Crowd Computing - Use Of Conversational Interfaces To Improve Task Clarity

Sercinho Banda
s.m.banda@student.tudelft.nl
TU Delft
Delft, the Netherlands 4194020

Alex Jeleniewski
a.j.jeleniewski@student.tudelft.nl
TU Delft
Delft, the Netherlands 4701925

Cees Jol
c.f.jol@student.tudelft.nl
TU Delft
Delft, the Netherlands 4653904

Stef Kaptein
s.b.m.kaptein@student.tudelft.nl
TU Delft
Delft, the Netherlands 4683919

Tom Saveur
t.saveur@student.tudelft.nl
TU Delft
Delft, the Netherlands 4657144

ABSTRACT

This paper discusses a method of improving perceived task clarity in micro task crowd sourcing. To try and improve the task clarity a new method of performing crowd sourcing tasks is discussed, namely using a conversational interface. Using an experiment it was tested whether or not having a conversational interface will improve perceived task clarity compared to a standard web interface. In which way a Conversational Interface might help improve task clarity was also tested.

CCS CONCEPTS

• **Information systems** → Chat; **Crowdsourcing**.

KEYWORDS

crowd sourcing, crowd computing, conversational interfaces, task clarity

ACM Reference Format:
Sercinho Banda, Alex Jeleniewski, Cees Jol, Stef Kaptein, and Tom Saveur. 2021. Crowd Computing - Use Of Conversational Interfaces To Improve Task Clarity. In *TU Delft '21: Crowd Computing, June 11, 2021,Delft, NL.* ACM, New York, NY, USA, 7 pages. https://doi.org/CS4145

TU Delft '21, June 11, 2021, Delft, NL

ACM ISBN CS4145…$0.00
https://doi.org/CS4145

1 INTRODUCTION

Microtask crowdsourcing enables data collection at large scale to solve problems that machines have yet to overcome. Think of problems like labeling data for the training of machine learning algorithms, or providing opinions based on a given context. The actors involved in this system are the task requesters (or work providers) who design Human Intelligence Task (HIT) which are published in batches on crowdsourcing platforms. The crowd workers (or employees) then perform these HITs in exchange for monetary rewards [6]. Before the crowd workers receive any rewards, their work is reviewed by the requesters. Should the quality of the work submitted by the crowd workers not meet the standards of the requester, then their work could be rejected which results in a loss of pay, time and reputation for the worker [16]. Rejecting a lot of data is also not beneficial for the requester, since the task will take longer to achieve the required number of workable data if a lot of results are rejected.

However, it should be noted that the lacking quality of the work submitted by the crowd worker can also be attributed to a poor task design by the requester [10, 12]. As a result, the performance of crowd workers is unfairly penalized due to unclear task instructions and unclear task expectations [9]. Furthermore, crowd workers interact with the crowdsourcing platform by means of a worker interface. Currently, most online crowdsourcing tasks are viewed and completed using standard HTML-based interfaces [19]. An alternative to the standard Web interface is a conversational interface (CI) that uses a text-based conversational agent to help crowd workers interact with the crowdsourcing system in a natural way [2]. According to Mavridis et al. [15], microtasks performed with conversational interfaces have a higher user satisfaction while producing similar quality and task execution time as Web interfaces. Also, individuals of diverse backgrounds are familiar with the design and functionality of conversational interfaces due to its wide use in messaging applications [2].

1.1 Original Contribution

We will test our hypothesis: *"Conversational flow positively influences the perceived task clarity in microtask crowdsourcing".*

To test this, our first goal is to increase our understanding *if* conversational interfaces can improve the perceived task clarity of crowd workers compared to a classical interface. Our second goal is find out *how* such interfaces could help. Hence, we will answer the following research questions:

- "Can the use of text-based conversational agents improve the perceived task clarity in microtask crowdsourcing?"
- "*How* can the use of text-based conversational agents improve the perceived task clarity in microtask crowdsourcing?"

2 BACKGROUND AND RELATED WORKS

Conversational agents, also known as chatbots, are computer programs that interact with humans to simulate human conversation [1]. The concept of using conversational agents is not new, but their abilities and application potential have expanded with advances in artificial intelligence, natural language processing, and cognitive systems [4]. Despite recent advances, most conversational agents are still unable to ask meaningful follow-up questions to refine the user's intent. Therefore, most works regarding conversational agents focus on a simple goal-oriented task, characterized by a structured conversation with predictable user input [7]. Furthermore, the behavior of the conversational agent is dictated by the design decisions made concerning what actions the conversational agent can take and how to implement these actions [3]. However, based on how conversational agents extract the information they can be categorized in three different types [7]:

- Passive: In this scenario, the conversational agent does not ask any questions, but instead it leaves query refinement to the user.
- Active: In this scenario, the conversational agent does ask questions to help the user make their query more specific.
- Pro-Active: In this scenario, the conversational agent makes query suggestions that transcend the scope of the original query and in doing so the agent can change and/or expand the available information need.

For our experiment, we will be using chatbots that fall under the first two categories, namely passive and active. We also define an extra category, a dummy bot which acts as a step in between the normal web interface and the active/passive bots. The dummy bot is a bot without any intelligence, it is just a different way of providing the same information as the normal web interface.

Conversational Interfaces include text-based dialogue systems, voice-based conversational assistants, embodied virtual agents, and social robots [8]. Compared to the traditional graphical user interface, the graphical elements in conversational interfaces are generally minimal, but they have a more humanlike interaction which provides them with an advantage [17]. Another advantage of conversational interfaces is that they can ease complex tasks by filtering information [18]. Previous research within the HCI community focused on people's experiences with conversational agents, highlighting the gaps between user's expectations and their actual experiences [5, 14]. Other research focused on user needs and user satisfaction in the context of specific tasks such as information retrieval [20, 21]. Based on the findings of prior research, conversational agents can positively impact the user experience. Therefore, this paper will investigate the potential of text-based conversational agents to improve task clarity in microtask crowdsourcing.

3 CONVERSATIONAL INTERFACE DESIGN

We created three types of CIs, all with various levels of intelligence and with various strategies to help the worker. Thus, there is only one independent variable; the conversation style of the interface. In this section each of the Conversational Interfaces will be described; starting with the *"dummy"* CI, after which the *active* and the *passive* CI will be explained.

To create the various CIs, Botpress[1] was used. Using this program allowed for fairly straightforward implementations of a Conversational Agent. This allows the user to ask questions and the agent to gauge the level of understanding of the user. In turn, the user could get more clarification on the task. The bots we implemented can be found on our Github repository[2].

3.1 Dummy CI

The *dummy* CI is the most basic form of CI: it is just the normal questionnaire that we give on Toloka, but asked in a CI fashion. This forms the baseline to see if having a CI alone is enough to improve clarity.

The CI will start by explaining the task the worker has to perform, after which it goes on straight to the questions.

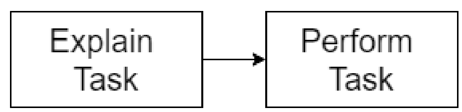

Figure 1: The Dummy CI Implementation

As can be seen from the implementation in figure 1, the worker is unable to ask questions to the bot, nor has the CI been given any Natural Language Understanding (NLU) skills. This means that it does not even know what the worker says whenever they enter something; it will just assume that whatever the worker enters, is the answer to the question.

The dummy CI is used as a baseline for the effect of a CI. We can compare it to the web interface to see if a CI in itself has effect on task clarity. In addition, we can see how the different types of CI effect the task clarity compared to the dummy interface.

3.2 Active CI

Thanks to the build in NLU, the active CI is able to answer questions that the worker has while it is in the appropriate state. Also, it asks the worker a check question before moving on to the real questions to ensure that the worker understands the problem.

[1]Botpress: **https://botpress.com/**
[2]The full implementation can be found at: **https://github.com/Tomaati/CS4145-Project**

It will start by explaining the task in the same way as the *dummy* CI does, but before moving on to the task it will ask the worker whether or not he/she understands everything; if this is not the case it will start the QnA phase, where the worker can ask questions to the CI. After the worker has no questions it will continue to the Check phase, where the CI will ask a simple question to the worker, and if he/she gets it correct the task will continue, else the CI will give a short and different explanation on the task and asks if the worker has any questions and move to the QnA or Check phase again based on the answer of the worker.

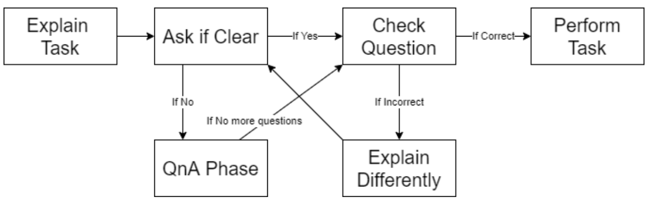

Figure 2: The Active CI check implementation

From the implementation in figure 2 it becomes clear what the difference is between the *"dummy"* CI proposed in section 3.1, since it will actively ask and poke the worker to gain a better understanding of their knowledge. It also shows the flow when a worker does not understand the task, which explains how it might help the worker understand the task better.

One thing to note is that in the *Active* CI the worker is unable to ask questions during the questions we ask for our task, this can only be done in the *Passive* CI, which will be described next.

The active CI is used to determine whether asking for a confirmation (before answering) improves the accuracy of the answers.

3.3 Passive CI

The main idea of the *Passive* CI is that the worker is able to ask questions to the CI *during* the task execution, instead of the CI taking the initiative. This was the most complicated CI since it will need to keep track of several *intents* of the worker. The bot aims to see the intent of the user: the worker might be answering the question, or they could be asking for clarification.

All of these factors combined lead to a CI that was quite complex to write. The CI works by keeping track of the current question being asked to the worker, and according to that information, finds the answers from their understanding. The CI asks a question to the worker, and if the worker has a question, the CI will move to the QnA phase, where the CI will answer the questions and return to the task.

Another difficult problem was figuring out what types of questions the worker could have. We ended up picking a mixture of questions about the person, and about the task as a whole, which resulted in questions like the following:

- What if someone has multiple middle names?
- Do all persons have a middle name?
- Should the names be capitalized?
- Where is PERSON from?
- Was PERSON very famous?
- When was PERSON born?
- Is PERSON still alive?
- Does PERSON have well-known relatives?

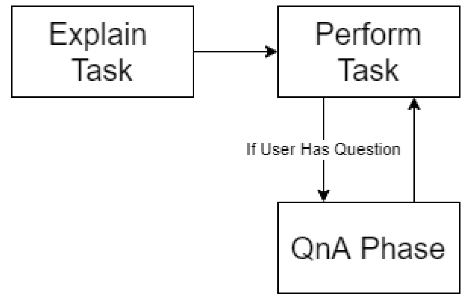

Figure 3: The Passive CI Implementation

The passive CI is closest to a simulation of a real conversation. Using this CI it can be determined whether the simulation of a real conversation improves the accuracy of the answers.

4 EXPERIMENTS

In order to test our hypotheses, an experiment was performed. The experiment was performed using the method called *"between-groups"*.

The platform we used for the experiments is Toloka[3]. Toloka is a crowdsourcing platform that allows for creating questionnaires. In this section, we will discuss the experiment design and its execution.

4.1 Setup

In the experiment, tasks were performed by four different groups: a group with a regular interface (so without a CI), a group with a so-called *"dummy"* CI, a group with a *passive* CI, and a group with an *active* CI. In all cases, the tasks are exactly the same, to avoid having any unwanted variables. In the CI groups, the tasks are presented in a conversational style, and crowd workers can ask for clarification through this interface. The regular interface group uses the standard interface provided by Toloka.

Like stated in section 3, the "Without CI" group acts as our baseline for *if* the CI could improve task clarity, while the "Dummy CI" acts as a baseline for *how*.

4.2 Independent Variables

To only measure our intended variables, we defined our independent variables. We will discuss all independent variables for each part of the experiment here, a more detailed description of the various types of CI can be found in section 3.

[3]Toloka: **https://toloka.yandex.com/**

Without CI. This group is the current implementation of normal crowd computing tasks, where the questions are asked as a regular survey.

Dummy CI. This version changes the conversational flow from a survey to a CI.

Active CI. In an active CI, the CI will poke and ask questions to the user. This group is made to answer the question on *how* a conversational interface could help improve task clarity.

Passive CI. In a passive CI, the CI will go through the explanation. Only once the user asks a question does the CI elaborate. This group is used to check how this stacks up to the Active CI in terms of improved task clarity.

4.3 Task type

It is difficult to improve task clarity on a task that is already clear. So, for this experiment, we used an *Interpretation and Analysis (IA)* type of task. Mainly because this task scores lowest when comparing various types of tasks based on task and goal clarity [10]. In tasks such as labeling or annotating, the task at hand is quite clear and clarification is often not needed. The IA task seems to be more difficult to understand than others, so in our theory, we could see the biggest improvement here.

4.4 Task definition

The IA task is a broad term that encapsulates multiple types of subtasks. More specifically, to execute our experiment, we used a *middle name lookup* task. We ask about a famous person's middle name, such as George Bush. Some context is provided around the person, to make sure that the task is not ambiguous. This can be a very hard task because there are often multiple people with such names, and the task does not always request the most famous person. The task difficulty means that the worker might be wanting to ask some questions, which allows us to test if our bot answering the questions improves task clarity. In short, we think this is a very good task to use for our experiment, as it gives us many chances to apply the features of our CIs.

The task consists of the following six questions:

(1) **What is the middle name of George Bush?** George Bush was the president of the United States in 2008.
(2) **What is the middle name of Michael Jordan?** Michael Jordan is an Artificial Intelligence researcher.
(3) **What is the middle name of Charles Davis?** Charles Davis was an American football player in 1976.
(4) **What is the middle name of Johann Bach?** Johann Bach was a composer in 1703.
(5) **What is the middle name of Richard Nixon?** Richard Nixon is a footballer.
(6) **What is the middle name of Kevin Johnson?** Kevin Johnson was an American Football player in 1995.

4.5 Dependent variables

To measure the performance of the different groups, we measure the task clarity, as well as the speed and accuracy of the worker. We also test the understanding of the worker using questions with known answers to evaluate the performance differences between the active and the passive bot.

4.5.1 Task Clarity. The task clarity can be defined by the factors: goal and role clarity. The goal and role clarity can be measured using a survey on a 5-point Likert scale [13]. All groups have to answer these questions at the end of the task.

4.5.2 Speed and Accuracy. When a task is difficult to understand, there are two possible consequences. First, a worker takes more time on a task, and second, the accuracy will be lower. The accuracy is a measure of the number of answers that are in the domain of the answer for the task. A good score implies that the worker most likely understood the question.

A worker might try to look up information about the task at hand, or he/she might try to answer the question although they do not fully understand it. So, when a task is more clear, a worker spends less time figuring out the task and is able to more accurately answer the task.

4.5.3 Testing Understanding of the Worker. Questions with known answers are often used to filter between crowd workers. By asking these types of questions, workers that fail them can be filtered out since they weren't paying enough attention, or do not understand what they are supposed to do [11]. In this experiment, we can use these questions to get a better understanding of how a passive and active CI influences the understanding/attention of a worker.

4.6 Implementation

The three bots described in 3 were deployed using Docker on Heroku, following the tutorial on the Botpress documentation[4]. Each CI was embedding on a Netlify site[5]. Links to the bots were provided on Toloka. After completing the task on the CI, the bots refer the user to a Google Form, to assess task clarity on the Likert scale. Finally, after completing the survey, the Google Form provides the crowd worker a code that they can enter in Toloka to finish the task.

4.7 Quality Control

Since this experiment is about the perceived task clarity it is important that the data gathered from the workers is high quality. This data can then used to give more insight into the perceived task clarity next to the Likert scale.

In order to ensure the highest quality of results and to mitigate cheating, we created a certain experimental flow. Namely, after completing all questions, the workers are given a link to the form, which contain the questions with answers on a scale from 1-5, the Likert scale. At the end of this form, they receive a code; which they need to enter in Toloka. If this code is correct we can safely assume that the participant tried to solve the task entirely.

Time is also tracked between each question. If the answers are wrong and the time between each question is short, we can safely

[4]Deploying a Botpress bot on Heroku: **https://botpress.com/docs/deploy/hosting**
[5]Embedding a Botpress bot on a website: **https://botpress.com/docs/channels/web**

assume that the worker was not motivated to solve the task, and just entered random answers. In the case of this, the result will be invalidated, and the worker will not receive any payment.

5 RESULTS AND EVALUATION

We evaluate our results with the metrics we defined in subsection 4.5. That is, we measure the speed and accuracy of a worker, and we compare the task clarity through the responses of the questionnaire that we provided for each task.

5.1 Accuracy

The worker accuracy for all experiments is listed in Table 1 and they were measured by comparing the answers provided by the workers with the expected value. The overall worker accuracy was the highest for the group that performed the tasks without a conversational interface (μ= 52.5). Also, when comparing the accuracy of the groups that performed the tasks with a CI, it is noticeable that the accuracy is lowest for the group that was actively supported by the conversational agent (μ= 28.57). Furthermore, the group where information elicitation about the clarity of the instruction was performed passively scored slightly higher (μ=34.17) than the group that used the dummy CI (μ=33.33).

Table 1: The worker accuracy (μ± σ: mean and standard deviation, unit: percentage) for all experiments.

Worker Accuracy			
Without CI	Dummy CI	Active CI	Passive CI
52.5 ± 24.31	33.33 ± 23.09	28.57 ± 20.02	34.17 ± 20.73

5.2 Execution time

The average execution times of the four groups are shown in Table 2. As can be seen in Table 2, the lowest average execution time of 397 seconds was achieved by the workers that did not use a CI. On the other hand, the tasks performed with the help of a passive conversational agent had the highest average execution time of 512 seconds. Also, the experiment with the dummy CI has the best execution time of all the experiments that used a CI. Furthermore, the average execution time of the groups that used a conversational agent was at least 100 seconds higher than their counter part. Meaning that if these results are to be believed than it was more expansive to perform the tasks with a conversational agent.

Table 2: The worker execution time (μ± σ: mean and standard deviation, unit: seconds) for all experiments.

Worker Execution Time			
Without CI	Dummy CI	Active CI	Passive CI
397 ± 178,61	450 ± 239,59	500,71 ± 240,23	512,4 ± 222,73

5.3 Clarity

In this section, the results of the questionnaires that were answered by the workers at the end of the task are outlined.

5.3.1 *CI vs No CI.*

The first four questions of the questionnaire compare the regular interface (without CI) to an interface with a CI.

Rate how confident you were in your answers to the tasks. The results shows no difference between workers confidence in answers to the tasks. No CI: 57.9% 5/5, 89.5% at least 4/5. CI: 60% 5/5, 81.5% at least 4/5.

Rate the perceived difficulty of the tasks. Workers of the CI groups rated the difficulty of tasks as more difficult. 75.3% rated the task as 4/5 or 5/5 difficult, while for the group without the CI, only 47.4% gave these ratings. Possible, the use of the CI made the task more difficult, as it introduced a new element for crowdworkers: they may not be familiar with this type of interface for performing tasks.

Do you feel that a more detailed description would have helped you with answering the tasks correctly? The group with the CI indicated that a detailed description would have helped slightly more than the group without the CI: 72,3% vs 63,1% indicated this at a 4/5 or 5/5 rating.

Do you feel that a more detailed description would have helped you with answering the tasks quicker? No statistical difference was found for this question, as 63,1% (no CI) vs 64,6% (CI) indicated that a more detailed description would have helped with answering the tasks quicker.

In general, the results indicate that for our experiment, a CI does not improve task clarity.

5.3.2 *Dummy vs Active vs Passive CI.*

The last two questions of the questionnaire test the three bot types against each other.

To what extent did the chat bot aid you in completing the tasks correctly? No significant differences: 68.2% for the dummy CI, 77.3% for the passive CI, 76.2% for the active CI.

To what extent did the chat bot influence the task completion time? No significant differences: 77.2% for the dummy CI, 72.7% for the passive CI, 81% for the active CI.

The results indicate that there is no difference in task clarity between the types of CI.

5.4 Evaluation

Overall, the crowdworkers who worked with the bots performed worse than the crowdworkers with a standard interface. We hypothesize that this is because of a few reasons:

- The bot was developed in a rather short amount of time, namely, a few weeks, and thus might not have performed as well as it would need to, to improve task clarity.
- The crowdworkers might not be used to working with a CI. Thus, they performed the task faster on the standard interface.

- The sample size of our experiment was rather low (20 workers per group), which means that there is quite some deviation in the results.

6 LIMITATIONS AND FUTURE WORK

Due to the time constraints of this project there are still some limitations of our experiments, as well as some future work that could be done given more time.

6.1 Conversational Interfaces

As mentioned in section 3 botpress was used to create the interfaces, and while the program is useful for quick prototyping we came to the conclusion that the use case might be different from our experiment. It was designed to be more of a chatbot that handles hotel reservations or books flights, instead of being an interface that a worker can interact with. This meant that some things did not go as planned. The most difficult part to program was the *passive* CI, since there was no clear documentation on how to write a QnA during a Question phase. In hindsight, we would have chosen a different framework that would fit our use case a little better.

6.1.1 NLU. The current implementation of the NLU is decent, but could be better. This is something that we had to work around due to the time constraints, since it is difficult to think of everything that a user might ask/tell. But should we have had more time, improving the NLU would important, since having good NLU will most likely result in better scores for the CIs.

6.2 Worker Validation

Currently we mainly used the code given at the end of the Google Forms described in 4 to validate that the worker did their tasks, and to ensure that they could not still have it copy/pasted we also ensured that each worker could only perform one of our tasks. But we did not really do a lot of validation for our data from the CI, we only used time to know whether or not they tried it. And we realized that for the small subset we had this worked fairly well, but there were some users that just waited for a bit, entered random answers and got the correct code from the Google Forms to enter in Toloka. These were of course ignored after manual review, but having to review every crowd worker took a lot of time. To scale up this experiment, more automatic validation would be needed. Like checking whether or not the answer contains a valid name, instead of just random characters.

7 CONCLUSION

In this paper, we tried to see if and how using a conversational interface could improve the task clarity in a crowdsourcing task. We designed, implemented, and evaluated an experiment to test our hypotheses on Toloka using Botpress. We designed three bots: a dummy, an active, and a passive bot. In our results, using a conversational interface did not improve task clarity. Also, the type of bot did not have influence on task clarity.

To improve task clarity, the interface should be intuitive and straightforward for the workers. Future work could be done about what makes a conversational interface useful to crowdworkers. Also, future work could be done about what kind of bot could improve task clarity: passive, active, or pro-active.

ACKNOWLEDGMENTS

With thanks to Sihang Qiu and Andrea Mauri for providing us with feedback during the project.

REFERENCES

[1] Bayan Abu Shawar and Eric Atwell. 2007. Chatbots: are they really useful? *In LDV-Forum* 22 (2007), 29–49. Issue 1.
[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* (2018), 52–61. https://doi.org/10.1145/3176349.3176380
[3] Eva Bittner, Sarah Oeste-Reiß, and Jan Marco Leimeister. 2019. Where is the Bot in our Team? Toward a Taxonomy of Design Option Combinations for Conversational Agents in Collaborative Work. *Proceedings of the 52nd Hawaii International Conference on System Sciences* 6 (2019), 284–293. https://doi.org/10.24251/hicss.2019.035
[4] Ana Paula Chaves and Marco Aurelio Gerosa. 2018. Single or multiple conversational agents? An interactional coherence comparison. *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), 1–13. https://doi.org/10.1145/3173574.3173765
[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* (2017), 1–12. https://doi.org/10.1145/3098279.3098539
[6] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philppe Cudre-Mauroux. 2015. The dynamics of micro-task crowdsourcing: The case of amazon mturk. *Proceedings of the 24th international conference on world wide web* (2015), 238–247. https://doi.org/10.1145/2736277.2741685
[7] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, Damien Anderson, and Sylvain Daronnat. 2020. Conversational Strategies: Impact on Search Performance in a Goal-Oriented Task. *In ACM CHIIR 3rd Conversational Approaches to Information Retrieval Workshop (CAIR)* (2020), 1–7.
[8] Justin Edwards and Elaheh Sanoubari. 2019. A Need for Trust in Conversational Interface Research. *In 1st International Conference on Conversational User Interfaces (CUI 2019)* (2019), 1–3. https://doi.org/10.1145/3342775.3342809
[9] Ujwal Gadiraju, B Fetahu, and C Hube. 2016. Crystal clear or very vague? Effects of task clarity in the microtask crowdsourcing ecosystem. *Trust In Crowd Work Workshop at WebSci'16 Hannover* (2016). https://doi.org/10.475/123_4
[10] 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* (Prague, Czech Republic) *(HT '17)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/3078714.3078715
[11] Anja S. Göritz, Kathrin Borchert, and Matthias Hirth. 2021. Using Attention Testing to Select Crowdsourced Workers and Research Participants. *Social Science Computer Review* 39, 1 (2021), 84–104. https://doi.org/10.1177/0894439319848726 arXiv:https://doi.org/10.1177/0894439319848726
[12] Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: A study of a large crowdsourcing marketplace. *Proceedings of the VLDB Endowment* 10, 7 (2017). https://doi.org/10.14778/3067421.3067431
[13] Ankur Joshi, Saket Kale, Satish Chandel, and Dinesh Pal. 2015. Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology* 7 (01 2015), 396–403. https://doi.org/10.9734/BJAST/2015/14975
[14] 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* (2016), 5286–5297. https://doi.org/10.1145/2858036.2858288
[15] 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* (2019), 243–251. https://doi.org/10.1145/3320435.3320439
[16] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a hit: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. *Proceedings of the 2016 CHI conference on human factors in computing systems* (2016), 2271–2282. https://doi.org/10.1145/2858036.2858539
[17] 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* (2017), 492–497.

https://doi.org/10.1145/3027063.3027077

[18] David G Novick. 2002. " Conversational " Dialogues in Direct-Manipulation Interfaces. *In Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems* (2002), 1–5.

[19] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Estimating Conversational Styles in Conversational Microtask Crowdsourcing. *In Proceedings of the ACM on Human-Computer Interaction* 4 (2020), 1–23. Issue CSCW1. https://doi.org/10.1145/3392837

[20] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L A 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* (2017), 2187–2193. https://doi.org/10.1145/3027063.3053175

[21] Jennifer Zamora. 2017. Rise of the Chatbots: Finding A Place for Artificial Intelligence in India and US. *In Proceedings of the 22nd international conference on intelligent user interfaces companion* (2017), 109–112. https://doi.org/10.1145/3030024.3040201