Communicating Context to the Crowd for Complex Writing Tasks

Niloufar Salehi[1], Jaime Teevan[2], Shamsi Iqbal[2], Ece Kamar[2]
[1]Stanford University, [2]Microsoft Research
niloufar@cs.stanford.edu, {teevan, shamsi, eckamar}@microsoft.com

ABSTRACT
Crowd work is typically limited to simple, context-free tasks because they are easy to describe and understand. In contrast, complex tasks require communication between the requester and workers to achieve mutual understanding, which can be more work than it is worth. This paper explores the notion of *structured communication*: using structured microtasks to support communication in the domain of complex writing. Our studies compare a variety of communication mechanisms with respect to the costs to the requester in providing information and the value of that information to workers while performing the task. We find that different mechanisms are effective at different stages of writing. For early drafts, asking the requester to state the biggest problem in the current write-up is valuable and low cost, while later it is more useful for the worker if the requester highlights the text that needs to be improved. These findings can be used to enable richer, more interactive crowd work than what currently seems possible. We incorporate the findings in a workflow for crowdsourcing written content using appropriately timed mechanisms for communicating with the crowd.

Author Keywords
Crowdsourcing; communication; context.

ACM Classification Keywords
H.5.3 Group and Organization Interfaces: Collaborative computing, Computer supported cooperative work.

INTRODUCTION
Crowdsourcing is primarily used for short, independent microtasks that do not require background or a particular skillset. However, there is a recent shift towards employing the crowd to complete complex tasks [27, 6], such as writing [12], programming [23], and fact checking [19]. As the crowd's engagement with tasks becomes more complex and interrelated, workers can no longer complete tasks independent of the *context*, or the collection of conditions surrounding the task. For example, a task's context can consist of the knowledge space that the task is situated in or the requester's implicit preferences.

To understand how to effectively transfer context within a crowd work setting, we explore several transfer mechanisms within the domain of writing. In recent years, writing has emerged as a challenging domain for crowd work [3, 12], perhaps because it is a notoriously difficult task to start [8, 14] and thus could benefit from initial crowd-based assistance. Moreover, writing is an ideal setting for studying context transfer because it is a common yet complicated process that relies on contextual knowledge of the domain, the audience, and the preferred writing style [10]. This makes it particularly challenging for crowd workers to assist in writing tasks without access to the often implicit and hazy information inside the requester's head. For example, a requester who needs the crowd to write a summary of their research may not realize until they see the first draft that they would like it to use formal language. This type of context transfer is only possible through communication and iteration.

Despite its importance in enabling and improving crowd work, context transfer is not well supported on current micro-task platforms such as Amazon Mechanical Turk (AMT). In contrast, on expert crowdsourcing platforms such as Upwork, requesters and workers engage in lengthy one-on-one discussions. In a series of interviews with 3 expert writers on Upwork we found that requesting and integrating context is a challenging yet crucial part of their work that is necessary to achieve mutual understanding and produce higher quality work. However, open-ended discussions tend to be costly and time consuming, with important information often coming later in the process than needed. These interviews suggest that communication costs, or the efforts associated with transferring work to the crowd should be balanced with the value that they add.

In this paper we propose supporting the communication of task context between requesters and crowd workers using short, structured tasks via an approach we call: *structured communication*. We explore the design space of structured communication by developing and studying mechanisms that vary in the added value to the worker to accomplish the task and also in the costs imposed on the requester. For example, it may be easy for a requester to highlight aspects of a crowd-written paragraph that do not make sense and

 ACM 978-1-4503-4335-0/17/03…$15.00
DOI: http://dx.doi.org/10.1145/2998181.2998332

need to be improved, but hard for a crowd worker to understand how to improve the highlighted text.

After a discussion of related work, we describe five structured communication mechanisms for transferring context between a requester and crowd workers for a writing task. We developed these mechanisms iteratively using insight from expert writers on Upwork. We then present two studies designed to quantify the value and cost of each mechanism. Study 1 measures the added value to crowd workers in completing the tasks, and Study 2 measures the cost of each mechanism for requesters. We find different mechanisms perform well at different stages in the writing process. Finally we propose and test a workflow for crowdsourcing written content that uses appropriately timed mechanisms for communicating with the crowd. Such efficient communication may be critical in designing future crowd powered systems that can more effectively leverage the potential of the crowd.

RELATED WORK

We build on prior research designing crowd-powered systems that incorporate the knowledge of the crowd to tackle complex problems.

Importance of Context for Crowd-powered Systems

Microtask platforms are becoming an increasingly common means of distributing human computation. Recently, crowd-powered systems have shifted beyond piecemeal labor, toward enabling more complex roles that handle interrelated parts and demand awareness of context. For instance, Mobi enables workers to view the current state of global constraints while performing tasks [17, 28], and Turkomatic [20] asks workers to take responsibility in dividing tasks into subtasks, a process that had previously been done by the requester or through automatic workflows (e.g., find-fix-verify [2]). With the growth of these complex crowd-powered systems, suboptimal work is sometimes performed because individual workers lack global awareness, for instance during taxonomy creation [5].

To address these shortcomings, recent work has begun to explore methods for communicating context in a distributed manner. Some approaches use visual indicators to manage awareness between concurrent workers [17, 22], assign roles to different workers to mitigate conflicts [1, 16], or utilize hybrid crowd-machine systems to scaffold a big picture view through small contributions [12]. Another approach with significant benefits is enabling the requester to give timely feedback to workers [7]. These approaches build on traditional models of apprenticeship [24].

As the crowd takes on more complex roles in these systems, context sharing becomes a necessity. However, it is still unknown how requesters can effectively communicate context to the crowd (e.g., provide feedback), and when this context is useful. Prior work has shown that synchronous delivery of feedback has more of an impact on crowd work because workers are engaged with the task and are motivated to revise the work [7, 13]. However, providing prompt feedback places a burden on the providers of feedback [7]. Communicating with the crowd takes time and effort. We believe that structured communication mechanisms can reduce the effort by making the task of providing guidance as simple as a microtask for the requester, while ensuring that the guidance is timely for the worker, thereby increasing its value.

Challenges of Effective Communication

Supporting effective communication within groups has been a long-standing challenge within HCI research [11]. Specifically, researchers have explored the domain of communication within computer-mediated work groups [15]. Much of this research has focused on co-located groups or groups with the opportunity to have long-term communication, for example via telepresence [26]. However, communication is particularly challenging for distributed crowd workers who communicate exclusively online. While traditional work groups share common ground and have some context of the global view [25], this is often not the case with current crowdsourcing systems. Lack of communication limits the capability of crowdsourcing systems since workers cannot collaborate or receive guidance on their work [18].

Based on these patterns we explore the design space of structured communication within a crowdsourcing system, focusing on the domain of writing. We propose designing mechanisms for structured communication that will enable us to transition from open-ended, free-form context transfer to a more thoughtful and controlled process. The goal of the structured mechanisms is to reduce the cost of providing context for the requester, while ensuring that the information provided has value for the worker and enables them to produce content that is useful for the requestor.

DESIGNING STRUCTURED COMMUNICATION

We start by describing our process that led to the design of several different structured communication mechanisms. Following an iterative design process, we started by exploring how requesters share context with workers on an expert crowdsourcing platform, Upwork. We learn from strategies developed by these workers for transferring context in writing tasks. In the following sections we describe these mechanisms and study their value and cost.

How Writers on Upwork Acquire Context

We conducted semi-structured interviews with three experienced and highly rated crowd workers who perform writing tasks on Upwork. All three were female and held graduate degrees, they had each worked an average of 906 hours on writing projects on Upwork. We asked these writers about their experiences, focusing mostly on how requesters transfer the context necessary to complete a task.

Learning the context of a writing task is one of the most difficult, yet important parts of expert crowd work.
Upwork writers told us that their work requires extra effort to understand the context of the task at hand, which is often not provided in the original task description. For example, some tasks require research to understand a new domain before writing about it. Workers also need to know logistic requirements such as how long the text should be, who the audience is, and what the preferred tone is. Requesters, especially novices, may leave out this information or have difficulty abstracting out of their own contextual knowledge to the extent of making that knowledge explicit and transferring it to someone else. Moreover, much of this contextual knowledge is not well defined or fully known before hand and varies from one project to the next:

> *"Overall there's this thing that you can't really put your finger on, but you need it to be able to write"*

Experienced requesters use templates, rules, and samples to communicate context.
Professional publications use official style guide manuals to establish and enforce certain styles of writing. Similarly, Upwork writers told us that experienced requesters use templates, rules, and samples to communicate context. They may also ask writers to submit a writing sample before accepting them for a task.

Writers communicate back and forth with requesters to understand the context of the task.
Requesters may be inexperienced in working with crowd workers and may struggle to transfer the context of a task to a worker. Upwork writers told us that sometimes requesters "*don't really know what they want"* or have difficulty providing the necessary information. In these cases it is very important for workers to communicate with requesters.

Professional Upwork writers manage communication with requesters. For example they make sure that everything is clear before spending time on the task. Early in the process they ask questions, and later they send in-progress work for feedback. These communication habits are shaped by experience and previous failures. But not all workers have the experience to manage communication. In this work we seek to support effective context transfer by initiating and supporting structured communication between the worker and requester at appropriate times.

Writers find the communication process costly and also take into account the cost of communication for the requester
Discovering information needs and communicating them is a challenging task, as one worker put it:

> "*These things are usually not specific and figuring them out is really hard*"

Writers were also conscious of the cost of communication:

> *"Communication is very important but you don't want to bombard the requester with questions"*

There is a delicate balance between doing low-quality work due to not understanding the requester's needs, and communicating too often, wasting time on both sides. While experienced crowd workers learn to achieve the balance, others may fall either way. Therefore, while having more context generally helps, it is important to take into account the cost of communicating that context.

Method
Based on the findings from our interviews we engaged in an iterative design process to create mechanisms for structured communication. We chose three researchers' personal websites and for each person we asked two different Turkers to write a one-paragraph bio about them. We then provided guidance to other Turkers using different mechanisms with the final goal of making the bios suitable for a conference website. For instance, we asked Turkers to gather examples of good conference bios or compare two paragraphs written about the same person and merge them.

A common method that Upwork writers used to communicate with requesters was to ask questions. Following this approach, we explored the design space of enabling the crowd to ask questions. A key factor to the success of this mechanism is composing good questions. To understand the types of questions workers might want to ask requesters, we performed an initial data collection on AMT. We posted 120 tasks instructing a Turker to read one of our earlier paragraphs and author four different questions to ask the task requester that could help improve that paragraph. We then hired an expert worker to read all 480 questions, analyze, and categorize them.

Structured Communication Mechanisms
We designed the following structured mechanisms (Figure 1) for communicating context between the requester and the crowd. The goal of the mechanisms is enabling improved crowd work with minimal overhead for task requesters.

1 and 2. Ask Questions (Q&A) – i) General (Q&A general), and ii) Specific (Q&A specific):

In our pilot study we found that it was very difficult for crowd workers to author good clarification questions for the requester without guidance. For example, many questions focused on irrelevant details or asked for more information to add to the paragraph. To learn more about how to author good questions we hired an expert crowd worker to review and categorize the questions that Turkers had created in our pilot study. The resulting categories were: (1) general and (2) specific questions.

General questions were those that could be asked about any paragraph regardless of the topic, for example: *Is the goal to educate about the issue, or to elicit a reaction?* The expert created a library of general questions, containing the sub-categories: structure, goal, audience, and voice. Specific questions were questions about the content of the

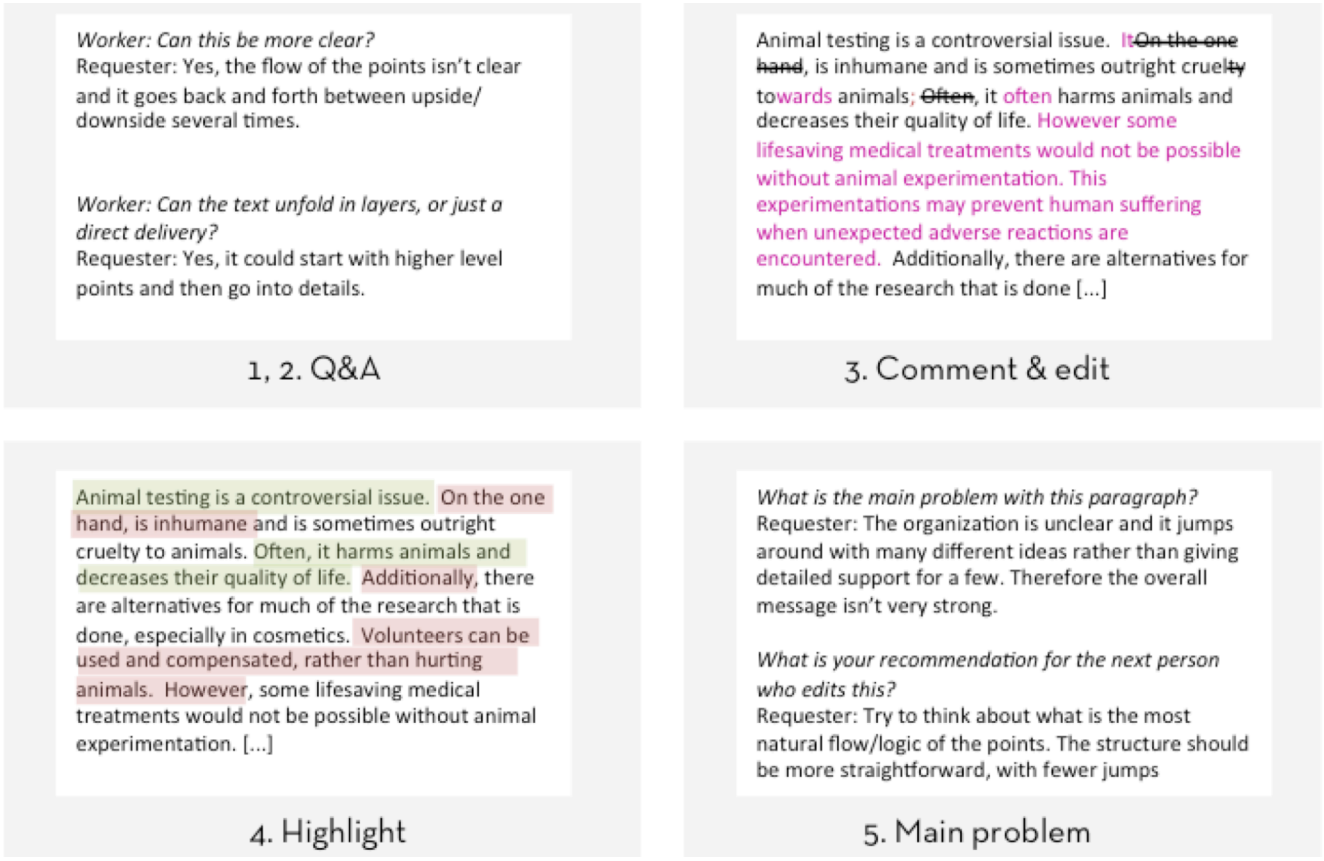

Figure 1. Structured mechanisms for communicating with the crowd. We compare these mechanisms based on the value that they add for the worker and their relative cost for the requester.

paragraph and often asked for more information or for clarification, for example: *What are the differences (if any) between communication and interaction?* Specific questions cannot be gathered in a library and reused, so our expert created a list of examples for specific questions as guidance for future crowd workers.

3. Free-form comment and edit (Comment & edit):

Another common practice on Upwork was sharing drafts and receiving comments and edits. This free-form method of communication is very strong because it gives the requester freedom to point out problems and areas of improvement either by leaving a comment or by fixing the problem directly. Moreover, unlike asking questions that require a certain level of abstraction, this form of communication is linked to tangible text, making it easier to communicate about complicated matters effectively. We implemented this method using the "suggestion" mode in Google docs that enables users to leave comments and make direct edits while tracking changes in another color.

In addition to these two methods that are frequently used on Upwork, we designed two other mechanisms. Our goal was to find methods that transfer important information with low costs by placing limitations on the requester.

4. Highlight strengths and weaknesses (Highlight):

In our early iterations we experimented with asking two different Turkers to write paragraphs from the same writing prompt, and have a third Turker merge the two. However, we found that merging two paragraphs was a very challenging task for crowd workers because they had difficulty deciding what parts of each paragraph were better. Based on this observation we designed highlighting as the third method for communication.

Highlighting is specifically designed as a method that minimizes the cost for the requester. In this method the requester receives text and can mark parts that they like in green and parts that they dislike in red. By limiting the requester's interactions, the goal is to reduce the effort that they need to put into the task and also design for scenarios where this interaction would happen on a device with limited input, such as a mobile phone or smart watch.

5. Identify the main problem and fix it (Main problem):

This mechanism seeks to find a middle ground between asking questions and free-form commenting and editing by providing some of the flexibility of the later while giving structure to the task. This mechanism follows our proposal that limiting the requester's interactions can result in less

costly interactions that are just as beneficial. For example, in the comment/edit scenario, a requester may waste time fixing low level problems that are cut out or changed in future edits. Whereas with this method the requester only spends time on the main issue that requires attention.

We implement this mechanism by showing the requester the current paragraph and asking them to respond to two predefined questions:

1- What is the main problem with this paragraph?
2- What is your recommendation for the next person who edits it?

In practice we found that requesters responded to these questions with 1-2 sentence responses. Thus this method was successful at creating a structure that reduced overhead for the requester.

To learn more about the effectiveness of each of these mechanisms as well as the imposed costs, we conducted two studies. The first study looks at the value of the communication for crowd workers, and the second measures the cost for the requester.

STUDY 1: VALUE OF STRUCTURED COMMUNICATION

The goal of Study 1 is to measure how valuable these five structured communication mechanisms are for crowd workers improving a paragraph to meet a requester's needs.

Method

First, we created a corpus of 66 mediocre initial paragraphs on a fixed number of topics as starting points. We then applied each of the five mechanisms to the paragraphs to collect additional context on how to make the text closer to the requester's intentions; we refer to this information as *feedback*. To ensure consistency in the quality of the feedback, we asked a single person to play the role of the requester and apply each mechanism across all paragraphs. Finally, we provided the original paragraph together with the feedback to crowd workers to edit and measured the change induced by each mechanism. All four of these steps are described in more detail below.

We ran the study on Amazon Mechanical Turk (AMT). Our tasks were open to Turkers who: Lived in Canada or the US, had successfully completed at least 500 tasks on the platform, and had an approval rating of 98% or higher. We used these qualifications to ensure that our crowd workers were fluent in English and would produce high quality results. We compensated Turkers according to the Dynamo guidelines for academic requesters[1] that requires payment of at least the federal minimum wage in the United States ($7.25/h at the time of writing).

[1] http://guidelines.wearedynamo.org

[2] http://humansystems.arc.nasa.gov/groups/tlx/

Prompt: *Does technology make us more alone?*

Step 1: Create list of ideas:

- The way a person uses technology has the capacity make that person as alone or social as they want to be.
- downside: some people become isolated by technology, always having their nose pointed at a screen
- downside: children who are taught to use technology may not learn to socialize effectively with people face to face
- upside: technology enables people to find activities and events to participate in
- upside: technology brings the world closer because news travels so quickly
- I think people who want to be alone would find a way to do so with, or without, technology.

Step 2: Write paragraph from list:

When considering if technology makes us more alone, the answer cannot be easily figured out because the way an individual uses technology has the capacity to make that person as alone or as social as they want to be. It's true that some people become isolated by technology and always have their nose pointed at a screen, however technology can enable people to find activities and events to participate in which could lead to friendships. Another drawback is that children who are taught to use technology may not learn to socialize effectively with people face to face, but again technology will also bring the world closer because news travels so quickly. In the end I think people who want to be alone would find a way to do so with, or without technology.

Figure 2. An example of a bullet point list of ideas created for the initial paragraph, and the resulting paragraph.

Creating the Initial Paragraphs

We used a two-step process to create the initial paragraphs, illustrated in Figure 2. In Step 1 we asked a crowd worker to create a list of ideas in response to a writing prompt. We used predetermined prompts to minimize topical variation. To encourage workers to engage with the topic, we provided a list of three topics to select from, chosen based on online writing resources and Turkers' votes:

1- What is your position on medical researchers and cosmetic companies performing experiments on animals?
2- What are the benefits of working on AMT?
3- Does technology make us more alone?

In Step 2 we asked a different crowd worker to create a cohesive paragraph based on the list of ideas created in Step 1. This two-step process enabled us to simulate a real-world writing request from a hypothetical task requester and the initial crowd written response.

Sixty crowd workers performed Step 1, creating lists that covered a variety of stances for each prompt. In Step 2, we hired 120 workers who each wrote a single paragraph based on one list of ideas. We gathered two paragraphs per list so that we could examine paragraphs of varying quality that started with the same initial list of ideas.

Our initial paragraphs varied significantly in quality. To reduce this variation, we evaluated paragraphs to identify ones that were of reasonable quality but had room for improvement. We hired an expert crowd worker from AMT with a degree in English, who we call the *reviewer*, to assign a score from 1 to 5 to all 120 paragraphs. Of the 60 idea lists created in Step 1, 33 resulted in paragraph pairs where one had a rate of 3/5 and the other 2/5. We selected these 66 paragraphs for further study.

Collecting Structured Communication Mechanisms

In the next phase we needed to annotate all 66 paragraphs with feedback from our five structured communication mechanisms (e.g., answering questions or highlighting segments). To do this we hired the same reviewer to read all 66 paragraphs and for each paragraph respond using one of five communication mechanisms in random order. Having the same person provide all of the feedback ensured consistency in the quality of the feedback and allowed us to observe variation in how the feedback is used versus what feedback is provided.

Improving the Paragraphs

After collecting the feedback, we posted a task on Amazon Mechanical Turk for each paragraph (N=66) and each communication mechanism (x5), instructing the Turker to edit the paragraph based on the requester's feedback. Each task consisted of one paragraph (created by the crowd worker) and feedback provided via one of the communication mechanisms. Our process resulted in 330 edited paragraphs.

To evaluate the changes in the paragraphs, we hired the reviewer again to perform a comprehensive assessment. We asked the reviewer to read all 396 (66 initial + 330 edited) paragraphs blind to condition and in random order and rate each paragraph on a scale of 1 to 5 across four criteria: writing quality, organization, word choice, and mechanics. In this study, our goal was to find communication mechanisms that were more successful at transferring the preferences and contextual information of the requester to the crowd. Therefore, we rely on final assessments from the same person who provided the feedback.

Results

The reviewer rated all 396 paragraphs across four different criteria (writing quality, organization, word choice, and mechanics); we calculate the average of all four criteria as the overall quality of the paragraph. Based on this score we divide the 66 initial paragraphs into three categories:

- **Low quality**: Paragraphs with a score of 2, N=13
- **Average quality**: Paragraphs with a score greater than 2 and less than 3, N=26
- **High quality**: Paragraphs with a score of 3, N=15

This enables us to understand how paragraphs of different quality changed after crowd workers edited them.

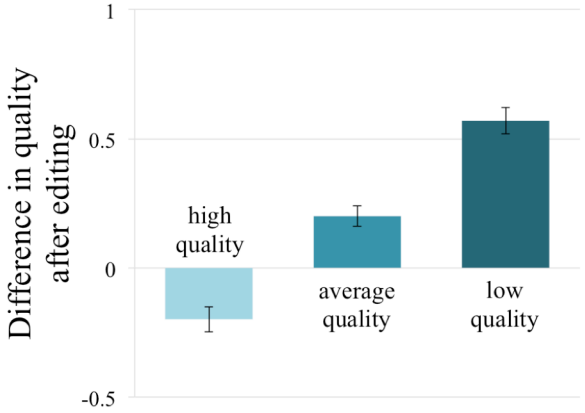

Figure 3. The effect of editing 330 paragraphs after receiving one round of feedback from the reviewer. Paragraphs that were initially low quality were easier to improve whereas paragraphs that were initially high quality were frequently made worse due to miscommunication.

Low and average quality paragraphs improved after revising based on communication with the reviewer. However, high quality paragraphs became worse.

Communication enabled crowd workers to improve low quality paragraphs (Figure 3). A Wilcoxon Signed-Ranks Test for paired ordinal measures indicated that the scores of edited paragraphs were significantly higher than the score of initial paragraphs, Z=-6.43, $p < 0.01$. Average paragraphs also improved after one round of feedback and editing, Z=-4.72, $p < 0.01$. However, editing reduced the quality of high quality paragraphs, Z=-3.68, $p < 0.01$.

Crowd workers had a harder time improving high quality work than low quality work. In fact without adequate guidance and skill, revising high quality work may actually make it worse. Therefore, we derive our first design implication for effective mechanisms for communication:

Design Implication 1: *Structured communication mechanisms with the crowd are most effective when the content quality is low or average. When content quality is high, limited communication can be counter-productive.*

Therefore we focus our attention on improving low and average quality paragraphs.

Different structured communication mechanisms had different effects on the paragraphs; the initial quality was important in determining the effect.

All of the communication mechanisms that we designed improved the low quality and average quality paragraphs, however the effect sizes differed. To measure the value of each mechanism we measured the difference in quality scores before and after the paragraph was edited (Figure 4).

For low quality paragraphs the *Main problem* mechanism created the highest average score differences. A Mann-Whitney U Test for ordinal measures showed that using the *Main problem* mechanism had a significantly greater effect

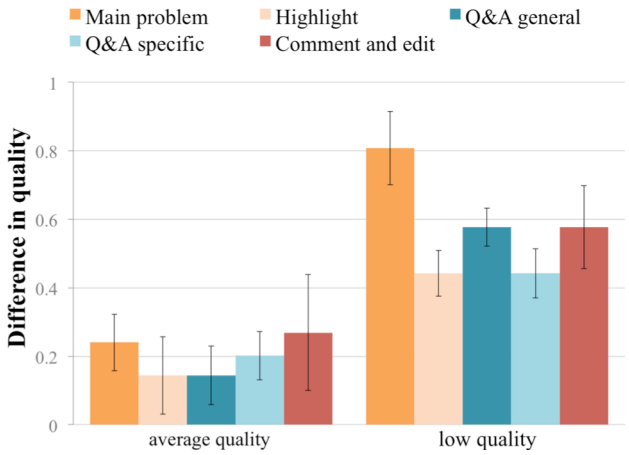

Figure 4. The effect of different structured communication mechanisms on the paragraphs. Low quality paragraphs benefited the most from the reviewer pointing out the paragraph's main problem and how to solve it. We did not find any significant differences for average paragraphs.

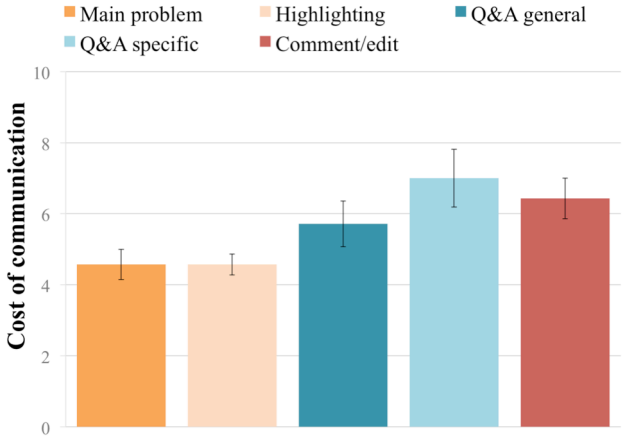

Figure 5. The cost of each structured communication mechanism is the sum of hard work and mental demand scores reported by requesters.

than the *Highlight* mechanism, U=50.5, $p < 0.05$. The *Main problem* mechanism was also significantly better than the *Q&A specific* mechanism, U=42.5, $p < 0.05$.

We argue that the *highlight* and *Q&A specific* mechanisms are least valuable for low quality paragraphs because they are closely tied to the content of the paragraph and when the paragraph is poor, struggle to provide information to improve it. Rather, in these cases, mechanisms that give the reviewer more freedom to transfer information are effective at enabling them to communicate major issues.

The average quality paragraphs were more difficult to improve and while we did see improvements, we did not find any significant differences between the different mechanisms for communication.

Design Implication 2: *Different structured communication mechanisms are effective at different stages. Early on, when content quality is poor, the requester needs to communicate major issues. Later, for average quality content, the different mechanisms have relatively similar added value.*

STUDY 2: COST OF STRUCTURED COMMUNICATION

Study 1 provided insights on the value added by communicating with the requester using the different mechanisms. In Study 2, we measure the cost these mechanisms impose on the requester.

Method

In study 1 we asked the same reviewer to provide feedback for all paragraphs because we favored consistency. In that study, the reviewer acted as a proxy for the requester to analyze how that feedback would affect crowd workers. Here we seek to learn how different requesters perceive the *costs* of each communication mechanism. Therefore we asked 7 researchers to provide feedback as realistic requesters who may use our system. We asked researchers to communicate with potential crowd workers on 7 different paragraphs that we randomly selected from our corpus of paragraphs in the previous study. By doing so we ensured that requesters read paragraphs written by crowd workers that were relatively consistent in terms of length and tone.

We asked each researcher to guide crowd workers to improve the paragraph that they were assigned to. Each requester completed 5 microtasks providing feedback using each of the mechanisms for communication in random order. They then answered questions based on the NASA TLX[2] test about mental demand and hard work of the task.

Results

Consistent with our pre-study interviews we found that *commenting and editing* and *answering specific questions* were the most costly forms of communication (Figure 5). *Main problem* and *highlighting* were the least costly.

Balancing the costs with the value added by each mechanism is important when deciding which mechanism to use. For example, *highlighting* is equally as capable as free-form *commenting and editing* for fixing problems in average paragraphs. However, highlighting is the superior choice for communication due to lower cost.

Design implication 3: *The cost of providing feedback via a structured communication mechanism is not always correlated to the value it provides.*

For instance, in the previous study we found that answering specific questions adds significantly less value to low quality paragraphs than identifying the main problem. However, a Wilcoxon Signed-Ranks Test for paired ordinal measures shows that answering specific questions is also significantly more costly for the requester (Z=1.92, p <0.05). On the other hand, free-form commenting and

[2] http://humansystems.arc.nasa.gov/groups/tlx/

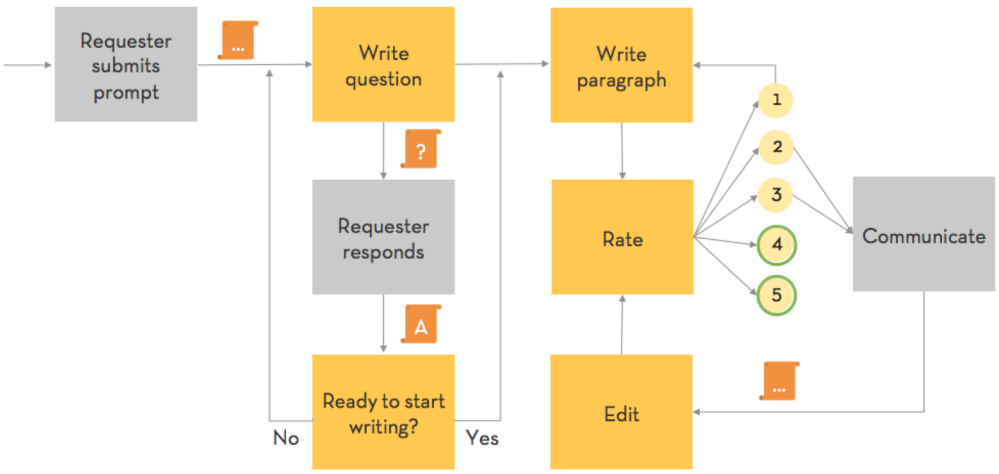

Figure 6. Full content creation workflow. The original prompt submitted by the requester goes through a series of questions and answers until the crowd is ready to write. After the paragraph is written the requester rates it, and communicates with the crowd through the mechanism that our system chooses. The crowd edits their work based on this new information. Yellow squares show tasks completed by crowd workers and orange squares show information moving between the requester and the crowd.

editing is significantly more costly than highlighting (Z=2.24, p <0.05), but we found no distinguishable difference in the value that they added to the low quality paragraphs. This suggests that giving requesters more freedom in communication does not necessarily result in more value. In practice we found that when given the option to provide free-form comments and edits requesters often engaged in low-level fixes instead of communicating major problems.

We did not find any significant differences in the costs associated with communicating on low quality vs. high quality paragraphs.

WORKFLOW DESIGN

Relying on the findings from our two studies we designed a workflow to support structured communication between a requester and crowd workers. We base our design on three key findings from our studies. First, structured communication yields information that can improve the paragraph. Second, the effect size of improvement is dependent on the current state of the paragraph and the communication mechanism. Third, communication mechanisms have different costs associated with them.

Our workflow goes through three main stages (Figure 6):

1) Validating and improving the starting point
2) Writing
3) Iterating

In this section we will explain each of these stages in turn.

Validating and Improving the Starting Point

Not all prompts that are submitted to our system will be ready to write about. Some will require more information or clarification; others may be impossible to write about, for example they may require domain specific knowledge that crowd workers do not have. The first stage of our workflow gathers more information about the prompt and validates it before moving to the writing stage.

We found that asking questions (both general and specific) were valuable for acquiring new information and asking for clarification. Therefore our workflow incorporates asking questions into the first stage, before writing has started. This is consistent with our initial interviews with Upwork writers who told us that they ask questions early on.

Sometimes asking questions led to the production of new content that the requester had not yet thought of. For example, in a number of cases the requester had written a bullet point response in favor of a new concept or idea, and a crowd worker asked them whether they would also like to include potential downsides, leading the requester to think of and respond with new information. This new information could also be in the form of instructions for the crowd. For example a crowd worker would ask whether the requester wanted their writing to have an emotional tone. The requester's response added new context for the writing task.

We found that it is better to gather information early on because adding new information later can hurt the writing. Therefore, the goal in the first stage of our workflow is to prepare the prompt for writing. This happens through asking a series of questions from the requester.

In our early deployments we left it to crowd workers to identify whether the prompt was ready to start writing or needed more information. This strategy was not effective because the task was not well defined. Turkers identified all prompts as ready to start writing and did not ask any questions. Therefore, we altered the task design to *require* asking questions. In this new design, a Turker would read a new writing prompt and write four questions for the requester, two general and two specific to the paragraph. They were then instructed to rate each question based on

how important it was to know the answer to be able to write a paragraph. Each Turker rated their questions on a scale of 1 to 5. In order to reduce the cost of answering unimportant questions, we only forwarded questions to the requester that were rated 3/5 or higher.

A critical decision in the first stage is identifying when we have gathered enough information to move on to the next stage: writing the paragraph. The Turker who has just written questions is in the best place to identify this as they have already read and thought thoroughly about the prompt. Therefore, after submitting their questions we ask the Turker to choose one of two options: *"after the requester answers these questions the prompt will be ready to start writing"* or *"this prompt is not clear, we will need to continue asking questions before we can start to write"*.

If a Turker chooses the first option, our workflow will gather the answers from the requester and move on to the next stage. If a Turker chooses the second option, our workflow will post a new question generation task after it receives the requester's responses to the first series of questions. This will give another Turker the chance to ask for more information. If a prompt goes through two question generation tasks and remains unclear, our workflow identifies a failure case and stops the process.

Writing and Iteration

The second stage of our workflow posts a task on AMT to write a paragraph based on the validated prompt and additional information gathered in the first stage.

Once the paragraph is written, our workflow enters the third stage: iteration. From our previous findings we know that it is important to know the current state to choose the best communication mechanism. Therefore, each iteration starts with the requester evaluating the current paragraph:

- If the requester rates the paragraph as 1/5, the quality is too low to be worth editing. The workflow returns to the second stage and gathers a new paragraph.
- If the requester rates the paragraph as 2 or 3/5, our workflow identifies the best communication mechanism and after one round of communication with the requester, asks a crowd worker to edit the paragraph based on this new information.
- If the requester rates the paragraph as 4 or 5/5, the workflow is complete.

Choosing the Best Form of Communication

Our workflow has to make a critical decision on which communication mechanism to use. This is a tradeoff between how effective that communication will be and how costly it is. To make this decision we maintain a table that keeps record of the *improvement score* for each communication mechanism and initial paragraph score. Our workflow uses this table to choose the least costly mechanism with the highest improvement score considering the current state of the paragraph.

For example, assume a requester completes a highlighting communication on a paragraph rated 3/5. Then a Turker edits the paragraph based on this new information and the requester evaluates the resulting paragraph 4/5. The improvement score for highlighting a paragraph of rate 3/5 will be +1. Therefore, each iteration adds a data point to our improvement score table that the workflow will use to make future decisions. At the beginning we pre-populated the improvement score table with data from study 1.

Offloading Communication

In study 1 we hired an expert crowd worker to act as a proxy for the requester. We showed that an expert crowd worker can successfully communicate with the crowd to improve a paragraph's writing and organization. Based on this finding we propose offloading parts of the communication role to the crowd. An expert in English writing communicates with the crowd to improve a paragraph's writing. Eventually additional expertise can be utilized. For example, a librarian fact checker can ensure that all facts are accurate; and a domain expert can guide the content within the writing process.

DEPLOYMENT

We used an instantiation of our content-creation workflow to write paragraphs for 10 academic researchers in response to the question: *If your research were incorporated into the real world, what would life look like in 10 years?*

Method

We evaluated our workflow with 10 researchers acting as requesters. We report on our observations of how they utilized the workflow. The workflow aims to maximize benefits gained from communication while minimizing costs to the requester. To do so we rely on lessons learned from our previous studies. Further research is required to verify that these benefits persist within the workflow.

Each researcher submitted a bullet point list of how they envision the future. Our workflow relied on the researcher to guide the content of the paragraph, and an expert in English writing to guide the writing, flow, and organization. This means that first the paragraph would go through a number of iterations, each time asking the researcher to evaluate the content and communicate with the crowd if needed. Once the paragraph's content achieved a high evaluation from the researcher, it would move on to the next stage. In this stage the writing expert would evaluate its writing, flow, and organization. Similarly, after a few iterations the paragraph's writing would reach a high evaluation and the workflow would be complete.

Results

The content creation workflow created 10 paragraphs for researchers on their vision of the future. Researchers' prompts received an average of three questions in the initial stage. These questions ranged from asking for more information about their research, to asking about their preferred tone and language. During the iteration phase

each paragraph went through one or two additional rounds of communication with the researcher or expert writer until it reached a 4/5 or higher rating from the requester.

Researchers were satisfied with the results and were willing to have their names associated with the final paragraphs if published publicly on the web. However, 2 researchers made revisions first. Researchers found the cost of communication with our system relatively low, reporting that they took on average 10.6 minutes to complete the task. They rated the mental demand of the task 3.2/5 and the hard work involved 2.2/5. All researchers reported that they would be willing to use the system again in the future for writing tasks aimed at a general audience such as a blog post or the abstract or introduction of a paper, but not for academic proposes or professional writing. They found that the paragraph matched what they would have written themselves in terms of content, but not writing style. Two researchers told us that they found value in hearing their ideas from someone else's perspective, helping them reframe their own ideas.

LIMITATIONS

In this paper we have focused our measure of cost on the cognitive costs to the requester. However, other important factors also contribute to communication costs. One is the cost for workers. For example, in our workflow, crowd workers generate more questions than needed and then rate the best questions to be asked from the requester. More efficient methods can reduce these costs. A requester's situation may also affect communication costs. In some scenarios the cost of communication may be more critical. For example, when supporting users with limited input on mobile devices or smart watches, or people facing accessibility issues. Future research will investigate ways to evaluate these costs more accurately and adjust accordingly.

In this work our design and application of structured communication is based on our experience in the domain of writing, future work will examine how to adapt our main design implications to other domains and other communication mechanisms. For example within a visual design project, an instantiation of our system could communicate with the requester to circle parts of a design that they like or dislike, or to communicate the main problem with a design and how to fix it. Future systems that aim to support communication in other domains will need to design mechanisms relevant to those domains, and measure their relative added value and costs at different stages of the process.

DISCUSSION

We have looked at the impact of structured communication on crowdsourcing writing, where the requester provides the initial ideas, and crowd workers turn those ideas into a coherent paragraph(s). We presented the results from two controlled studies on the cost and value of different mechanisms, and identified the mechanisms that required the lowest effort to produce the most value. Using these findings, we created a workflow that produced reasonable text with minimal effort on the part of the requester because it identified the most effective form of communication given the state of the text.

Our structured communication methods attempt to minimize the cost of communication for the requester. However, our studies showed that as a result, our mechanisms can be limited at extracting certain information. For example, in the Q&A mechanism, Turkers were often unable to come up with good clarification questions. When Turkers did ask strong questions, the context transfer happened sooner and easier as requesters found it relatively easy to answer a question. To help Turkers with generating questions, we tried different methods such as having the crowd author two paragraphs, then asking a Turker to compare them and ask questions to learn which one the requester prefers. We also tried more structured approaches such as fixing the style of the question and asking Turkers to fill in the blanks, for example: *which phrasing do you prefer __ or __?* We found the most effective method to be fixing the category of the question (general vs. specific) and providing libraries and examples. Future work will find new methods and techniques for generating valuable questions.

Our studies highlight the importance of understanding the current quality of the text and how close it is to the requester's goal. This can help with deciding whether to initiate more rounds of communication and deciding on the most effective form of communication. Sometimes, providing limited information to edit a paragraph that was already quite good ended up hurting it. However, we found that evaluation is challenging because of the expertise required and the inherent subjectivity of the task. For example, Turkers struggled to accurately evaluate writing quality. Prior research has also found that novices may have a hard time evaluating work [4]. Our model tackles this challenge by leveraging the strengths of different actors. The requester's judgment is very important for evaluating the content, while an expert writer (the reviewer) can evaluate and guide the writing quality. Furthermore, a crowd of workers who match the intended audience may do a better job at evaluating if the content is understandable.

Our approach raises further questions about the roles that different actors can fulfill within a system. How can we identify the best person to take on a task? What roles should the requester take on? What about the crowd and experts from the crowd? For example, in our initial conception of crowd writing, we assumed that the requester would review quality, but our studies suggested that an expert crowd worker can also evaluate work. Ensemble is an example of a system that supports complementary strengths of actors within a system [16]. Actors' roles within a system can also be dynamic and change to adapt to a new situation. For example, the requester may want to proxy the role of the

crowd by completing microtasks at a later time with a fresh pair of eyes or from a different device such as a smart watch. Future research will identify the strengths of different actors and how to best utilize them.

CONCLUSION

This paper explored ways to support the communication of context between requesters and crowd workers. To understand the strengths and trade-offs of various communication mechanisms, we performed three controlled studies on over 400 different paragraphs written by Turkers. We used different mechanisms to get feedback on each paragraph and had Turkers iterate on each paragraph based on the feedback that they received. We found that communication with the requester helped improve the quality of a paragraph, and that the effectiveness of different mechanisms for communication was related to the initial quality of the paragraph. Our findings can be used to enable rich, interactive crowd work that accomplishes tasks that are more complex than what seem feasible in crowdsourcing today and require the transfer of contextual information. We demonstrated this by exploring a workflow for crowdsourcing written content from a list of bullet points using appropriately timed structured communication mechanisms.

ACKNOWLEDGEMENTS

We thank Matthew Kelleher, the expert crowd worker who contributed to this project, as well as crowd workers on AMT and Upwork for their work. We are also grateful to the selfsourcing team at Microsoft Research, including Carrie Cai, Nick Greer, Rhema Linder, Ke Tran, and Rajan Vaish, for their valuable input.

REFERENCES

1. Paul André, Robert E. Kraut, and Aniket Kittur. "Effects of simultaneous and sequential work structures on distributed collaborative interdependent tasks." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 139-148. ACM, 2014.
2. Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. "Soylent: a word processor with a crowd inside." In Proceedings of the 23nd annual ACM symposium on User interface software and technology, pp. 313-322. ACM, 2010.
3. Carrie J. Cai, Shamsi T. Iqbal, and Jaime Teevan. "Chain reactions: The impact of order on microtask chains." In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI'16). ACM*, vol. 6. 2016.
4. Michelene TH Chi, Paul J. Feltovich, and Robert Glaser. "Categorization and representation of physics problems by experts and novices." Cognitive science 5, no. 2 (1981): 121-152.
5. Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. "Cascade: Crowdsourcing taxonomy creation." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1999-2008. ACM, 2013.
6. Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. "Toward a Learning Science for Complex Crowdsourcing Tasks." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2623-2634. ACM, 2016.
7. Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. "Shepherding the crowd yields better work." In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pp. 1013-1022. ACM, 2012.
8. Linda Flower. "Writer-based prose: A cognitive basis for problems in writing." College English (1979): 19-37.
9. Adam Fourney, Ben Lafreniere, Parmit Chilana, and Michael Terry. "InterTwine: creating interapplication information scent to support coordinated use of software." In Proceedings of the 27th annual ACM symposium on User interface software and technology, pp. 429-438. ACM, 2014.
10. Nick Greer, Jaime Teevan, and Shamsi T. Iqbal. An introduction to technological support for writing. Technical Report. Microsoft Research Tech Report MSR-TR-2016-001, 2016.
11. Jonathan Grudin. "Computer-supported cooperative work: History and focus." Computer 5 (1994): 19-26.
12. Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. "The Knowledge Accelerator: Big Picture Thinking in Small Pieces." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2258-2270. ACM, 2016.
13. R.J. Havighurst. Human Development and Education.
14. Suzanne Hidi and Pietro Boscolo. "Motivation and writing." Handbook of writing research (2006): 144-157.
15. Andrea B. Hollingshead, Joseph E. McGrath, and Kathleen M. O'Connor. "Group task performance and communication technology a longitudinal study of computer-mediated versus face-to-face work groups." Small group research 24, no. 3 (1993): 307-333.
16. Joy Kim, Justin Cheng, and Michael S. Bernstein. "Ensemble: exploring complementary strengths of leaders and crowds in creative collaboration." In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, pp. 745-755. ACM, 2014.
17. Juho Kim, Haoqi Zhang, Paul André, Lydia B. Chilton, Wendy Mackay, Michel Beaudouin-Lafon, Robert C.

Miller, and Steven P. Dow. "Cobi: A community-informed conference scheduling tool." In Proceedings of the 26th annual ACM symposium on User interface software and technology, pp. 173-182. ACM, 2013.

18. Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. "The future of crowd work." In Proceedings of the 2013 conference on Computer supported cooperative work, pp. 1301-1318. ACM, 2013.
19. Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. "Integrating on-demand fact-checking with public dialogue." In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, pp. 1188-1199. ACM, 2014.
20. Anand Kulkarni, Matthew Can, and Björn Hartmann. "Collaboratively crowdsourcing workflows with turkomatic." In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pp. 1003-1012. ACM, 2012.
21. Anand Kulkarni, Philipp Gutheim, Prayag Narula, David Rolnitzky, Tapan Parikh, and Björn Hartmann. "Mobileworks: Designing for quality in a managed crowdsourcing architecture." Internet Computing, IEEE 16, no. 5 (2012): 28-35.
22. Walter S. Lasecki, Juho Kim, Nicholas Rafter, Onkur Sen, Jeffrey P. Bigham, and Michael S. Bernstein. "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, pp. 1925-1934. ACM, 2015.
23. Thomas D. LaToza,, W. Ben Towne, Christian M. Adriano, and André Van Der Hoek. "Microtask programming: Building software with a crowd." In*Proceedings of the 27th annual ACM symposium on User interface software and technology*, pp. 43-54. ACM, 2014.
24. Jean Lave and Etienne Wenger. Situated learning: Legitimate peripheral participation. Cambridge university press, 1991. Longmans, Green and Co, 1955.
25. Gary M. Olson and Judith S. Olson. "Distance matters." *Human-computer interaction* 15, no. 2 (2000): 139-178.
26. Irene Rae, Gina Venolia, John C. Tang, and David Molnar. "A framework for understanding and designing telepresence." In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 1552-1566. ACM, 2015.
27. Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S. Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa Valentine, and Michael S. Bernstein. "Expert crowdsourcing with flash teams." In Proceedings of the 27th annual ACM symposium on User interface software and technology, pp. 75-85. ACM, 2014.
28. Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. "Human computation tasks with global constraints." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 217-226. ACM, 2012.