning*, we followed the same procedure used in previous SCIM-C studies (e.g., [21]) to compare the difference between the participant's score of the historical interpretation in the post-test and the pre-test. The interpretations were graded by two graders who were trained with the same materials used in previous SCIM-C studies and blind to the crowdsourcing techniques and crowdsourced tasks. The same grading rubric (see Appendix A for details) from prior SCIM-C studies was also used for grading. Interrater reliability was determined by comparing the graders' responses (binary yes/no) to the 12 scoring rubric questions across 60 interpretations from the pilot studies and calculating Cohen's Kappa. Cohen's Kappa ranges from 0.0 (agreement is no better than chance) to 1.0 (perfect agreement), and is appropriate for measuring interrater reliability for categorical data. The graders had a Kappa score of 0.89, indicating high reliability.

To measure the *quality of the summary*, we used the score of the summary. The summaries were graded by the same two graders who were blind to the crowdsourcing techniques. They used the rubric developed from previous work and guidelines gathered from school writing centers, and approved by a history professor, Historian B (see Appendix B for details). Interrater reliability was determined by comparing the graders' responses (yes or no) to the rubric questions across all unrevised summaries. The graders had a Kappa score of 0.83, indicating high reliability.

Quality was divided into three categories: low (0–3), medium (4–6), and high (7–10) based on a 10- point scale. A high quality summary contains no or minor issues that do not affect reading. A medium quality summary misses some important information, detail or context. A low quality summary misses much of important information, detail and/or context. These categories were nominal labels to help make sense of the scores, but we used raw scores for all of the following data analyses.

To measure the *quality of the tones*, we compared each worker's response with a gold standard response provided prior to the study by a history professor, Historian A. Specifically, we measured the Cohen's weighted kappa between the crowd's response and the gold standard response.

To measure the *quality of the tags*, we compared each worker's response with gold standard response also provided prior to the study by Historian A. Specifically, we measured the precision and recall of the tags created by the crowd.

To measure the *quality of connections* (i.e., theme ratings), similarly, we measured the Cohen's weighted kappa between the crowd's response and the gold standard response provided by Historian A.

To measure how the work quality is affected by the crowdsourcing technique, we calculated the difference between revised and original work as quality change, if applicable (there was no revision for the Baseline condition).

We also measured the crowd's efficiency in analyzing documents in terms of time and attempts as attrition. *Time* describes how long it takes for a task to be completed and is an indicator of

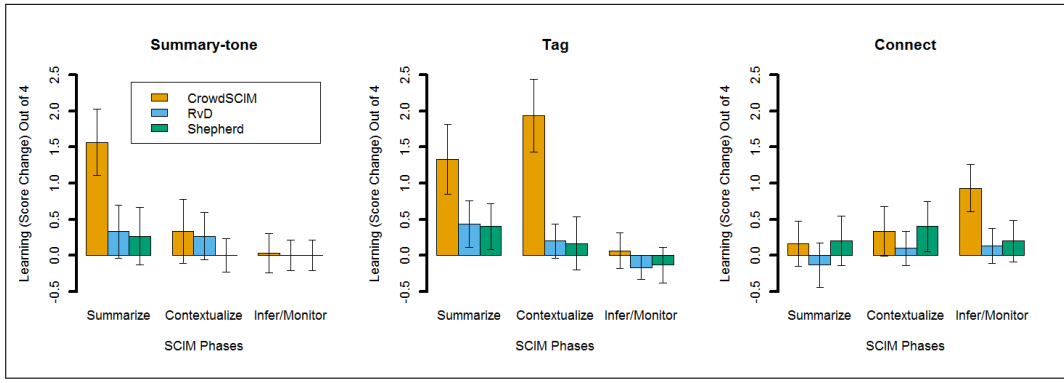

Fig. 5. Individual phase learning across the crowdsourcing techniques

Table 1. Learning (score change) of different tasks across all crowdsourcing techniques

*: $p < 0.05$

Task and Technique	*Baseline*		*CrowdSCIM*		*RvD*		*Shepherd*	
	mean	sd	mean	sd	mean	sd	mean	sd
Summary-tone	0.07	1.2	**1.9***	2.3	0.60	1.7	0.30	1.4
Tag	0.17	1.6	**3.3***	2.5	0.50	1.5	0.43	1.7
Connect	0.13	1.7	**1.4***	1.7	0.10	1.2	0.80	1.8

how much effort the task requires. *Attempts* describes how many workers accept and return a HIT before it is completed and is an indicator of the perceived task difficulty.

5 RESULTS

5.1 Learning: Only CrowdSCIM improves learning

The mean of the pre-test scores was 1.4 (sd=1.2) out of a maximum 12 points. The learning (score change between pre-test and post-test) of each task for each of the three crowdsourcing techniques is shown in Table 1 and Figure 5. For the Baseline technique, there is almost no learning; average learning scores for the Summary-tone, Tag and Connect tasks are 0.07, 0.17 and 0.13, respectively. For the CrowdSCIM technique, average learning scores for the Summary-tone, Tag, and Connect tasks are 1.9, 3.3, and 1.43, respectively. For the RvD technique, average learning scores for the Summary-tone, Tag, and Connect tasks are 0.60, 0.47, and 0.10, respectively. For the Shepherd technique, average learning scores for the Summary-tone, Tag, and Connect tasks are 0.27, 0.43, and 0.80, respectively.

A two-way ANOVA showed a significant main effect of crowdsourcing technique ($F(3, 354)=26$, $p<0.01$), insignificant main effect of crowdsourced tasks ($F(2, 354)=2.5$, $p=0.08$), and a significant interaction effect ($F(6, 348)=8.5$, $p=0.01$) on learning. Since there was a significant interaction effect and we were interested in how these crowdsourcing techniques affect the quality based on the 5 measures, we ran a one-way ANOVA's for each of the tasks. To control the overall Type I error level (α_E) as 0.05, we used Bonferroni's adjustment for each ANOVA, whose Type I error level (α_I) became 0.017.

For the Summary-tone task, a one-way ANOVA showed a significant effect of crowdsourcing technique ($F(3, 116)=7.5$, $p<0.01$). Post-hoc Tukey tests indicated that learning of CrowdSCIM was

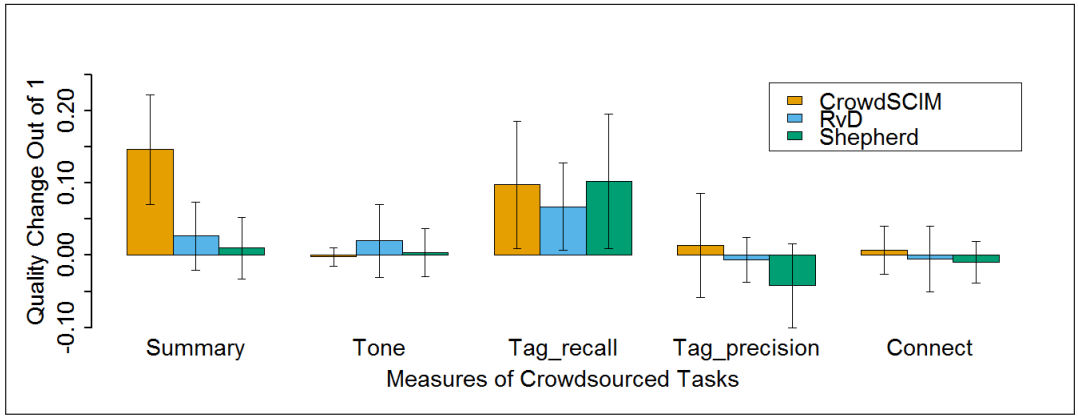

Fig. 6. Quality change of the crowdsourcing techniques for each crowdsourced task

Table 2. Quality change of the crowdsourcing techniques for each crowdsourced task

*: $p < 0.05$; Out of maximum 1.0 (Normalized)

Task	*Summary*		*Tone*		*Tag (rec.)*		*Tag (pre.)*		*Connect*	
Technique	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
CrowdSCIM	**0.15***	0.21	0.00	0.03	0.10	0.25	0.01	0.20	0.01	0.09
RvD	0.03	0.13	0.02	0.14	0.07	0.17	-0.01	0.09	0.00	0.13
Shepherd	0.01	0.12	0.00	0.19	0.10	0.26	-0.04	0.16	-0.01	0.08

significantly higher than other three techniques (all $p \leq 0.01$) and no significant difference among other three techniques.

For the Tag task, a one-way ANOVA showed a significant effect of crowdsourcing technique ($F(3, 116)=19$, $p<0.01$). Post-hoc Tukey tests indicated that learning of CrowdSCIM was significantly higher than other three techniques (all $p<0.01$) with no significant differences among other three techniques.

For the Connect task, a one-way ANOVA showed a significant effect of crowdsourcing technique ($F(3, 116)=4.5$, $p<0.01$). Post-hoc Tukey tests indicated that learning of CrowdSCIM was significantly higher than Baseline and RvD (both $p \leq 0.01$) with no significant differences among other three techniques.

To better understand what abilities workers learned, we divided overall learning into SCIM phases: Summarize, Contextualize, and Infer/Monitor, as shown in Figure 5. For CrowdSCIM, we can see that CrowdSCIM almost always creates learning gains, especially when task and phase are aligned, and never hurts learning (i.e., post-test worse than pre-test). In addition, CrowdSCIM in the Tag task also helped workers learn the Summarize ability. In contrast, RvD and Shepherd show much smaller learning gains and can actually hurt learning in some cases, e.g., Infer/Monitor in the Tag task for both RvD and Shepherd.

5.2 Quality: Only CrowdSCIM improves summary quality

Since there were five quality measures in the three tasks (summary, tone, recall of tag, precision of tag, and connect), we first normalized all the measures to a 0–1 scale and then conducted a two-way ANOVA. The two-way ANOVA showed a significant main effect of crowdsourcing technique

(F(3, 580)=3.9, p=0.01), a significant effect of crowdsourcing task (the five measures) (F(4, 580)=7.0, p<0.01), and a significant interaction effect (F(12, 580)=1.9, p=0.03). Since there was a significant interaction effect and we were interested in how these crowdsourcing techniques affect the quality based on the five measures, we ran a one-way ANOVA for each of the measures. To control the overall Type I error level (α_E) as 0.05, we again used Bonferroni's adjustment for each ANOVA, whose Type I error level (α_I) became 0.01. The overall quality change of each crowdsourcing technique for each crowdsourced task is shown in Figure 6 and Table 2.

5.2.1 Summary: Similar original quality but CrowdSCIM brings quality change. The mean summary quality of the original work was 4.0 (sd=2.5) out of maximum 10 points. This mean corresponds to the lowest score in the medium quality category that still contains some important information, detail, and context.

The mean summary quality change was 1.5 for CrowdSCIM (sd=2.1), 0.3 for RvD (sd=1.3), and 0.1 for Shepherd (sd=1.2). A one-way ANOVA showed a significant main effect of crowdsourcing technique (F(3, 116)=7.3, p<0.01) on summary quality change. Post-hoc Tukey tests showed the summary quality change of CrowdSCIM was significantly higher than other techniques (all p≤0.01) and no significant difference among other techniques.

5.2.2 Tone: Similar original quality without quality change. The mean tone rating quality of the original work was 0.54 (sd=0.30) out of maximum 1. The Cohen's kappa of 0.54 is generally considered "moderate agreement" with the gold standard [32].

The mean quality changes in tone rating were 0.00 for CrowdSCIM (sd=0.03), 0.02 for RvD (sd=0.14), and 0.00 for Shepherd (sd=0.09). A one-way ANOVA showed no significant main effect (F(3, 116)=0.41, p=0.75) on quality change in tone rating.

5.2.3 Tag: Similar original quality without quality change. The mean recall of the original work was 0.59 (sd=0.28) out of maximum 1. The mean precision of the original work were 0.61 and 0.28 (out of maximum 1).

The mean quality changes for recall were 0.10 for CrowdSCIM (sd=0.25), 0.07 for RvD (sd 0.17), and 0.10 for Shepherd (sd=0.26). A one-way ANOVA showed no significant main effect (F(3, 116)=1.7, p=0.17) on quality changes for recall. The mean quality changes for precision were 0.01 for CrowdSCIM (sd=0.20), -0.01 for RvD (sd=0.09), and -0.04 for Shepherd (sd=0.16). A one-way ANOVA showed no significant main effect (F(3, 116)=0.91, p=0.44) of crowdsourcing technique on quality change for precision.

5.2.4 Connection: Similar original quality without quality change. The mean connection quality (theme rating) of the original work was 0.65 (sd=0.26) out of maximum 1. The kappa value 0.65 is generally considered "substantial agreement" with the gold standard [32].

The mean quality changes for connection were 0.01 for CrowdSCIM (sd=0.09), 0.00 for RvD (sd=0.13), and -0.01 for Shepherd (sd=0.08). A one-way ANOVA showed no significant main effect (F(3, 116)=0.20, p=0.90) on quality change for connection.

5.3 Efficiency: Different efficiency but similar attrition

5.3.1 Time: Baseline requires the least time while CrowdSCIM needs the most. Except that the Baseline did not include intervention nor revision, all techniques contained pre-test, task, intervention, revision, and post-test. The time spent on each stage for each technique is shown in Table 3. Since we were interested in how these crowdsourcing techniques affect the efficiency of the tasks, we ran a two-way ANOVA for each of the task-related activities (task, intervention and revision). To control the overall Type I error level (α_E) as 0.05, we again used Bonferroni's adjustment for each ANOVA whose Type I error (α_I) became 0.017.

Table 3. Time spent at different work stages across different crowdsourcing techniques

*: $p < 0.05$; Out of maximum 1.0 (Normalized)

Stage	*Pre-test*		*Task*		*Intervention*		*Revision*		*Post-test*		*Total*	
Technique	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Baseline	6.4	4.4	6.0	3.9	N/A		N/A		6.8	7.5	19	12
CrowdSCIM	5.8	3.5	7.1	6.6	**10***	6.9	1.8	3.1	7.5	7.3	33	18
RvD	6.6	5.9	6.0	3.6	2.8	1.8	1.5	1.7	6.6	5.4	24	13
Shepherd	6.0	3.8	6.2	4.6	1.9	1.4	1.9	2.0	6.3	4.3	23	10

The mean time required to complete the pre-test was 6.2 minutes (sd=4.5).

Overall, it took 6.4 minutes (sd=4.8) to complete a task. For each task, it took 7.3 minutes (sd=5.1) to complete the Summary-tone task, 6.3 minutes (sd=5.6) to complete the Tag task, and 5.4 minutes (sd=3.3) to complete the Connect task. A two-way ANOVA showed a significant main effect of crowdsourced task ($F(2, 348)=4.8$, $p=0.01$) on time spent on the task. Post-hoc Tukey tests showed it took significantly more time to finish the Summary-tone than the Connect task ($p=0.01$).

In general, it took 5.1 minutes (sd=5.7) to complete any of the three learning interventions. Across the three techniques, it took, on average, 10 minutes (sd=6.9) to complete the CrowdSCIM intervention, 2.8 minutes (sd=1.8) to complete the RvD intervention and 1.9 minutes (sd=1.4) to complete the Shepherd intervention. A two-way ANOVA showed a significant main effect of crowdsourcing technique ($F(2, 261)=120$, $p<0.01$) on time spent on the intervention. Post-hoc Tukey tests showed it took significantly more time to finish the CrowdSCIM intervention than the other two interventions for each of the crowdsourced tasks (all $p<0.01$).

On average, it took 1.7 minutes (sd=2.4) to complete the revision of a crowdsourcing technique. For each of the three techniques, it took 1.8 minutes (sd=3.1) for CrowdSCIM, 1.5 minutes (sd=1.7) for RvD and 1.9 minutes (sd=2.0) for Shepherd. Two-way ANOVA showed a significant main effect of crowdsourced task ($F(2, 261)=4.6$, $p=0.01$) on time spent on revision. Post-hoc Tukey tests showed it took significantly more time to finish the Summary-tone revision than the Connect revision ($p=0.01$).

The mean time required to complete the post-test was 6.8 minutes (sd=6.3).

5.3.2 Attrition (via attempts): Extra work takes extra time but similar attrition rate. The mean attempts required to complete a task were 3.2 (sd=2.5). A two-way ANOVA showed no significant main effect of crowdsourced tasks nor techniques on attempts before a task is completed.

6 DISCUSSION

6.1 Learning: CrowdSCIM supports learning while other techniques do not

The results of the pre-test scores (only 1.4 out of 12, on average) showed that crowd workers generally lacked sufficient historical thinking skills to write a strong historical interpretation for the given primary source. The learning results for the Baseline condition further suggested that an instructor's intervention may be necessary for Incite users to learn historical thinking. Example historical interpretations are shown in Appendix D.

The CrowdSCIM results showed significant learning gains when the crowdsourced task and the SCIM phase were aligned (e.g., Summary-tone with Summarize, Tag with Contextualize, and Connect with Infer/Monitor). These results indicate that the iterative design process we employed through pilot studies was both effective and necessary. Both the first pilot study and the Baseline learning results from our evaluation showed that merely doing the crowdsourced tasks did not

help with learning. Further, the second and third pilot studies showed that simply applying SCIM-C from the classroom to crowdsourced settings would not work, either. Only by decomposing the SCIM-C technique into micro-tasks could the learning gains be realized. Finally, the results of comparison with RvD and Shepherd support the intuition that a domain expertise-related intervention is more effective than task-related interventions in achieving learning gains in historical thinking skills.

Looking at the most-learned ability for each task, CrowdSCIM improved worker learning by 1-2 points (out of maximum 4) for the corresponding phase in SCIM. That amount of learning is comparable to previous SCIM-C studies [21] in which the learning gains were 1.26 for Summarize, 1.48 for Contextualize, and 0.61 for both Infer and Monitor combined after three 2.5-hour tutorials across three instructional episodes. Although our study recruited novice workers from AMT and participants of prior SCIM-C studies were school students, the learning gains were comparable across these two very different participant pools, settings, and time frames.

CrowdSCIM in the Tag task also helped workers learn the Summarize ability. This suggests that there is a strong correlation between the two phases and abilities. In contrast, the results of RvD and Shepherd showed insignificant learning, and these techniques could even hurt learning in some cases.

6.2 Quality: Crowd's work quality is moderate and CrowdSCIM improves summary quality

The quality results for summarization tasks showed that the participants in general were able to generate summaries of middling quality (4 out of 10). CrowdSCIM was able to improve the average summary quality to 5.5 (see Appendix D for an example). RvD and Shepherd did not improve the summary quality, in contrast to previous work [13, 57]. Unfamiliarity with historical primary sources might make it difficult for these workers to write a very good summary.

The results of tone rating showed there was a moderate baseline agreement between the crowd and the expert historian. For the Connect task, the baseline crowd results were even better, showing substantial agreement with the expert. Because the baseline performance for these tasks is already reasonably good, improvement may not be necessary for some use cases.

The recall results of the Tag task showed the baseline crowd was able to tag about 60% of the expert's tags. Due to the large number of responses and tags to analyze, our calculations only recognized exact matches; i.e., minor differences such as "Va." and "Va" were considered different. Therefore, this number should be seen as a lower bound because the recall could be higher with alias handling techniques. For historians, recall is generally more important than precision because they are accustomed to false positives, whereas relevant primary sources are rare and missing one is costly.

Aside from the Summary task, none of the three crowdsourcing techniques improved quality results for any of the other tasks: tone, tag, or connect. Two possible explanations for CrowdSCIM's lack of effect are that the SCIM phases are too abstract for workers to immediately transfer to micro-tasks, or that there is a ceiling effect caused by solid initial performance. Notably, none of the learning techniques hurt work quality, either.

6.3 Efficiency: Extra work takes extra time but attrition is similar

As expected, Baseline is the most efficient technique for the crowdsourced task, followed by Shepherd and RvD, and finally CrowdSCIM. CrowdSCIM's learning intervention took significantly longer (10 min) than RvD (3 min) or Shepherd (2 min). However, the revision step took the same amount of time for all three techniques (about 3 minutes). Further, across all SCIM phases and task types, the total task times were suitable for crowd work (10 minutes or less). We primarily included

a revision step in all conditions to help quantify the advantage of each learning technique, but future work could explore omitting the initial task completion in CrowdSCIM to improve efficiency.

Although the intervention and revision time of CrowdSCIM was longer, the attrition rate of CrowdSCIM was similar other techniques. This seemed to suggest that CrowdSCIM intervention provided some extra attraction to keep the attrition rate as the same level as others.

6.4 Trade-offs: No one technique works best for all situations

Based on the results, there is no one technique among the four we evaluated that can serve all purposes, but depending on the requester's goals, some approaches work significantly better than others. In general, if the task design is *learning*-oriented, CrowdSCIM is the clear winner, because workers show the highest learning gains while producing work of similar or better quality, although this approach is slower than the others. If the design is *quality*-oriented, CrowdSCIM should be used for summary tasks, and Baseline for the other tasks, since all approaches perform similarly, but Baseline is fastest. If the design is *efficiency*-oriented, Baseline is the fastest, but workers will not learn anything, and summary quality will be degraded. Further, the "add-on" design of CrowdSCIM makes it easy to switch between Baseline and CrowdSCIM if the requester's needs change frequently.

7 BROADER IMPLICATIONS

7.1 Implications for historical education and research

Historical documents are critical sources for both scholarly research and learning in the domain of history [46, 47], and teaching students to think like a historian is one of the main goals in history education [23, 37, 52]. While we evaluated CrowdSCIM with paid crowds to support rapid iteration and scaling up, we expect that crowds of traditional history students could also benefit from using CrowdSCIM. SCIM-C is designed to be a sequential process that a student should follow from Summarize to Corroborate, but our experiments suggest that different phases in historical thinking may be learned or improved individually.

CrowdSCIM may also offer a useful supplement to the classroom teaching. For example, it may be used as a first pass, allowing the instructor to focus on learning material that CrowdSCIM does not provide. Or it may be used in a targeted way to improve one specific ability of historical thinking that a student or teacher identifies as weaker than the others by working on corresponding tasks with CrowdSCIM. Finally, in settings where the ratio of number of students per expert is high, such as in MOOCs or citizen science projects, CrowdSCIM may provide a scalable way for students to learn historical thinking with minimal intervention from experts.

Our quality results also show that the crowd can already do a reasonable job for most crowdsourced tasks, although there is often room for improvement. For example, the summary captures key information of the primary source, so it can help the historian quickly decide whether a source is worth extra attention. Taking our test documents as examples, the average length of the documents is 253.2 words and the average length of a crowdsourced summary is 49.9 words. This suggests a historian could save about 80% of the reading time while searching for relevant documents. In addition, the tagging results show that crowds have moderate to substantial agreement with a historian in identifying documents that are relevant to the historian's topics of interest.

While Shepherd and RvD have demonstrated significant value in other task domains, our results suggest they are not well-suited for promoting learning or improving output quality in historical research. Why not? While neither Shepherd or RvD was designed for historical research, both were previously evaluated with writing or summarization tasks, so it is perhaps most surprising that summarization was the only task where CrowdSCIM yielded significantly better (vs. similar)

work quality. This result suggests that writing about historical primary sources creates unique challenges. We propose two reasons why CrowdSCIM is better-suited for learning (and, in the case of summaries, doing) analysis of historical documents. One is that the reflective questions in step 2 provide scaffolding to help workers engage in deeper thinking and stimulate higher-level cognitive processes. Specifically, the questions provide structure in the form of a "specific cognitive strategy" [21] that reduces the initial complexity of an open-ended process. They also problematize the task by drawing workers' attention to issues they might not normally consider [21]. A second reason may be that the practice interpretation in step 3 helps workers to synthesize and internalize their new expertise before attempting to transfer it to a new application (i.e., the revision). This mental organization may make the expertise more readily available for the post-test and beyond.

7.2 Implications for crowdsourcing research and practice

Our results demonstrate possibilities for a better "future of crowd work," [29]. Instead of doing repetitive, low-payment tasks, crowd workers can learn and develop new skills to steadily handle more complex and creative tasks and improve work quality and payment. When learning can improve the work quality, as in the Summary task, learning may be seen as "training" directly related to the work, and the requester could pay for the training as in a traditional job market. When the learning does not improve the work quality, as in the Tone, Tag, and Connect tasks, learning may be seen as "education" not directly related to the work, and the requester may choose not to pay for it, but rather provide it as free education. At the same time, crowd workers can also decide if they want to do more tasks to get paid or spend the time with skill development. Previous work (e.g., [34, 41]) suggests that learning domain knowledge (factual knowledge) may hinder work quality, but our results show learning domain expertise (such as analytical skills and thinking strategies) may help with some task types, such as writing a summary, without impeding work quality for other types of tasks.

While the main focus of our study is using crowdsourcing to support historical research, the CrowdSCIM workflow may be generalize to other domains focused on sensemaking and analysis of primary source documents. Historians have been described as "detectives searching for evidence among primary sources to a mystery that can never be completely solved" [2], which shares similarities with other investigative domains such as journalism, law enforcement, and political fact-checking.

To adapt CrowdSCIM for other domains, we envision a generalized workflow comprising 1) an unmodified initial text analysis task, 2) a scaffolded learning intervention, 3) a practice task, and 4) a revised attempt at the initial task. Given our focus in this paper on supporting historical research, CrowdSCIM's learning intervention in step 2 was adapted from SCIM-C's historical thinking prompts. For other domains, however, these prompts could be substituted for alternative domain-specific reflective questions most relevant to the given task. For example, a CrowdSCIM variant for supporting crowdsourced journalism might provide reflective questions derived from ethnographic studies of expert practice for workers to learn to review user-generated content for newsworthy themes [38, 48]. A crowdsourced fact-checking effort could ask questions based on verification principles to help crowds learn to research politically oriented claims and assess the reliability of their sources [31, 44]. In a law enforcement context, novice workers could learn to analyze police reports for patterns of interest, guided by reflective questions about motive, opportunity, and lack of alibi [17, 18].

Our experiences with CrowdSCIM and historical documents also suggest some caveats in adapting our approach for other domains. First, although reflective questions for the target task may

already exist, they were likely developed for a different audience, such as students or junior practitioners, and will likely require iterative design to repurpose for novice crowd workers in a microtasking context. Our pilot studies showed that SCIM-C, while effective in traditional classrooms, required extensive modularization for crowds. Second, we recommend aligning each task with the most relevant subset of reflective questions; with CrowdSCIM, proper alignment made the difference between productivity gains and learning losses.

While future work is needed, we anticipate that CrowdSCIM's specific orientation towards historical thinking and working with primary sources, as well as its more general approach to decomposing complex thought processes into just-in-time learning interventions, may be applicable to these and other domains sharing similar processes.

8 CONCLUSION

As crowdsourcing markets become more popular and pervasive, researchers and practitioners have begun to seriously consider what future crowd work should ideally look like, suggesting some potential trade-offs such as learning and productivity. One the one hand, we would like to help workers learn and develop new skills. On the other hand, learning and skill development do not come without cost (e.g., immediate productivity or money). In this paper, we investigated potential trade-offs between learning domain expertise and productivity for historical research, a domain that has seen little attention from crowdsourcing researchers. We adapted a technique from educational research to create a crowdsourcing workflow, CrowdSCIM, that allows novice crowd workers to learn historical thinking skills while completing useful historical research tasks. Results from our experiments showed that CrowdSCIM was effective at helping workers learn domain expertise while producing work of equal or higher quality compared to baseline and prior work conditions. We also use CrowdSCIM as an example to discuss broader implications for future crowd work in terms of training and education and implications for history education and research.

A SCIM-C SCORING RUBRIC (BASED ON [21])

Summarizing (1 point each)

(1) Does the response indicate the subject of the source?
(2) Does the response indicate the audience for the source?
(3) Does the response indicate the author of the source?
(4) Does the response include specific details from the source?

Contextualizing (1 point each)

(1) Does the response indicate when the source was produced?
(2) Does the response indicate where the source was produced?
(3) Does the response indicate why the source was produced?
(4) Does the response indicate the immediate or broader context?

Inferring/Monitoring (1 point each)

(1) Does the response include explicit and/or implicit inferences?
(2) Does the response include inferences based on omissions?
(3) Does the response indicate the need for information beyond the source?
(4) Does the response evaluate the usefulness or significance of the source?

B SUMMARY SCORING RUBRIC

(1) The worker wrote an original summary. The worker did not plagiarize.
(2) The worker wrote a summary that contains all important information in the document.

(3) The worker wrote a summary that has balanced coverage. The summary does not focus on some specific parts of the original document.
(4) The worker wrote a summary without adding personal opinions nor emotions.
(5) The worker wrote a summary with sufficient information and details from the original document.
(6) The worker did not have spelling and grammar mistakes.
(7) The worker wrote the right amount (3 sentences).

C SCIM QUESTIONS USED IN CROWDSCIM (BASED ON [21])

Summary-tone (Summarize)

(1) What type of historical document is the source?
(2) What specific information, details and/or perspectives does the source provide?
(3) What is the subject and/or purpose of the source?
(4) Who was the author and/or audience of the source?

Tag (Contextualize)

(1) When and where was the source produced?
(2) Why was the source produced?
(3) What was happening within the immediate and broader context at the time the source was produced?
(4) What summarizing information can place the source in time and place?

Connect (Infer/Monitor)

(1) What interpretations, inferences, perspectives or points of view may be drawn from or indicated by the source?
(2) What inferences may be drawn from absences or omissions in the source?
(3) What additional evidence beyond the source is necessary to answer the historical question?
(4) How useful or significant is the source for its intended purpose in answering the historical question?

D SAMPLE PARTICIPANT RESPONSE

Table 4. Sample historical interpretations of CrowdSCIM across pre-test and three crowdsourced tasks

Pre-test	"The farming communities are having a real tough time. Most of the crops have died out over the last season, due to weather related catastrophes. Merchants have began cutting off credit, so farmers are really going to start feeling the pinch. A family member also seems to have gotten a good job as a personal chauffeur." — P287 with 1 point for describing detail in Summarize
Summary-tone	"Estella writes to her sister about life on the farm in 1911. She relates details about the tough weather and pest conditions that led to a worrisome set of circumstances for all the farmers. The situation was dire enough that merchants were cutting off credit to farm families. She was also hopeful, though, that the current rains might change the conditions in their favor so that they might have a good crop of wheat. She also writes of her daughter who is headed for a job in NYC and another mutual acquaintance who was changing his job from chauffeur to pilot." — P310 with 3 points in Summarize and 1 point for time in Contextualize
Tag	"This letter from Estella Stigebower to her sister Ella Roesch, written on August 3rd, 1911, indicates conditions on the plains were very harsh during the early 20th century. Writing from Marion, Nebraska, Stigebower responds to her sister's previous letter by indicating she is sorry to hear the crops were a total failure. Her letter also demonstrates that crop failure was experienced by many people in the plains, largely as a result of storms and plagues of grasshoppers and army worms. Crops had failed to such an extent that merchants were forced to cut off credit to farmers. Finally, Stigebower offers a glimpse of how other family members outside of the plains have been fairing, indicating they have been able to get jobs and not mentioning any specific hardship they have had." — P332 with full points in Summarize and Contextualize
Connect	"Conditions of life in farming communities during this time period were difficult. Crops were failing not only for this family, but for many families- suggesting that many families were unable to support themselves properly. The lack of help from the town (cutting the credit system) also suggests that the wealthier families had no desire to help the other citizens. Although this point of view is missing, it seems that like today, the wealthier businesses only wanted to help themselves and not those who needed assistance. If the article included information about other businesses, it would add to the perspective of "the other side"." — P359 with 1 point for details in Summarize and 2 points for inference and monitoring in Infer/Monitor

ACKNOWLEDGMENTS

We wish to thank Paul Quigley, Daniel Newcomb, Andy Kapinos, Ashleigh Grubb, and Mike Grooms, the study participants, and the anonymous reviewers of this paper. This research was supported in part by grant DH50013-15 from the US National Historical Publications and Records Commission.

Table 5. An example of an improved summary using CrowdSCIM

Original Summary	Revised Summary (CrowdSCIM)
"Estella has written a letter to her sister and family, catching them up on what's going on in her life. She's concerned about her family's failed crops and has seen similar issues in her area" — P312 with score 4/10	"In the Midwest, pre-WWI, Estella has written a letter to her sister and family, catching them up on what's going on in her life. She's concerned about her family's failed crops and has seen similar issues in her area where corn fields were destroyed by grasshoppers and hail storms. She also talks about other people, including Lillie and Gus, who was working as a chauffeur for $125/month and hoping to run an airship soon." — P312 revised work with score 9/10 and extra context (location and time), details (grasshoppers and hail storms), and coverage (relatives)

REFERENCES

[1] [n. d.]. DIYHistory | Transcribe. https://diyhistory.lib.uiowa.edu/

[2] [n. d.]. Feature Article - Historical Thinking, Winter 2010- Teaching with Primary Sources | Teacher Resources - Library of Congress. http://www.loc.gov/teachers/tps/quarterly/historical_thinking/article.html

[3] [n. d.]. Incite | Crowdsourced Analysis of Omeka Documents | Virginia Tech. http://incite.cs.vt.edu/

[4] [n. d.]. Smithsonian Digital Volunteers. https://transcription.si.edu/

[5] Keith C. Barton and Linda S. Levstik. 2004. *Teaching history for the common good.* Routledge.

[6] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 313–322. https://doi.org/10.1145/1866029.1866078

[7] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: Crowdsourcing Taxonomy Creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1999–2008. https://doi.org/10.1145/2470654.2466265

[8] D. Coetzee, Seongtaek Lim, Armando Fox, Björn Hartmann, and Marti A. Hearst. 2015. Structuring Interactions for Large-Scale Synchronous Peer Learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 1139–1152. https://doi.org/10.1145/2675133.2675251

[9] Robin George Collingwood and Willem J. van der Dussen. 1993. *The idea of history.* Oxford University Press on Demand.

[10] Allan Collins, John Seely Brown, and Susan E. Newman. 1988. Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. *Thinking: The Journal of Philosophy for Children* 8, 1 (1988), 2–10.

[11] Mira Dontcheva, Robert R. Morris, Joel R. Brandt, and Elizabeth M. Gerber. 2014. Combining Crowdsourcing and Learning to Improve Engagement and Performance. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3379–3388. https://doi.org/10.1145/2556288.2557217

[12] 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, New York, NY, USA, 2623–2634. https://doi.org/10.1145/2858036.2858268

[13] 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, New York, NY, USA, 1013–1022. https://doi.org/10.1145/2145204.2145355

[14] B. Farrimond, S. Presland, J. Bonar-Law, and F. Pogson. 2008. Making History Happen: Spatiotemporal Data Visualization for Historians. In *Second UKSIM European Symposium on Computer Modeling and Simulation, 2008. EMS '08.* 424–429. https://doi.org/10.1109/EMS.2008.42

[15] Elena L. Glassman, Juho Kim, Andrés Monroy-Hernández, and Meredith Ringel Morris. 2015. Mudslide: A Spatially Anchored Census of Student Confusion for Online Lecture Videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1555–1564. https://doi.org/10.1145/2702123.2702304

[16] Elena L. Glassman, Aaron Lin, Carrie J. Cai, and Robert C. Miller. 2016. Learnersourcing Personalized Hints. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1626–1636. https://doi.org/10.1145/2818048.2820011

[17] Nitesh Goyal and Susan R. Fussell. 2015. Designing for Collaborative Sensemaking: Leveraging Human Cognition For Complex Tasks. *arXiv preprint arXiv:1511.05737* (2015).

[18] Nitesh Goyal and Susan R. Fussell. 2016. Effects of Sensemaking Translucence on Distributed Collaborative Analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 288–302. https://doi.org/10.1145/2818048.2820071

[19] Michael D. Greenberg, Matthew W. Easterday, and Elizabeth M. Gerber. 2015. Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition (C&C '15)*. ACM, New York, NY, USA, 235–244. https://doi.org/10.1145/2757226.2757249

[20] Stuart Greene. 1994. Students as authors in the study of history. In *Teaching and Learning in History*. Routledge, 137–170.

[21] David Hicks and Peter E. Doolittle. 2008. Fostering Analysis in Historical Inquiry Through Multimedia Embedded Scaffolding. *Theory & Research in Social Education* 36, 3 (July 2008), 206–232. https://doi.org/10.1080/00933104.2008.10473373

[22] David Hicks, Peter E. Doolittle, and E. Thomas Ewing. 2004. The SCIM-C strategy: expert historians, historical inquiry, and multimedia. *Social Education* 68, 3 (April 2004), 221–226.

[23] Cynthia Hynd, Jodi Patrick Holschuh, and Betty P. Hubbard. 2004. Thinking like a historian: College students' reading of multiple historical documents. *Journal of Literacy Research* 36, 2 (2004), 141–176. http://jlr.sagepub.com/content/36/2/141.short

[24] Michael J. Jacobson, Chrystalla Maouri, Punyashloke Mishra, and Christopher Kolar. 1995. Learning with Hypertext Learning Environments: Theory, Design, and Research. *J. Educ. Multimedia Hypermedia* 4, 4 (Dec. 1995), 321–364. http://dl.acm.org/citation.cfm?id=227170.227173

[25] Michael J. Jacobson and Rand J. Spiro. 1995. Hypertext Learning Environments, Cognitive Flexibility, and the Transfer of Complex Knowledge: An Empirical Investigation. *Journal of Educational Computing Research* 12, 4 (June 1995), 301–333. https://doi.org/10.2190/4T1B-HBP0-3F7E-J4PN

[26] Charlene Jennett, Laure Kloetzer, Daniel Schneider, Ioanna Iacovides, Anna Cox, Margaret Gold, Brian Fuchs, Alexandra Eveleigh, Kathleen Methieu, Zoya Ajani, and others. 2016. Motivations, learning and creativity in online citizen science. *Journal of Science Communication* 15, 3 (2016). http://oro.open.ac.uk/47008/

[27] Juho Kim, Robert C. Miller, and Krzysztof Z. Gajos. 2013. Learnersourcing Subgoal Labeling to Support Learning from How-to Videos. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)*. ACM, New York, NY, USA, 685–690. https://doi.org/10.1145/2468356.2468477

[28] Juho Kim and others. 2015. *Learnersourcing: improving learning with collective learner activity*. Ph.D. Dissertation. Massachusetts Institute of Technology. http://dspace.mit.edu/handle/1721.1/101464

[29] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[30] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. CrowdForge: Crowdsourcing Complex Work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. ACM, New York, NY, USA, 43–52. https://doi.org/10.1145/2047196.2047202

[31] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. 2014. Integrating On-demand Fact-checking with Public Dialogue. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 1188–1199. https://doi.org/10.1145/2531602.2531677

[32] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. https://doi.org/10.2307/2529310

[33] Edith Law, Krzysztof Z. Gajos, Andrea Wiggins, Mary L. Gray, and Alex Williams. 2017. Crowdsourcing As a Tool for Research: Implications of Uncertainty. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1544–1561. https://doi.org/10.1145/2998181.2998197

[34] Doris Jung-Lin Lee, Joanne Lo, Moonhyok Kim, and Eric Paulos. 2016. Crowdclass: Designing Classification-Based Citizen Science Learning Modules. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*. https://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/view/14027

[35] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. 2010. Exploring Iterative and Parallel Human Computation Processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. ACM, New York, NY, USA, 68–76. https://doi.org/10.1145/1837885.1837907

[36] Kurt Luther, Nathan Hahn, Steven P. Dow, and Aniket Kittur. 2015. Crowdlines: Supporting Synthesis of Diverse Information Sources through Crowdsourced Outlines. In *Third AAAI Conference on Human Computation and Crowdsourcing*. http://www.aaai.org/ocs/index.php/HCOMP/HCOMP15/paper/view/11603

[37] Nikki Mandell. 2008. Thinking like a Historian: A Framework for Teaching and Learning. *OAH Magazine of History* 22, 2 (April 2008), 55–59. https://doi.org/10.1093/maghis/22.2.55

[38] Ville J. E. Manninen. 2017. Sourcing practices in online journalism: an ethnographic study of the formation of trust in and the use of journalistic sources. *Journal of Media Practice* 18, 2-3 (Sept. 2017), 212–228. https://doi.org/10.1080/14682753.2017.1375252

[39] Andrea L. McNeill, Peter E. Doolittle, and David Hicks. 2009. The effects of training, modality, and redundancy on the development of a historical inquiry strategy in a multimedia learning environment. *Journal of Interactive Online Learning* 8, 3 (2009), 255–269.

[40] Piotr Mitros. 2015. Learnersourcing of Complex Assessments. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale (L@S '15)*. ACM, New York, NY, USA, 317–320. https://doi.org/10.1145/2724660.2728683

[41] Vineet Pandey, Amnon Amir, Justine Debelius, Embriette R. Hyde, Tomasz Kosciolek, Rob Knight, and Scott Klemmer. 2017. Gut Instinct: Creating Scientific Theories with Online Learners. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 6825–6836. https://doi.org/10.1145/3025453.3025769

[42] Jennifer Rutner and Roger Schonfeld. 2012. *Supporting the Changing Research Practices of Historians*. Technical Report. Ithaka S+R, New York. http://sr.ithaka.org/?p=22532

[43] John W. Saye and Thomas Brush. 2002. Scaffolding critical reasoning about history and social issues in multimedia-supported learning environments. *Educational Technology Research and Development* 50, 3 (Sept. 2002), 77–96. https://doi.org/10.1007/BF02505026

[44] Ivor Shapiro, Colette Brin, Isabelle Bédard-Brûlé, and Kasia Mychajlowycz. 2013. Verification as a Strategic Ritual. *Journalism Practice* 7, 6 (Dec. 2013), 657–673. https://doi.org/10.1080/17512786.2013.765638

[45] Jakub Šimko, Marián Šimko, Mária Bieliková, Jakub Ševcech, and Roman Burger. 2013. Classsourcing: Crowd-Based Validation of Question-Answer Learning Objects. In *International Conference on Computational Collective Intelligence*. Springer, 62–71. http://link.springer.com/chapter/10.1007/978-3-642-40495-5_7

[46] Peter N. Stearns, Peter C. Seixas, and Sam Wineburg. 2000. *Knowing, teaching, and learning history: National and international perspectives*. NYU Press.

[47] Bill Tally and Lauren B. Goldenberg. 2005. Fostering historical thinking with digitized primary sources. *Journal of Research on Technology in Education* 38, 1 (2005), 1–21. http://www.tandfonline.com/doi/abs/10.1080/15391523.2005.10782447

[48] Peter Tolmie, Rob Procter, David William Randall, Mark Rouncefield, Christian Burger, Geraldine Wong Sak Hoi, Arkaitz Zubiaga, and Maria Liakata. 2017. Supporting the Use of User Generated Content in Journalistic Practice. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3632–3644. https://doi.org/10.1145/3025453.3025892

[49] Sarah Weir, Juho Kim, Krzysztof Z. Gajos, and Robert C. Miller. 2015. Learnersourcing Subgoal Labels for How-to Videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 405–416. https://doi.org/10.1145/2675133.2675219

[50] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. 2016. AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale (L@S '16)*. ACM, New York, NY, USA, 379–388. https://doi.org/10.1145/2876034.2876042

[51] Sam Wineburg. 1999. Historical thinking and other unnatural acts. *The Phi Delta Kappan* 80, 7 (1999), 488–499. http://www.jstor.org/stable/20439490

[52] Sam Wineburg. 2010. Thinking like a historian. *Teaching with primary sources quarterly* 3, 1 (2010), 2–4. https://www.weteachnyc.org/media/filer_public/24/4a/244ab1eb-4540-4ca1-a60e-5ef96326b365/research_history.pdf

[53] Sam Wineburg. 2010. Thinking Like a Historian. *Library of Congress* (2010). http://www.loc.gov/teachers/tps/quarterly/historical_thinking/article.html

[54] Samuel S. Wineburg. 1991. On the Reading of Historical Texts: Notes on the Breach Between School and Academy. *American Educational Research Journal* 28, 3 (Sept. 1991), 495–519. https://doi.org/10.3102/00028312028003495

[55] Samuel S. Wineburg and Suzanne M. Wilson. 1991. Subject matter knowledge in the teaching of history. *Advances in research on teaching* 2 (1991), 305–347.

[56] Anbang Xu, Huaming Rao, Steven P. Dow, and Brian P. Bailey. 2015. A Classroom Study of Using Crowd Feedback in the Iterative Design Process. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 1637–1648. https://doi.org/10.1145/2675133.2675140

[57] Haiyi Zhu, Steven P. Dow, Robert E. Kraut, and Aniket Kittur. 2014. Reviewing Versus Doing: Learning and Performance in Crowd Assessment. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 1445–1455. https://doi.org/10.1145/2531602.2531718

Received April 2018; revised July 2018; accepted September 2018