Improving Worker Engagement Through Conversational Microtask Crowdsourcing

Qiu, Sihang; Gadiraju, U.K.; Bozzon, Alessandro

DOI
10.1145/3313831.3376403

Publication date
2020

Document Version
Accepted author manuscript

Published in

Citation (APA)

Important note
To cite this publication, please use the final published version (if applicable). Please check the document version above.
Improving Worker Engagement Through Conversational Microtask Crowdsourcing

Sihang Qiu  
Delft University of Technology  
Delft, The Netherlands  
s.qiu-1@tudelft.nl

Ujwal Gadiraju  
L3S Research Center,  
Leibniz Universität Hannover  
Hannover, Germany  
gadiraju@L3S.de

Alessandro Bozzon  
Delft University of Technology  
Delft, The Netherlands  
a.bozzon@tudelft.nl

ABSTRACT
The rise in popularity of conversational agents has enabled humans to interact with machines more naturally. Recent work has shown that crowd workers in microtask marketplaces can complete a variety of human intelligence tasks (HITs) using conversational interfaces with similar output quality compared to the traditional Web interfaces. In this paper, we investigate the effectiveness of using conversational interfaces to improve worker engagement in microtask crowdsourcing. We designed a text-based conversational agent that assists workers in task execution, and tested the performance of workers when interacting with agents having different conversational styles. We conducted a rigorous experimental study on Amazon Mechanical Turk with 800 unique workers, to explore whether the output quality, worker engagement and the perceived cognitive load of workers can be affected by the conversational agent and its conversational styles. Our results show that conversational interfaces can be effective in engaging workers, and a suitable conversational style has potential to improve worker engagement. Our findings have important implications on workflows and task design with regard to better engaging workers in microtask crowdsourcing marketplaces.

Author Keywords
Microtask crowdsourcing; conversational interface; conversational style; user engagement; cognitive task load.

CCS Concepts
•Information systems → Chat; Crowdsourcing; •Human-centered computing → Empirical studies in HCI;

INTRODUCTION
There has been a gradual rise in the use of conversational interfaces aiming to provide seamless means of interaction with virtual assistants, chatbots, or messaging services. There is a growing familiarity of people with conversational interfaces owing to the widespread proliferation of mobile devices and messaging services such as WhatsApp, Telegram, and Messenger. Today, over half the population on our planet has access to the Internet with ever-lowering barriers of accessibility. This has led to flourishing paid crowdsourcing marketplaces like Amazon’s Mechanical Turk (AMT) or Figure-Eight (F8), where people around the world can participate in online work with an aim to earn their primary livelihood, or as a secondary source of income.

Recent work by Mavridis et al. [29] has explored the suitability of conversational interfaces for microtask crowdsourcing by juxtaposing them with standard Web interfaces in a variety of popularly crowdsourced tasks. The authors found that conversational interfaces were positively received by crowd workers, who indicated an overall satisfaction and an intention for future use of similar interfaces. The tasks executed using the conversational interfaces took similar execution times as those using the standard Web interfaces, and yielded comparable output quality. Although these findings suggest the use of conversational interfaces as a viable alternative to the existing standard, little is known about the impact of conversational microtasking on the engagement of workers. Previous works have studied the nature of tasks that are popularly crowdsourced on AMT, showing that tasks are often deployed in large batches consisting of similar HITs (human intelligence tasks) [1, 8]. Long and monotonous batches of HITs pose challenges with regards to engaging workers, potentially leading to sloppy work due to boredom and fatigue [6]. There is a lack of understanding of whether conversational microtasking would either alleviate or amplify the concerns surrounding worker engagement. In this work, we aim to fill this knowledge gap.

We conducted a study on AMT, involving 800 unique workers across 16 different experimental conditions to address the following research questions.

#RQ1: To what extent can conversational agents improve the worker engagement in microtask crowdsourcing?

#RQ2: How do conversational agents with different conversational styles affect the performance of workers and their cognitive load while completing tasks?

We deployed batches of different types of HITs; information finding, sentiment analysis, CAPTCHA recognition, and image classification tasks on the traditional web interface and...
three conversational interfaces having different conversational styles (4 task types × 4 interface variants).

We first investigated the effect of conversational interfaces with different conversational styles on quality related outcomes in comparison to the traditional web interfaces. We addressed **RQ1** by using two measures of worker engagement; (i) worker retention in the batches of tasks, and (ii) self-reported scores on the short-form user engagement scale [32, 44]. We addressed **RQ2** by considering different conversational styles within conversational interfaces that workers interact with, and by using the NASA-TLX instrument to measure cognitive load after workers complete the tasks they wish to. Our results show that conversational interfaces have positive effects on worker engagement, as well as the perceived cognitive load in comparison to traditional web interfaces. We found that a suitable conversational style has the potential to engage workers further (in specific task types), although our results were inconclusive in this regard. Our work takes crucial strides towards furthering the understanding of conversational interfaces for microtasking, revealing insights into the role of conversational styles across a variety of tasks.

**RELATED WORK**

**Conversational Agents**

Conversational interfaces have been argued to have advantages over traditional graphical user interfaces due to having a more human-like interaction [30]. Owing to this, conversational interfaces are on the rise in various domains of our everyday life and show great potential to expand [43]. Recent work in the HCI community has investigated the experiences of people using conversational agents, understanding user needs and user satisfaction [4, 5, 27]. Other works have studied the scope of using conversational agents in specific domains. Vandenberghe introduced the concept of bot personas, which act as off-the-shelf users to allow design teams to interact with rich user data throughout the design process [40]. Others have studied the use of conversational agents in the domains of complex search [2, 22, 41] or food tracking [14]. These works have shown that conversational agents can improve user experiences and have highlighted the need to further investigate the use of conversational agents in different scenarios. In contrast to existing works, we explore the use of conversational agents in improving worker engagement in microtask crowdsourcing.

**Crowdsourced Conversational Interfaces**

Prior research has combined crowdsourcing and the conversational agent for training the dialogue manager or natural language processing component [24]. Lasecki et al. designed and developed Chorus, a conversational assistant able to assist users with general knowledge tasks [26]. Conversations with Chorus are powered by workers who propose responses in the background, encouraged by a game-theoretic incentive scheme. Workers can see the working memory (chat history) and vote on candidate responses on a web-based worker interface. Based on Chorus, an improved conversational assistant named Evorus was proposed. It can reduce the effort of workers by partially automating the voting process [19]. The same authors also developed a crowdsourced system called Guardian, which enables both expert and non-expert workers to collaboratively translate Web APIs into a dialogue system format [20].

A conversational agent called Curious Cat was proposed to combine the crowdsourcing approach from a different perspective [3]. While most crowdsourced conversational agents provide information to users according to their requests, the Curious Cat was designed as a knowledge acquisition tool, which actively asked data from users. In our study, we propose a conversational agent that serves in a conversational interface for workers, to perform different types of popular crowdsourcing tasks.

**Conversational Styles**

Previous works have already shown that conversational styles have played an important role in human lexical communication. Lakoff suggested that conversational styles could be classified into four categories from the least relationship between participants to the most relationship between participants: 1) Clarity, an ideal mode of discourse; 2) Distance, a style that does not impose others; 3) Deference, a style giving options; and 4) Camaraderie, direct expression of desires [25]. Based on Lakoff’s system, Tannen performed a systematic analysis on conversational style using the conversations recorded from a Thanksgiving dinner [36, 37]. Tannen concluded several important features, and accordingly distinguished conversational style from Involvement (overlapping with Lakoff’s camaraderie strategy) to Considerateness (overlapping with Lakoff’s distance strategy).

Researchers have also attempted to apply theories pertaining to conversational styles in the field of human computer interaction. [34] studied the preferred conversational style for a conversational agent. Their results suggested that users preferred the agent whose style matched their own. A similar conclusion was drawn from an analytical study of information seeking conversation conducted by [38] using the MISC dataset [39]. [21] compared survey response data quality acquired from the web platform and chatbot. Particularly, they performed the experiment using formal and casual styles. The chatbot using “casual” conversational style in their study tried to establish relationship with users, where we can find linguistic features from both Tannen’s “High-Involvement” and “High-Considerateness”. The chatbot using “formal” style is akin to the “clarity style” (showing no involvement and the least relationship with the user) as summarized by Lakoff [25].

Our work uses features and linguistic devices from Tannen’s “High-Involvement” and “High-Considerateness” styles to design the conversation for microtask crowdsourcing, and we conduct experiments to see the effects of using different conversational styles.

**Worker Engagement**

Crowdsourcing microtasks can often be monotonous and repetitive in nature. Previous works have attempted to tackle the issues of boredom and fatigue manifesting in crowdsourcing marketplaces as a result of long batches of similar tasks that workers often encounter. A variety of methods to retain and engage workers have been proposed. [33] suggested introducing
is presented in sequence. Each new question contains a brief transition sentence (e.g. "Good! The next one."), the question number (helping workers find and edit previous questions), and the content itself (which can contain any HTML-based task type). Furthermore, the conversational interface supports two modes of input from workers; in the form of free text and multiple choices. When the expected input form of the answer is free text (e.g. in character recognition or audio transcription tasks), the worker must type the answer in the text area of the conversational interface. When the task includes multiple-choice answers, the worker can either type the answer as free text (exactly the same value as one of the options), or simply click the corresponding UI button.

4) After the worker has answered 10 questions, the conversational agent gives a break to relieve workers from the monotony of the batch of tasks. During the break, the conversational agent may send a “meme” or a joke for amusement, and then remind workers that they can stop answering and submit answers whenever they want.

5) When a worker decides to stop task execution, or when no more pending questions are available, the conversational agent sends a list of answers provided by the worker, for review. The worker is then allowed to review one or more previous answers and make any preferred edits.

6) The conversational agent then uploads the worker’s final answers to the server. Once it confirms the answers have been successfully uploaded, a Task Token is given to the worker.

7) By pasting the Task Token on AMT, the worker can claim the corresponding monetary compensation, proportional to the number of answered questions.

Figure 1. The workflow of conversational microtask crowdsourcing.

---

1 https://qiushihang.github.io/csbot
Table 1. Design criteria for conversation styles of the agent.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>High-Involvement</th>
<th>High-Considerateness</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1. Rate of speech</td>
<td>fast</td>
<td>slow</td>
</tr>
<tr>
<td>C2. Turn taking</td>
<td>fast</td>
<td>slow</td>
</tr>
<tr>
<td>C3. Introduction of topics</td>
<td>without hesitation</td>
<td>without hesitation</td>
</tr>
<tr>
<td>C4. Use of syntax</td>
<td>simple</td>
<td>complex</td>
</tr>
<tr>
<td>C5. Directness of content</td>
<td>direct</td>
<td>indirect</td>
</tr>
<tr>
<td>C6. Utterance of questions</td>
<td>frequent</td>
<td>rare</td>
</tr>
</tbody>
</table>

Conversational Styles: Involvement or Considerateness

Tannen’s analysis of conversational style [36] is based on an audio-taped conversation at a Thanksgiving dinner that took place in Berkeley, California, on November 23, 1978. Tannen found that, among the 6 participants present, 3 of them were New Yorkers and shared a conversational style. Tannen named the style of New Yorkers “High-Involvement”, which can be characterized as follows: “When in doubt, talk. Ask questions. Talk fast, loud, soon. Overlap. Show enthusiasm. Prefer personal topics, and so on.” The conversational style of non-New Yorkers was called “High-Considerateness”, and can be characterized as follows: “Allow longer pauses. Hesitate. Don’t impose one’s topics, ideas, personal information. Use moderate paralinguistic effects, and so on”. We selected Tannen’s classification of style to define conversational styles of agents, since recent work has shown its suitability in understanding styles in human-human conversations, and also in human-agent conversations [34]. Moreover, Tannen’s classification has served as the basis for aligning the style of an end-to-end voice-based agent with that of an interlocutor [17].

Tannen identified four main features of the conversational style, namely topic, pacing, narrative strategies, and expressive paralinguistics [37]. Based on these features and some linguistic devices used in the conversation of the Thanksgiving dinner, we created the following criteria to design conversation consistent with the High-Involvement and High-Considerateness styles for the conversational agent, as shown in Table 1. The criteria can be organised into two categories:

1) Pacing (C1, C2): Since the conversational agent communicates with the worker by typing text instead of via voice utterances, we use typing speed and the pause before sending a bubble (message) to simulate the rate of speech and the pause before turn taking. The High-Involvement style has a faster rate of speech and turn taking. Hence, we set a 1 ms delay per character to simulate typing speed (C1), and 100 ms pause before sending a bubble for simulating turn taking (C2). As for the High-Considerateness style, which corresponds to a slower pace, we set a 2 ms delay per character and a 200 ms pause before animating the bubble.

2) Content (C3, C4, C5, C6): The conversational agent corresponding to the High-Involvement style introduces a new topic to the worker (for instance, telling workers how to answer questions, how to edit answers, and how to submit answers) without hesitation (C3). On the contrary, we use some words or paralanguage such as “Well...” and “Hmm...” to simulate the hesitation of the High-Considerateness conversational agent (C3). Furthermore, the conversational agent of High-Involvement style uses less syntax (C4) and chats directly (C5), while the agent of High-Considerateness style uses relatively complex syntax (C4) and tends to express ideas/topics in an indirect or polite way (C5). Tannen also emphasized the importance of asking questions for the High-Involvement style [37]. Therefore, we use the frequency of questioning as one of the criteria (C6) for conversation design.

Based on the content criteria described above, we created templates of conversation for microtask crowdsourcing, as shown in Table 2.

Aligning Conversational Styles

Previous studies suggest that there is no such thing as the best conversational style, since a style needs to be adapted to the interlocutor [34, 38]. We therefore estimate the conversational style of the worker, and investigate whether aligning the style of the conversational agent with the conversational style of the worker can positively effect quality related outcomes in the tasks being completed.

To estimate the conversational style of the worker, a basic strategy could be to analyze features of the worker’s replies and classify the replies using these features. Note that the conversational style of a worker must be estimated and aligned before the worker starts answering questions, since replies given during the actual task execution are in essence answers to the crowdsourcing tasks, rather than natural conversation. Therefore, the conversational style of the agent should be aligned right after the “opening greetings”, “time requirement” and “task instruction” interactions (in Table 2). However, such conversational elements are typically not rich enough to enable feature extraction and style classification. In this study, we therefore give workers dual options of conversational styles to select from (Figure 2), and then adapt the style of the conversational agent according to the worker selection.

We estimate the conversational style of workers as follows: 1) For each interaction, we provide one or two options that lead the worker to the next interaction (we call these actual options). These options serve the purpose of ensuring progressivity in the interaction [11, 35]. Note that actual options are invisible.
to workers. The only actual option corresponding to “opening greetings” is go ahead, while the only actual option of “time requirement” is understand the time requirement. For the “task instructions” interaction, there are two actual options: show instructions and skip instructions, where the former elucidates how to answer the crowdsourcing question and the latter directly leads the worker through to the task execution stage. 2) As actual options are invisible to workers, we create two visible options (referring to High-Involvement and High-Considerateness respectively) for each actual option. To proceed, workers select a single response from the provided visible options. 3) As a result of these three interactions, we obtain three specifically selected responses from each worker. If two or more replies refer to a High-Involvement style, we consider the conversational style of the worker to be that of High-Involvement, and vice versa.

On determining the conversational style of the worker, the style of the conversational agent is spontaneously aligned with that of the worker.

**EXPERIMENTAL DESIGN**

The main goal of our study is to investigate the impact of the conversational interface on the output quality, worker engagement, and cognitive task load, we therefore consider the traditional web interface (Web) for comparison, wherein the input elements are default HTML-based question wid-
gets provided by AMT. This will allow us to analyse our results in the light of recent findings from work by Mavridis et al. [29]. Another important objective is to study the effect that different conversational styles have on the performance of workers, completing microtasks through conversational interfaces. We thereby set up three different conversational interfaces; one with a High-Involvement style (Con+I), a High-Considerateness style (Con+C), and an aligned style (aligning the style of the agent with the estimated style of the worker, Con+A). The conversational interface with High Involvement or High Considerateness (namely, Con+I or Con+C) initiates with its corresponding conversational style and maintains it through all interactions, while the conversation interface with style alignment (Con+A) initiates with either High Involvement or High Considerateness randomly, and adjust its conversational style after conversational style estimation.

In terms of the task types, we consider two input types (free text and multiple choices) and two data types (text and images), resulting in a cross-section of 4 different types of tasks (as shown in Table 3): Information Finding, Sentiment Analysis, CAPTCHA Recognition, and Image Classification [12].

<table>
<thead>
<tr>
<th>Input type</th>
<th>Text</th>
<th>Imagery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free text</td>
<td>Information Finding</td>
<td>CAPTCHA Recognition</td>
</tr>
<tr>
<td>Multiple choices</td>
<td>Sentiment Analysis</td>
<td>Image Classification</td>
</tr>
</tbody>
</table>

**Information Finding (IF).** Workers are asked to find a given store on Google Maps and report its rating (i.e., the number of stars). The information corresponding to stores is obtained from a publicly available Yelp dataset².

**Sentiment Analysis (SA).** Workers are asked to read given reviews of restaurants from the Yelp dataset, and judge the overall sentiment of the review.

**CAPTCHA Recognition (CR).** Workers are asked to report the alphanumeric string contained in a CAPTCHA generated by Claptcha³, in the same order as they appear in the image.

**Image Classification (IC).** Workers are asked to analyse images pertaining to 6 animal species (butterfly, crocodile, dolphin, ²Yelp Open Dataset. [https://www.yelp.com/dataset](https://www.yelp.com/dataset)
³ [https://github.com/kuszaj/claptcha](https://github.com/kuszaj/claptcha)
panda, pigeon, and rooster) selected from Caltech101 Dataset [9]. They are tasked with determining which animal a given image contains, and selecting the corresponding option.

Our experimental study is therefore composed of 16 experimental conditions (4 task types × 4 interfaces).

**Task Design**

The task is organised in four steps: a demographic survey, the microtask, the User Engagement Scale Short Form (UES-SF), and the NASA Task Load Index form (NASA-TLX).

The demographic survey consists of 6 general background questions. The microtask contains 5 mandatory questions and 45 optional questions. When a worker completes the 5 mandatory questions, the conversational agent asks the worker whether he/she wants to continue, while the traditional Web interface features a button named I want to answer more questions that prompts additional questions when clicked. During task execution, both the Web interface and conversational agent induce a small break after 10 consecutive questions. During the breaks, the conversational agent (as well as the Web interface) show a “meme” for amusement. The rationale behind such a micro-diversion is to ensure that worker responses are not affected by boredom or fatigue [6, 33], making our experimental setup robust while measuring worker engagement across different conditions. Thereafter, the conversational agent periodically reminds workers that they can stop anytime and asks the worker if he/she wants to continue. Similarly on the Web interface, a click on the I want to answer more questions button prompts a meme and 10 more questions. Workers could quit at any point after the mandatory questions by entering ‘stop task’ in the conversational interfaces or clicking a stop button on the Web interface; this could be used by workers to exit the tasks and claim rewards for work completed.

Next, workers are asked to complete the short-form of the User Engagement Scale (UES-SF) [31, 32]. The UES-SF contains four sub-scales with 12 items, comprising a tool that is widely used for measuring user engagement in various digital domains. Each item is presented as a statement using a 7 point Likert-scale from “1: Strongly Disagree” to “7: Strongly Agree”. We chose the UES-SF since it has been validated in a variety of HCI contexts, and to date, it is the most tested questionnaire that measures user engagement. UES-SF perfectly fits our context of online crowdsourcing. With a total of only 12 items, it is easy to motivate workers to respond. Finally, workers are asked to complete the NASA Task Load Index (NASA-TLX) questionnaire, where workers rate their feelings about the task workload\(^4\). The questionnaire has six measurements (questions) about Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration respectively. We use the NASA-TLX due to considerable evidence of its robustness in measuring the cognitive task load (across 6 dimensions) of users accomplishing given tasks, which aligns with the goal of our study [16].

**Worker Interface**

Both the Web interface and the conversational interface are designed and implemented on top of AMT (see Figure 3). For both interfaces, the demographic survey, UES-SF and NASA-TLX are created using default HTML-based questions widgets provided by AMT; using the Crowd HTML Element.

The element crowd-radio-group including several crowd-radio-buttons is used for creating all the background questions from the demographic survey. The worker can select only one crowd-radio-button from the crowd-radio-group. The element crowd-slider is used for creating all the questions from UES-SF and NASA-TLX, since corresponding responses are on an integer scale ranging from 1 to 7 (UES-SF) or from 0 to 100 (NASA-TLX).

In recent work that has explored conversational interfaces for microtasking, the interface was built on top of platforms such as Telegram [29], demanding extra effort from crowd workers to register an account (if they were not registered users before) and redirecting workers to the social platform from the crowdsourcing platform. To overcome this limitation and fairly compare the conversational interface with the traditional web interface, we designed and implemented the conversational agent purely based on HTML and Javascript. Thus, it can be perfectly embedded on the AMT task page without any restrictions. Finally, a crowd-input element is placed below the conversational agent for entering the Task Token received on completion of the tasks.

The only difference between the interfaces (traditional Web versus conversational) is in the interaction with the user and how input is received. The Web interface contains either crowd-input or crowd-radio-group, respectively for free text and multiple choices, whereas the conversational interface uses textarea (shown at the bottom) and bubble-like buttons for each. As shown in Figure 3, we developed a rule-based conversational agent based on chat-bubble\(^5\).

**Experimental Setup**

Each experimental condition (modeled as a batch of HITs) consists of 50 questions and we recruit 50 unique workers to answer these 50 questions. Each worker is asked to complete at least 5 mandatory questions. Across the 16 experimental conditions, we thereby acquired responses from 16 × 50 = 800 unique workers in total.

When a worker successfully completes the demographic survey, UES-SF, NASA-TLX and at least 5 mandatory questions, the worker immediately receives 0.5$. The reward for the optional questions is given to workers through the “bonusing” function on AMT. We estimated the execution time and paid workers 0.01$ per optional task as a bonus for the image tasks (Image Classification and CAPTCHA Recognition), 0.02$ per optional task for the text tasks (Information Finding and Sentiment Analysis). On task completion, we instantly bonused workers the difference required to meet an hourly pay of 7.25$ based on the total time they spent on tasks (including the time for breaks). The instructions clearly explained rewards for each optional task; workers knew of the base reward and

---

\(^4\)NASA-TLX: Task Load Index. https://humansystems.arc.nasa.gov/groups/TLX/

\(^5\)https://github.com/dmitrizzle/chat-bubble
bonuses at the onset, ensuring that there was no unnatural financial uncertainty other than what is typical on AMT.

Quality Control
To prevent malicious workers from executing the crowdsource
tasks, we only accept participants whose overall HIT ap-
proval rates are greater than 95%. Using Javascript and track-
ing worker-ids, we also ensure that each worker submits at
most one assignment across all experimental conditions, to
avoid learning biases due to repeated participation.

Evaluation Metrics
The dependent variables in our experiments are output quality,
worker engagement, and cognitive task load. We use pairwise
independent tests to test for statistical significance (expected
\( \alpha = 0.05 \), two-tailed, corrected to control for Type-I error
inflation in our multiple comparisons). We use the Holm-
Bonferroni correction to control the family-wise error rate
(FWER) [18].

Output quality, is measured in terms of the judgment accu-
curacy of workers. It is measured by comparing the workers’
responses with the ground truth. Thus, a given worker’s ac-
curacy is the fraction of correct answers provided by the worker
among all the provided answers. In case of Information Find-
ing tasks, the stars provided by workers should exactly match
the stars from Google Maps. For the other task types, the
workers’ answers (string) should be identical to the ground
truth (case insensitive).

Worker engagement, is measured using 2 popular approaches:
1) the worker retention, i.e. the number of answered optional
questions, and the proportion of workers answering at least one
optional question; and 2) the UES-SF overall score (ranging
from 1 to 7; the higher the UES score is, the more engaged the
worker is).

Cognitive task load, is evaluated by unweighted NASA-TLX
test. Through the scores (ranging from 0 to 100: higher score
means the heavier task load) of the TLX test, we study if and
how conversational interfaces affect perceived cognitive load
for the executed task.

RESULTS
Worker Demographics
Of the unique 800 workers, 37.8% were female and 62.2% were male. Most workers (89.8%) were under 45 years old.

72.7% of workers reported that their education levels were
higher than (or equal to) Bachelor’s degree. 37.9% of the
workers claimed AMT as their primary source of income,
while about half of the workers (55.8%) reported that AMT
was their secondary source of income.

Distribution of Conversational Styles
We estimated the conversational style of workers across all the
conversational interface conditions using the method proposed
in Figure 2. The number of workers whose conversational
styles were estimated as High Involvement and High Consider-
ateness are shown in Figure 4.

As we described earlier, the conversational agent maintains a
High-Involvement and High-Considerateness styles in Con+I
and Con+C conditions respectively, while in Con+A conditions
the conversational agent initiates with either High-
Involvement or High-Considerateness style randomly. Figure
5 shows the number of workers whose conversational styles
were estimated as High Involvement and High Considerateness
respectively in conversational interfaces with style alignment
**Output Quality**

**Main result:** In terms of output quality, conversational interfaces have no significant difference (min $p = 0.09$) compared to the traditional web interface, and there is no significant difference across conversational styles.

Table 4 shows the mean and standard deviation of workers accuracy across all experimental conditions. Since the Image Classification task is objective and simple, we obtained high-accuracy (98%-100%) results across the 4 different interface conditions.

Pairwise independent t-tests revealed no significant difference in the output qualities across four interfaces (conversational styles) within each task type. This aligns with the findings from previous work [29]. For Image Classification tasks, the worker accuracy across all interfaces and conversational styles is higher than other types of tasks due to the relative simplicity.

**Table 4. Worker accuracy ($\mu \pm \sigma$; mean and standard deviation) and $\rho$-values across different task types and interface conditions.**

<table>
<thead>
<tr>
<th>Task type</th>
<th>Web (vs. Con+IC,A)</th>
<th>Con+I (vs. Con+CA)</th>
<th>Con+C (vs. Con+A)</th>
<th>Con+A</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF</td>
<td>0.66 ± 0.29</td>
<td>0.63 ± 0.3</td>
<td>0.58 ± 0.3</td>
<td>0.65 ± 0.29</td>
</tr>
<tr>
<td></td>
<td>(p = 0.69, 0.2, 0.88)</td>
<td>(p = 0.37, 0.81)</td>
<td>(p = 0.26)</td>
<td></td>
</tr>
<tr>
<td>SA</td>
<td>0.62 ± 0.27</td>
<td>0.54 ± 0.29</td>
<td>0.62 ± 0.26</td>
<td>0.64 ± 0.27</td>
</tr>
<tr>
<td></td>
<td>(p = 0.18, 0.99, 0.74)</td>
<td>(p = 0.16, 0.09)</td>
<td>(p = 0.74)</td>
<td></td>
</tr>
<tr>
<td>CR</td>
<td>0.72 ± 0.16</td>
<td>0.69 ± 0.14</td>
<td>0.67 ± 0.19</td>
<td>0.75 ± 0.12</td>
</tr>
<tr>
<td></td>
<td>(p = 0.33, 0.13, 0.23)</td>
<td>(p = 0.48, 0.02)</td>
<td>(p = 0.01)</td>
<td></td>
</tr>
<tr>
<td>IC</td>
<td>1.0 ± 0.03</td>
<td>0.98 ± 0.09</td>
<td>0.99 ± 0.04</td>
<td>0.98 ± 0.07</td>
</tr>
<tr>
<td></td>
<td>(p = 0.19, 0.39, 0.09)</td>
<td>(p = 0.41, 0.95)</td>
<td>(p = 0.29)</td>
<td></td>
</tr>
</tbody>
</table>

**Worker Engagement**

**Worker Retention.**

**Main result:** Conversational interfaces lead to significantly higher worker retention in multiple-choice tasks compared to the traditional web interface. Particularly, a High-Involvement style corresponds to significantly higher worker retention across all task types compared to the web interface.

Figure 6 shows a violin plot representing the number of optional tasks completed by workers. In this figure, each “violin” represents the distribution of workers in each of the experimental conditions. The width of the violin at any point, represents the number of workers who answered the corresponding number of optional questions. The distribution does not meet any assumptions for parametric tests. Thus, we use the Wilcoxon Rank-Sum test (expected $\alpha = 0.05$, two-tailed, corrected by Holm-Bonferroni method) to test the significance of the pairwise difference. Results are shown in Table 5. We found that across all task types, the number of optional tasks completed by workers using the Web interface was significantly lower than that in the conversational interface with Involvement style (RQ2). Compared with Web, the conversational interface with style alignment (Con+A) also shows significantly higher worker retention except in the Information Finding task, while the Considerateness style shows significantly higher worker retention in multiple-choice tasks (RQ2). We found that the workers using conversational interfaces were generally better retained than the Web workers in multiple-choice tasks, and none of the Web workers completed all the available optional questions in the Information Finding task (RQ1).

**Table 5. The worker retention ($\mu \pm \sigma$; mean and standard deviation, unit: the number of optional tasks completed by workers) and $\rho$-values across different task types and interface conditions.**

<table>
<thead>
<tr>
<th>Task type</th>
<th>Web (vs. Con+IC,A)</th>
<th>Con+I (vs. Con+CA)</th>
<th>Con+C (vs. Con+A)</th>
<th>Con+A</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF</td>
<td>4.18 ± 8.66</td>
<td>15.47 ± 18.0</td>
<td>7.55 ± 12.4</td>
<td>9.4 ± 13.63</td>
</tr>
<tr>
<td></td>
<td>(p = 1.8e-4, 0.02, 5.1e-3)</td>
<td>(p = 0.08, 0.17)</td>
<td>(p = 0.67)</td>
<td></td>
</tr>
<tr>
<td>SA</td>
<td>5.4 ± 11.99</td>
<td>11.78 ± 15.26</td>
<td>8.63 ± 10.32</td>
<td>14.92 ± 16.09</td>
</tr>
<tr>
<td></td>
<td>(p = 2.7e-5, 8.3e-5, 2.3e-6)</td>
<td>(p = 0.6, 0.29)</td>
<td>(p = 0.09)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(p = 1.3e-3, 2.3e-3, 2.8e-4)</td>
<td>(p = 0.98, 0.22)</td>
<td>(p = 0.19)</td>
<td></td>
</tr>
<tr>
<td>IC</td>
<td>28.6 ± 17.97</td>
<td>28.6 ± 17.97</td>
<td>20.29 ± 19.08</td>
<td>25.61 ± 20.52</td>
</tr>
<tr>
<td></td>
<td>(p = 1.2e-6, 6.1e-5, 2.9e-5)</td>
<td>(p = 0.07, 0.67)</td>
<td>(p = 0.34)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6 lists the number and percentage of the workers who answered at least one optional question. While only 26%-32% of workers decided to answer at least one optional question in the Web condition, 60%-84% of the workers operating with the conversational agents answered at least one optional question. This result also suggests a higher degree of retention associated with the conversational interface.

**Table 6. The number of workers (percentages) who completed at least one optional question across all task types and the four interfaces.**

<table>
<thead>
<tr>
<th>Task type</th>
<th>Web</th>
<th>Con+I</th>
<th>Con+C</th>
<th>Con+A</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF</td>
<td>16 (32%)</td>
<td>34 (68%)</td>
<td>30 (60%)</td>
<td>32 (64%)</td>
</tr>
<tr>
<td>SA</td>
<td>13 (26%)</td>
<td>40 (80%)</td>
<td>39 (78%)</td>
<td>41 (82%)</td>
</tr>
<tr>
<td>CR</td>
<td>14 (28%)</td>
<td>34 (68%)</td>
<td>34 (68%)</td>
<td>35 (70%)</td>
</tr>
<tr>
<td>IC</td>
<td>13 (26%)</td>
<td>42 (84%)</td>
<td>38 (76%)</td>
<td>37 (74%)</td>
</tr>
<tr>
<td>Overall</td>
<td>56 (28%)</td>
<td>150 (75%)</td>
<td>141 (70.5%)</td>
<td>145 (72.5%)</td>
</tr>
</tbody>
</table>

**User Engagement Scale (UES-SF).**

**Main result:** Input and data types can significantly affect the UES-SF score, while interfaces and conversational styles were found to have no significant impact.

Table 7 lists the UES-SF scores across all the experimental conditions. Pairwise independent t-tests (expected $\alpha = 0.05$, two-tailed, corrected by Holm-Bonferroni method) between web and conversational interfaces (RQ1) with different conversational styles (RQ2) show that the UES-SF scores have no significant difference across four interfaces (conversational styles) within each task type.

However, as shown in Table 8 ($p$-values), between-task pairwise independent t-tests (expected $\alpha = 0.05$, two-tailed, corrected by Holm-Bonferroni method) revealed that the overall Perceived Usability of image-based tasks (CAPTCHA Recognition and Image Classification) is significantly higher than text-based tasks (Information Finding and Sentiment Analysis). In terms of overall Aesthetic Appeal, Reward Factor and Overall UES score, the scores of multiple-choice tasks (Sentiment Analysis and Image Classification) are higher than free-text tasks (Information Finding and CAPTCHA Recognition) with statistical significance.
Figure 6. A violin plot representing the number of optional questions answered by workers across different task types and different interfaces, where the black dots represent the mean value. A violin plot is a hybrid of a box plot and a kernel density plot, revealing peaks in the data that cannot be visualized using box plots.

### Table 7. The UES-SF score ($\mu \pm \sigma$; mean and standard deviation) of all task types with four interfaces.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Information Finding</th>
<th>Sentiment Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Web</td>
<td>Con+I</td>
</tr>
<tr>
<td>Focused attention</td>
<td>4.12±1.40</td>
<td>4.39±1.44</td>
</tr>
<tr>
<td>Perceived usability</td>
<td>3.71±1.67</td>
<td>3.70±1.61</td>
</tr>
<tr>
<td>Aesthetic appeal</td>
<td>4.23±1.46</td>
<td>4.29±1.29</td>
</tr>
<tr>
<td>Reward factor</td>
<td>4.35±1.23</td>
<td>4.44±1.53</td>
</tr>
<tr>
<td>Overall</td>
<td>4.10±0.85</td>
<td>4.21±0.85</td>
</tr>
</tbody>
</table>

**Table 8. p-values of between-task statistical tests of UES-SF score.**

<table>
<thead>
<tr>
<th>Categories</th>
<th>IF vs. SA</th>
<th>IF vs. CR</th>
<th>IF vs. IC</th>
<th>SA vs. CR</th>
<th>SA vs. IC</th>
<th>CR vs. IC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focused attention</td>
<td>0.11</td>
<td>0.90</td>
<td>0.11</td>
<td>0.85</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Perceived usability</td>
<td>0.21</td>
<td>0.20</td>
<td>0.01</td>
<td>0.17</td>
<td>0.85</td>
<td>0.16</td>
</tr>
<tr>
<td>Aesthetic appeal</td>
<td>0.60</td>
<td>0.54</td>
<td>0.54</td>
<td>0.51</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>Reward factor</td>
<td>0.94</td>
<td>0.73</td>
<td>0.73</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>Overall</td>
<td>7.3e-6*</td>
<td>9.4e-2</td>
<td>3.2e-8*</td>
<td>2.4e-2</td>
<td>0.14</td>
<td>6.6e-4*</td>
</tr>
</tbody>
</table>

* = statistically significant (corrected t-test)

### DISCUSSION

**Main result:** We found no significant difference in NASA-TLX scores across different interfaces (web vs. conversational interface and between conversational styles).

To answer RQ2, we calculated and listed unweighted NASA-TLX scores in Table 9. According to pairwise independent t-tests (expected $\alpha = 0.05$, two-tailed, corrected by Holm-Bonferroni method), the NASA-TLX scores have no significant difference across four interfaces (conversational styles) within each task type. However, the conversational interface with aligned style has the potential to reduce the cognitive task load for Information Finding task compared with the web interface (no significance, $p = 0.033$, which is less than 0.05 but higher than corrected $\alpha$).

**Important to study how conversational styles and alignment can improve worker experience and satisfaction.**

Through our experiments, we found that workers preferred using High-Considerateness style while conducting Information Finding and Sentiment Analysis tasks. In contrast, we found that workers tend to use High-Involvement style while completing CAPTCHA Recognition and Image Classification tasks. This suggests that workers are likely to exhibit an involved conversational style when they are relatively more confident, or the tasks are less difficult (RQ2). The results of style alignment further show that workers’ conversational styles are mainly affected by task types rather than initial styles of the agent. We note that Information Finding and Sentiment Analysis tasks are typically more complex [42] in comparison to CAPTCHA Recognition and Image Classification. This calls for further exploration of the impact of task complexity on task outcomes within conversational microtask crowdsourcing.

In terms of the effect of conversational styles on worker retention, there was no significant difference between the different
We found that workers tend to exhibit different conversational styles. A possible explanation can be the maximum limit (45) of the available optional tasks that a worker can answer, as we found that many workers who conducted image-based tasks (i.e. CAPTCHA Recognition and Image Classification) on the conversational interfaces with High-Involvement and style alignment completed all the available 45 optional tasks. Our findings regarding the impact of conversational style on worker retention suggests that a High-Involvement conversation style can provide workers with engagement stimuli for long-term retention (RQ2). As 45 optional tasks limit the scale of worker retention in this study, using an unlimited number of optional questions to analyze the impact on worker retention should be considered in the future research.

Our results showed significant differences between image-based tasks and text-based tasks with regard to UES-SF scores. This is potentially due to the complexity of the tasks (the two text-based tasks are more taxing than the two image-based tasks). The results also suggest that the input type (free text vs. multiple choices) have a principal impact on the UES-SF scores, which weaken the effect of different interfaces and conversational styles. The influence of task complexity and its mediating interaction with conversational styles should be considered in the imminent future.

There was no significant difference in NASA-TLX scores of workers between web and conversational interfaces. As Information Finding and Sentiment Analysis are more demanding than the CAPTCHA Recognition and Image Classification, the results of NASA-TLX also suggest that the task complexity has an impact on the perceived cognitive load. We aim to study this further and tease out the interaction between task complexity and cognitive load in conversational microtasking.

**Design Implications for HCI**

We found that workers tend to exhibit different conversational styles due to the effect of task complexity. However, our results of aligning conversational styles of the agent with that of the workers suggest that giving the conversational agent a High-Involvement style can generally improve the worker retention in conversational microtask crowdsourcing.

A healthy relationship between workers and requesters is critical to the sustainability of microtask marketplaces. It is in the interest of requesters to take steps to ensure this. By adopting conversational interfaces, requesters can improve worker engagement, particularly in less complex tasks as suggested by our findings, allowing workers to complete more work, earn more money, and foster good faith in the requester-worker long term relationship. These constitute important design implications that task requesters can consider while optimizing for worker engagement in long batches of HITs. Distilling the complex interactions between task difficulty, conversational styles and quality related outcomes in conversational microtasking can help make crowdsourcing systems more engaging and effective. The HCI community is uniquely suited to further explore the impact of conversational styles on quality related outcomes in microtask crowdsourcing, and we believe our work presents an important first step in this direction. Accurately estimating the general or preferred conversational styles of individuals, so as to adapt conversational styles of agents can bear great dividends in domains beyond conversational microtoking.

In this study, we have designed and developed a web-based conversational agent that is able to execute crowdsourcing microtasks of common task types (including Information Finding, Sentiment Analysis, Optical Character Recognition, Image Classification, Audio Transcription, Survey, etc.). Furthermore, since the conversational interface is purely HTML-based, elements used in traditional web interfaces can be easily ported into conversational interfaces. Therefore, the overheads of designing and implementing conversational interfaces can be easily reduced, which is a small price to pay for an increase in worker engagement. The code corresponding to our conversational agent is available publicly on the companion page, alongside data for the benefit of the community.

**Caveats and Limitations**

Our findings with respect to the impact of conversational interfaces on worker engagement across different task types suggest that different conversational styles of the agent can affect the worker retention, albeit not consistently. Moreover, further experiments that decouple the impact of task difficulty are needed to fully uncover the impact of conversational styles in conversational microtask crowdsourcing. Having said that, our findings are an important first step towards optimizing novel conversational interfaces for microtask crowdsourcing.

**Influence of Monetary Incentives.** Workers earned monetary rewards across all conditions in our study. Monetary rewards have been shown to incentivize workers to complete more work [7]. However, we ensured that the pay per unit time (reward) is identical across all conditions and task types: making comparisons across conditions in our study valid and meaningful. Our long-term goal through conversational microtasking is to improve engagement, help workers overcome fatigue or boredom and reduce task abandonment [15].

**Implementing Conversational Interfaces.** For task requesters, it can be difficult to adapt some types of tasks to conversational interfaces (such as drawing free-form boundaries around objects). However, as research in conversational microtasking advances, so will the support for requester assistance in realizing such interfaces with ease. Requesters can further consider the trade-off between implementation costs and the benefits of increased worker engagement.

**CONCLUSIONS**

In this paper, we studied how workers engagement can be affected by conversational interfaces and conversational styles. We conducted online crowdsourcing experiments to study whether the worker engagement can be affected by the conversational interface (RQ1). We used post-task surveys to test workers’ user engagement and cognitive load while completing tasks using conversational interfaces with different conversational styles (RQ2). We show that the use of conversational interfaces can improve the perceived worker engagement, and that the adopted conversational style can also have an effect on worker retention.
REFERENCES


