Crafting Policy Discussion Prompts as a Task for Newcomers

BRIAN MCINNIS, Cornell University, USA
GILLY LESHED, Cornell University, USA
DAN COSLEY, Cornell University, USA

Inspired by policy deliberation methods and iterative writing in crowdsourcing, we developed and evaluated a task in which newcomers to an online policy discussion, before entering the discussion, generate prompts that encourage existing commenters to engage with each other. In an experiment with 453 Amazon Mechanical Turk (AMT) crowd workers, we found that newcomers can often craft acceptable prompts, especially when given guidance on prompt-writing and balanced opinions between the comments they synthesize. However, crafting these prompts had little effect on the quality of comments they posted to a simulated discussion forum following the prompt task, as measured by the reasoning and topic coherence of comments. Our results inform best practices and pose questions for the design of discussion systems, both in general and for online policy discussion in particular.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**;

Additional Key Words and Phrases: Online discussion, crowdsourcing, deliberation, newcomer, meta talk

ACM Reference Format:
Brian McInnis, Gilly Leshed, and Dan Cosley. 2018. Crafting Policy Discussion Prompts as a Task for Newcomers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 121 (November 2018), 23 pages. https://doi.org/10.1145/3274390

1 INTRODUCTION

Research about online policy discussion routinely finds that people tend to talk *past* rather than *with* each other about their different perspectives on an issue [13]. One route toward more effective discussion is to include meta-talk in the discussion [62]. Taken from the policy deliberation literature, meta-talk refers to talk about the current state of a discussion, such as its tone [4, 10], opportunities for consensus [54, 61], or points of conflict [62]. In face-to-face policy deliberations, which often include a professional facilitator, this moderator will actively listen to the conversation for opportunities to insert meta-talk, often in the form of a discussion prompt related to the ongoing conversation [41, 50].

This strategy is hard to directly transfer to online policy discussions because the scale of both audience and discussion is far larger than individual moderators can manage [16, 36]. In this paper, we explore whether the work of crafting reflective meta-talk discussion prompts could be distributed to the community. More specifically, we focus on newcomers as a resource for crafting these discussion prompts.

Newcomers are often seen as a problem for groups to manage, with needs around socialization and mentoring [8, 17, 25] and often different perspectives than existing members [12]. Different perspectives, however, can be useful: as naïve outsiders to a group, they can raise questions or observations that other members may have forgotten or willingly ignore [1, 8, 9, 37]. Further, Kraut

Authors' addresses: Brian McInnis, Cornell University, USA, bjm277@cornell.edu; Gilly Leshed, Cornell University, USA, gl87@cornell.edu; Dan Cosley, Cornell University, USA, drc44@cornell.edu.

2573-0142/2018/11-ART121 $15.00
https://doi.org/10.1145/3274390

et al. [33] call on system designers to leverage these potential benefits as newcomers investigate an online community. As such, newcomers' fresh perspectives and minimal social constraints may make them well-suited for creating meta-talk-based prompts.

To this end, we designed a task that newcomers complete before they enter a discussion. In the task, they are asked to create a prompt for group discussion based on a pair of comments with different perspectives on a policy issue [55]. The design is inspired by existing human-computer interaction research about listening systems, notably *Reflect*, a micro-task workflow designed to interrupt the way that people reply to each other in an online discussion by segmenting the reply-and-response cycle into micro-tasks modeled after active listening principles [34]. It also draws on elements of text summarization in iterative crowd-writing systems, e.g., *Turkit* [39]. It goes beyond these, however, by asking participants not just to restate or summarize existing text, but instead to craft new questions to further an ongoing discussion.

A task that asks newcomers to listen to others' comments may also have beneficial effects on newcomers' later participation in the discussion. Kriplean et al. [34] offer an untested provocation that, "[...] listening interfaces help establish an empathetic normative environment," proposing that, "if the interface can encourage some users to listen, others may follow, helping to establish constructive communicative norms" [34, pg. 2]. Thus, we also ask how performing the task affects newcomers' comments in the policy discussion: are they more explicitly reasoned [62] or coherent with existing topics [44], as these measures indicate an effort to talk with, rather than past, others [22, 40, 58].

To address these questions, we conducted a controlled experiment in which Amazon Mechanical Turk (AMT) workers participated in a HIT that included a simulated online policy discussion about the AMT participation agreement. We briefed participants about the policy issue, surveyed them about their initial perspectives and relevant backgrounds, and exposed them to one of several versions of the onboarding task design that varied in terms of the degree of disagreement in the comments they were asked to design prompts for, whether they were creating a new prompt or improving an existing one, and the amount and structure of the task instructions. After completing the onboarding task, participants entered a simulated discussion forum where they could read other comments about the policy and post their own comments. Before exiting the HIT, participants were polled once more about their perspectives on the policy issue.

To evaluate the discussion prompts, we created measures inspired by the style of prompts that teachers develop for social studies lessons that involve classroom-based deliberation (see [23, 43, 55]). We found that participants were often able to produce acceptable discussion prompts, such that the prompt identifies commonalities and points of comparison between the two comments and asks a genuine question about the differences. They were better able to do this when given more instructions on how to write them and comment pairs with more disagreement; and in variations of the task where they were improving an existing prompt, when given more open-ended prompts to work on. However, completing the onboarding task had no effect on the reasoning or topicality of comments participants later made to the discussion forum. In fact, participants in task variations without instructions contributed longer forum comments than those who received instructions, raising questions about how instructions affected the way participants allocated their effort between the prompt creation and discussion commenting portions of the study.

Taken together, these results suggest that newcomers may indeed be able to create discussion prompts that community moderators might use to further a policy discussion. Giving clear instructions dramatically increased people's ability to effectively complete the task, adding weight to the importance of clear task design in the practice of crowd work, while the findings about the value of difference and open-endedness can inform the design of other crowd-based systems to support synthesis and reflection. Finally, our lack of evidence for beneficial effects of a meta-talk task on

newcomers' later contributions to the discussion raises questions for future work that hopes to more effectively onboard newcomers to discussion communities.

2 BACKGROUND

2.1 Meta-talk

2.1.1 Definition. A policy deliberation is a group-based discussion activity where a group is challenged to carefully weigh the diverse views of its members on a policy issue [7]. Policy deliberation scholars use a number of discourse analysis methods to study how people talk with each other during a deliberation [3]. Particularly relevant to our goal of leveraging newcomers' insights to pose questions for an existing deliberation are the problem-talk and meta-talk concepts from Stromer-Galley's method for *Measuring Deliberation's Content* [62]. This method defines an effective deliberation as incorporating six elements: reasoned opinion expression, sourcing, disagreement, equality, topic, and engagement. These elements are identified through discourse analysis that monitors the discussion for shifts in topic and for the presence of specific forms of talk: problem-talk, meta-talk, process-talk, and social-talk.

Whether phrased as a reasoned opinion or a rhetorical question, when people make statements that advance a claim, such as adding facts or arguments to the discussion, they are contributing *problem-talk* to the deliberation. By contrast, *meta-talk* is "talk about the talk" that clarifies and identifies potential consensus or conflicts related to the problem-talk, "[...] that attempts to step back and observe what the participant thinks has happened or is happening and why it's happening" [62, pg. 12].

As mentioned earlier, newcomers may be especially able to contribute certain kinds of meta-talk to a discussion. Because of their lack of investment in and history with a group, they may be more likely than existing members to provide new framings of existing arguments, raise issues a group has avoided but that need to be addressed, or ask genuine questions that bridge competing opinions, especially if they come in without strong opinions of their own. There are potential downsides to the idea of newcomers working to generate meta-talk as well—they might pose tired, old topics that are well-settled, or inadvertently trip over taboos or negative relationships—but we still see this as an idea worth pursuing.

2.1.2 Online Discussion System Examples. Other online discussion systems have explored the value of encouraging meta-talk, though usually among existing members. Several such systems have focused on negative interactions and disrespect between existing participants, as meta-talk is often observed calling out such behavior [4, 10, 49]. For instance, *Mediem* [54] is described as a "deep dialogue discussion forum" that provides several mechanisms for reflective interaction during a group discussion. Using a "Conversation Thermometer", participants rate and reflect on the quality of their discussion at specific moments.

In a more specific attempt to introduce meta-talk—and one that might be especially apropos newcomers—the *Reflect* platform was designed to scaffold an active listening exchange between speakers and listeners [34]. Reflect augments an existing discussion forum by incorporating a micro-discussion about a comment, primarily to clarify statements made by a speaker. In the Reflect workflow, a listener briefly summarizes "what they heard" in a comment, highlighting specific sentences and phrases. The speaker then responds to the summarized points by clarifying and responding to the listeners' interpretations. Although aimed more at supporting one-on-one interaction than the kinds of group processes meta-talk generally targets, Reflect was a real inspiration toward our task design, and we use a similar process as a baseline onboarding activity that emphasized listening to others' comments without the meta-talk aspects of our design.

2.2 Crafting a Policy Discussion Prompt

Having argued for the general idea of asking newcomers to contribute meta-talk to a discussion, we now turn to the more specific question of what sort of meta-talk to contribute. Meta-talk is often provided in formal policy deliberations by professional moderators [27]. Moore [50] observes that moderators often "[follow] from the front," prompting conversation with a *good question*, then getting out of the way, "follow[ing] the group as it unfolds its own discourse on the issue at hand" [50, pg. 4].

Several strands of research have looked at criteria and sub-components that make for effective meta-talk discussion prompts. For example, from the discourse analysis perspective, Schiffrin [59] operationalizes meta-talk as having a few common linguistic components: a *reference* (e.g., phrases, sentences) from the existing discourse, which are accompanied by *logical operators* that evaluate a reference or compare a set of references, as well as verbs that *request* an action related to the reference(s), such as clarify, tell, or argue.

There are close parallels between Schiffrin's formulation and aspects of teacher training in social studies [19, 23] aimed at the craft of creating policy deliberation prompts [43, 55]. Social studies teachers often use these prompts in their classroom practice to facilitate policy deliberations among students. Parker [55] recommends that teachers craft a policy deliberation prompt with a formula similar to Schiffrin's: (1) introduce a common problem and how it is personally relevant to the members of a discussion group, (2) logically compare a set of alternative solutions, and (3) request that the group make a decision. These are illustrated in the following example prompt, about whether teachers should reveal their own views on a policy to their students [55, pg. 14]:

> "(**1. Common problem**) You are learning a number of ways to identify controversies that are at the core of the topics you are going to teach and also a number of ways to help your students study those controversies. (**2. Alternative solutions**) Do you believe that teachers who engage students in the study of controversial issues should reveal their own positions on those issues? Or is it better for teachers to keep their opinions to themselves? (**3. Request**) I wonder if we can come to a consensus on this."

This prompt both illustrates the formula and calls out a larger issue that needs to be considered in crafting policy prompts and in moderating deliberations, around the risks of biasing the discussion through both a prompt's content and how it is presented. Rhetorical questions, for instance, are considered to be problem-talk that advances a position rather than genuine questions that characterize meta-talk [62, pg. 25], while subtleties of both the verbal and body language of a moderator posing the prompt can influence deliberation participants' behavior [60].

This line of work around crafting effective prompts to support meta-talk informed our task and experimental design. The elements of listening to the existing discourse, finding points of comparison and connection to call out, and asking genuine questions related to them are directly embedded in our coding scheme for evaluating the quality of prompts, the task instructions we developed, and some variations of the task interface.

2.3 Onboarding Newcomers to an Online Policy Discussion

As previously mentioned, during a face-to-face deliberation, a trained moderator will use discussion prompts to encourage the group toward deep consideration of a policy issue [41, 50]. In online discussions, a team of moderators might work together to facilitate large audiences, using specialized systems that help the team to promote dialogue among many participants [16, 18]. Whether in face-to-face policy deliberations [7] or online policy discussions [36], newcomers typically do not perform this type of facilitation.

In fact, there are often few expectations of newcomers to an online policy discussion. Existing members will greet newcomers [25], listening to their concerns and providing feedback [5, 6]. In online collaborative work, existing members will offer newcomers training and tasks [8, 17] to help them feel useful and invested through peripheral, but legitimate participation [21, 65]. These and other tactics are used to encourage newcomers to become regular members, as an online community needs new members to grow [33]. However, this work can tax existing members [17].

Rather than relying entirely on existing members, Kraut et al. [33] argue that the design of an online system should help people "[...] make a decision about joining and to respond to the common moves that newcomers use when forming impressions of the community" [33, pg. 3]. During what is referred to as the *Investigatory* phase of group socialization, a newcomer will collect information about the group to predict whether it will fit their needs [38, 51, 52]; such investigation in online discussions often involves reading a potentially large number of comments to make sense of the perspectives already in the discussion [32, 56, 67].

While investigating a group, a newcomer might raise questions that help existing members to recognize new ideas from older discussion points (called *Newcomer Innovation*) [9, 37]. We suggest that an onboarding activity of crafting a discussion prompt around existing comments might support newcomers' investigation while encouraging such innovation, explicitly providing occasion for both listening and for raising questions. The activity might also elicit feelings of investment, by tasking a newcomer with work that has value to the community.

2.4 Leveraging Ideas from Crowd-Writing to Build Policy Discussion Prompts

One issue with asking newcomers to craft discussion prompts as an onboarding task is that it is effortful: crafting a discussion prompt can take a moderator a substantial amount of time [55]. As newcomers are not likely to be willing to invest too much time or effort in a new group, we surveyed the crowd-writing literature to consider task designs that might ease this burden. Examples of crowd-writing systems we considered include *CrowdForge* [31], *Knowledge Accelerator* [20], *Mechanical Novel* [29], *Soylent* [2], *Storia* [28], and *TurkIt* [39]. These crowd-writing systems demonstrate how a complex writing task, such as crafting a policy discussion prompt [55], might be structured as a sequence of briefer tasks.

In particular, we focus on ideas from iterative writing workflows. Little et al. describe the process of iterative writing through crowd work as having three main general sub-tasks: writing, improving, and evaluating pieces of text [39]. Providing clear instructions and criteria for the write and improve tasks implicitly supports self-evaluation and generally improves performance [31], while mitigating risks of low-quality work or unfair rejection of work in micro-task markets [45]. Explicit support for helping people self-assess their work has also been shown to yield better quality writing [14]. Support for evaluating others' work can also help guide workers to make better suggestions for improvement or recognize that no improvement is necessary [20].

Another consideration is that workers performing improvement tasks are likely influenced by the text they are improving. A piece of text offered for improvement implicitly communicates information about both stylistic norms [20, 67] and the thinking of previous individual contributors. In a policy deliberation context, it is possible that the same kinds of biases that Parker [55] was concerned about in prompts crafted by teachers could be present in prompts crafted by crowds, raising questions about how the specific positions expressed in a prompt might interact with future contributors' own positions. In the context of our meta-talk task, we imagine that a worker asked to improve a discussion prompt that represented a position they are opposed to might change the sentiment, the question, or even the topic of the prompt. We also expect this to happen more generally in iterative writing tasks: people asked to engage with text that contradicts their own positions might find it hard to do so. For instance, in *Wikum*, a crowd-writing system designed to

summarize comment threads, participants assigned to summarize a political discussion reported feeling that, "[...] summarizing content they disagreed with took more effort" [67, pg. 2090].

Our task design addresses these considerations in several ways. First, in some conditions, we provide explicit criteria around listening, comparing, and questioning as described above, to allow us to study the effects of providing clear guidance. In others, rather than asking workers too write the prompt as a whole, we ask them to write separate sentences to address each criterion, with the thought that focusing attention on each of the three sub-tasks would help. Second, we surveyed each participant before and after the task to consider how having a neutral versus a non-neutral initial position on the topic may affect their performance. Third, we vary the positions expressed in the comments people are crafting a prompt around to explore how positional agreement affected people's ability to successfully complete the task.

3 STUDY DESIGN: CRAFTING DISCUSSION PROMPTS ABOUT AMT POLICY

3.1 Research Questions

Based on the above discussion, we designed an experiment to address a number of questions around the use of meta-talk tasks as a tool for onboarding newcomers to an online policy discussion.

RQ1. To what extent can newcomers to a discussion effectively create meta-talk-based discussion prompts?

- **1a**. How do people perform when writing a new policy discussion prompt versus improving an existing one?
- **1b**. How does the structure of the task affect performance? How much do instructions and explicit subdivision of the task into sub-tasks improve people's ability to complete the task?
- **1c**. How does the position toward the policy proposal of the selected comments affect performance, in terms of both their relationship to each other and to a newcomer's own position toward the proposal?
- **1d**. Apart from task structure and policy position, when and why do people tend to perform well or poorly at the task?

RQ2. To what extent does crafting meta-talk discussion prompts affect newcomers' subsequent contribution to the discussion, in terms of their engagement with others, topicality, reasoning, and effort, compared to performing baseline onboarding tasks?

In the experiment, Amazon Mechanical Turk (AMT) workers were invited, as newcomers, to an online policy discussion about the AMT participation agreement. Previous studies have shown that requesters rejecting work is a thorny issue [26, 42] and that AMT workers have proposed to amend the policy to provide them with either partial payment for the work they had completed or a second chance to fix their work [45]. We designed a study that simulated newcomers entering an online policy discussion around the Partial Payment proposal.

3.2 Onboarding Activity Conditions

3.2.1 Variations of the main meta-talk prompt task. To address our questions about newcomers' ability to write prompts, we developed a number of variations on an onboarding task that involved crafting a meta-talk discussion prompt based on comments drawn from a policy discussion. Participants in the meta-talk conditions were assigned to perform a *Prompt Task* (Write or Improve) with a particular *Design* (No Structure, Instructions, or Scaffolded) and *Content* (Biased or Balanced)—a 2x3x2 factorial between-subject design. Figure 1 presents a screenshot of the interface in the Write-Instructions-Biased condition, with labels indicating each experimentally varied component. Below we discuss in more detail how they varied.

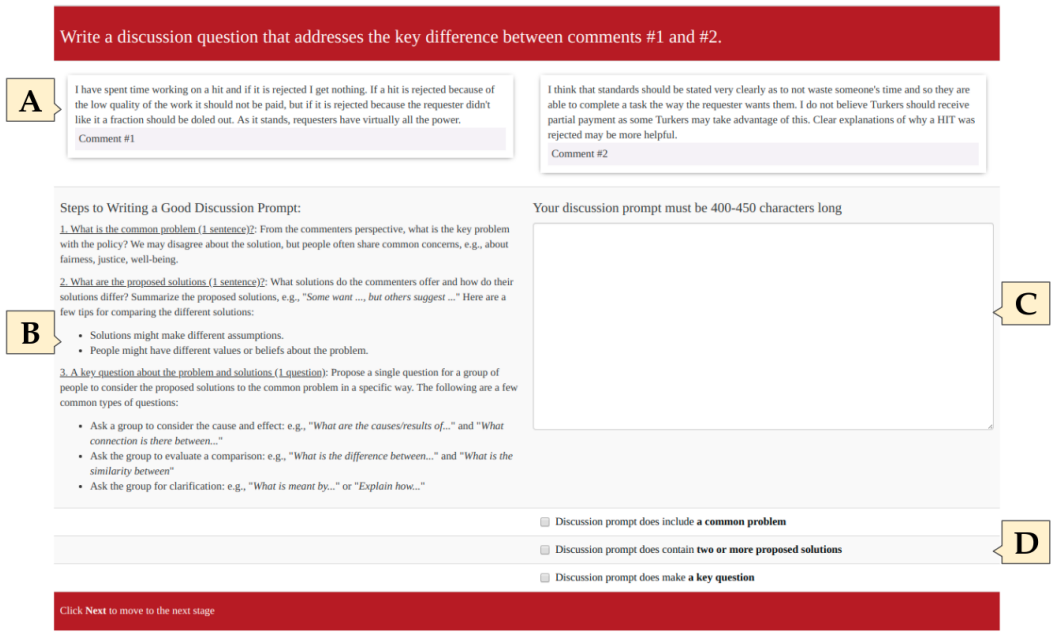

Fig. 1. Screen shot of the Write-Instructions interface. **A**. Participants were provided two comments as a reference. In the Improve task condition, the comments were accompanied by an existing prompt to improve. **B**. Steps to writing a good discussion prompt, provided in the Instructions and Scaffolded task layout conditions. **C**. Text area with a 400-450 character restriction. In the Scaffolded condition, the interface included three additional text boxes for each of the steps to writing a good discussion prompt. **D**. Self-assessment of the discussion prompt via check box items, which varied by the presence or absence of Instructions.

Prompt Task (Write, Improve): In the Write condition, participants were presented with an empty text box and were asked to "Write a discussion question that addresses the key difference between comments #1 and #2." In the Improve condition, participants were presented with one of three prompts that had been crafted by participants in the Write-Scaffolded-Balanced condition, then asked to "Improve the discussion question to address the key difference between comments #1 and #2." In both conditions, we restricted the submission text box to prompts between 400-450 characters, to reduce the effect of prompt length on acceptability, by imposing a standard. As we did not test an *Evaluate* condition, four researchers reviewed prompts from the Write-Scaffolded-Balanced condition to select the following by consensus for the *Improve* condition:

- *Prompt 1*: Defining a task clearly so workers will comprehend the requirements is the common problem. One solution proposes clearer standards for work, while another suggests both objective standards and partial payment. How can diverse requesters and tasks be made to hold to an objective standard of task clarity? Defining different levels of task difficulty could be helpful to workers in deciding if they will work on any particular task.
- *Prompt 2*: The concern seems to be the lack of appeal for workers, stemming from lack of communication between workers and requesters. One commenters suggests partial pay for rejected HITs, but the other says that better communication as to HIT guidelines and rejection reasons would solve the issue. What can be done on the Mturk platform to better facilitate communication about the requeters HITs from the task description to their acceptance/rejection?

- *Prompt 3*: There are concerns about the power balance between turkers and requesters, and how rejected hits are dealt with. Some would like rejected hits to be accompanied with a clear explanation for the rejection, while others feel they should be compensated for what they correctly did. Should rejected hits be viewed as an opportunity to learn from and improve ones turking, or should requesters seek to improve their own hits?

Task Design (No Structure, Instructions, Scaffolded): We varied the Write and Improve prompt task design in three ways. Participants in the No Structure condition were provided with the basic instructions listed in the Prompt Task description (above). Participants in the Instructions condition were also presented with a set of *Steps to Writing a Good Discussion Prompt* based on existing guidelines [24, 55] (shown in Figure 1B):

(1) *What is the common problem* (1 sentence)? From the commenters perspective, what is the key problem with the policy? We may disagree about the solution, but people often share common concerns, e.g., about fairness, justice, well-being.
(2) *What are the proposed solutions* (1 sentence)? What solutions do the commenters offer and how do their solutions differ? Summarize the proposed solutions, e.g., "Some want ..., but others suggest ..." Here are a few tips for comparing the different solutions: Solutions might make different assumptions, people might have different values or beliefs about the problem.
(3) *A key question about the problem and solutions* (1 question) Propose a single question for a group of people to consider the proposed solutions to the common problem in a specific way. The following are a few common types of questions: e.g., "What are the causes/results of..." "What connection is there between..." and "What is meant by..."

Participants in the Scaffolded condition saw these instructions, but instead of writing the discussion prompt as a whole, they assembled it by writing 1-2 sentences for each step, using separate textboxes for each step. The Scaffolded task was inspired by similar crowd-writing tasks [28, 34, 67]. In the Instructions and Scaffolded conditions participants were presented with three checkboxes corresponding to each of the steps to writing a good discussion prompt (Figure 1D), inspired by the observation by Dow et al. [14] that tasking participants to self-assess their work improved performance. In the No Structure condition participants were presented with a checkbox with the message: "Discussion prompt does not need further improvement."

Task Content (Biased, Balanced): The two specific reference comments on which participants were asked to base their discussion prompt could have a significant effect on performance. We therefore varied these two comments, drawing them from a set of three where two supported and one opposed Partial Payment:

- *Support-PP1*: "I feel that the requester should communicate more with the turker. I agree with the partial payment, there should be some kind of compensation given for the time spent on certain types of hits. With a transcription, there could be some small errors that the requester was not happy about, but the overall time spent was too great to ignore."
- *Support-PP2*: "I have spent time working on a hit and if it is rejected I get nothing. If a hit is rejected because of the low quality of the work it should not be paid, but if it is rejected because the requester didn't like it a fraction should be doled out. As it stands, requesters have virtually all the power."
- *Oppose-PP*: "I think that standards should be stated very clearly as to not waste someone's time and so they are able to complete a task the way the requester wants them. I do not believe Turkers should receive partial payment as some Turkers may take advantage of this. Clear explanations of why a HIT was rejected may be more helpful."

The Balanced condition included exposure to a support-oppose pair; the Biased condition included exposure to a support-support pair. Twenty participants were assigned to each Content pair variation in the Meta-talk conditions.

3.2.2 Baseline onboarding task conditions. Two other activities served as baselines to compare the effects of doing the meta-talk activity on subsequent contributions to the discussion forum.

AMT Policy Baseline Task. Participants assigned to this baseline condition were presented with an excerpt from the AMT Participation Agreement and were asked to propose three sentences to delete. After selecting three sentences, the participant was asked to respond to the following prompt: "Why are these the right sentences to remove from the AMT Participation Agreement?"

Active Listening Baseline Task. Participants assigned to this baseline condition were presented with a pair of existing comments drawn from the set of three used for the meta-talk task. Using an interface modeled after the *Reflect* platform [34], participants were asked to highlight a phrase in each comment that captures the commenter's position on the policy and then asked to write what they "[...] hear this commenter saying."

3.3 Recruitment and Procedure

The HIT description recruited Turkers to test the user interface of a novel online discussion forum system. We did not restrict access to the HIT (e.g., to Turkers from specific countries or with specific levels of experience); however, a majority of participants were U.S.-based Turkers who spoke English as a first language. All participants who completed the HIT were rewarded $4.50 for their time; the average time to completion was 30 minutes, which equates to an hourly rate of $9.

We received IRB-approved informed consent from all participants. While the HIT was active, we monitored Turkopticon [26] and TurkerNation [42] closely to listen for any problems related to the HIT, in addition to any concerns we received directly through the AMT interface. We also maintained an active "dummy HIT" to compensate Turkers who encountered a technical error with the system. Finally, we indicated that our research group is not associated with Amazon and that the purpose of the experiment was for research. We applied these actions based on recommended best practices for ethical conduct of academic research with crowd workers [46].

Prior to entering the experiment, participants were informed that they would be participating in a discussion about the policy topic "what should happen when a HIT is rejected?" and that they would have an opportunity to add their voice to the discussion. We informed participants that the intent of the discussion is to help resolve a lack of consensus among Turkers around two proposals to address the question: partial payment and second chance. Before starting the experiment, participants were asked to rate their initial preference toward both the partial payment and second chance proposals on a 5-item scale, from strongly disagree to strongly agree. They were also asked basic demographic questions and questions about their Turking experience, variables we used as control characteristics in our statistical analyses.

Participants were then randomly placed into one of the fourteen conditions to complete before entering the discussion forum. The discussion interface was modeled after RegulationRoom, a platform for civic engagement in public rulemaking [57]. The interface included two side-by-side panels (Figure 2). The left panel included a summary of the AMT Participation Agreement policy and proposals to amend it. The language for the summary, as well as the partial payment and second chance proposals associated with the deliberation "What should happen when a HIT is rejected?", were developed through a prior study [44].

The right panel included a discussion forum interface and a text box where participants could enter their comments. To populate the discussion thread, we selected 20 comments contributed in the prior study, half in favor of and half opposed to partial payment. We added time stamps to

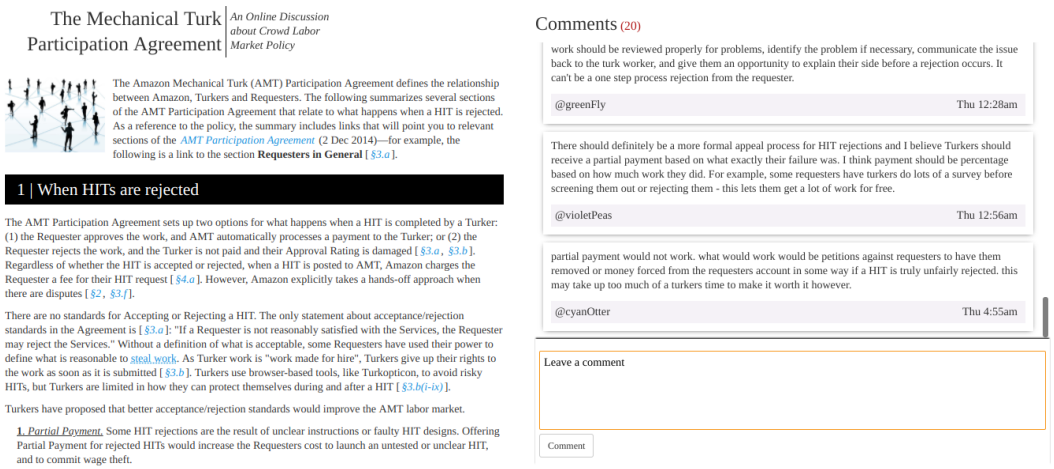

Fig. 2. Screen shot of the online policy discussion interface.

make the discussion appear recent and assigned participants a pseudonym (a concatenated color and animal, e.g., *@purpleOstrich*) to each comment in the thread. The interface did not include voting or other mechanisms for engaging with the content (e.g., reply, like, or share) as our focus was specifically on commenting behavior.

Participants received the following instructions when transitioning to the discussion forum: "You are about to begin the Discussion portion of the task. After 1 minute you will be able to leave the discussion and move onto the post-survey; however, you are welcome to remain in the online discussion as long as the HIT permits. Payment for this HIT is entirely based on your pre- and post-survey responses." Although participants were only required to spend one minute testing the discussion forum interface, the average dwell time was 4 minutes, 35 seconds (SD 3:37).

4 DATA ANALYSIS

4.1 Coding for Prompt Task and Discussion Comment Quality

We evaluated the discussion prompts based on criteria from the *Steps to writing a good discussion prompt* presented in section 3.2. Evaluations were conducted by two researchers, each training on a sample of 120 prompts and testing on a held out set of 65. Table 1 presents the Cohen's kappa scores for inter-rater reliability at each level of the evaluation criteria.

- **Listen**: (binary) The discussion prompt correctly identifies the common problem.
- **Compare**: (binary) The discussion prompt presents a comparison of the different views.
- **Question Type**: ("None", "Rhetorical", "Genuine") The type of question asked, if any.
- **Acceptable**: (binary) Acceptable discussion prompts ("True") are those that include Listening, Comparing, and a Genuine question.

The discussion comments were coded for their expressed position (e.g., agreement or disagreement with partial payment and second chance) and whether they presented some form of reasoning for their position. Because a comment might include a mix of positions, we coded for both partial payment and second chance separately. We also coded for three types of commenting behaviors that signify engagement with a policy discussion [62]. To identify whether a comment was Topic Coherent, we coded for whether the topic of the comment added to the topics present in the first three comments closest to the text box [44, 63]. We coded whether a comment included a genuine

or rhetorical question, or no question [62]. Finally, although the discussion forum interface did not include a "reply" feature, we coded for whether participants replied to another commenter by mentioning a user's pseudonym (e.g., *@purpleOstrich*), statements like "I agree with you," and evidence of linguistic entrainment (adopting the language of others).

	Train	*Test*	*Observations*	
Prompt Task Evaluation				
Listen	0.82	0.92	215	(66.9%)
Compare	0.92	0.90	211	(65.7%)
Question Type	0.81	0.85	229	(71.3%)
Comment position				
Partial Payment	0.86	0.84	190	(39.0%)
Second Chance	0.92	0.91	146	(29.9%)
Reasoned Opinion	0.85	0.72	236	(48.7%)
Discussion Comments				
Topic Coherent	0.81	0.78	187	(38.3%)
Question Type	1.00	0.85	23	(4.7%)
Reply	0.75	0.71	24	(4.9%)

Table 1. Report of the Cohen's Kappa score for inter-rater reliability, along with number and percentage of comments in which each code was observed. Two researchers trained with the coding scheme on 120 prompts and 120 comments until an acceptable level of inter-rater reliability was reached (Cohen's Kappa ≥ 0.7) and then tested on a held out set of 65 prompts and 65 comments. A total of 321 discussion prompts were crafted during the meta-talk onboarding task and 484 comments were posted to the discussion forum.

4.2 Control Variables

To account for characteristics about the participant that might relate to performance of the prompt task or to participation in the discussion forum, we incorporated a series of control variables surveyed before a participant was exposed to the experimental conditions.

As the bias of a moderator is an important concern when crafting a policy discussion prompt [55, 60], we pre-surveyed participants on their initial preferences toward the partial payment and second chance proposals. To evaluate the influence of having a position versus not having a position, we coded a binary measure of whether participants rated their preference as "Neutral" or took a stance (either "Agree" or "Disagree") about partial payment.

In group situations, social sensitivity—a measure of how well a person works with others—is associated with positive performance on group discussion-based tasks. While we did not survey for social sensitivity, women tend to score better on the measure than men, and groups with a higher proportion of female participants tend to exhibit a higher collective intelligence than other groups when discussion is involved in the problem-solving [66]. For this reason, we include the participant's stated gender identity (Not disclosed, Female, Male). Because workers have different levels of investment in AMT and this might affect their performance or their position on partial payment, we also surveyed participants about their time spent and daily income from AMT.

At the discussion comment level, longer comments might be more likely to include characteristics that are valuable in policy deliberation such as reasoned opinions and topic coherence. For that reason, we also control for the character length of comments.

4.3 Statistical Models

To evaluate how prompt task, design, and content affect prompt acceptability, we used a standard logistic regression to measure the binary variable of prompt acceptability, as well as each individual element of the acceptability criteria (i.e., Listen, Compare, Question, Genuine question). Model significance was evaluated using the log-likelihood ratio test to compare the goodness-of-fit of a model incorporating just the control characteristics with a model that also includes experimental variables. This procedure provides both a higher threshold than comparison with just the intercept and consistency in treatment of the control characteristics through the modeling process.

The model coefficients are interpreted as the expected change that each independent variable contributes to the logits of the response variable. Throughout the findings, we exponentiate the logits to present the odds ratios, which can be interpreted as the change in the response variable expected from a one-unit increase of an independent variable, holding all others constant. For example, in the Write Prompt condition Question model (Table 4), the independent variable "Gen: Female" indicates that female participants are 3.28 times ($p \leq 0.001$) more likely than male participants to write prompts that meet the Question criteria.

As a few participants contributed more than one comment, we treated Participant as a mixed-effects nesting variable to account for non-independence when examining the effect of the onboarding activity on later participation in the discussion forum. Mixed-effects logistic regressions were used to predict the presence of both a reasoned opinion and topic coherence at the comment level. As with the standard logistic regression to model prompt acceptability, we exponentiated the logit estimate for each coefficient to present the odds ratios for each response variable.

As each model incorporates several independent and control variables, to account for the effect of multiple comparisons, we applied a Tukey-based post-hoc comparison of the estimated marginal means (EMMs) at each level of a significant factor variable. Through this procedure, the effect and significance of a variable are averaged over the effect and significance of other variables in a model to provide a more accurate assessment of its impact.

5 FINDINGS

5.1 Descriptive Overview

Nearly half of those who started the task (988 participants) completed it (453). Those who completed were on average 34 years of age (SD 10 years) and about half identified as female (226 male, 219 female, 8 not disclosed). A majority of participants were located in the United States (85%) and speak English as a first language (92%). They reported earning about $15 per day (SD $13), doing tasks a few days per week for about 5 hours (SD 1.5) on average.

Table 2 shows the breakdown of participants by condition. Almost everyone who completed the task posted one or more comments in the following discussion (431 commenters, 484 total comments). Participants spent 4 minutes, 35 seconds in the discussion on average (SD 217 seconds), posting comments that averaged 297 characters (SD 226). Of the 338 participants in the meta-talk conditions, 321 attempted to write a prompt. To report quotes from the meta-talk task prompts, we assigned participants a unique ID ranging from P1-P321.

5.2 RQ1. Effects on Acceptability of Discussion Prompts

5.2.1 RQ1a. Prompt task. Table 3 reports on the overall and specific acceptability criteria for prompts in the Write and Improve conditions. In the Write conditions, approximately 38% of the Write prompts were rated overall acceptable ("True"), i.e., showed evidence of listening to points of connection, comparing the two comments, and asking a genuine question about them. In the

Baseline	Participants		
AMT Policy	56		
Active Listening	59		
Total	115		
Meta-talk	*Prompt Task*		
Design	Write	Improve	*Total*
No Structure	58	57	115
Instructions	50	56	106
Scaffolded	59	58	117
Total	167	171	338

Table 2. Breakdown of participants assigned to each of the onboarding conditions.

	Acceptable		(A) *Listen*		(B) *Compare*		(C) *Question*		
Prompt Task	True	False	True	False	True	False	None	Rhetorical	Genuine
Write	62	101	99	64	104	59	55	21	87
Improve	80	78	116	42	107	51	39	24	95

Table 3. Prompt acceptability by task (i.e., Write, Improve), both overall and on specific evaluation criteria.

	Acceptable		*Listen*		*Compare*		*Question*		*Genuine*	
Write Prompt	OR (SD)		OR (SD)		OR (SD)		OR (SD)		OR (SD)	
(Intercept)	0.10 (0.5)	***	0.62 (0.4)		0.66 (0.4)		0.73 (0.4)		3.97 (0.7)	*
D: Instructions	6.70 (0.5)	***	3.51 (0.4)	**	1.77 (0.4)		3.98 (0.5)	**	1.28 (0.7)	
D: Scaffolded	4.23 (0.5)	**	5.59 (0.4)	***	1.40 (0.4)		4.53 (0.4)	***	1.06 (0.6)	
C: Balanced	1.58 (0.4)		0.62 (0.4)		2.38 (0.4)	*	0.45 (0.4)	.	1.50 (0.5)	
Control Characteristics										
PP: Neutral	2.09 (0.4)	.	1.25 (0.4)		1.81 (0.4)		2.15 (0.5)		1.64 (0.6)	
Daily Earn	0.89 (0.2)		0.97 (0.2)		1.03 (0.2)		0.86 (0.2)		1.03 (0.3)	
Gen: Female	1.35 (0.4)		1.67 (0.4)		0.93 (0.4)		3.28 (0.4)	**	0.47 (0.5)	
Gen: ND	1.64 (1.1)		2.13 (1.2)		1.99 (1.2)		1.80 (1.2)		—	
Goodness-of-fit $\chi^2(3, 163)$	=22.2	***	=20.4	***	=8.3	*	=18.0	***	=0.8	

Table 4. Acceptable prompts in the Write task condition, by Design (i.e., No Structure, Instructions, Scaffolded) and Content (i.e., Biased, Balanced). To evaluate the goodness-of-fit of each model to the data, we used the log-likelihood ratio test to compare a model with just the Control Characteristics against the full model and report the χ^2 statistic. Note that including the condition variables did not significantly improve the fit of the *Genuine* model to the data. p-value significance codes: 0.0001 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Improve conditions, 50% of the prompts were rated acceptable, meaning that half of the initially acceptable prompts fed through an Improve task *declined* in quality.

5.2.2 RQ1b. Task design. Tables 4 and 5 report models of how the task design and content relate to performance on the Write and Improve tasks, respectively. When tasked to Write a prompt, the design of the task affected acceptability. Prompts written in the Instructions and Scaffolded task designs were 6.7 and 4.23 times, respectively, more likely to be Acceptable than those written in the No Structure design. The story is similar for the Listen and Question evaluation criteria: providing

Improve Prompt	*Acceptable* OR (SD)		*Listen* OR (SD)		*Compare* OR (SD)		*Question* OR (SD)		*Genuine* OR (SD)	
(Intercept)	0.37 (0.4)	*	1.22 (0.4)		1.59 (0.4)		1.42 (0.4)		1.15 (0.5)	
D: Instructions	1.66 (0.4)		1.75 (0.4)		0.83 (0.5)		1.32 (0.5)		3.31 (0.7)	
D: Scaffolded	0.56 (0.4)		0.89 (0.4)		0.47 (0.4)	.	0.80 (0.5)		0.89 (0.6)	
P: Prompt 2	3.98 (0.4)	**	2.53 (0.4)	*	1.94 (0.4)		2.38 (0.5)	.	5.61 (0.6)	**
P: Prompt 3	5.09 (0.4)	***	3.28 (0.4)	**	3.20 (0.4)	**	2.69 (0.5)	*	10.20 (0.6)	***
Control Characteristics										
PP: Neutral	0.95 (0.4)		0.73 (0.4)		1.03 (0.4)		0.77 (0.4)		1.17 (0.6)	
Daily Earn	0.82 (0.2)		0.92 (0.2)		1.11 (0.2)		0.95 (0.2)		0.64 (0.2)	*
Gen: Female	1.12 (0.3)		1.34 (0.4)		1.03 (0.4)		1.84 (0.4)		0.62 (0.5)	
Gen: ND	—		—		—		—		—	
Goodness-of-fit $\chi^2(4, 158)$=21.51 ***			=9.49 *		=9.81 *		=6.39		=18.04 **	

Table 5. Acceptable Prompts in the Improve task condition, by Design (i.e., No Structure, Instructions, Scaffolded) and Prompt (i.e., which specific prompt was being improved). As in Table 4, we report the χ^2 statistic from a goodness-of-fit log-likelihood ratio test of each model. Note that participants in the Improve-Scaffolded task were less likely to meet the Compare criteria, though this was not significant in a Tukey-based post-hoc comparison. p-value significance codes: 0.0001 '***' 0.001 '**' 0.01 '*' 0.05 '.'

instructions and scaffolding for the task contributed to people's ability to meet the criteria. We did not find an effect of task design on prompt acceptability in the Improve prompt task.

5.2.3 RQ1c. Content and position toward the policy proposal. The models in tables 4 and 5 also show the relationship between content and prompt acceptability in the Write and Improve tasks, respectively. When tasked to Write a prompt, the effect of having an initial position toward the policy on prompt acceptability was marginally significant, such that people who came in with pre-existing positions were less likely to craft an acceptable prompt (Table 4, "PP: Neutral"). Whether the reference comments were balanced or biased did not affect overall prompt acceptability, but it did affect participants' ability to write a prompt that compares them (Table 4, "C: Balanced"). When the comment positions were balanced with one in support and the other in opposition to Partial Payment, participants were 2.38 more likely to write a prompt that includes a comparison between the comments than if the comments are biased toward support.

In the Improve task, the content of the prompt to improve affected acceptability (Table 5). Specifically, improvements to Prompt 2 and Prompt 3 were 3.98 and 5.09 times, respectively, more likely to be acceptable than improvements to Prompt 1, suggesting there was something about the content of Prompt 1 that induced people to deviate from the instructions; we discuss this below.

5.2.4 RQ1d. When and why people tend to perform well or poorly. In both the Write and Improve tasks, some control characteristics are significant predictors of acceptability. Female participants are more likely to meet the Question criteria in the Write task, though this effect is not significant in the Improve task. Participants who report earning less per day on AMT were more likely to transform the provided prompt in the Improve task from a genuine to a rhetorical question than participants who earn more.

We next look more closely at the specific errors people made. In the Write task, participants in the No Instructions condition rarely wrote acceptable prompts (only 16%), so we focus on the unacceptable prompts written by people in the Instructions and Scaffolded conditions with the goal of discovering kinds of errors people made even when given useful instruction. To do this, we

manually reviewed the text of the unacceptable prompts in those conditions. This was not a formal content analysis coding process; we simply read through the unacceptable prompts and looked for representative examples of the errors that would be useful for other researchers and system designers to think about.

One pattern was the presence of personal opinion. Some participants ignored the separate components to writing a prompt and just offered an opinion. Examples include (P142) "[...] I think amazon should have some rules in places that help protected it's turkers [...]" and (P308) "[...] small tasks with set answers need to be checked before issuing a rejection. We could propose an automatic system when the tasks are big enough to catch mistakes." Many other participants provide some acceptable prompt components, but also a personal opinion, as in the next example that identifies a common problem and explicitly compares the comment pair, yet offers a personal opinion where there should be a question in the last sentence (P261):

> "Both the comments adressing HIT rejections. Comment #1 says reason to get paid for rejected work, based on the quality of the work. Where comment #2 says Requesters must communicate with turkers upon rejection, and make a partial payment for the time spent on the work. Whether there is payment made or not made, but rejections without fair reasons or just rejections made without looking quality of work, would really hurt a turker."

Other prompts imply an opinion by the way their question was phrased. The following are rhetorical questions posed by otherwise acceptable prompts:

- (P19) "How do we know why or how our work was rejected if the final say is all up to the requester?"
- (P87) "How often are undue rejections received for lengthy time consuming work any way?"
- (P116) "But doesn't doling out partial payment create more hassel for the requestor?"
- (P319) "Does the requestor not see it as fair for the worker to recieve either a [SC or PP]?"

Errors in the Improve task were largely due to the question in the provided prompt. By reviewing the less acceptable improvements, we realized that the question in Prompt 1 could be seen as being rhetorical or narrow: "*How can diverse requesters and tasks* ***be made to hold to*** *an objective standard of task clarity?*" The bold words imply holding Requesters accountable, rather than other means of improving task clarity. This meant that participants exposed to Prompt 1 had the additional challenge of converting the question to a less pointed form, such as (P111) "How can the requesters improve their standards of clarity?" or to a more open and genuine question (P309) "What are some ways in which a requester can help workers complete their HITs and get better quality work?" The question also makes an assumption that task clarity is the *right* solution, which invited opposition, such as (P276) "Will stronger guidelines really help?" and (P209) "If the requester states their instructions were clear but still rejected the work, who is at fault then?"

5.3 RQ2. Effects of a Meta-talk task on Discussion Comments

We now turn to our second main research question, about whether performing the prompt construction task affects participants' subsequent commenting behavior. As mentioned earlier, most (431 of 453) participants posted a comment, and we found no significant difference in the likelihood to comment by condition, $\chi^2(7, N=459) = 8.172$, $p = 0.31$.

We had planned to look at a range of response variables, including whether participants asked genuine questions, engaged directly with other posters, provided reasoned opinions, and posted comments that were topically coherent with the existing discussion. However, fewer than 5% of the comments included questions or replies (Table 1), so we could not build meaningful statistical models predicting those variables. The response variables we were left with—reasoned opinion and

topic coherence—were not significantly predicted by task or design (Table 6). Specifically, longer comments are more likely to include a reasoned opinion and our attempts to model factors that predict topically-coherent comments were largely unsuccessful.

	Reasoned Opinion		*Topic Coherence*	
Discussion Comments	OR (SD)		OR (SD)	
(Intercept)	0.80 (0.27)		0.47 (0.29)	**
T: Improve	1.04 (0.24)		1.08 (0.24)	
D: Instructions	0.61 (0.31)		1.66 (0.31)	
D: Scaffolded	0.93 (0.30)		0.88 (0.31)	
Control Characteristics				
Acceptable: Listen	1.46 (0.27)		1.06 (0.27)	
Comment Length	2.09 (0.15)	***	1.08 (0.12)	
Goodness-of-fit: $\chi^2(5, 367)$	=41.39	***	=5.78	

Table 6. Reasoned Opinion and Topic Coherence by exposure to the Prompt task (i.e., Write, Improve), Design (i.e., No Structure, Instructions, Scaffolded) and by the prompt task Listen evaluation criterion.
p-value significance codes: 0.0001 '***' 0.001 '**' 0.01 '*' 0.05 '.'

	Write		*Improve*	
Discussion Comments	Exp. Char. (SD)		Exp. Char. (SD)	
(Intercept)	292.83 (0.11)	***	309.87 (0.10)	***
B: Active Listening	0.92 (0.14)		0.93 (0.13)	
D: No Structure	1.35 (0.14)	*	1.29 (0.13)	*
D: Instructions	1.06 (0.15)		0.91 (0.13)	
D: Scaffolded	0.87 (0.14)		1.00 (0.13)	
Control Characteristics				
PP: Neutral	1.02 (0.10)		1.01 (0.09)	
Daily Earn	1.04 (0.05)		1.06 (0.04)	
Gen: Female	1.16 (0.09)	.	1.03 (0.08)	
Gen: ND	1.43 (0.30)		1.08 (0.33)	
Goodness-of-fit: $\chi^2(4, 282)$	=12.73	**	=10.57	*

Table 7. Total expected discussion comment length ("Exp. Char.") by exposure to the Baseline (i.e., AMT Policy, Active Listening) and Meta-talk design conditions.
p-value significance codes: 0.0001 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Comment length can be taken as an indicator of effort, and longer comments, because they have more exposition, are also more likely to contain elements such as reasoned opinion. Thus, we next looked at what variables might predict comment length. To do this, we applied a negative binomial regression, which is appropriate for modeling count variables that are over-dispersed, meaning that the variance is greater than the mean. We exponentiated the coefficients from the negative binomial regression to present the incident rate ratios associated with each independent variable and expected comment length.

In the best-fitting models we could construct, the Write-No Structure and the Improve-No Structure tasks led to 1.35 and 1.29 times longer discussion comments, respectively, than the

Baseline AMT Policy task (Table 7, "D: No Structure" line). A Tukey-based pairwise comparison between No Structure and the other design conditions showed that participants were more likely to write longer comments when exposed to the Write-No Structure versus the Write-Scaffolded design (1.53 times, $p \leq 0.01$), or when exposed to the Improve-No Structure condition versus the Improve-Instructions condition (1.42 times, $p \leq 0.03$).

Our interpretation of the findings about comment length is that effort on the task may have impacted effort in the discussion forum. The more involved meta-talk conditions, with instructions to consider and multiple steps to complete, may have consumed more of participants' time and effort, reducing the effort they put into commenting later.

6 DISCUSSION AND LIMITATIONS

In the related work, we define meta-talk as "talk about the talk" that aims to revisit points of conflict in an existing discussion. We argue that newcomers to a discussion could create opportunities for meta-talk by crafting policy discussion prompts that encourage existing members to talk about their different perspectives. Further, we leveraged ideas from crowd-writing (see section 2.4) to test parts of a task designed to help newcomers to build policy discussion prompts from two existing comments or to improve a prompt drafted by another newcomer.

The experiment to test this addressed two primary research questions (see section 3.1). Our first question was *to what extent can newcomers to a discussion effectively create meta-talk-based discussion prompts?* We split this question into four parts to examine how task structure (RQ1a), design (RQ1b), and policy position (RQ1c) relate to task performance, as well as when and why people tended to perform poorly (RQ1d). Inspired by the potential for a micro-task to affect discussion norms [34], we also asked *to what extent does crafting meta-talk discussion prompts affect newcomers' subsequent contribution to the discussion?* (RQ2)

To briefly summarize our findings, (RQ1a) writing and improving policy discussion prompts is a difficult task, but (RQ1b) one that newcomers can perform reasonably well with appropriate structuring (and some caveats described below). While we found little effect of having an initial policy position (RQ1c), we did find that complementary perspectives in the comment pair improved their ability to complete the Write task and aspects of the question in the draft prompt affected performance on the Improve task. Our analysis of common errors (RQ1d) revealed that the participants, being informed and opinionated, engaged with the task material by inserting their own experience and by challenging assumptions implicit in the comment pair or the draft prompt provided for the task. However, our hope that a newcomer onboarding meta-talk task would lead to more engaged comments in the discussion (RQ2) was not supported. Here we unpack how these findings contribute to the design of meta-talk-based tasks and to future work on onboarding newcomers to discussion communities.

6.1 Provide clear instructions and scaffolding to produce acceptable prompts

We were encouraged by people's performance on the *Write* task; almost 40% of the composed prompts met the goals for a meta-talk-based prompt including hearing, comparing, and writing a genuine question to encourage other discussants to engage with each other. This is not an easy task; the criteria are rigorous, and the costs of failure might be detrimental to a group policy discussion, much as a biased [60] or untrained facilitator [16, 41] can negatively impact the group. We found that clear guidance helped many participants accomplish the task despite its difficulty; those who received support for the task in terms of instructions and scaffolding were much more likely to write acceptable prompts (50%) than those who did not (16%).

Providing clear instructions aligns with best practices drawn from related crowd-writing systems [20, 28, 34] and effective crowd task design more broadly [30, 46]. However, designing good

instructions is not easy, and in our case it took several iterations before arriving at the designs presented here. Looking back, the designs share many characteristics with rubrics and peer feedback systems designed for MOOCs [35]: the instructions align closely with the sub-criteria; they give definitions, considerations, and examples to illustrate each of the criteria; and the interface provides people with a self-assessment tool toward the criteria.

More generally, our results emphasize the magnitude of the need for effective instructions, as participants were three times better at crafting acceptable prompts given appropriate guidance. Taking the time to iteratively pilot task instructions is important, in terms of both quality of outputs and worker fairness, by reducing time-wasting and pay-reducing rejections induced by ambiguous instructions [26, 45].

6.2 Selecting appropriate content in comparison and synthesis tasks

Our results also demonstrate that for complex synthesis and iteration tasks, designers should pay careful attention to the balance of perspectives exhibited by the content presented. For instance, in the *Write* condition, participants found it easier to effectively compare two comments whose positions were farther apart than two comments that were roughly in agreement. In this experiment, we hand-chose each pair of comments to support experimental control, but in a real system, topic modeling and sentiment analysis could be used to select comment sets that are more likely to have topical agreement, but differing opinions. Social network analysis techniques might also be used to mitigate the concern raised in the background section around newcomers inadvertently encouraging interaction between people with a negative relationship [4, 48].

Several crowd-writing systems also include a process for selecting sets of comments that exhibit different properties. For example, *Knowledge Accelerator* [20] implements a process for filtering a list of information down to a unique set with a comparison task where workers "are asked to sample random items from the data in order to create a set of non-matching items [...]". The task continues until "a worker's familiarity with the distribution gives them a sense that [the items in the list] represent substantively different topics" [20, pg. 2264]. We imagine this pattern, called "Open-ended Set Sampling," could be adapted to filter a thread of comments for exemplars of each policy perspective to identify the points of conflict. For other discussion goals, such as constructing meta-talk about opportunities for consensus or summarizing the arguments toward particular policy proposals, the open-ended set sampling could reduce a list of comments to those that exhibit a similar position (e.g., support for partial payment), but with different reasoning.

6.3 Consider evaluating prompts as a way to facilitate discussion

Our finding that many prompts declined in quality in the *Improve* task raises questions about deciding both what prompts are worth iterating on and when to stop iterating. Many crowd-writing systems set a fixed number of iterations [2, 20, 39] or apply an agreement-based scheme [11, 15, 64] as the stopping criteria for an iterative process.

In the case of "acceptable" policy discussion prompts, deciding when to stop is harder even though the high level desirable properties of listening, comparing, and questioning are well-defined. As we saw, one of the prompts in the Improve condition had a question that, although considered acceptable in our initial coding, in retrospect felt more rhetorical and implied a more concrete policy solution than the others. Other work has shown that people opposed to a point of view often engage with it indirectly [44, 53], and we saw that participants exposed to that prompt tended to ask rhetorical questions that challenged its assumptions. Possible ways to reduce that tendency would be to ask workers to provide their reasoning to justify a proposed improvement [15] or to introduce an active listening step to facilitate writer-and-improver dialogue [34], which may have an added benefit of introducing newcomers to each other [33].

We did not evaluate the effectiveness of the policy discussion prompts in fomenting actual meta-talk. To do so would involve organizing a group to consider a prompt and then studying their discussion [3]. However, simply inserting a prompt into an ongoing discussion might not trigger the reflection or innovative thinking intended by newcomer-crafted meta-talk [9, 37, 62]. Instead, we might draw inspiration from the ways that crowd-writing systems manage progress toward a complex objective. For example, the *Mechanical Novel* implements a process to crowd-write a fictitious story, yet allowing the plot to deviate by periodically *reflecting* on current progress and then *revising* the intermediary goals. This process of "[...] looping between reflecting on progress to identify a goal and revising based on that goal, allows [a group of workers] to converse with their work and evaluate options by trying them out" [29, pg. 235]. In a similar way, the process of evaluating several newcomer prompts could be formalized as a periodic group activity for existing members, which would provide them with a way to reflect on and revisit older topics by considering prompts generated from new perspectives.

6.4 Limitations and Open Questions

The most salient limitation of this study is that we traded off the ecological validity of real discussions for experimental control. This leads to questions about how to apply and extend these findings when generalizing from Turkers to discussion newcomers, from this task design to actual forum designs, and from this one-time HIT to ongoing policy discussions.

We believe that choosing the AMT participation agreement as the policy context mitigates many of the concerns about how Turkers might be different from newcomers to policy discussions more generally. For instance, it is possible that having some background knowledge of the topic would improve people's ability to construct effective prompts, and newcomers to a discussion will often have a pre-existing interest in the topic. Choosing Turkers and the AMT participation agreement is representative of this situation—so much so that the question of whether Turkers could construct prompts for an arbitrary discussion topic as part of a generic crowd-powered workflow is still open. We think it also reduces concerns that Turkers' paid status affects their motivation toward the task relative to volunteer newcomers, since Turkers do have a stake in the participation agreement. Still, pay matters, and whether newcomers to a policy discussion would be willing to spend the time required to do the onboarding tasks is also an open question, although cases like Reflect [34] suggest the answer is sometimes yes.

Another limitation is that the experimental design did not integrate the onboarding task and the discussion forum, which were posed as different stages/sub-tasks in the larger context of a HIT and had different visual designs and no direct incorporation of materials from one task to the other. We suspect this reduced possible carryover of socialization effects from doing the onboarding task into the discussion. More generally, in this experiment we did not consider how to integrate the prompts later into the discussion or how participants would interact with the prompts because our focus was on the support of prompt creation and newcomers' initial forum behavior. The thoughts around prompt evaluation above are a start, but the question of how to use newcomers' inputs in a facilitator's, moderator's, or group's larger policy discussion process is still unresolved.

Finally, our choice of a controlled, simulated, one-shot discussion leaves open questions about more natural and continued engagement. Because much participation in online deliberations comes from one-time contributors [47], we think many of the findings are likely to hold up for many participants in real discussions. Still, the canned and one-shot nature of both the underlying comments that seeded the forum and participants' own experience could have masked effects of the onboarding task on newcomer participation that would show up only in more interactive situations. (In fact, a few participants noted to us that the discussion lacked interaction between commenters).

The general future work implication is obvious—integrate into full workflows and test in the field. However, we believe the results show promise around newcomers to a discussion being able to craft prompts that according to theories of deliberation should be useful to the group.

7 CONCLUSION

In online policy discussions, people often talk past rather than with each other [13]. In this paper, we consider the role that newcomers might play in interrupting this pattern by introducing meta-talk prompts to an ongoing discussion. Meta-talk is referred to as "talk about the talk" that seeks to address points of conflict already in a discussion [62], which we relate to the type of policy discussion prompts that social studies teachers prepare for their students [55]. Drawing on the crowd-writing systems literature, we develop and evaluate key parts of a task designed to support newcomers in constructing policy discussion prompts. Our findings suggest implications for the design of crowd-writing systems, demonstrating the benefits of clear instructions and calling attention to the role that opinion and bias might play in performing judgment tasks. For online policy discussion, the findings offer considerations about the design of mechanisms that introduce newcomers to an ongoing policy discussion and existing members to new perspectives.

ACKNOWLEDGMENTS

We thank all the Turker participants, whose thoughtful effort made this research possible. Early in the research, Poppy McLeod played an instrumental role in helping to develop the experimental design and recommending procedures for the study. During the revision cycle, we worked with Molly Feldman to more critically consider how the research findings relate to challenges in crowdsourcing and human computation. The analysis was supported by a dedicated team of research assistants: Claire Han, Anita Sharma, Simran Shinh and Varun Talwar. The research was also supported by the National Science Foundation (HCC 1314778); it was conducted while Dan Cosley was serving at the NSF and does not necessarily reflect the views of the NSF.

APPENDIX: A METHODOLOGICAL NOTE ABOUT PLATFORM POLICY SHIFTS

The AMT participation agreement changed during the course of the study, which in principle could have affected the results. We conducted the experiment during three periods, recruiting for the *Write* and *Baseline* conditions from October 3–8, 2017 and the *Improve* conditions between November 5–6 and 14–15, 2017. Amazon updated the Mechanical Turk participation agreement on October 17, 2017.[1] Most relevant to this study, the updates add a formal requirement for Requesters and Turkers to be professional and courteous toward each other and a way for both Turkers and Requesters to report alleged violations of the participation agreement. These and other shifts in the policy suggest that Amazon will no longer take a "hands-off" approach to managing Mechanical Turk. For the study, it means that participants exposed to the *Improve* task may have had a different understanding or experience of a HIT rejection than those in the *Write* task, although in practice we suspect that like many terms of service agreements, most participants paid little attention to it.

REFERENCES

[1] Blake E. Ashforth, David M. Sluss, and Alan M. Saks. 2007. Socialization tactics, proactive behavior, and newcomer learning: Integrating socialization models. *Journal of Vocational Behavior* 70, 3 (2007), 447–462. https://doi.org/10.1016/j.jvb.2007.02.001

[1] We also noticed that the link to the AMT Participation Agreement has digitally moved to a new website—switching from https://www.mturk.com/worker/participation-agreement to the new link https://www.mturk.com/mturk/conditionsofuse. For CSCW researchers studying online platform policies, such incidental changes are important to note as they can make it more challenging to use research tools like the WayBackMachine to study platform policy shifts over time.

[2] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2015. Soylent: a word processor with a crowd inside. *Commun. ACM* 58, 8 (2015), 85–94.

[3] Laura W. Black, Stephanie Burkhalter, John Gastil, and Jennifer Stromer-Galley. 2010. Methods for analyzing and measuring group deliberation. *Sourcebook of political communication research: Methods, measures, and analytical techniques* (2010), 323–345.

[4] Laura W. Black, Howard T. Welser, Dan Cosley, and Jocelyn M. DeGroot. 2011. Self-Governance Through Group Discussion in Wikipedia. *Small Group Research* 42, 5 (10 2011), 595–634. https://doi.org/10.1177/1046496411406137

[5] Moira Burke, Robert E. Kraut, and Elisabeth Joyce. 2010. Membership Claims and Requests: Conversation-Level Newcomer Socialization Strategies in Online Groups. *Small Group Research* 41, 1 (2 2010), 4–40. https://doi.org/10.1177/1046496409351936

[6] Moira Burke, Cameron Marlow, and Thomas Lento. 2009. Feed me: motivating newcomer contribution in social network sites. *Proceedings of the SIGCHI conference on* (2009). http://dl.acm.org/citation.cfm?id=1518847

[7] Stephanie Burkhalter, John Gastil, and Todd Kelshaw. 2002. A conceptual definition and theoretical model of public deliberation in small face-to-face groups. *Communication Theory* 12, 4 (2002), 398–422.

[8] Boreum Choi, Kira Alexander, Robert E. Kraut, and John M. Levine. 2010. Socialization tactics in wikipedia and their effects. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10*. ACM Press, New York, New York, USA, 107. https://doi.org/10.1145/1718918.1718940

[9] Hoon-Seok Choi and Leigh Thompson. 2005. Old wine in a new bottle: Impact of membership change on group creativity. *Organizational Behavior and human decision* 98, 2 (2005), 121–132. https://doi.org/10.1016/j.obhdp.2005.06.003

[10] Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication* 64, 4 (8 2014), 658–679. https://doi.org/10.1111/jcom.12104

[11] Peng Dai and Daniel Sabey Weld. 2010. Decision-Theoretic Control of Crowd-Sourced Workflows. *Twenty-Fourth AAAI Conference on Artificial Intelligence* (7 2010). https://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/viewPaper/1873

[12] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*. 307–318.

[13] Richard Davis. 1999. *The web of politics: The Internet's impact on the American political system*. Oxford University Press.

[14] Steven P. Dow, An 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.

[15] Ryan Drapeau, Lydia B. Chilton, Jonathan Bragg, and Daniel S. Weld. 2016. MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy. *Human Computation and Crowdsourcing* (2016).

[16] Dmitry Epstein and Gilly Leshed. 2016. The Magic Sauce: Practices of Facilitation in Online Policy Deliberation. *Journal of Public Deliberation* 12, 1 (2016).

[17] Casey Fiesler, Shannon Morrison, R. Benjamin Shapiro, and Amy S. Bruckman. 2017. Growing Their Own. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. ACM Press, New York, New York, USA, 1375–1386. https://doi.org/10.1145/2998181.2998210

[18] Ali Gürkan, Luca Iandoli, Mark Klein, and Giuseppe Zollo. 2010. Mediating debate through on-line large-scale argumentation: Evidence from the field. *Information Sciences* 180, 19 (2010), 3686–3702. https://doi.org/10.1016/j.ins.2010.06.011

[19] Carole Hahn. 1998. *Becoming political: Comparative perspectives on citizenship education*. Suny Press.

[20] Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. 2016. The Knowledge Accelerator: Big Picture Thinking in Small Pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2258–2270.

[21] Aaron Halfaker, Oliver Keyes, and Dario Taraborelli. 2013. Making peripheral participation legitimate. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*. ACM Press, New York, New York, USA, 849. https://doi.org/10.1145/2441776.2441872

[22] Daniel Halpern and Jennifer Gibbs. 2013. Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior* 29, 3 (2013), 1159–1168. https://doi.org/10.1016/j.chb.2012.10.008

[23] Diana E. Hess. 2009. *Controversy in the classroom: The democratic power of discussion*. Routledge.

[24] Diana E. Hess and Paula McAvoy. 2014. *The political classroom: Evidence and ethics in democratic education*. Routledge.

[25] Gary Hsieh, Youyang Hou, Ian Chen, and Khai N. Truong. 2013. Welcome!: Social and psychological predictors of volunteer socializers in online communities. In *Computer Supported Cooperative Work*. 827–838. http://dl.acm.org/citation.cfm?id=2441870

[26] Lilly C. Irani and M. Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 611–620. http://dl.acm.org/citation.cfm?id=2470742

[27] Christopher F. Karpowitz and Chad Raphael. 2014. *Deliberation, democracy, and civic forums: Improving equality and publicity*. Cambridge University Press.
[28] Joy Kim and Andres Monroy-Hernandez. 2016. Storia: Summarizing social media content based on narrative theory using crowdsourcing. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1018–1027.
[29] Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S. Bernstein. 2016. Mechanical Novel: Crowdsourcing Complex Work through Reflection and Revision. In *Computer Supported Cooperative Work and Social Computing*. ACM, 233–245. https://doi.org/10.1145/2998181.2998196
[30] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1301–1318.
[31] 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.
[32] Mark Klein. 2011. How to harvest collective wisdom on complex problems: An introduction to the mit deliberatorium. *Center for Collective Intelligence working paper* (2011).
[33] Robert E. Kraut, Moira Burke, John Riedl, and Paul Resnick. 2010. Dealing with Newcomers. *Evidence-based Social Design: Mining the Social Sciences to Build Online Communities* 1, 1 (2010), 42.
[34] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew Ko. 2012. Is this what you meant?: promoting listening on the web with reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1559–1568.
[35] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. 2013. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction* 20, 6 (12 2013), 1–31. https://doi.org/10.1145/2505057
[36] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31, 2 (4 2014), 317–326. https://doi.org/10.1016/j.giq.2013.11.005
[37] John M. Levine, Hoon-Seok Choi, and Richard L. Moreland. 2003. Newcomer innovation in work teams. *Group creativity: Innovation* (2003).
[38] John M. Levine and Richard L. Moreland. 1994. Group Socialization: Theory and Research. *European Review of Social Psychology* 5, 1 (1 1994), 305–336. https://doi.org/10.1080/14792779543000093
[39] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. 2010. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 57–66.
[40] Rousiley C. M. Maia and Thaiane A. S. Rezende. 2016. Respect and Disrespect in Deliberation Across the Networked Media Environment: Examining Multiple Paths of Political Talk. *Journal of Computer-Mediated Communication* 21, 2 (3 2016), 121–139. https://doi.org/10.1111/jcc4.12155
[41] Jane Mansbridge, Janette Hartz-Karp, Matthew Amengual, and John Gastil. 2006. Norms of deliberation: An inductive study. (2006).
[42] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 224–235.
[43] Paula McAvoy and Diana Hess. 2013. Classroom Deliberation in an Era of Political Polarization. *Curriculum Inquiry* 43, 1 (1 2013), 14–47. https://doi.org/10.1111/curi.12000
[44] Brian McInnis, Dan Cosley, Eric Baumer, and Gilly Leshed. 2018. Effects of Comment Curation and Opposition on Coherence in Online Policy Discussion. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*. ACM, Sanibel Island, Florida, 347–358.
[45] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around Rejection, Mistrust, Risk, and Workers' Experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2271–2282.
[46] Brian McInnis and Gilly Leshed. 2016. Running user studies with crowd workers. *Interactions* XXIII, 5 (2016), 50. http://dl.acm.org/citation.cfm?id=2968077
[47] Brian McInnis, Elizabeth Murnane, Dmitry Epstein, Dan Cosley, and Gilly Leshed. 2016. One and Done: Factors affecting one-time contributors to ad-hoc online communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 609–623.
[48] Rony Medaglia and Yang Yang. 2017. Online public deliberation in China: evolution of interaction patterns and network homophily in the Tianya discussion forum. *Information, Communication & Society* 20, 5 (5 2017), 733–753. https://doi.org/10.1080/1369118X.2016.1203974
[49] Hamideh Molaei. 2014. The prospect of civility in Indonesians' online polarized political discussions. *Asian Journal of Communication* 24, 5 (9 2014), 490–504. https://doi.org/10.1080/01292986.2014.917116

[50] Alfred Moore. 2012. Following from the front: Theorizing deliberative facilitation. *Critical Policy Studies* 6, 2 (7 2012), 146–162. https://doi.org/10.1080/19460171.2012.689735
[51] Richard L Moreland and John M Levine. 1982. Socialization in small groups: Temporal changes in individual-group relations. *Advances in experimental social psychology* 15 (1982), 137–192.
[52] Richard L Moreland and John M Levine. 1988. Group dynamics over time: Development and socialization in small groups. (1988).
[53] Sean A. Munson and Paul Resnick. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1457–1466.
[54] Tom Murray, Lynn Stephens, Beverly Park Woolf, Leah Wing, Xiaoxi Xu, and Natasha Shrikant. 2013. Supporting Social Deliberative Skills Online: The Effects of Reflective Scaffolding Tools. Springer, Berlin, Heidelberg, 313–322. https://doi.org/10.1007/978-3-642-39371-6{_}36
[55] Walter C. Parker. 2006. Public Discourses in Schools: Purposes, Problems, Possibilities. *Educational Researcher* 35, 8 (11 2006), 11–18. https://doi.org/10.3102/0013189X035008011
[56] Jennifer Preece and Ben Shneiderman. 2009. The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer* (2009).
[57] Cynthia R. Farina, Dmitry Epstein, Josiah B. Heidt, and Mary J. Newhart. 2013. RegulationRoom: Getting more, better civic participation in complex government policymaking. *Transforming Government: People, Process and Policy* 7, 4 (2013), 501–516.
[58] Ian Rowe. 2015. Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms. *Journal of Broadcasting & Electronic Media* 59, 4 (10 2015), 539–555. https://doi.org/10.1080/08838151.2015.1093482
[59] Deborah Schiffrin. 1980. Meta-Talk: Organizational and Evaluative Brackets in Discourse. *Sociological Inquiry* 50, 3-4 (7 1980), 199–236. https://doi.org/10.1111/j.1475-682X.1980.tb00021.x
[60] Paolo Spada and James Raymond Vreeland. 2013. Who Moderates the Moderators? The Effect of Non-neutral Moderators in Deliberative Decision Making. *Journal of Public Deliberation* 9, 2 (2013). http://www.publicdeliberation.net/jpdhttp://www.publicdeliberation.net/jpd/vol9/iss2/art3
[61] Kim Strandberg and Janne Berg. 2013. Online Newspapers' Readers' Comments - Democratic Conversation Platforms or Virtual Soapboxes? *Comunicação e Sociedade* 23, 1 (2013), 132. https://doi.org/10.17231/comsoc.23(2013).1618
[62] Jennifer Stromer-Galley. 2007. Measuring deliberation's content: A coding scheme. *Journal of public deliberation* 3, 1 (2007).
[63] Jennifer Stromer-Galley and Anna M. Martinson. 2009. Coherence in political computer-mediated communication: analyzing topic relevance and drift in chat. *Discourse & Communication* 3, 2 (5 2009), 195–216. https://doi.org/10.1177/1750481309102452
[64] Vasilis Verroios and Michael S. Bernstein. 2014. Context trees: Crowdsourcing global understanding from local views. In *Second AAAI Conference on Human Computation and Crowdsourcing*.
[65] Etienne Wenger and Jean Lave. 2002. Legitimate peripheral participation in communities of practice. *Supporting lifelong learning* (2002).
[66] Anita W. Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science (New York, N.Y.)* 330, 6004 (10 2010), 686–8. https://doi.org/10.1126/science.1193147
[67] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2082–2096.

Received April 2018; revised July 2018; accepted September 2018