

# Can Crowds Customize Instructional Materials with Minimal Expert Guidance? Exploring Teacher-guided Crowdsourcing for Improving Hints in an AI-based Tutor

KEXIN BELLA YANG, Carnegie Mellon University  
TOMOHIRO NAGASHIMA, Carnegie Mellon University  
JUNHUI YAO, Carnegie Mellon University  
JOSEPH JAY WILLIAMS, University of Toronto  
KENNETH HOLSTEIN, Carnegie Mellon University  
VINCENT ALEVEN, Carnegie Mellon University

AI-based educational technologies may be most welcome in classrooms when they align with teachers' goals, preferences, and instructional practices. Teachers, however, have scarce time to make such customizations themselves. How might the crowd be leveraged to help time-strapped teachers? Crowdsourcing pipelines have traditionally focused on content generation. It is an open question how a pipeline might be designed so the crowd can succeed in a revision/customization task. In this paper, we explore an initial version of a teacher-guided crowdsourcing pipeline designed to improve the adaptive math hints of an AI-based tutoring system so they fit teachers' preferences, while requiring minimal expert guidance. In two experiments involving 144 math teachers and 481 crowdworkers, we found that such an expert-guided revision pipeline could save experts' time and produce better crowd-revised hints (in terms of teacher satisfaction) than two comparison conditions. The revised hints however, did not improve on the existing hints in the AI tutor, which were carefully-written but still have room for improvement and customization. Further analysis revealed that the main challenge for crowdworkers may lie in understanding teachers' brief written comments and implementing them in the form of effective edits, without introducing new problems. We also found that teachers preferred their own revisions over other sources of hints, and exhibited varying preferences for hints. Overall, the results confirm that there is a clear need for customizing hints to individual teachers' preferences. They also highlight the need for more elaborate scaffolds so the crowd can have specific knowledge of the requirements that teachers have for hints. The study represents a first exploration in the literature of how to support crowds with minimal expert guidance in revising and customizing instructional materials.

CCS Concepts: • **Human-centered computing** → **Collaborative content creation**; **Computer supported cooperative work**; **Empirical studies in HCI**; • **Applied computing** → **Computer-assisted instruction**; **Interactive learning environments**.

Additional Key Words and Phrases: Expert-facilitated Crowdsourcing, Teachersourcing, Human Computation, Learning at Scale, AI in Education

---

Authors' addresses: Kexin Bella Yang, [kexiny@cs.cmu.edu](mailto:kexiny@cs.cmu.edu), Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania, 15213; Tomohiro Nagashima, [tnagashi@cs.cmu.edu](mailto:tnagashi@cs.cmu.edu), Carnegie Mellon University; Junhui Yao, [junhuiy@cs.cmu.edu](mailto:junhuiy@cs.cmu.edu), Carnegie Mellon University; Joseph Jay Williams, [joseph@cs.cmu.edu](mailto:joseph@cs.cmu.edu), University of Toronto, 27 King's College Cir, Toronto, Ontario, M5S; Kenneth Holstein, [kenneth.holstein@gmail.com](mailto:kenneth.holstein@gmail.com), Carnegie Mellon University; Vincent Alevan, [aleven@cs.cmu.com](mailto:aleven@cs.cmu.com), Carnegie Mellon University.

---

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

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

2573-0142/2021/4-ART119

<https://doi.org/10.1145/3449193>

**ACM Reference Format:**

Kexin Bella Yang, Tomohiro Nagashima, Junhui Yao, Joseph Jay Williams, Kenneth Holstein, and Vincent Alevan. 2021. Can Crowds Customize Instructional Materials with Minimal Expert Guidance? Exploring Teacher-guided Crowdsourcing for Improving Hints in an AI-based Tutor. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 119 (April 2021), 24 pages. <https://doi.org/10.1145/3449193>

**1 INTRODUCTION**

AI-based educational software is increasingly used in K-12 classrooms, enabling teachers to facilitate more personalized learning experiences [36, 41, 50, 71]. The scalability of these platforms makes it easy to reach teachers and students across a broad range of educational contexts [30, 58]. However, these complex systems are rarely designed to be readily customizable by teachers, students, or other educational stakeholders [30, 54]. With the diversity of textbooks, curricula, and teacher pedagogies, which frequently change over time, the instructional behavior of AI-based software can easily become misaligned with educational practice in a given context [30]. Thus, allowing teachers to customize and revise aspects of AI-based tutoring software may be essential for wider adoption and for sustaining their relevance longer term [7, 51–53, 55]. For example, in a study that investigated the needs of K-12 teachers who use AI tutors in their classrooms, Holstein et al. [30] found that teachers may discontinue the use of the AI tutors if there is a misalignment between instructional materials in the software and the curriculum in use, especially if there are no convenient ways for teachers to customize the content. The current paper takes inspiration from a prior study [33] that uncovered a critical teacher need: when using an AI tutor in classrooms, teachers often find that particular on-demand hints are insufficiently helpful for their students. In turn, teachers desire a simple way to revise these hints, for the benefit of future students who reach a similar impasse. In co-design sessions with teachers, teachers envisioned mechanisms to provide rapid feedback to the AI-tutor, even during an ongoing class session, as soon as a complaint about a particular hint arises, with minimal interruption of their teaching activity [33].

Despite the need described above, prior work has not found a scalable, time-efficient way to enable stakeholders (teachers and students) to customize instructional materials in AI-based software [30, 33, 54]. The current state-of-the-art method for generating educational content is through manual authoring by domain experts or instructional designers [7]. While involving teachers in materials authoring processes can be effective [23, 28], one challenge is that teachers, like other experts, tend to have very limited time for these activities [33, 38, 51, 57].

Crowdsourcing holds the potential to be a cost-effective and scalable way to meet this need [1, 15, 40, 72]. In the current study, we focus on sourcing hint improvement to generic crowds. We explore a first version of a teacher-guided crowdsourcing pipeline, aiming at improving and customizing instructional materials so they better meet expert stakeholders' (teachers) preferences, while minimizing the time they put in. Minimizing expert involvement is an important issue not just in education, but in other crowdsourcing task domains as well [18, 21], as the cost and scarcity of expertise is a central motivation for employing crowdsourcing in many contexts [8, 35, 59, 67, 68]. Prior studies have shown settings in which crowdworkers generate reasonable quality instructional materials for K-12 math, where the goal is typically to create instructional resources where none existed previously (e.g., [1, 10, 73]). Less is known about whether crowdworkers could succeed in content revision, where they must improve upon existing materials. On the one hand, content revision tasks may be easier for crowdworkers than generating content from scratch, since workers have access to the original materials as a strong starting point for their task, plus any guidance on how to revise the given materials. On the other hand, revision tasks could be harder for crowdworkers than content generation, as producing consistent, high-quality revisions requires crowdworkers to fully comprehend the original materials and contextualize expert guidance (e.g.,

brief specs for the revisions), which may not be easy for a non-expert. Revision tasks could be especially hard if the original materials are already of high quality, as in our case.

It is not known - and worth finding out - if generic crowds have sufficient relevant expertise to customize instructional materials with no guidance other than brief specs. Their past knowledge of the given application domain (middle school mathematics) may or may not be sufficient. In our specific case, an alternative to crowdworkers is not easily available because teachers, as argued, are busy, and instructional designers are expensive. As well, it may not be appropriate to source it to K-12 students, as there might not be enough educational value in the task [19, 26, 74].

In sum, we contribute a study that experiments with an expert-guided crowdsourcing pipeline for adaptive instructional content revision. The research addresses a practical need and provides insight into the broader question of how minimal expert guidance can be, in scaffolding complex crowdwork. Specifically, we find that two forms of minimal expert guidance are effective only to a degree, and contribute an analysis of where additional scaffolding may be most needed. We outline a number of specific challenges that future expert-guided crowd-revision pipelines may need to overcome. The work also yields new insight into teacher needs regarding customization of instructional materials, which could be applied to create a more effective expert-guided revision crowdsourcing pipeline in and beyond math education contexts.

## 2 RELATED WORK

### 2.1 Expert-facilitated Crowdsourcing

There are at least two broad categories of crowdsourcing tasks: simple, context-free microtasks, like transcribing or image labeling, or complex tasks that can be broken down [10, 40, 42, 45, 47]. Complex tasks that do not lend themselves easily to decomposition are often done by domain experts [12, 43, 44, 61, 66]. Crowds may need concrete expert feedback to perform well enough on these tasks [21, 63, 64]. Prior research explored strategies for integrating experts into crowdsourcing processes, from offering high-level “inspirations” [9], to expert-generated task-specific feedback [21], to an expert-informed checklist [34]. As an example, Chan et al. studied expert facilitation of crowd innovation. In their pipeline, the crowd was encouraged to come up with creative ideas to solve a social dilemma. Experts monitored the crowd and offered high-level “inspirations” to guide ideation. They found experienced facilitators increased the quantity and creativity of workers’ ideas, while novice facilitators reduced them [10].

Studies [18, 21, 68] have also explored the design space of scaffolding methods when it is desirable to minimize costly or scarce expert input. Suzuki et al. [68] explored the idea of micro-internship, to lower the threshold for the crowd to develop skills and minimize mentors’ effort: *Atelier* guides mentor-intern pairs in breaking down tasks into milestones, and solving problems together. It can help interns maintain forward progress and absorb best practice. Mentors still spend on average 5.3 hours in their study, which is much higher than desired in our case. In *Shepherd*, Dow et al. [21] tested how feedback can improve the quality of crowdwork. They found, on crowd’s finished tasks (writing customer reviews), both self-assessment and external (expert) assessment yield overall better work than control condition, while crowdworkers who receive expert feedback revise their work more.

In general, two differences distinguish our pipeline from prior expert-facilitated crowdsourcing studies:

**1) Content generation versus content revision.** In prior related crowdsourcing pipelines, including in those that, like our study, involve writing tasks (for example, writing customer reviews in *Shepherd* [21], or help request emails in *IntroAssist* [34]), the tasks often involve content *generation*, rather than content *revision*. This nuanced difference may require different pipeline designs. Content

generation generally asks crowdworkers to author the content wholesale according to a prompt or guideline, while content revision asks them to comprehend original materials, contextualize the prompt (expert guidance in our case) by carefully weighing what to keep and what to revise, and perform the desired revision.

**2) Minimal expert guidance from K-12 teachers.** Given our project goal to create a teacher-guided crowdsourcing pipeline for use within regular teaching practice with AI-based tutoring software, the pipeline could only ask K-12 teachers for minimal and rapid feedback on existing materials. Specifically, we asked teachers to spend *only 15 - 20 seconds* to give feedback on each set of hints. The fact that our experts are busy teachers renders pipelines that require a large amount of expert time, (e.g., iterative communication flow between experts and crowd [15]) infeasible in our context.

## 2.2 Crowdsourcing in Education

Crowdsourcing can help create educational materials [16, 29, 65], provide real-world educational experiences [11, 17], exchange complementary knowledge [25] and provide feedback and evaluations for learners [20, 37, 49]. For example, Aleahmad et al. [1] crowdsourced the generation of worked examples in mathematics (problems and step-by-step solutions) to volunteers using an open authoring tool. They asked the volunteers to tailor the worked examples to given student profiles that contained students' hobbies and skill proficiency. They found, interestingly, that math teachers wrote the best problem statements but that amateurs wrote the best solutions. Close to our domain, Chen et al. [10] crowdsourced math word problems and hints using a pipeline with built-in scaffolding of: 1) tagging components of crowdworkers' answers to a given template and 2) a faded, step-by-step tutorial. Whitehill and Seltzer [73] crowdsourced math tutorial videos that explain logarithms. They scaffolded crowds with examples of good video explanations, explicit guidelines, and contrasting examples of handwriting quality (given crowd may present their hand-written notes in the tutorial videos). Their crowdsourced tutorial videos lead to comparable learning gains as a popular Khan Academy video on logarithms. Most of these studies seek to contribute new content, rather than to improve on or customize existing content (which can be a higher bar for evaluation). Similarly trying to improve quality explanation in educational software, Williams et al. [74] introduced AXIS that asks learners to generate, revise and evaluate explanations as they solve a problem.

Our pipeline involves teachers as expert facilitators to guide crowdwork, rather than merely as a comparison group like [1]. Success stories from prior work bode well for open authoring and crowdsourcing of instructional materials. However, it remains an open question what level and kinds of expert (teacher) guidance can be effective in supporting crowd revision tasks.

## 3 METHODS

### 3.1 Research Questions

Our work broadly concerns the following questions, with specific sub-research questions for each of the two studies.

- **RQ1.** Can an expert-guided crowd pipeline lead to quality revisions that improve on the original artifacts, while saving experts' time?
- **RQ2.** How does an expert-guided crowd revision condition compare to other comparison conditions?
- **RQ3.** What expert guidance is needed for crowdworkers to perform the revision tasks?
- **RQ4.** Can customization of instructional materials in AI-based educational software improve teachers' satisfaction?

- **RQ5.** What challenges might arise in such an expert-guided crowd revision workflow, and how might they be overcome?

### 3.2 Study Context: AI-based tutoring software

The context of our study is an AI-based tutoring system, Lynnette (Fig. 1), which offers guided practice to middle-school students in basic equation solving. The study focuses on two-step equations (e.g.,  $2x + 4 = 0$ , solve for  $x$ ). Lynnette provides step-by-step guidance, in the form of adaptive hints, correctness feedback, and error-specific messages. AI-tutors are increasingly being used in K-12 classrooms to help teachers more effectively personalize instruction [62].

Our workflow focuses on improving the *on-demand hints* in the software, as shown in Fig. 1. When using the AI tutor, students can ask for hints when they get stuck. Hints of this type are given by the system at the student's request [60]. The hint suggests what to do next and explains why that is a good thing to do in terms of underlying problem-solving principles. Hints of this type are assumed to help students enhance their understanding of key concepts and principles (e.g., [2, 4, 5]) and reduce floundering during problem solving [3]. To reduce students' wheel-spinning or unproductive struggle [6], the existing hints in this particular system have the answer to each step as the last hint level, sometimes referred to as "bottom-out hint".

The AI-tutor's existing hints were authored by an experienced researcher in the area of intelligent tutoring systems (ITS), whose goal was to create hints that are context-sensitive (adaptive to students' current input in the systems), concise, emphasizing underlying algebra rules and helpful for learning. With these hints, this particular AI-tutor (Lynnette) has been scientifically proven to improve students' equation-solving skills through several classroom studies [31, 48, 70]. However, classroom studies reveal that teachers using the AI tutor still often wish they could make edits to customize the existing hints [32, 33], suggesting certain teacher needs are left unmet by them, and opportunities for further hint improvement and customization.

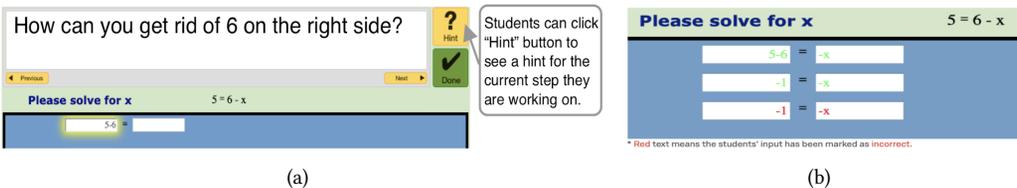


Fig. 1. Left: An example hint level in the math AI tutor (Lynnette). Right: An example problem state.

### 3.3 Study Overview

We developed *Timic*, an expert-guided crowdsourcing pipeline for instructional materials improvement and customization (Fig. 2). To answer our research questions, we conducted two studies in which teachers guided the crowd to improve AI-tutor's existing hints, in a *generic* (Study 1) or *customized* way (Study 2). Specifically, in Study 1 crowdworkers revised the hints based on feedback from *multiple* experts (thus their revision is *generic*), while in Study 2, the crowd revised the hints based on feedback from a *single* expert (such that their revision is *customized*). In the *Timic* pipeline, experts and crowdworkers participated by completing surveys online (implemented in Qualtrics). We next introduce the four stages in the pipeline.

#### 3.4 Stage 1: Guide

In Stage 1 (Fig. 3, left), teachers acted as expert facilitators and provided ratings and comments on the existing hints, to be used in later stages as guidance for the crowd to perform the revisions. To

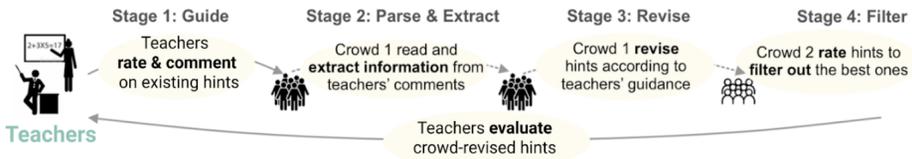


Fig. 2. *Timic* - Teacher-guided crowdsourcing workflow for instructional materials (AI-tutor hints) revision.

familiarize teachers with the AI tutor, they were shown how the tutor works (partly shown in Fig. 1, left), and asked to try out the system for themselves. Teachers then read, rated, and commented on five sets of AI-tutor's existing hints presented in a random sequence, corresponding to 5 math problem states (Fig. 1, right). The five problem states and hints were selected to exemplify major teachers' critiques documented in previous classroom studies [32]. Subsequently, teachers wrote their own preferred version hint for each problem state, and completed four open-ended questions about their preferences for how AI-tutor hints should be authored.

Fig. 3. Left: Stage 1. teachers rated(a) and commented(b) on existing hints; Right: In stage 2, crowdworkers read teachers' guidance and listed changes teachers want to see and merit to keep in the hints to be revised

### 3.5 Stage 2: Parse and Extract

In Stage 2, crowdworkers read teachers' guidance, which included comments on how the hints can be improved and/or ratings of the hints effectiveness in helping students learn equation solving. Based on the guidance, crowds then listed out changes to make and merits to keep in existing hints (Fig. 3, right).

### 3.6 Stage 3: Revise

Stage 3 happened in the same online survey as Stage 2, in the same Human Intelligence Task (HIT). In this stage, MTurkers wrote or revised hints, for each of the five math problem states (Fig. 4, left). Researchers approved the HIT completed in good faith, and rejected HITs that did not contain any proper hints. Unqualified hints included random input, copy-pasting materials, or *mere* answers to the questions without *any* explanation (i.e. it was acceptable for the hint to contain a bottom-out answer as the last level, if it included some math explanation or guidance.)

### 3.7 Stage 4: Filtering

Similar to community upvote [21], we implemented a stage for quality control of crowdwork. In this stage, new MTurkers rated the *crowd-generated* or *crowd-revised* hints on three dimensions: 1)

### Stage 3: Revise

Please improve the hints, based on **teachers' comments**, to make them more effective for students' equation learning.

**First hint level**

**Second hint level**

**Third hint level**

**Fourth hint level (Optional)**

**Fifth hint level (Optional)**

### Stage 4: Filter

	Extremely bad 1	Moderately bad 2	Slightly bad 3	Neither good nor bad 4	Slightly good 5	Moderately good 6	Extremely good 7
How <b>mathematically correct/ accurate</b> are the hints?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How <b>easy to understand</b> are the hints?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Overall, how <b>effective</b> are the hints in <b>helping students learn</b> equation solving?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Fig. 4. Left: Interface for crowd to write/revise hints; Right: Metric used for hint evaluation in 2 studies.

mathematical correctness/ accuracy, 2) ease of understanding, and 3) anticipated effectiveness in helping students learn (Fig. 4, right). The design of this question sequence was inspired by prior crowdsourcing research [24, 27, 39]). We started with more objective questions with verifiable answers (e.g., rating hints' correctness), to give workers more familiarity with the materials. This might in turn help them to do better, subsequently, on the more subjective portion (rating hints' effectiveness) [10, 39]. A seven-point Likert scale was adopted, with the minimum and maximum value being 1 and 7. Additionally, crowdworkers were asked to justify why they rated a certain set the highest, as research shows this might reduce gaming behaviors [22, 28]. Each hint received evaluations from 8-12 crowdworkers. We assigned the mean of all MTurkers' ratings to each set of generated hints as overall evaluation. We then Researchers then selected *one* set of highest-rated hints from each crowd-produced condition, for expert evaluation. For the filtered hints selected, no full copy-paste behaviors were observed.

### 3.8 Initial Scaffolds for Crowdworkers

We embedded the following scaffolds in the survey, before crowdworkers started the revision tasks.

#### AI tutor tutorial and simulated system demo.

To familiarize workers with the AI tutor, we provided a tutorial that walked them through how students interact with the tutor (partially shown in Fig 1, left). Furthermore, crowdworkers were then asked to try out the system themselves (embedded in the survey). Crowdworkers were instructed to trigger hints in the AI-tutor, and answer a question about the tutor's hints (i.e. "Please write the first three words of any second-level hint").

**Goals and task-specific principles.** Crowdworkers were instructed about the twofold goal of revising hints: they should not only help learners make progress with the current step in equation solving, but also help learners build general equation-solving skills. Workers were also instructed that the hints they write should be *clear, concise* and *easy to read*.

## 4 STUDY 1- GENERIC HINT IMPROVEMENT

### 4.1 Experiment Design, Participants, Procedure

Study 1 addresses **RQ1**, **RQ2** and **RQ3**, and aims to answer the following sub research questions: **RQ1**. Can an expert-guided crowd revision pipeline improve upon AI-tutor's existing hints, while saving experts' time?

**RQ1.1**. Does it take less time for experts to guide MTurkers to revise hints than to make the

revisions themselves?

**RQ1.2.** In terms of expert satisfaction, do crowdsourced generic revised hints (with minimal expert guidance) improve the AI-tutor's existing hints?

**RQ2.** How does an expert-guided crowd revision condition compare to other comparison conditions?

**RQ2.1.** Is crowd revision (with minimal experts guidance) better than the crowd's from-scratch generation, in terms of expert satisfaction?

**RQ2.2.** Is crowd revision (with minimal experts guidance) better than average expert-revised hints, in terms of expert satisfaction?

**RQ3.** What minimal expert guidance is needed for crowdworkers to perform the revision tasks?

**RQ3.1.** Does the combination of expert ratings and expert comments on the to-be-revised hints, facilitate crowd revisions more than expert ratings only?

**Participants.** For Stage 1 (Guide), we recruited 76 math teachers from an online teacher forum who had taught equation solving from an online teacher forum. Among these teachers, 24 had taught equation solving at the middle school level, 7 at the high school level, and 22 at multiple levels including middle school, high school, higher education and others. More than 80% had more than 4 years of teaching experience, with the majority having 10 years or more. Teachers were paid a \$6 USD Amazon Gift Card for the 10 - 20 minute survey, and were entered in a raffle to win an additional \$40 USD. The 76 teachers produced 53 valid responses. We excluded responses that did not contain valid hints, for example copy-pasting given materials, or random input such as "0".

We recruited 41 (for Stage 2 and 3) and 105 (for Stage 4) crowdworkers from Amazon Mechanical Turk. They were paid \$3 USD for a HIT completed, a survey that we estimated would take 15 - 25 minutes. (The estimate turned out to be accurate: the survey for Stage 2 took MTurkers on average 21.5 minutes, and the survey for stage 3 took them on average 18.4 minutes.) We restricted participation to US-based, adult MTurkers with a HIT approval rate of 98% or above and more than 1000 HITs approved. Beyond basic literacy, we required the crowdworkers to have some familiarity with middle school (grade 6 - 8) math algebra (which we assumed most adult MTurkers in the US would have). Whether this knowledge would be sufficient to complete the revision task was an open question our study aims to find out.

For pipeline evaluation (arrow at the bottom of Fig. 2), we recruited new math teachers ( $n = 87$ ) from the same teacher forum to rate the revised hints, using the metrics in Fig. 4 (right). The teachers' demographic was similar to that recruited for Stage 1. 27 of them teach equation solving in middle school, 22 at high school, 5 in higher education, and 33 at multiple levels. 80.5% of teachers had more than 4 years of teaching experience, with 37% of them ( $N = 32$ ) having taught for 10 years or more.

**Methods.** Study 1 had 2 experimental conditions and 3 comparison conditions. The conditions were: **1) From-Scratch Generation Condition (baseline comparison)** where MTurkers received no existing hints or teacher guidance, and wrote hints from scratch; **2) Rating Only Revision Condition (experimental)** where MTurkers received multiple teachers' ratings of exiting AI-tutor hints; **3) Rating + Comment Revision Condition (experimental)** where MTurkers received multiple teachers' ratings and comments on AI-tutor's existing hints. We also had two conditions that did not go through any pipeline stage or involve MTurkers, but primarily served as comparison groups: **4) Randomly Selected Teacher-written Hints (comparison)** that contained hints written by randomly selected teacher experts, and **5) AI-tutor's existing Hints Condition (comparison)** which contained the original hints from the AI-tutor.

Study 1 followed the four-stage workflow. The main manipulation took place in stages 2 and 3 of the workflow, where we adopted a within-subjects design: Every crowdworker completed five

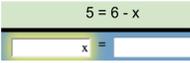
Problem state and AI-Tutor's Existing Hints	Example of one set of high-quality crowd-revised hints	Example teachers comments on crowd-revised hints	Teacher rating
 <p><b>Original hints:</b></p> <ol style="list-style-type: none"> <li>1. You have constants on both sides: 5 on the left and 6 on the right. How can you have all constants on the left and no constants on the right?</li> <li>2. How can you get rid of 6 on the right side?</li> <li>3. Subtract 6 from both sides.</li> <li>4. Write 5 - 6 on the left.</li> </ol>	<ol style="list-style-type: none"> <li>1. You have constants on both sides: 5 on the left and 6 on the right. We want to get all the constants on one side. How do we do this?</li> <li>2. Let's get the constants together on one side of the equation: what should you do with the 6?</li> <li>3. Let's get that X alone on the right side. Remember that whatever you do on the right side you also have to do on the left.</li> </ol> <p><i>(Rating + Comment Revision condition)</i></p>	<p>T13: "Again, I would have started by moving the variable to the left, but these are good, reasonable, clear hints"</p> <p>T52: "Third hint is perfect. This is how I state it in my class"</p>	<p>5.77 out of 7</p>

Table 1. Example of highly-rated crowd-revised hints and expert feedback.

problem states, and for each problem state, we randomly assigned the given crowdworker to one of three conditions 1, 2, or 3, taking advantage of Qualtrics' randomization feature and branch logic. With a total of 41 crowdworkers (stages 2 and 3), randomization ensured that each condition had 12 - 14 crowdworkers. In stage 3, 105 workers rated all crowd-produced hints to filter the top-rated ones for expert evaluation.

**Measures.** As dependent measures, we used the mean of the teacher ratings of the hints' effectiveness in helping students learn, as a proxy for teachers' satisfaction of hints, on the basis that a key goal for hints in tutoring software is to help students learn the equation solving. Given our pipeline goal is to increase teacher satisfaction, teachers' ratings were used as a proxy for hint quality. For each of the five problem states, teachers rated five sets of hints (in the five conditions) in a random sequence. The hints in condition 1,2 and 3 were the 15 filtered ones (5 problem states  $\times$  3 crowd conditions). Teachers could optionally leave feedback on hints they rated. Table one shows one set of crowd-revised hints and example feedback experts left on it, as well as teachers' average rating.

## 4.2 Results

To see if giving guidance to crowdworkers took experts less time than doing the revisions themselves (**RQ1.1**), we calculated the time it took for teachers to rate, comment on, and revise the hints, using the log data from the survey platform. Averaging across the five problem states, it took teachers 37.2 seconds to read and rate a set of hints for a given problem state, an additional 57.6 seconds to comment on the hints, and an additional 145.6 seconds to rewrite them (Fig. 5, left). We measured the time taken for each activity (read+rate, comment, and re-write) separately, using the time logged by Qualtrics (the survey platform). Separate time measurement was possible because each activity was in a separate page in the survey. Our experiment showed it took teachers less time to guide workers to do revision tasks than to do the revisions themselves ( $M = 37.2$  vs.  $57.6$  vs.  $145.6$ ,  $SD = 9.36$  vs.  $9.86$  vs.  $23.22$ ,  $df = 52$ ,  $p < 0.001$ ), regardless of whether the guidance was in form of rating, commenting, or rating and commenting combined.

We performed a one-way ANOVA comparing teacher satisfaction for hints in five conditions. There is a significant main effect of among hints in five conditions,  $F(4, 1911) = 18.91$ ,  $p < 0.001$ . To

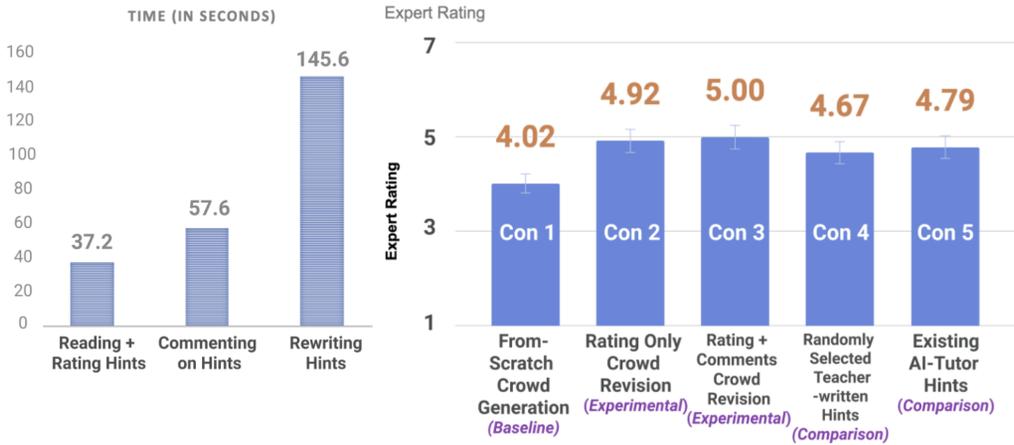


Fig. 5. Left: Average time taken for teacher experts to rate, comment or write hints in seconds; Right: Expert evaluation result of Study 1

answer **RQ1.2**, we conducted post hoc paired t-tests<sup>1</sup> comparing AI-tutor's existing hints (condition 5) with crowd-revised hints (condition 2 and 3). Results showed, although crowdworkers' revised hints were better-rated than the original ones, the improvement was not significant after Bonferroni correction (adjusted alpha level: 0.005), for either the *Rating Only Revision* condition ( $M = 4.92$  vs.  $4.79$ ,  $SD = 1.64$  vs.  $1.65$ ,  $df = 86$ ,  $p = 0.29$ ), or the *Rating + Comment Revision* condition ( $M = 5.00$  vs.  $4.79$ ,  $SD = 1.74$  vs.  $1.65$ ,  $df = 86$ ,  $p = 0.08$ ).

Con1: From-scratch Generation	Con2: Rating Only Crowd Revision	Con3: Rating + Comments Revision	Con4: Teacher-written hints	Con5: Existing AI-tutor Hints
$4.02 \pm 1.88$	$4.92 \pm 1.64$	$5.00 \pm 1.74$	$4.67 \pm 1.83$	$4.79 \pm 1.65$

Table 2. Mean and Standard Deviation (SD) for each hint condition evaluated in Study 1

To answer **RQ2.1**, we conducted post hoc tests comparing crowd-revised hints (conditions 2 and 3) to the crowd's from-scratch generation (condition 1). The crowd-revised hints were better than hints crowd generated from scratch, for either forms of the minimal guidance the crowd received: the *Rating Only Revision* Condition ( $M = 4.92$  vs.  $4.02$ ,  $SD = 1.64$  vs.  $1.88$ ,  $df = 86$ ,  $p < 0.0001$ ) and *Rating + Comment Revision* Condition ( $M = 5.00$  vs.  $4.02$ ,  $SD = 1.74$  vs.  $1.88$ ,  $df = 86$ ,  $p < 0.0001$ ). It's worth noting that for the crowd-revised hints evaluated, no full copy-paste behaviors were observed and substantial revisions were made in each set. To answer **RQ2.2**, we compared *randomly selected teacher-written hints* (condition 4) with the two crowd revisions (conditions 2 and 3). Surprisingly, the crowd-revised hints in the *Rating + Comment Revision* condition were higher-rated than hints generated by randomly selected math teachers ( $M = 4.67$  vs.  $5.00$ ,  $SD = 1.83$  vs.  $1.74$ ,  $df = 86$ ,  $p < 0.001$ ). The crowd-revised hints in *Rating Only Revision* condition were rated more highly than teacher-written hints, but the difference was not significant ( $M = 4.67$  vs.  $4.92$ ,  $SD = 1.83$  vs.  $1.64$ ,  $df = 86$ ,  $p = 0.06$ ). To investigate if experts' comments facilitated revisions by the crowd over and above experts' ratings only (**RQ3.1**), we compared

<sup>1</sup>Throughout the paper, post hoc statistical tests are two-tailed independent sample t-tests. We assume equal variance in t-tests if the population are the same population (teachers), and unequal variance if not.

the teacher satisfaction of crowd-revised hints in condition 2 and 3, and there was no significant difference observed ( $M = 4.92$  vs.  $5.00$ ,  $SD = 1.73$  vs.  $1.64$ ,  $df = 86$ ,  $p = 0.48$ ).

### 4.3 Discussion of Study 1 Result

Our results show that it took teachers less time to provide minimal guidance to the crowd (in the form of ratings and comments) than to revise the hints themselves. While such minimal guidance, together with the hints to be revised, helped crowdworkers produce revised hints that were better than when they generated hints from scratch, this minimal guidance was not adequate for the crowd to improve on AI-tutor's existing hints (of relatively good quality already). Nor did we see that the addition of expert *comments* could scaffold the workers over and above expert *ratings*.

We did, however, find that with one form of expert guidance (comments+ratings), crowd-revised hints (filtered) were of higher quality than randomly-selected expert-written hints. We see two possible explanations for this result: 1) crowdworkers, given multiple teachers' comments, were able to tailor the hint content toward a wider range of preferences of the teacher evaluators, whereas a randomly selected math teacher did not have this advantage, and/or 2) the filtering stage in the crowd pipeline effectively selected out high-quality crowd-revised hints. The findings suggest that when there is a practical, real-world need to improve materials to satisfy experts, adopting a minimally-expert-guided pipeline with quality control (e.g., filtering) may be a better solution than randomly choosing one member from the group of experts to revise the materials, both in terms of revision quality (i.e., expert satisfaction) and time efficiency (i.e., time savings for experts).

### 4.4 Analyzing teacher preferences - how much do they differ?

One factor that might limit the effectiveness of crowdsourced *generic* hint revisions is possible variability in teachers' preferences. If teachers have different or conflicting preferences for hint content, then, under the generic approach to hint revision that was the subject of Study 1, a crowdworker might receive conflicting comments regarding how to improve the hints for a given problem state. As a result, the crowdworker might be unsure where to go with the revisions<sup>2</sup>. We therefore analyzed whether teachers demonstrate varying preferences (RQ4.1), to help answer whether customization in AI-based educational software can increase teachers' satisfaction (RQ4).

#### RQ4.1 Are teachers' preferences on hints in AI-tutor homogeneous or diverse?

We first analyzed whether teachers' hint preferences converge or diverge, using their survey responses ( $n = 51$ ) regarding four aspects of math hints for an AI-tutor namely: 1) the proper length of hints, 2) how conceptual or procedural hints should be, 3) whether hints should contain answers, and 4) whether hints should adapt to students' strategy. As shown in Fig. 6 (upper part), teachers were relatively aligned in their preference for conceptual hints over procedural ones, and their preference for hints that adapt to student strategies versus not. They had more divergent preferences on the two other dimensions (hints' length, whether hints contain bottom-out answers) (Fig. 6, lower part).

Furthermore, from teachers' comments on the AI-tutor's existing hints, we found that teachers not only have *different* views on how hints should be authored, but sometimes *conflicting* ones. For example, whereas some teachers prefer hints to be *longer* and give more detailed explanations (e.g., T13, T29), some prefer them to be *shorter* as they think students do not like to read much (e.g., T6, T11); whereas many teachers prefer hints that use mathematically *accurate terminologies* (e.g., T17,

<sup>2</sup>Due to randomization and limitation in Qualtrics, we were unable to track which five comments each crowdworker received.

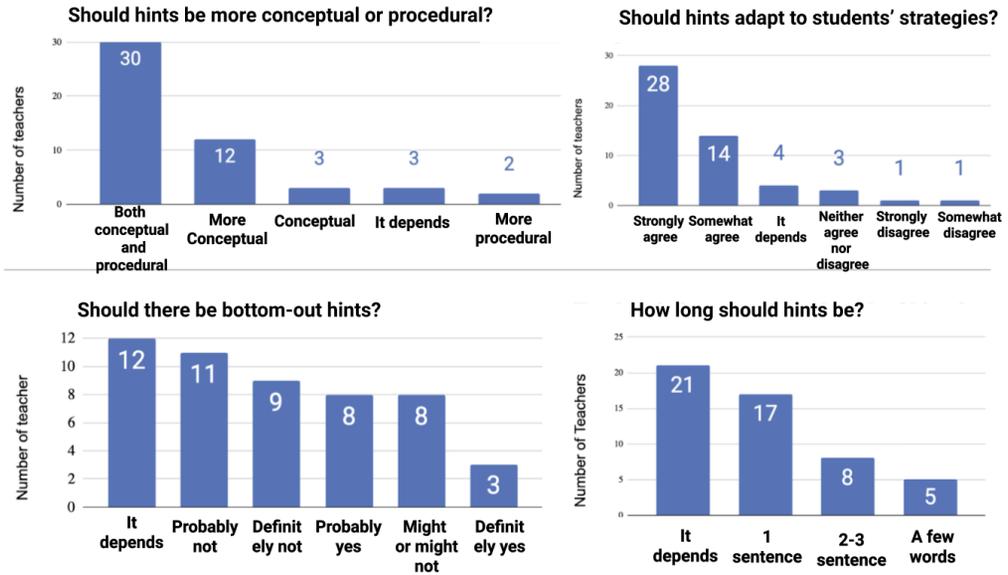


Fig. 6. Teachers' preferences on hints: preference that align (upper) and those that differ (lower)

T22), many prefer *simpler* language (e.g., T9,T20); whereas many teachers think having answers in hints *does not* help students develop transferable skills (e.g., T24, T47), some think it *does* (e.g., T23).

Additionally, some teachers had idiosyncratic preferences that seem hard to satisfy without customization. For example, they can be very specific about how long hints should be (“*The first hint should probably be 2-3 sentences at most. Any more than that & most students won’t even bother to read it. The rest of the hints should be 1 sentence.*” (T17)), or how conceptual they should be (“*The first hint should be more conceptual, but the second hint should be more procedural but with some conceptual in it and the last hint pure procedural.*” (T2)). These results show teachers can have different, highly specific and even conflicting preferences regarding math hints in AI-based educational software. Thus, teacher customization of instructional materials may be important for satisfying the varying and nuanced preferences of end-users (teachers) in this context.

## 5 STUDY 2- CUSTOMIZED HINT IMPROVEMENT

### 5.1 Motivation

Since we found teachers can have varied, highly specific, or even conflicting hint preferences, non-customized, generic fixes (i.e., fixes that try to address comments from multiple teachers at the same time) may not cater to everyone’s taste. We conducted study 2 to investigate if “*customized*” hints by crowdworkers (i.e., hints revised based on a single teacher’s guidance) can lead to greater teacher satisfaction. This experiment targeted the following research questions:

**RQ1.** Can an expert-guided crowd revision pipeline improve upon an AI-tutor’s existing hints, while saving experts’ time?

**RQ1.3.** In terms of teacher satisfaction, do crowdsourced customized revisions improve the AI-tutor’s existing hints?

**RQ4.** Can customization of instructional materials in AI-based educational software improve teachers’ satisfaction?

**RQ4.2.** In terms of teacher satisfaction, how do teachers' own revised hints compare to the AI-tutor's existing hints?

## 5.2 Experiment Design, Participants, Procedure

**Participants.** In this experiment, hints were customized for a subset ( $n = 18$ ) of the math teachers in Study 1. These teachers were selected as they had indicated in Study 1 that they would be willing to participate in a follow-up study. We also recruited 335 new US-based, adult MTurkers to customize hints (Stage 2 and 3) and performed quality control (Stage 4), each paid \$3 USD for one HIT completed, for a survey estimated to take 15 - 25 minutes. Post-study analysis showed the estimation was accurate (Stage 2 takes 22.4 minute and Stage 3 takes 19.1 minute on average).

**Procedure.** We assigned anonymous IDs to the 18 participating teachers, and labelled the data with these IDs, to make sure each teacher would later receive hint revisions customized specifically for them (i.e., crowd-revised hints based *only* on their guidance). The workflow largely emulates Study 1. The main difference was that in Study 2, each worker received only *one* teacher's guidance (in the form of rating + comments), instead of *multiple* as in Study 1. The customized revisions similarly went through a filtering stage, where each set of hints was rated by 8-12 MTurkers on their correctness, ease of understanding, and effectiveness in supporting student learning (Fig. 4). We averaged the MTurkers' ratings for each set of hints. We then selected 90 top-rated customized hint sets (5 problem states  $\times$  18 teachers), and sent them to the 18 teachers for evaluation. We received 12 teacher responses. For each problem state, each teacher rated six sets of hints in a random sequence. The six sets were: three of the conditions in Study 1 (Conditions 1,2 and,3), 4) the given teacher's own revised hints (comparison), 5) the AI-tutor's Existing Hints (comparison) and 6) the new crowd-customized hints condition (experimental). Our main goal was to compare teachers' satisfaction with the crowd-customized hints, the AI-tutor's existing hints and their own revised hints (i.e., hints sets 4, 5, and 6).

## 5.3 Results

Fig. 7 shows teachers' average rating on all sources of hints. **RQ1.3, RQ4.2** involve comparison among three sources of hints (customized, AI-tutor's existing hints, teachers' own revised hints). We conducted a one-way ANOVA and observed a significant main effect among the sources of hints,  $F(2, 162) = 12.53, p < 0.001$ . A post hoc test showed that, in the eyes of teachers for whom the hints were customized, the customized hints did not improve on the AI-tutor's existing hints ( $M = 4.25$  vs.  $4.36, SD = 1.77$  vs.  $1.68, df = 11, p = 0.74$ ).

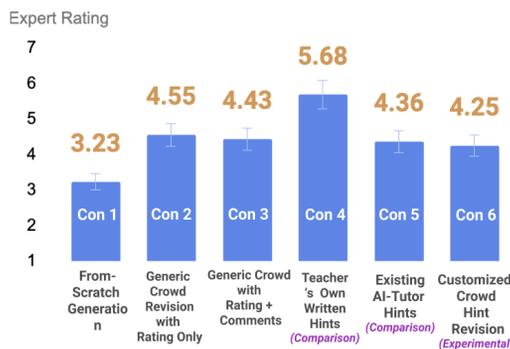


Fig. 7. Teacher evaluation for hints in Study 2

In a second pairwise test, we found that teachers perceive their *own* hints to be better than the AI-tutor's existing hints (**RQ4.2**) ( $M = 5.68$  vs.  $4.36$ ,  $SD = 1.22$  vs.  $1.68$ ,  $df = 11$ ,  $p < 0.001$ ). Additional comparisons showed that teachers perceived their own revised hints to be better than *any* of the other five sources of hints ( $p < 0.001$ ). Analyzing the 24 teacher comments left on their own revised hints, we found no evidence that they recognized *which* sets of hints were revisions made by themselves two months prior. This finding was not unexpected, as hints were presented in random sequence, and could bear considerable resemblance to each other.

#### 5.4 Study 2 Discussion

Study 2 found that teachers perceived hints written by themselves to be better and more satisfying than the AI-tutor's existing hints, but that they did not perceive the crowd-produced customized hints as improving on original hints. This finding indicates that there is room for improvement in the AI-tutor's existing hints. It suggests that customizing hints for teachers may, in principle, achieve higher satisfaction with the hints in principle, although customizing by crowds with minimal guidance, as in our study, did not achieve this goal. Interestingly, although teachers preferred their own written hints to *any* other sources of hints, their comments on their *own* edits from 2 months prior were not uniformly positive: About half of them (45%) contained some degree of critique to their own edits (Appendix A.2).

### 6 CHALLENGES IN REVISING MATH AI-TUTOR HINTS

Our two studies did not find that minimal guidance can help crowds to improve on the AI tutor's existing hints, regardless of whether the guidance was used for *generic* improvement (study 1) or *customized* improvement (study 2). In order to guide future iterations on the crowd pipeline design, we wanted to understand what might lead to this result and what challenges exist in the workflow (cf. [46]) (**RQ5**). To do so, we analyzed experts' feedback on the crowd's revisions.

**Qualitative Content Analysis.** Two HCI graduate students conducted qualitative content analysis [14] to systematically describe and categorize experts' opinions on the AI tutor's existing hints and the crowd-revised hints, and to see whether the experts' initial critiques were properly addressed. Researchers coded 692 teacher comments, including 130 from the Rating Only Revision condition (Study 1), 153 from the Rating + Comment Revision condition (Study 1), 29 from the Customized Revision condition (Study 2), and 380 on the AI tutor's existing hints.

Coding for each condition followed the same procedure, we here illustrate the analysis process taking the *Customized Revision condition* as an example. Firstly, coders reviewed each entry multiple times and inductively synthesized teachers' comments into higher-level codes. Coders iteratively reviewed existing codes for common, prominent themes that emerged, and went through a data reduction process (e.g., by dropping codes that appeared only once in the dataset). This process yielded a total of 29 recurring mid-level themes. Researchers then combined similar codes and converged on high-level themes, which were reviewed and agreed upon by each coder. Across all conditions, 7-9 high-level themes representing teachers' most prominent opinions for each hint source emerged.

**Data Triangulation.** Triangulation refers to the use of multiple methods or data sources to develop a comprehensive understanding of phenomena, and is a qualitative research strategy to test validity through the convergence of information from different sources [14, 56, 69]. We compared teachers' main opinions across different conditions, to explore what complaints are *common* across multiple conditions and what are *specific* to one. Our goal was to see which complaints crowdworkers *successfully resolved*, which complaints crowdworkers *failed to resolve*, and what types of *new problems* crowdworkers introduced, if any. The results of this analysis are shown in Tables 3 and Table 4, organized by condition.

Hint Source	Experts' Complaints
<b>Teachers think hint should</b>	
<b>Existing Tutor Hints And Crowd-sourced Revision</b>	<b>O1.</b> Ask <i>conceptual questions</i> and <i>scaffold</i> students to reach answers on their own.
	<b>O2.</b> Explain <i>why</i> certain moves are appropriate.
	<b>O3.</b> Mention the <i>ultimate</i> goal and intent for solving this problem.
	<b>O4.</b> Use more proper, mathematically <i>accurate language</i> .
	<b>O5.</b> Address the <i>highlighted step</i> students are currently on.
	<b>O6.</b> Use <i>notations</i> students and teachers are familiar with.
	<b>O7.</b> Use <i>simple language</i> to avoid overwhelming struggling students.

Table 3. Triangulation result - experts' complaints that persist.

Hint Source	Experts' complaints
<b>Teachers do not like hints that:</b>	
<b>Crowd-sourced Revision</b>	<b>Rating Only Revision</b> <b>O8.</b> <i>Skip steps</i> or stuff too many steps in one level because it might overwhelm students.
	<b>O9.</b> Seem <i>hasty and rushed</i> (in particular the later levels of the sequence).
	<b>Rating + Comment Revision</b> <b>O10.</b> Stay at an abstract level. They want hints to be gradually specific and more <i>procedural</i> .
	<b>O11.</b> <i>Miss important information</i> which may lead to confusion or misconception.
<b>Customized Revision</b>	<b>O12.</b> Contain <i>typo, inaccurate</i> or <i>incorrect</i> expressions in hints.
	<b>O13.</b> Are unclear or do not make sense.
<b>Teachers think hints should:</b>	
<b>Existing Hints in AI-tutor</b>	<b>O14.</b> delay giving answers after more scaffolding, or not give answers at all.
	<b>O15.</b> Be more <i>flexible</i> and <i>contextualized</i> , and allow students to choose answer sequences, skip steps or use other valid strategies.

Table 4. Triangulation result - experts complaints specific to each source of hint.

### 6.1 Complaints that the crowd failed to resolve

O1 - 7 in Table 3 are complaints that the crowd were *unsuccessful* in fixing (as indicated by complaints that persist and exist on both the original and crowd-revised hints). These include, teachers want hints to ask conceptual questions and scaffold students (O1), mention the ultimate goal(O3), use mathematically accurate language (O4), and notations teachers and students are familiar with (O5). These usually demand domain knowledge as well as ability to contextualize the complaints. Despite their general familiarity with the task domain, the crowd may lack (or fail to recall) the necessary content and pedagogical knowledge needed to address these complaints.

### 6.2 Complaints that the crowd resolved

O14 - 15 in Table 4 were complaints that the crowd appears to have been successful in resolving (i.e., teacher complaints that exist *only* on the original materials but not on crowd-revised materials). As

an example, a common teacher complaint on the AI tutor's existing hints is that they gave answers at the last hint level (so-called "bottom-out hints") (O14). Crowdworkers were quite successful in resolving this issue, by removing last hint level and/or rewording to avoid giving answers directly.

### 6.3 Complaints that were newly introduced by the crowd

The O8 - O13 in Table 4 were "new" issues that experts found in crowd-revised hints, (i.e., complaints that exist *only* in the crowd-revised hints, but not in the AI tutor's existing hints). These represent areas where crowdworkers may have made hints "worse", in teachers' eyes, than the original ones. The most prevalent newly-introduced critiques related to hints' *clarity* (O13), *consistency* (O8, O11), *concreteness* (O10), and *mathematical accuracy* (O12).

## 7 DISCUSSION

### 7.1 Findings from the Two Experimental Studies

Our two studies tested the effectiveness of an expert-guided crowd revision pipeline, with the goal of increasing teacher satisfaction with an AI tutoring system's hints, while minimizing teachers' time input. We found that generating minimal guidance for crowds took teachers less time than revising the hints themselves, indicating that minimally-guided crowdsourcing has the potential to save teachers time. With minimal guidance, the crowdsourced pipeline generates hint revisions that (in terms of teacher satisfaction) were better than two comparison conditions: (1) hints written from scratch by the crowd (with filtering), and (2) hints written by randomly selected teachers. However, the crowd's hint revisions did not improve upon the original hints.

In addition, we found that teachers strongly preferred their own revisions over revisions by the crowd or those by AI-tutor researchers, and they rated only their own revisions as improving on the original. However, teachers did not view revisions made by other teachers as improvements, even if they viewed their own revisions as improvements, suggesting that improvement in the eyes of experts may require customization to their (differing) preferences.

Further analysis revealed that teachers had highly stringent and specific requirements for hints, which made it hard for crowds to revise hints according to teachers' specifications without introducing new problems. The crowd faced challenges in understanding the experts' brief written comments and in implementing them in the form of effective edits, without raising new concerns among teachers. Teachers indicated that hints revised by crowdworkers tend to introduce new issues in terms of clarity, inconsistencies (e.g., skipping steps), and mathematical inaccuracies. Future pipelines targeting revision tasks should explore how additional scaffolding might mitigate such issues.

In sum, our results indicate a clear need to customize AI tutors' hints to individual teachers' preferences, as well as the potential for a crowd pipeline to save teachers' time. Our results provide a modicum of support for minimally-guided crowd revision, given that it improves upon both fully-crowd-generated hints or hints generated by randomly-selected teachers. However, our findings highlight the need for more elaborate scaffolds in the pipeline that help supplement the crowd's prior math knowledge with specific knowledge of teachers' requirements for adaptively-provided hints. The pipeline we have described is experimental, but part of the larger goal is to embed it economically in the regular infrastructure of classrooms using AI-based tutoring software. For example, we envision that such teacher-guided crowd pipelines for hint customization could be useful for educational technology companies who serve diverse teachers and students across a wide range of instructional content.

## 7.2 Challenges in Expert-guided Crowd-Revision Pipelines

Prior research shows that crowds can contribute high quality educational content, even in task domains that are similar to ours, (e.g., math word problems [1, 10], hints [10] and math tutorial videos [73]). Prior work also shows that crowds can succeed in writing tasks with expert guidance [9, 21, 34]. Most of these workflows focus on content generation tasks. This distinction between revision and generation is rarely made in the literature on crowdsourcing, but our experiments show that revision tasks may present unique challenges in designing minimally-guided crowd pipelines. Below, we outline three main challenges that we identified for such minimally-guided crowd pipelines targeting revision tasks, and provide insight regarding how future pipelines might overcome these challenges.

**7.2.1 Challenge 1: Communicating Experts' Nuanced Preferences to the Crowd.** One major challenge is to sufficiently communicate experts' preferences to the crowd. Experts' nuanced preferences may not be adequately conveyed through minimal guidance – at least, not in a manner easily digestible by crowdworkers. One design challenge for future work is to capture experts' explicit and implicit beliefs, and effectively communicate them to the crowd, without taking up too much of experts' time. One possible approach is the design of brief expert "preference profiles" that capture essential dimensions relevant to crowdworkers' tasks (a mockup is shown in appendix A.1). The current study provides a starting point for understanding what dimensions might exist in a profile for teacher hint revision, by uncovering features that teachers care about particularly strongly (see Sections 4.4 and 6). If experts' preferences turn out to be highly context-specific or likely to evolve over time, such profiles could benefit from being easily updated or expanded.

**7.2.2 Challenge 2: Help Crowd Develop Relevant Knowledge for Revision Tasks.** Another key challenge is helping the crowd learn the domain and pedagogical content knowledge needed to produce effective revisions. Explaining the underlying rationales behind the problem-solving procedures requires different knowledge than solving the problem itself. In our study, we see considerable variance in crowdworkers' abilities to generate such educational explanations. While some crowdworkers seemed comfortable (e.g., "Fun study. I enjoyed it." (C21)), others considered the task very difficult (e.g., "This was really hard, I am terrible at math. I hope I did okay" (C3)). In light of comments like these, one design challenge is to help crowdworkers recall or learn relevant domain knowledge and pedagogical content knowledge. While not the focus of the current work, future pipelines might explore how to effectively help novice workers acquire at least a baseline level of such knowledge in a relatively short amount of time. For example, future research could consider displaying contrasting examples of hint features (e.g., conceptual and procedural hints), having crowdworkers review experts' examples [18], or embedding a checklist of experts' gold standard [34], as learning or performance aid. The expert preferences uncovered in this work can provide an excellent foundation for creating instructional materials for the crowd.

**7.2.3 Challenge 3: Prevent Crowdworkers from Introducing New Issues.** Finally, it can be challenging for crowdworkers to avoid *introducing new issues*, in the process of addressing expert critiques. Unlike in content-generation tasks where crowdworkers generate materials wholesale, in revision tasks crowdworkers need to carefully weigh what to keep and what to change in existing materials. In a revision pipeline, successful task performance requires the crowd to 1) comprehend the existing materials, 2) comprehend and contextualize expert guidance regarding desired changes, and 3) address the expert-identified issues without inadvertently removing merits or raising new issues. One risk, for example, is that crowdworkers may over-correct in their revisions. In our work, one critique from experts of the existing hints is that they are not adequately *conceptual*. However, some crowdworkers may have over-corrected as they tried to address this critique. A few experts

subsequently complained that the revised hints were too *abstract*, and should be more *procedural* (“*Hints should be more specific by the third [hint] level*” (T49)). Future pipeline designs could explore methods to prevent crowdworkers from introducing new issues on clarity, correctness, consistency and concreteness as in our study.

To tackle these challenges, more scaffolding and practice is likely needed, but providing the crowd with (immediate) feedback may be especially helpful. For example, one possible approach may be to build an adaptive tutoring system that provides training for the crowd. Such a tutoring system could include immediate feedback on the requirements for generic material revision, and could also potentially embed expert preferences (e.g., taking inspirations from the teachers’ preferences uncovered in the current studies as a starting point) for *customized* material revision.

### 7.3 Implications for Future Work

We briefly summarize the implications of our findings for future work in CSCW. We argue that (1) future *expert-guided* crowdsourcing tasks should seek to concisely capture experts’ explicit and implicit preferences; (2) tasks that crowdsource explanations should help the crowd gain relevant pedagogical content knowledge needed to produce useful explanations, not just content knowledge; and (3) for tasks targeting *revision / customization*, scaffolds must be designed to prevent the crowd from inadvertently introducing new issues in the process of addressing old ones, or from rendering existing materials incoherent through their revisions.

The materials that our pipeline targets are math hints in AI-based tutoring systems, which are adaptive and contextual to students’ input (e.g., contrary to static, paper-based math worksheet). We expect that the insights presented in this work will generalize beyond our specific context (i.e. the revision/customization of adaptive mathematical hints in AI-based tutoring software), and can also inform the design of expert-guided crowdsourcing pipelines for short, contextual materials in other educational software and feedback systems (e.g., [13]). Much as intelligent tutoring systems provide students with adaptive help in the form of on-demand hints, these feedback systems may similarly provide users with context-sensitive help, such as contextual feedback when users are writing emails or coding a program.

## 8 CONCLUSIONS

In sum, our work contributes an empirical study to test the feasibility of a teacher-guided crowdsourcing pipeline for improving and customizing instructional materials, with the goal of increasing teachers’ satisfaction while minimizing their time input. To our knowledge, our study is the first to explore the effectiveness of minimal forms of expert guidance, to support generic crowdworkers in educational content revision/customization tasks. We experimented with two forms of minimal expert guidance explored in this work and contributed analysis and insight regarding where additional scaffolding is most needed. By delving into a case where a crowd revision pipeline *did not* ultimately improve upon the original materials, this research responds to recent calls in CSCW, HCI, and educational technology for greater reporting of null and negative results to avoid “file drawer” effects in the literature.

Our study shows that experts’ preferences for hints in educational software can vary considerably, and that there are opportunities to increase experts’ satisfaction by tailoring hints to their preferences. However, the design of such expert-crowd revision pipelines need to overcome a number of specific challenges, including 1) adequately capturing experts’ explicit and implicit preferences and communicating these to crowdworkers, 2) helping the crowd learn relevant content and pedagogical knowledge, and 3) ensuring coherence in revision tasks and preventing the crowd from introducing new issues. While we cannot say that overcoming these three challenges alone

will be *sufficient* to produce a successful crowd revision pipeline, our work suggests that these may be enabling conditions.

### **ACKNOWLEDGEMENT**

This work was supported in part by Grant #1822861 from the National Science Foundation (NSF) and Grant #N00014-18-1-2755 from the Office of Naval Research (ONR). Any opinions presented in this article are those of the authors and do not represent the views of the NSF or the ONR. We thank all the participating teachers and crowdworkers, and all the anonymous reviewers for their feedback in early version of the work.

**A APPENDIX**

**A.1 System mock-up showing teachers' preference in a multi-dimensional profile**

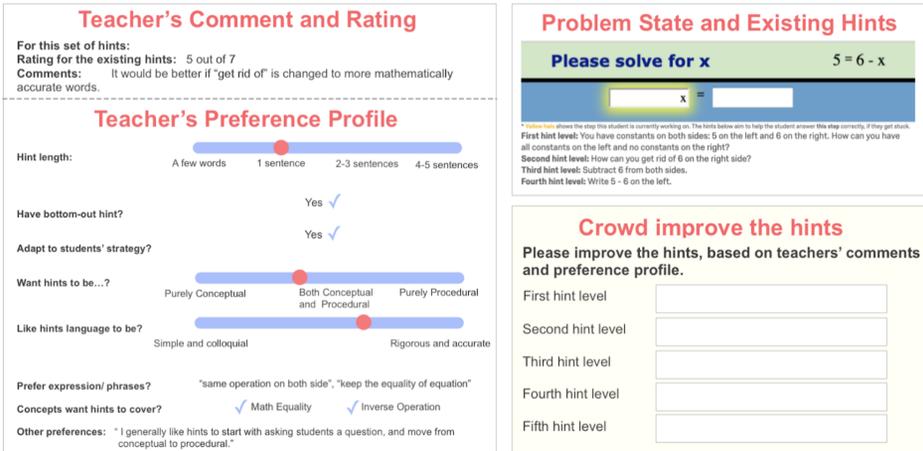


Fig. 8. Sample system showing teachers' preference in a multi-dimensional profile.

**A.2 Teachers' comments on hints edited by themselves 2 months prior**

Examples of Positive Comments	Examples of Negative Comments
"I like the use of questions"(T3-PS1). "It's not a hard problem. This is ample scaffolding" (T3-PS2). "Much better. Promotes understanding of the process and uses mathematical terminology" (T6-PS1). "The hint do a good job of defining vocabulary, when necessary. It also clearly separates hints by one step at a time"(T5-PS1). "Good guides, mathematical terminology, makes students think"(T8-PS1). "Good explanations, math terminology, guiding for students. Not doing for them" "Good questioning to guide the student, needs further steps, perhaps"(T8-PS5). "This is my favorite!"(T9-PS3) "This is exactly what I would have written and/or said if tutoring a student for this problem. This is clear and straightforward."(T18-PS3). "good understanding of what is happening"(T16-PS3) "This is very good. It helps move the student along and includes explanation"(T18-PS5). "clearest one so far"(T16-PS1).	"Equations are about balancing - it is missing property of equality"(T1-PS1). "The word "undoes" would need to be defined in another place"(T1-PS4). "Vocabulary definitions are missing"(T5-PS5). "I don't especially love the word unchanged since -4x and 7x didn't *change* their identify as terms...they didn't morph into something else, the first statement is still equivalent to the 2nd"(T3-PS5). "This is simple, but I'm not sure it helps the student get why they do this, or maybe what it means to simplify"(T8-PS2). "Good questioning to guide the student, needs further steps, perhaps"(T8-PS5). "you give the complete answer in hint 3 you may want to change that to the hint 4"(T9-PS5). "better not great"(T16-PS2). "The levels seem scattered. Some are written better than others"(T18-PS1). S: Well i heard someone mention alleles earlier but I thought dominant and recessive t" "This is confusing in tone and language"(T18-PS2). "I think the language is clunky here with, 'how do you move?'"(T18-PS4).

### A.3 Example of hints (original and crowd produced) for one problem state

**Problem State: 1**

**Hints for this problem state:**

Original hints in the tutoring system	Crowd written from scratch (Study 1 - Condition 1)
<p><b>First Level Hint:</b> You have constants on both sides: 5 on the left and 6 on the right. How can you have all constants on the left and no constants on the right?</p> <p><b>Second Level Hint:</b> How can you get rid of 6 on the right side?</p> <p><b>Third Level Hint:</b> Subtract 6 from both sides.</p> <p><b>Fourth hint level:</b> Write 5 - 6 on the left.</p>	<p><b>First Level Hint:</b> the variable x moves to left side. so it gets positive sign</p> <p><b>Second Level Hint:</b> it gives x=6-5</p> <p><b>Third Level Hint:</b> subtract 5 from 6 it gives 1</p>
Crowd Improved Hints with only Teachers Ratings (Study 1- Condition 2)	Crowd Improved Hints with Teachers Ratings and Comments (Study 1- Condition 3)
<p><b>First Level Hint:</b> keep only one type of term on each side of the equation. Either all variables on right and constants on left or all variables on left and constants on right</p> <p><b>Second Level Hint:</b> move x on the left side and 5 on right side</p> <p><b>Third Level Hint:</b> equation becomes x=6-5</p>	<p><b>First Level Hint:</b> You have constants on both sides: 5 on the left and 6 on the right. We want to get all the constants on one side. How do we do this?</p> <p><b>Second Level Hint:</b> Let's get the constants together on one side of the equation: what should you do with the 6?</p> <p><b>Third Level Hint:</b> Let's get that X alone on the right side. Remember that whatever you do on the right side you also have to do on the left.</p>

Fig. 9. Hints from different sources for one problem state, hints for the other four problem states can be found in Supplementary Materials.

### REFERENCES

- [1] Turadg Aleahmad, Vincent Alevan, and Robert Kraut. 2009. Creating a corpus of targeted learning resources with a web-based open authoring tool. *IEEE Transactions on Learning Technologies* 2, 1 (2009), 3–9.
- [2] Vincent Alevan. 2013. Help seeking and intelligent tutoring systems: Theoretical perspectives and a step towards theoretical integration. In *International handbook of metacognition and learning technologies*. Springer, 311–335.
- [3] Vincent Alevan and Kenneth R Koedinger. 2000. Limitations of student control: Do students know when they need help?. In *International conference on intelligent tutoring systems*. Springer, 292–303.
- [4] Vincent Alevan, Ido Roll, Bruce M McLaren, and Kenneth R Koedinger. 2016. Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 205–223.
- [5] JR Anderson. 1993. Rules of the mind Hillsdale. *New Jersey, Laurence Erlbaum Associates., 319p* (1993).
- [6] Joseph E Beck and Yue Gong. 2013. Wheel-spinning: Students who fail to master a skill. In *International conference on artificial intelligence in education*. Springer, 431–440.
- [7] Phyllis Blumenfeld, Barry J Fishman, Joseph Krajcik, Ronald W Marx, and Elliot Soloway. 2000. Creating usable innovations in systemic reform: Scaling up technology-embedded project-based science in urban schools. *Educational psychologist* 35, 3 (2000), 149–164.

- [8] Daren C Brabham. 2013. *Crowdsourcing*. Mit Press.
- [9] Joel Chan, Steven Dang, and Steven P Dow. 2016. Improving crowd innovation with expert facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1223–1235.
- [10] Yvonne Chen, Travis Mandel, Yun-En Liu, and Zoran Popović. 2016. Crowdsourcing accurate and creative word problems and hints. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.
- [11] Zhenyu Chen and Bin Luo. 2014. Quasi-crowdsourcing testing for educational projects. In *Companion Proceedings of the 36th International Conference on Software Engineering*. 272–275.
- [12] Ahmad Chettih, David Gross-Amblard, David Guyon, Erwann Legeay, and Zoltán Miklós. 2014. Crowd, a platform for the crowdsourcing of complex tasks.
- [13] Parmit K Chilana, Andrew J Ko, and Jacob O Wobbrock. 2012. LemonAid: selection-based crowdsourced contextual help for web applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1549–1558.
- [14] Ji Young Cho and Eun-Hee Lee. 2014. Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *The qualitative report* 19, 32 (2014), 1.
- [15] John Joon Young Chung, Joseph Jay Williams, and Juho Kim. 2018. Collaborative Crowdsourcing Between Experts and Crowds for Chronological Ordering of Narrative Events. (2018), 621–626.
- [16] Andrew Cross, Mydhili Bayyapunedu, Dilip Ravindran, Edward Cutrell, and William Thies. 2014. VidWiki: enabling the crowd to improve the legibility of online educational videos. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1167–1175.
- [17] 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 SIGCHI Conference on Human Factors in Computing Systems*. 3379–3388.
- [18] 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*. 2623–2634.
- [19] Shayan Doroudi, Joseph Williams, Juho Kim, Thanaporn Patikorn, Korinn Ostrow, Douglas Selent, Neil T Heffernan, Thomas Hills, and Carolyn Rosé. 2018. Crowdsourcing and Education: Towards a Theory and Praxis of Learnersourcing. International Society of the Learning Sciences, Inc.[ISLS].
- [20] 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*. 227–236.
- [21] 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*. 1013–1022.
- [22] 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*.
- [23] Mark Floryan and Beverly Park Woolf. 2013. Authoring Expert Knowledge Bases for Intelligent Tutors through Crowdsourcing. In *International Conference on Artificial Intelligence in Education*. Springer, 640–643.
- [24] Michael J Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. 2011. CrowdDB: answering queries with crowdsourcing. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. 61–72.
- [25] 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*. 1626–1636.
- [26] Elena L Glassman and Robert C Miller. 2016. Leveraging learners for teaching programming and hardware design at scale. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. 37–40.
- [27] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 203–212.
- [28] Neil T Heffernan, Korinn S Ostrow, Kim Kelly, Douglas Selent, Eric G Van Inwegen, Xiaolu Xiong, and Joseph Jay Williams. 2016. The future of adaptive learning: Does the crowd hold the key? *International Journal of Artificial Intelligence in Education* 26, 2 (2016), 615–644.
- [29] Thomas T Hills. 2015. Crowdsourcing content creation in the classroom. *Journal of Computing in Higher Education* 27, 1 (2015), 47–67.
- [30] Kenneth Holstein, Bruce M McLaren, and Vincent Alevan. 2017. Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. 257–266.
- [31] Kenneth Holstein, Bruce M McLaren, and Vincent Alevan. 2018. Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In *International conference on artificial intelligence in education*. Springer, 154–168.
- [32] Kenneth Holstein, Bruce M McLaren, and Vincent Alevan. 2019. Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics* 6, 2 (2019), 27–52.

- [33] Kenneth Holstein, Bruce M McLaren, and Vincent Alevan. 2019. Designing for complementarity: Teacher and student needs for orchestration support in ai-enhanced classrooms. In *International Conference on Artificial Intelligence in Education*. Springer, 157–171.
- [34] Julie S Hui, Darren Gergle, and Elizabeth M Gerber. 2018. IntroAssist: A Tool to Support Writing Introductory Help Requests. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [35] Nguyen Quoc Viet Hung, Duong Chi Thang, Matthias Weidlich, and Karl Aberer. 2015. Minimizing efforts in validating crowd answers. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. 999–1014.
- [36] Shiyan Jiang, Kexin Yang, Chandrakumari Suvarna, Pooja Casula, Mingtong Zhang, and Carolyn Rose. 2019. Applying Rhetorical Structure Theory to student essays for providing automated writing feedback. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*. 163–168.
- [37] Yuchao Jiang, Daniel Schlagwein, and Boualem Benatallah. 2018. A Review on Crowdsourcing for Education: State of the Art of Literature and Practice.. In *PACIS*. 180.
- [38] Paul A Kirschner. 2015. Do we need teachers as designers of technology enhanced learning? *Instructional science* 43, 2 (2015), 309–322.
- [39] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–456.
- [40] 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*. 43–52.
- [41] James A Kulik and JD Fletcher. 2016. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research* 86, 1 (2016), 42–78.
- [42] Anand Kulkarni, Matthew Can, and Björn Hartmann. 2012. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the acm 2012 conference on computer supported cooperative work*. 1003–1012.
- [43] Anand Kulkarni, Philipp Gutheim, Prayag Narula, David Rolnitzky, Tapan Parikh, and Björn Hartmann. 2012. Mobileworks: Designing for quality in a managed crowdsourcing architecture. *IEEE Internet Computing* 16, 5 (2012), 28–35.
- [44] Anand Kulkarni, Prayag Narula, David Rolnitzky, and Nathan Kontny. 2014. Wish: Amplifying creative ability with expert crowds. In *Second AAAI Conference on Human Computation and Crowdsourcing*.
- [45] Walter S Lasecki, Juho Kim, Nick Rafter, Onkur Sen, Jeffrey P Bigham, and Michael S Bernstein. 2015. Apparition: Crowdsourced user interfaces that come to life as you sketch them. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1925–1934.
- [46] Tianyi Li, Yasmine Belghith, Chris North, and Kurt Luther. 2020. CrowdTrace: Visualizing Provenance in Distributed Sensemaking. *Proceedings of IEEE VIS 2020* (2020).
- [47] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. 2010. TurkIt: human computation algorithms on mechanical turk. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*. 57–66.
- [48] Yanjin Long and Vincent Alevan. 2013. Supporting students’ self-regulated learning with an open learner model in a linear equation tutor. In *International conference on artificial intelligence in education*. Springer, 219–228.
- [49] Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P Dow. 2015. Structuring, aggregating, and evaluating crowdsourced design critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 473–485.
- [50] Wenting Ma, Olusola O Adesope, John C Nesbit, and Qing Liu. 2014. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of educational psychology* 106, 4 (2014), 901.
- [51] Camillia Matuk, Libby Gerard, Jonathan Lim-Breitbart, and Marcia Linn. 2016. Gathering requirements for teacher tools: Strategies for empowering teachers through co-design. *Journal of Science Teacher Education* 27, 1 (2016), 79–110.
- [52] Barbara Means, William R Penuel, and Christine Padilla. 2001. *The connected school: Technology and learning in high school*. ERIC.
- [53] Tomohiro Nagashima, Kexin Yang, Anna Bartel, Elena Silla, Nicholas Vest, Martha Alibali, and Vincent Alevan. 2020. Pedagogical Affordance Analysis: Leveraging Teachers’ Pedagogical Knowledge to Elicit Pedagogical Affordances and Constraints of Instructional Tools. (2020).
- [54] Benjamin D Nye. 2014. Barriers to ITS adoption: A systematic mapping study. In *International Conference on Intelligent Tutoring Systems*. Springer, 583–590.
- [55] Ronald Owston. 2007. Contextual factors that sustain innovative pedagogical practice using technology: An international study. *Journal of educational Change* 8, 1 (2007), 61–77.
- [56] Michael Quinn Patton. 1999. Enhancing the quality and credibility of qualitative analysis. *Health services research* 34, 5 Pt 2 (1999), 1189.
- [57] William R Penuel, Jeremy Roschelle, and Nicole Shechtman. 2007. Designing formative assessment software with teachers: An analysis of the co-design process. *Research and practice in technology enhanced learning* 2, 01 (2007), 51–74.

- [58] Kristen Purcell, Alan Heaps, Judy Buchanan, and Linda Friedrich. 2013. How teachers are using technology at home and in their classrooms. *Washington, DC: Pew Research Center's Internet & American Life Project* (2013).
- [59] Nguyen Quoc Viet Hung, Duong Chi Thang, Matthias Weidlich, and Karl Aberer. 2015. Erica: Expert guidance in validating crowd answers. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1037–1038.
- [60] Leena Razzaq and Neil T Heffernan. 2010. Hints: is it better to give or wait to be asked?. In *International Conference on Intelligent Tutoring Systems*. Springer, 349–358.
- [61] Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa Valentine, and Michael S Bernstein. 2014. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 75–85.
- [62] Steven Ritter, John R Anderson, Kenneth R Koedinger, and Albert Corbett. 2007. Cognitive Tutor: Applied research in mathematics education. *Psychonomic bulletin & review* 14, 2 (2007), 249–255.
- [63] D Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional science* 18, 2 (1989), 119–144.
- [64] Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research* 78, 1 (2008), 153–189.
- [65] Badariah Solemon, Izyana Ariffin, Marina Md Din, Rina Md Anwar, et al. 2013. A review of the uses of crowdsourcing in higher education. *International Journal of Asian Social Science* 3, 9 (2013), 2066–2073.
- [66] Michael Stonebraker, Daniel Bruckner, Ihab F Ilyas, George Beskales, Mitch Cherniack, Stanley B Zdonik, Alexander Pagan, and Shan Xu. 2013. Data curation at scale: the data tamer system.. In *Cidr*, Vol. 2013.
- [67] James Surowiecki. 2005. *The wisdom of crowds*. Anchor.
- [68] Ryo Suzuki, Niloufar Salehi, Michelle S Lam, Juan C Marroquin, and Michael S Bernstein. 2016. Atelier: Repurposing expert crowdsourcing tasks as micro-internships. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2645–2656.
- [69] Data Source Triangulation. 2014. The use of triangulation in qualitative research. In *Oncology nursing forum*, Vol. 41. 545.
- [70] Maaikje Waalkens, Vincent Aleven, and Niels Taatgen. 2013. Does supporting multiple student strategies lead to greater learning and motivation? Investigating a source of complexity in the architecture of intelligent tutoring systems. *Computers & Education* 60, 1 (2013), 159–171.
- [71] Xu Wang, Meredith Thompson, Kexin Yang, Dan Roy, Kenneth Koedinger, Carolyn Penstein Rosé, and Justin Reich. 2020. Practice-Based Teacher Education with ELK: A Role-Playing Simulation for Eliciting Learner Knowledge. (2020).
- [72] Daniel S Weld, Eytan Adar, Lydia Chilton, Raphael Hoffmann, Eric Horvitz, Mitchell Koch, James Landay, Christopher H Lin, and Mausam Mausam. 2012. Personalized online education—a crowdsourcing challenge. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.
- [73] Jacob Whitehill and Margo Seltzer. 2017. A Crowdsourcing Approach to Collecting Tutorial Videos—Toward Personalized Learning-at-Scale. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. 157–160.
- [74] 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*. 379–388.

Received June 2020; revised October 2020; accepted December 2020