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
CHI 2020 - Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems

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 $\times$ 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

micro-breaks into workflows to refresh workers, and showed that under certain conditions micro-breaks aid in worker retention and improve their accuracy marginally. Similarly, [6] proposed to intersperse diversions (small periods of entertainment) to improve worker experience in lengthy, monotonous microtasks and found that such micro-diversions can significantly improve worker retention rate while maintaining worker performance. Other works proposed the use of gamification to increase worker retention and throughput [10]. [28] studied worker engagement, characterized how workers perceive tasks and proposed to predict when workers would stop performing tasks. [7] introduced pricing schemes to improve worker retention, and showed that paying periodic bonuses according to pre-defined milestones has the biggest impact on retention rate of workers.

In this work, we measure worker engagement by using the proxy of worker retention, and a standardized questionnaire called the 'user engagement scale' (UES), introduced by [32]. The UES was recently used by [44] to study the impact of worker moods on their engagement in crowdsourced information finding tasks.

METHOD: CONVERSATIONAL INTERFACE FOR MICROTASK CROWDSOURCING

In this study, we design and implement conversational interfaces that enable the entire task execution process, while exploring the impact of different conversational styles on worker performance and engagement. The reader can directly experience interaction with the conversational interface on the companion page.[1]

Workflow of Conversational Microtasking

The conversational interface is designed to help workers in carrying out crowdsourcing tasks. The main building blocks of conversational microtasks are similar to those of traditional Web interfaces; they include initiating the conversation (starting the task execution), answering questions, and finally paying the workers. To assist the workers in task execution, the workflow of conversational microtask crowdsourcing, as realized in our study is depicted in Figure 1 and described below.

1) After a worker accepts the task and opens the task page, the conversational interface is initialized with opening greetings from the conversational agent. The worker can respond by selecting one of two options. During this step, the conversational agent prompts brief information about the task, such as the task name and the time limit. The goal of this step is twofold: to make users familiar with the conversational interface; and to estimate the conversational style of the worker. As explained later (in Section *Aligning Conversational Styles*), this step is needed to align the agent's conversational style with that of the worker.

2) If the worker asks for the task instructions after the opening greetings, the conversational agent prompts the task instructions. Otherwise, this step is skipped.

3) Next, the conversational agent presents tasks framed as questions to the worker. On answering a question, another one 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.

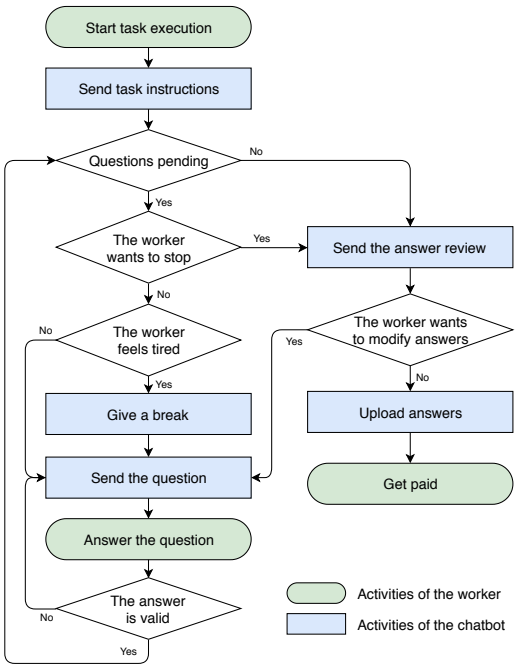

Figure 1. The workflow of conversational microtask crowdsourcing.

[1] https://qiusihang.github.io/csbot

Table 1. Design criteria for conversation styles of the agent.

Criteria	High-Involvement	High-Considerateness
C1. Rate of speech	fast	slow
C2. Turn taking	fast	slow
C3. Introduction of topics	w/o hesitation	w/ hesitation
C4. Use of syntax	simple	complex
C5. Directness of content	direct	indirect
C6. Utterance of questions	frequent	rare

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

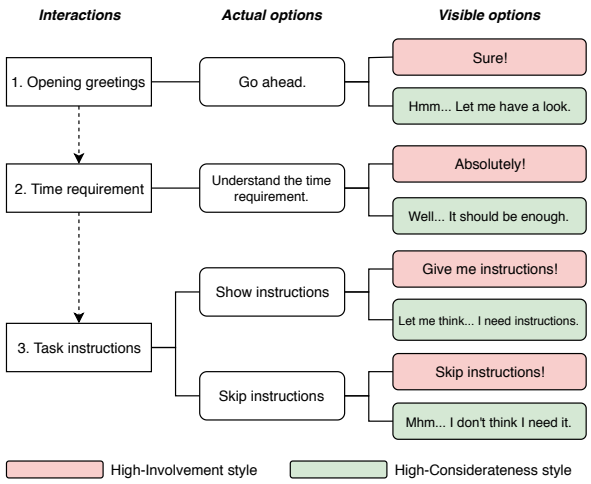

Figure 2. Options given to the worker for conversational style estimation and alignment.

(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

Table 2. Conversation templates for conversational agents with high-involvement and high-considerateness styles designed according to criteria distilled from Tannen's characterization of conversation styles (cf. Table 1).

Interactions	High-Involvement	High-Considerateness	Criteria
Opening greetings	Hey! Can you help me with a task called [TASK NAME]?	Thank you in advance for helping me with a task called [TASK NAME].	C4, C6.
Time requirements	You must complete this task within 30 minutes, otherwise I won't pay you :-)	I think 30 minutes should be more than enough for you to finish :-)	C5.
Task instructions	Here is the task instructions. Take a look!	I kindly ask you to have a look at the task instructions.	C4.
Introducing questions	Listen, the first question! / OK! The next one. / Here you go.	Good! Here is the first question. / Okay, I got it. Here is the next question. / Alright, this is the question you want to have a look again.	C4.
Completing mandatory questions	Hey, good job! The mandatory part has been done! I know you want to continue, right?	OK, you have finished the mandatory part of the task. Well... please let me know if you want to answer more questions.	C3, C6.
Receiving an invalid answer	Oops, I don't understand your answer. Do you forget how to answer the question? Just type "instruction".	Hmm... Sorry, I don't get it. Maybe you can type "instruction" to learn how to answer the question.	C3, C6.
Break	Are you feeling tired? If I'm driving you crazy, you can type "stop task" to leave me.	Well... alright, it seems that you have answered a lot of questions. No worries, you can type "stop task" if you don't want to continue.	C3, C5, C6.
Review	You have completed the task! Here are your answers: [ANSWERS]. Something wrong? Just edit the answer by typing its question number, or type "submit" to submit your answers.	Good job! The task has been completed. Here is the review of your answers: [ANSWERS]. Well... if you find something wrong here, please edit the answer by typing its question number. Otherwise, you can type "submit" to submit your answers.	C3, C4, C6.
Bye	Your task token is [TASK TOKEN]. I'm off ;)	Your task token is [TASK TOKEN]. Thank you! Your answers have been submitted. Nice talking to you. Bye!	C4.

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 widgets 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 3. Summary of task types.

Input type	**Text**	**Imagery**
Free text	Information Finding	CAPTCHA Recognition
Multiple choices	Sentiment Analysis	Image Classification

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 dataset[2].

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 Claptcha[3], 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,

[2] Yelp Open Dataset. https://www.yelp.com/dataset

[3] 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 $\times$ 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].

[4]NASA-TLX: Task Load Index. https://humansystems.arc.nasa.gov/groups/TLX/

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-button`s 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 $16 \times 50 = 800$ 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

[5]https://github.com/dmitrizzle/chat-bubble

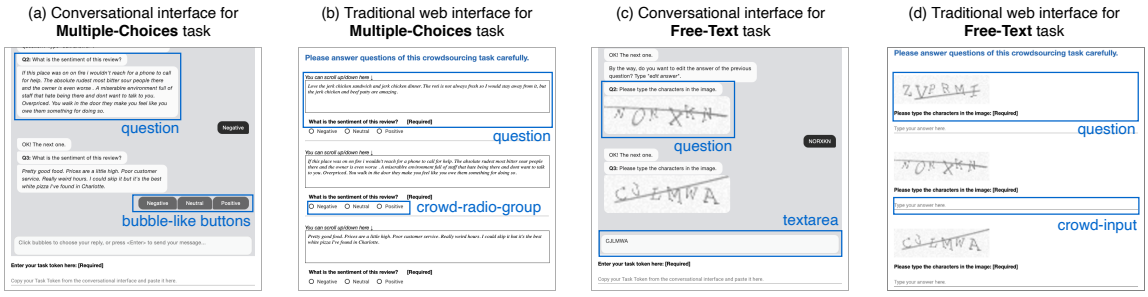

Figure 3. The comparison of conversational interfaces embedded on the user interface of AMT and traditional web interfaces using HTML elements provided by AMT, where the worker needs to provide a Task Token acquired from the conversational interface after the task is completed.

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 crowdsourcing tasks, we only accept participants whose overall HIT approval rates are greater than 95%. Using Javascript and tracking `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 accuracy of workers. It is measured by comparing the workers' responses with the ground truth. Thus, a given worker's accuracy is the fraction of correct answers provided by the worker among all the provided answers. In case of Information Finding 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 Considerateness are shown in Figure 4.

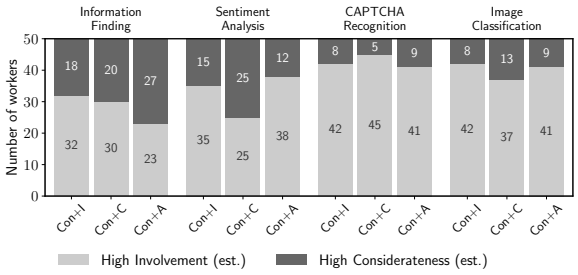

Figure 4. Distributions of estimated styles across all conditions.

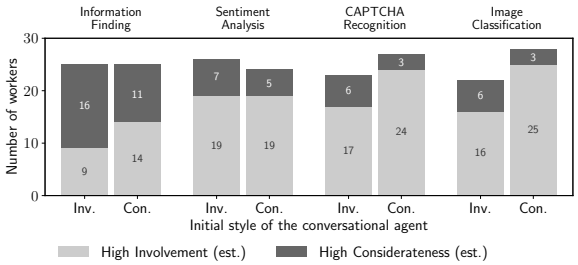

Figure 5. Distributions of estimated styles of conversational interfaces with style alignment by two initial styles.

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

(Con+A), across all task types with two initial conversational styles (High Involvement and High Considerateness).

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 the 16 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 p-values across different task types and interface conditions.

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

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 p-values across different task types and interface conditions.

Task type	**Web** (vs. Con+I,C,A)	**Con+I** (vs. Con+C,A)	**Con+C** (vs. Con+A)	**Con+A**
IF	4.18 ± 8.66 ($p =$ **1.8e-4***, 0.02, 5.1e-3)	15.47 ± 18.0 ($p = 0.08, 0.17$)	7.55 ± 12.14 ($p = 0.67$)	9.4 ± 13.63
SA	5.4 ± 11.99 ($p =$ **2.7e-5***, **8.3e-5***, **2.3e-6***)	11.78 ± 15.26 ($p = 0.6, 0.29$)	8.63 ± 10.32 ($p = 0.09$)	14.92 ± 16.09
CR	8.4 ± 16.75 ($p =$ **1.3e-3***, 2.3e-3, **2.0e-4***)	14.96 ± 16.73 ($p = 0.98, 0.22$)	15.14 ± 16.8 ($p = 0.19$)	21.37 ± 19.9
IC	8.7 ± 17.17 ($p =$ **1.2e-6***, **6.1e-5***, **2.9e-5***)	28.6 ± 17.97 ($p = 0.07, 0.67$)	20.29 ± 19.08 ($p = 0.34$)	25.61 ± 20.52

* = statistically significant (corrected Wilcoxon Rank-Sum test)

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 (with percentages) who completed at least one optional question across all task types and the four interfaces.

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

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.

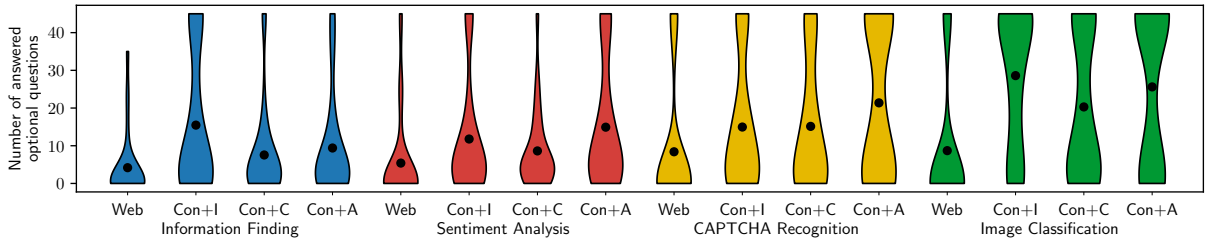

Figure 6. A violin plot representing the number of optional questions answered by workers across different task types and different interfaces, where the black dots represent the mean value. A violin plot is a hybrid of a box plot and a kernel density plot, revealing peaks in the data that cannot be visualized using box plots.

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

Categories	Web	Con+I	Con+C	Con+A	Overall	Web	Con+I	Con+C	Con+A	Overall
	Information Finding					**Sentiment Analysis**				
Focused attention	4.12 ± 1.40	4.39 ± 1.44	3.68 ± 1.45	3.81 ± 1.53	$\mathbf{3.98 \pm 1.51}$	4.12 ± 1.30	4.43 ± 1.20	4.07 ± 1.46	4.28 ± 1.38	$\mathbf{4.21 \pm 1.37}$
Perceived usability	3.71 ± 1.67	3.70 ± 1.61	3.86 ± 1.70	4.24 ± 1.61	$\mathbf{3.86 \pm 1.68}$	3.91 ± 1.83	3.86 ± 1.67	4.19 ± 1.63	4.42 ± 1.85	$\mathbf{4.08 \pm 1.78}$
Aesthetic appeal	4.23 ± 1.46	4.29 ± 1.29	4.10 ± 1.12	4.01 ± 1.58	$\mathbf{4.14 \pm 1.40}$	4.75 ± 1.28	4.67 ± 1.51	4.84 ± 1.31	4.86 ± 1.12	$\mathbf{4.76 \pm 1.35}$
Reward factor	4.35 ± 1.23	4.44 ± 1.53	4.41 ± 1.33	4.17 ± 1.49	$\mathbf{4.33 \pm 1.44}$	4.99 ± 1.23	4.90 ± 1.33	4.95 ± 1.31	5.05 ± 1.37	$\mathbf{4.95 \pm 1.36}$
Overall	4.10 ± 0.85	4.21 ± 0.85	4.01 ± 0.69	4.06 ± 1.00	$\mathbf{4.07 \pm 0.90}$	4.44 ± 0.87	4.46 ± 0.98	4.51 ± 0.90	4.65 ± 0.88	$\mathbf{4.50 \pm 0.97}$
	CAPTCHA Recognition					**Image Classification**				
Focused attention	3.83 ± 1.76	3.92 ± 1.61	3.93 ± 1.56	4.39 ± 1.55	$\mathbf{4.00 \pm 1.66}$	4.30 ± 1.45	4.21 ± 1.77	4.35 ± 1.62	4.16 ± 1.85	$\mathbf{4.23 \pm 1.70}$
Perceived usability	4.95 ± 1.57	4.71 ± 1.66	4.56 ± 1.64	4.74 ± 1.43	$\mathbf{4.71 \pm 1.62}$	4.41 ± 1.93	4.91 ± 1.53	4.90 ± 1.67	4.68 ± 1.78	$\mathbf{4.70 \pm 1.77}$
Aesthetic appeal	3.74 ± 1.71	4.10 ± 1.73	3.95 ± 1.81	3.94 ± 1.69	$\mathbf{3.92 \pm 1.76}$	4.73 ± 1.42	4.53 ± 1.56	4.75 ± 1.37	4.75 ± 1.65	$\mathbf{4.67 \pm 1.54}$
Reward factor	4.43 ± 1.71	4.42 ± 1.79	4.50 ± 1.68	4.25 ± 1.66	$\mathbf{4.38 \pm 1.74}$	4.97 ± 1.32	5.09 ± 1.73	4.87 ± 1.56	5.14 ± 1.60	$\mathbf{5.00 \pm 1.60}$
Overall	4.24 ± 1.27	4.29 ± 1.11	4.23 ± 1.12	4.33 ± 1.14	$\mathbf{4.25 \pm 1.20}$	4.60 ± 0.91	4.69 ± 1.20	4.72 ± 1.03	4.68 ± 1.19	$\mathbf{4.65 \pm 1.13}$

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

Categories	IF vs. SA	IF vs. CR	IF vs IC	SA vs. CR	SA vs. IC	CR vs. IC
Focused attention	0.11	0.90	0.11	0.17	0.85	0.16
Perceived usability	0.21	**2.8e-7***	**1.2e-6***	**1.8e-4***	**4.1e-4***	0.96
Aesthetic appeal	**8.0e-6***	0.16	**3.6e-4***	**1.1e-7***	0.53	**6.5e-6***
Reward factor	**9.4e-6***	0.73	**1.2e-5***	**2.8e-4***	0.75	**2.4e-4***
Overall	**7.3e-6***	9.4e-2	**3.2e-8***	**2.4e-2***	0.14	**6.6e-4***

* = statistically significant (corrected t-test)

Table 9. The unweighted NASA-TLX score ($\mu \pm \sigma$: mean and standard deviation) and p-values of all task types with four interfaces.

Task type	Web (vs. Con+I,C,A)	Con+I (vs. Con+C,A)	Con+C (vs. Con+A)	Con+A
IF	52.35 ± 20.75 ($p = 0.51, 0.12, 0.03$)	49.62 ± 20.39 ($p = 0.37, 0.13$)	46.05 ± 19.63 ($p = 0.51$)	43.4 ± 20.25
SA	50.27 ± 17.76 ($p = 0.95, 0.31, 0.17$)	50.02 ± 20.54 ($p = 0.37, 0.22$)	46.54 ± 18.26 ($p = 0.71$)	45.15 ± 18.85
CR	38.23 ± 19.56 ($p = 0.81, 0.74, 0.6$)	37.29 ± 20.26 ($p = 0.58, 0.78$)	39.54 ± 20.2 ($p = 0.4$)	36.14 ± 19.89
IC	43.38 ± 22.64 ($p = 0.46, 0.07, 0.44$)	40.22 ± 19.56 ($p = 0.23, 0.94$)	35.57 ± 19.11 ($p = 0.3$)	39.89 ± 21.94

Cognitive Task Load

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 α).

DISCUSSION

Aspects such as task complexity [42], task types, instructions [13] are instrumental in shaping crowd work [23]. However, previous work has shown that conversational interfaces can effectively benefit workers from different perspectives, such as satisfaction [29] and effort [26, 19]. Conversational interfaces are on the rise across different domains and it is 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

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 microtasking.

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 [42] 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

[1] Alan Aipe and Ujwal Gadiraju. 2018. SimilarHITs: Revealing the Role of Task Similarity in Microtask Crowdsourcing. In *Proceedings of the 29th on Hypertext and Social Media*. ACM, 115–122.

[2] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. 2018. SearchBots: User Engagement with ChatBots During Collaborative Search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. ACM, 52–61.

[3] Luka Bradeško, Michael Witbrock, Janez Starc, Zala Herga, Marko Grobelnik, and Dunja Mladenić. 2017. Curious Cat–Mobile, Context-Aware Conversational Crowdsourcing Knowledge Acquisition. *ACM Transactions on Information Systems (TOIS)* 35, 4 (2017), 33.

[4] Phil Cohen, Adam Cheyer, Eric Horvitz, Rana El Kaliouby, and Steve Whittaker. 2016. On the future of personal assistants. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1032–1037.

[5] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. What can i help you with?: infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 43.

[6] Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. 2015. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. In *Proceeding of The 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. ACM, 628–638.

[7] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, and Philippe Cudré-Mauroux. 2014. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

[8] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G Ipeirotis, and Philippe Cudré-Mauroux. 2015. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *Proceedings of the 24th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 238–247.

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2007. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer vision and Image understanding* 106, 1 (2007), 59–70.

[10] Oluwaseyi Feyisetan, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. 2015. Improving paid microtasks through gamification and adaptive furtherance incentives. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 333–343.

[11] Joel E Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for voice interface design. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. ACM, 26.

[12] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 218–223.

[13] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, 5–14.

[14] Bettina Graf, Maike Krüger, Felix Müller, Alexander Ruhland, and Andrea Zech. 2015. Nombot: simplify food tracking. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 360–363.

[15] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 321–329.

[16] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[17] Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2019. An End-to-End Conversational Style Matching Agent. *arXiv preprint arXiv:1904.02760* (2019).

[18] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[19] Ting-Hao Kenneth Huang, Joseph Chee Chang, and Jeffrey P Bigham. 2018. Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 295.

[20] Ting-Hao Kenneth Huang, Walter S Lasecki, and Jeffrey P Bigham. 2015. Guardian: A crowd-powered spoken dialog system for web apis. In *Third AAAI conference on human computation and crowdsourcing*.

[21] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 86, 12 pages. DOI:http://dx.doi.org/10.1145/3290605.3300316

[22] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM, 121–130.

[23] 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*. ACM, 1301–1318.

[24] Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan Houben. 2018. Human-Aided Bots. *IEEE Internet Computing* 22, 6 (2018), 36–43.

[25] Robin Tolmach Lakoff. 1979. Stylistic strategies within a grammar of style. *Annals of the New York Academy of Sciences* 327, 1 (1979), 53–78.

[26] Walter S Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P Bigham. 2013. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 151–162.

[27] Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5286–5297.

[28] Andrew Mao, Ece Kamar, and Eric Horvitz. 2013. Why stop now? predicting worker engagement in online crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[29] Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2019. Chatterbox: Conversational Interfaces for Microtask Crowdsourcing. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, 243–251.

[30] Robert J Moore, Raphael Arar, Guang-Jie Ren, and Margaret H Szymanski. 2017. Conversational UX design. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 492–497.

[31] Heather O'Brien. 2016. Theoretical perspectives on user engagement. In *Why Engagement Matters*. Springer, 1–26.

[32] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39.

[33] Jeffrey M Rzeszotarski, Ed Chi, Praveen Paritosh, and Peng Dai. 2013. Inserting micro-breaks into crowdsourcing workflows. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[34] Ameneh Shamekhi, Mary Czerwinski, Gloria Mark, Margeigh Novotny, and Gregory A Bennett. 2016. An exploratory study toward the preferred conversational style for compatible virtual agents. In *International Conference on Intelligent Virtual Agents*. Springer, 40–50.

[35] Tanya Stivers and Jeffrey D Robinson. 2006. A preference for progressivity in interaction. *Language in society* 35, 3 (2006), 367–392.

[36] Deborah Tannen. 1987. Conversational style. *Psycholinguistic models of production* (1987), 251–267.

[37] Deborah Tannen. 2005. *Conversational style: Analyzing talk among friends*. Oxford University Press.

[38] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. ACM, 42–51.

[39] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, Vol. 5.

[40] Bert Vandenberghe. 2017. Bot personas as off-the-shelf users. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 782–789.

[41] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2187–2193.

[42] Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. 2016. Modeling task complexity in crowdsourcing. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

[43] Xi Yang, Marco Aurisicchio, and Weston Baxter. 2019. Understanding Affective Experiences With Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 542.

[44] Mengdie Zhuang and Ujwal Gadiraju. 2019. In What Mood Are You Today?: An Analysis of Crowd Workers' Mood, Performance and Engagement. In *Proceedings of the 11th ACM Conference on Web Science, WebSci 2019, Boston, MA, USA, June 30 - July 03, 2019*. 373–382. DOI:http://dx.doi.org/10.1145/3292522.3326010